CONSIDERING STUDENTS, TEACHERS, AND WRITING ASSESSMENT

VOLUME 1, TECHNICAL AND POLITICAL CONTEXTS

Perspectives on Writing

Edited by
Diane Kelly-Riley
Ti Macklin, and Carl Whithaus

CONSIDERING STUDENTS, TEACHERS, AND WRITING ASSESSMENT: VOLUME 1, TECHNICAL AND POLITICAL CONTEXTS

PERSPECTIVES ON WRITING

The Perspectives on Writing series addresses writing studies in a broad sense. Consistent with the wide ranging approaches characteristic of teaching and scholarship in writing across the curriculum, the series presents works that take divergent perspectives on working as a writer, teaching writing, administering writing programs, and studying writing in its various forms.

The WAC Clearinghouse and University Press of Colorado are collaborating so that these books will be widely available through free digital distribution and low-cost print editions. The publishers and the series editors are committed to the principle that knowledge should freely circulate and have embraced the use of technology to support open access to scholarly work.

Recent Books in the Series

Amy Cicchino and Troy Hicks (Eds.), *Better Practices: Exploring the Teaching of Writing in Online and Hybrid Spaces* (2024)

Genesea M. Carter and Aurora Matzke (Eds.), *Systems Shift: Creating and Navigating Change in Rhetoric and Composition Administration* (2023)

Michael J. Michaud, *A Writer Reforms (the Teaching of) Writing: Donald Murray and the Writing Process Movement, 1963–1987* (2023)

Michelle LaFrance and Melissa Nicolas ((Eds.), *Institutional Ethnography as Writing Studies Practice* (2023)

Phoebe Jackson and Christopher Weaver (Eds.), *Rethinking Peer Review: Critical Reflections on a Pedagogical Practice* (2023)

Megan J. Kelly, Heather M. Falconer, Caleb L. González, and Jill Dahlman (Eds.), *Adapting the Past to Reimagine Possible Futures: Celebrating and Critiquing WAC at 50* (2023)

William J. Macauley, Jr. et al. (Eds.), *Threshold Conscripts: Rhetoric and Composition Teaching Assistantships* (2023)

Jennifer Grouling, *Adapting VALUEs: Tracing the Life of a Rubric through Institutional Ethnography* (2022)

Chris M. Anson and Pamela Flash (Eds.), *Writing-Enriched Curricula: Models of Faculty-Driven and Departmental Transformation* (2021)

Asao B. Inoue, *Above the Well: An Antiracist Argument From a Boy of Color* (2021)

Alexandria L. Lockett, Iris D. Ruiz, James Chase Sanchez, and Christopher Carter (Eds.), *Race, Rhetoric, and Research Methods* (2021)

CONSIDERING STUDENTS, TEACHERS, AND WRITING ASSESSMENT: VOLUME 1, TECHNICAL AND POLITICAL CONTEXTS

Edited by Diane Kelly-Riley, Ti Macklin, and Carl Whithaus

The WAC Clearinghouse
wac.colostate.edu
Fort Collins, Colorado

University Press of Colorado
upcolorado.com
Denver, Colorado

The WAC Clearinghouse, Fort Collins, Colorado 80523

University Press of Colorado, Denver, Colorado 80202

ISBN 978-1-64215-216-6 (PDF) 978-1-64215-217-3 (ePub) 978-1-64642-619-5 (pbk.)

DOI 10.37514/PER-B.2024.2166

Produced in the United States of America

Library of Congress Cataloging-in-Publication Data

Pending

Copyeditor: Andrea Bennett
Designer: Mike Palmquist
Cover Photo: Image by Rawpixel.com. Image ID 13081212.
Series Editors: Rich Rice and J. Michael Rifenburg
Consulting Editor: Susan H. McLeod
Associate Editors: Jonathan M. Marine, Johanna Phelps, and Qingyang Sun

The WAC Clearinghouse supports teachers of writing across the disciplines. Hosted by Colorado State University, it brings together scholarly journals and book series as well as resources for teachers who use writing in their courses. This book is available in digital formats for free download at wac.colostate.edu.

Founded in 1965, the University Press of Colorado is a nonprofit cooperative publishing enterprise supported, in part, by Adams State University, Colorado State University, Fort Lewis College, Metropolitan State University of Denver, University of Alaska Fairbanks, University of Colorado, University of Denver, University of Northern Colorado, University of Wyoming, Utah State University, and Western Colorado University. For more information, visit upcolorado.com.

Citation Information: Kelly-Riley, Diane, Ti Macklin, & Carl Whithaus (Eds.). (2024). *Considering Students, Teachers, and Writing Assessment: Volume 1, Technical and Political Contexts*. The WAC Clearinghouse; University Press of Colorado. https://doi.org/10.37514/PER-B.2024.2166

Land Acknowledgment. The Colorado State University Land Acknowledgment can be found at landacknowledgment.colostate.edu.

CONTENTS

FOREWORD

Kathleen Blake Yancey
Florida State University

Brian Huot
Kent State University

It's a truism to note that writing assessment has come into its own during the last several decades, and one of the factors propelling that growth is the *Journal of Writing Assessment (JWA)*. As the selected articles-now-chapters presented here suggest, writing assessment is both more and different than what it seems to be. While it can simply appear as a rudimentary exercise in evaluating writing, writing assessment, as the authors here have documented, researched, and theorized, is at least twofold: (1) an exercise of considerable sophistication and complexity operating within a context (2) that can overwhelm, and sometimes sabotage, the exercise itself. These twin observations informed our goal when we created *JWA*, a new journal focused on writing assessment that would circulate scholarship taking up questions about how to best assess writing as well as about the contextual factors, often invisible, that shape and, too often, mis-shape writing assessment. Put in the current vernacular, with *JWA* we hoped to make writing assessment—and its many dimensions—transparent.

The articles in our first issue of *JWA* made this goal visible. In "Moving Beyond Holistic Scoring through Validity Inquiry," for instance, Peggy O'Neill (2003) focused on validity, a key issue in writing assessment; her article is included here. Turning to context, George Hillocks (2003) addressed the impact of state assessments in his "How State Assessments Lead to Vacuous Thinking and Writing." Sandra Murphy did likewise, in her case looking not at the impact of writing assessment on students, but rather at its impact on teachers in one state; such teachers support students' writing development as they practice assessment within their classrooms.

That first issue of *JWA* concluded with an annotated bibliography; compiled by Peggy O'Neill, Michael Neal, Ellen Schendel, and Brian Huot (2003), it too spoke to *JWA*'s vision. Three bibliographic entries in particular articulate *JWA*'s goal and its importance while forecasting the kinds of research, theory, and practice published in *JWA* during the last 14 years, as sampled in this edited collection. The first bibliographic entry, Nicholas Lemann's 1999 book *The Big Test: The Secret History of the American Meritocracy*, details a social and cultural history of the SAT.

DOI: https://doi.org/10.37514/PER-B.2024.2166.1.2

Although the stated purpose of the SAT was to change the college admissions process by eliciting relevant information from college applicants so as to predict their success in college, it also clearly intended to shift college admissions from one based in legacies to one based in merit. The Lemann account also clarifies how the SAT both succeeded and failed in that intention, demonstrating that assessment, even when informed by the science of tests and measurements, is always contextualized, always enacting a policy, whether visible or not.

A second item in the bibliography, O. Palmer's College Board Report 42, "Sixty Years of English Testing," (1960) argues that the science informing the College Entrance Examination Board (CEEB) English testing contributes to such testing "as a scientifically defensible practice" (O'Neill at al., 2003). Again, here too science plays a role, not so much to forward a kind of democracy, however, but rather to defend the practices of a growing assessment industrial complex. In the CEEB model Palmer defends, both teachers and direct writing assessment are positioned as opponents of CEEB, as "resistant to the scientific progress achieved in English testing" (O'Neill et al., 2003). What teachers, rooted in the everyday of the human classroom, may have understood better than measurement experts is how writing assessment, regardless of the science, cannot be cleaved from the contexts and complications accompanying it. As important, seeing students day in and day out, teachers also understood how very contingent any decision based on assessment is.

In his 1994 "A Technological and Historical Consideration of Equity Issues Associated with Proposals to Change the Nation's Testing Policy," George Madaus seems to agree with teachers. Approaching what we might call the assessment problem philosophically in this third bibliographic entry, with a view informed by both phenomenology and practicality, Madaus observes that certain principles define assessment. All evaluations, he notes, rely on samples of behavior; all evaluations make inferences "about a person's probable performance relative to the domain" (Madaus, 1994); and all assessments render decisions by individual or institution. Moreover, the technologies don't operate apart from the culture of their origin. Instead, as

> products of a culture, they often extend, shape, and reproduce the same culture. The values that underlie testing are utilitarianism, economic competition, technological optimism, objectivity, bureaucratic control and accountability, numerical precision, efficiency, standardization, and conformity. (O'Neill et al., 2003)

It's worth noting that while such values, including standardization, conformity, and economic competition, may locate the US, its testing industry, and its schools, they are much less likely to be the values motivating teachers.

The articles in the two volumes of this edited collection carry these issues of assessment and context forward, especially as they have been raised and considered over time. In Volume 1, the collection's first section, *Technical Issues in the Assessment of Writing—Reliability and Validity,* speaks to issues articulated by both Peggy O'Neill and George Madaus, issues inherent in assessment that, as both O'Neill and Madaus demonstrate, are *not apart from* larger human issues, but are rather *a part of* them. The second section, *Politics and Public Policy of Large-Scale Writing Assessment,* calls to mind the article by George Hillocks and the history of college admissions provided by *The Big Test.* The third section, *Implications of Automated Scoring of Writing*, questions how the evolution of automated essay scoring extends the dangerous logic of a "true" score as valid and reliable across contexts. In Volume 2, the fourth section*, Theoretical Evolutions—Towards Fairness and Aspiring to Justice,* again calls to mind the equity issues and analysis developed by Madaus. And the fifth section, *Students' and Teachers' Lived Experiences,* evokes the line of inquiry pursued by Sandra Murphy. As astute readers have already noted, it's also fair to observe that in this set of correspondences between the introductory issue of *JWA* and the current collection's chapters, one section in the collection, *Implications of Automated Scoring of Writing*, is left out: our first issue of *JWA* did not provide for the important questions about writing assessment raised by digital technologies. Still, apprehending that they were on the horizon, we made a start in the very next issue, courtesy of Michael Williamson's (2003) "Validity of Automated Scoring: Prologue for a Continuing Discussion of Machine Scoring Student Writing."

All of which is *not* to say that we anticipated all of the rich writing assessment scholarship of the next decade and a half: our correspondences, of course, are not predictive. But it is to say that the chapters here extend and elaborate what we had hoped for in creating *JWA*, in the process refiguring continuing issues, sounding new notes, and pointing us to new futures. For example, one chapter argues that the divide between the educational measurement and the writing assessment communities might be bridged with a "unified field of writing assessment." The construct of writing, another chapter explains, can no longer ignore "the role of commonly available tools such as word processing software." And yet another chapter brings together science and the law in a shared inquiry into the results and subsequent effects of writing assessment, employing a disparate impact analysis framework contributing to a better, more human, more humane, and more equitable assessment. Threaded throughout are the technical issues and principles of writing assessment, the writing assessments themselves, and the contexts in which they are embedded.

Remembering the recent history of writing assessment, focused on assessments and their contexts as we prepare for a better writing assessment future, we

are very pleased to be learning from and with the authors included here. We feel confident that you will be as well.

REFERENCES

Hillocks, G. (2003). How state assessments lead to vacuous thinking and writing. *Journal of Writing Assessment, 1*(1), 5-21. https://escholarship.org/uc/item/33k2v0w5

Lemann, N. (1999). *The big test: The secret history of the American meritocracy*. Farrar, Straus and Giroux.

Madaus, G. F. (1994). A technological and historical consideration of equity issues associated with proposals to change the nation's testing policy. *Harvard Educational Review, 64*(1), 76.

Murphy, S. (2003). That was then, this is now: The impact of changing assessment policies on teachers and the teaching of writing in California. *Journal of Writing Assessment, 1*(1), 23-45. https://escholarship.org/uc/item/1fg1531r

O'Neill, P. (2003). Moving beyond holistic scoring through validity inquiry. *Journal of Writing Assessment*, *1*(1), 47-65. https://escholarship.org/uc/item/4qp611b4

O'Neill, P., Neal, M., Schendel, E., & Huot, B. (2003). An annotated bibliography of writing assessment. *Journal of Writing Assessment*, *1*(2), 73-78. https://escholarship.org/uc/item/78q9v569

Palmer, O. (1960). Sixty years of English testing. *College Board*, *42*, 8-14.

Williamson, M. M. (2003). Validity of automated scoring: prologue for a continuing discussion of machine scoring student writing. *Journal of Writing Assessment*, *1*(2), 85-104. https://escholarship.org/uc/item/8nv3w3w8

CONSIDERING STUDENTS, TEACHERS, AND WRITING ASSESSMENT: VOLUME 1, TECHNICAL AND POLITICAL CONTEXTS

INTRODUCTION TO VOLUME 1, TECHNICAL AND POLITICAL CONTEXTS

Diane Kelly-Riley
University of Idaho

Ti Macklin
Boise State University

Carl Whithaus
University of California, Davis

Considering Students, Teachers, and Writing Assessment, Volumes 1 and 2 focus on the increasing importance of students' and teachers' lived experiences within the development and use of writing assessments. Together, the pieces in these volumes reflect upon how writing assessment research has contributed to five major themes: (1) technical psychometric issues, particularly reliability and validity; (2) politics and public policies around large scale writing assessments; (3) the evolution of—and debates around—automated scoring of writing; (4) the major theoretical changes elevating fairness within educational measurement and writing assessment; and (5) the importance of considering the lived experiences of the humans involved in the assessment ecology.

The Journal of Writing Assessment (*JWA*) has been a primary scholarly forum that has chronicled this evolution. These two volumes examine key themes from scholarship published in *JWA* in the past twenty years. Each section is introduced by current scholars in writing assessment who provide a retrospective for the issues of the past and these authors comment on the ways in which these issues continue to unfold. As such, they also represent generations of scholars in conversation with each other providing a model necessary as we continue to navigate the unfolding complexities of writing assessment situated in society. That is this field, in particular, benefits from revisiting issues and controversies of the past to see how our responses informed the practices of the present.

Volume 1 explores the dynamic issues connected to reliability and validity and how writing assessment contributed to the evolution of these concepts, the

DOI: https://doi.org/10.37514/PER-B.2024.2166.1.3

shifting political context of writing assessment and the rise of automated scoring of writing. Volume 2 explores the evolutions in theory and practice related to fairness and writing assessment and then the ways in which the people who teach and learn in these spaces shape writing assessment practice.

TECHNICAL EVOLUTIONS

The first volume focuses on technical and political issues. The rise of local considerations in writing assessment emerges most fully in the articles published in *JWA* during the first two decades of the twenty-first century. Rather than excluding the lived experiences of students and teachers, *JWA* has taken the lead in documenting how contextual, situated, and localized forms of writing assessment may provide fuller—more valid, reliable, and fairer—pictures of students' writing. The history of this move valuing localized forms of writing assessment has not been fully told. This movement reaches back to Edward White's (1978) early advocacy for direct assessment of students' writing rather than a reliance on indirect forms of writing assessment. It also echoes—perhaps even amplifies—Kathleen Yancey's (1999) and others' work (Calfee & Perfumo, 1996; Elbow & Belanoff, 1997; Hawisher & Selfe, 1997; Herman et al., 1993; Herter, 1991) on writing portfolios in the 1990s reflecting on students' emerging knowledge about writing, their writing processes, and their development as writers. Since its inception, *JWA* has published scholarship from the unique angle of how local contexts inform writing assessments.

Revisions to the *Standards for Educational and Psychological Testing* (AERA, APA & NCME, 2014) shifted discussions around the core educational measurement constructs of validity and reliability and drove changes in these scholarly areas. The *Standards* govern much of the thinking about standardized assessments, particularly true within the psychometric and educational measurement sides of writing assessment. The *Standards* is a living document open to revision, and the changes between the 4th and 5th editions in 1999 shifted discussions within writing assessment away from a singular focus on the importance of reliability to an understanding that validity is the most important consideration in writing assessment systems and is situated in particular contexts. The published discussions between Richard Haswell (1998) and Pamela Moss (1998) foreground how debates around the concept of validity assumed an increasingly important role in writing assessment. Once the focus of validity changed, teachers had a clearer role in determining and contributing to meaningful assessment. The increased emphasis on validity enabled teachers to push back against the limitations of standardized tests, opening up a new area of research that involved local contexts and faculty expertise. *JWA*'s establishment in 2003 provided a

venue for writing teachers and educational researchers to explore the implications of considering local contexts on writing assessments.

PROGRAMMATIC IMPLICATIONS

A frequently told origin story of writing assessment in North American postsecondary education points toward 1874 and the addition of an extemporaneous writing sample in the Harvard entrance examination. Norbert Elliot details (2005) how the Harvard exam was used to place students into its curricula. More than half of the students required remedial coursework and additional support setting up the tension between assessment and instruction. Elizabeth A. Wright, Suzanne Bordenlon, and S. Michael Halloran (2020), however, offer a corrective to this historicizing of writing assessment. In "'Available Means' of Rhetorical Instruction," they take up Royster and Kirsch's call to explore "the lessons taught to those students unable to attend those schools for elite white men" (p. 245). Wright, Bordenlon, and Halloran point out how late 19th-century rhetorical education and writing instruction took place in a wide variety of secondary and postsecondary educational contexts including Catholic institutions, women's colleges, historically Black universities and colleges, as well as within the often repressive contexts of boarding schools for indigenous children (pp. 254-257). Thus, there is a broader history of the structures and lasting impacts of writing assessment yet to be explored.

Across all of these contexts, writing placement mechanisms grew more profoundly as standardized tests became more widely available. Many of these placement exams attempted to capture students' readiness to enter postsecondary study, but the means of the exams often did not correspond to the curricular realities in the classrooms. Haswell (2004) notes that the 1900's "saw testing firms grow ever more influential and departments of English grow ever more divided between using ready made goods, running their own placement examinations, or foregoing placement altogether." (para. 3) Faculty in English departments devised their own assessment systems. The English Equivalency Exam (EEE) was used by the California State University and Colleges between 1973 and 1981; later, it was replaced by the English Placement Test in 1977 developed by Edward White and his colleagues. Haswell and Elliot (2017) observe "the few scholars and test administrators who were using holistic scoring were using all their energies to confront the problems of cost and scoring reliability, as practical aspects of the large testing programs they were supervising." (White, 1993, p. 82) The EEE went beyond that, as White (1984) himself declared in his essay, "Holisticism." The method of holistic scoring may have achieved some pragmatic ends making "the direct testing of writing practical and relatively

reliable" (White, 1984, p. 408), and it may have achieved some indirect social ends, bringing "together English teachers to talk about the goals of writing instruction" (p. 408), but beyond that "it embodies a concept of writing that is responsible in the widest sense . . ." (p. 408). It was responsible for its product, which was responsible for its advertised use.

This move toward localization continued in the late 1980s when an area of research emerged from the lived experiences of teachers and students in composition courses in response to accreditation and accountability mandates. Moore, O'Neill, and Crow (2016) detail this extensive history of compositionists "using assessment to improve student learning before it was emphasized so much by accreditors . . . [because they] understood the link between learning assessment and teaching improvement before accreditors made the connection explicit" (p. 20). Many of these teacher-researchers struggled with the day-to-day implications of the theoretical constructs of validity and reliability. As they grappled with these constructs in their contexts, new practices and research paths emerged. Early examples are detailed by Moore et al. (2016) demonstrating the field of composition's historical response to external assessment mandates. The first was Elbow and Belanoff's (1997) portfolio system which replaced a mandated university proficiency exam. Another system, developed and implemented at Washington State University, included an entry-level Writing Placement Exam and junior Writing Portfolio developed by Richard Haswell and his colleagues (2001). This program entwined formative writing assessment with disciplinarily situated writing instruction across the entire undergraduate curriculum. At all levels, writing teachers were involved in the assessments, and a comprehensive writing center provided support for students, including required small group sessions concurrently supporting students in upper-division disciplinary writing courses for those who did not pass the mid-career assessment (Haswell, 2001).

The core educational measurement constructs of validity and reliability continued to undergo major reconceptualization. In 1999, major revisions to the *Standards for Educational and Psychological Testing* were jointly authored by the American Psychological Association (APA), American Education Research Association (AERA), and the National Council of Measurement in Education (NCME)—professional organizations which guide and govern best practices in assessment and measurement. In this revision, validity was cast as the most important consideration above all. Now, tests or assessments were no longer considered stand-alone entities that needed to adhere to standards of technical qualities of reliability or validity. Instead, a major philosophical understanding of assessment shifted to see these measurements in social contexts in which the uses and interpretations of scores must be considered in each and every setting. This was a revolutionary shift.

The late 1990's also saw significant educational reform in the US with assessment playing a key role in these public and political arenas. During this time, writing studies teachers pushed back against the standardized test movement which attempted to represent and measure writing ability through knowledge of grammar and other writing rules (Bloom et al., 1996). In standardized testing, multiple choice test items were used as a way to measure the quality of students' writing. Writing teachers and researchers resisted these indirect, decontextualized forms of evaluating students' writing abilities. From their positions in the classroom, compositionists knew this evaluation did not serve the instructional needs of either students or faculty, and they advocated for locally-developed assessment measures attentive to classroom contexts and actual student learning outcomes. Thus, portfolio assessment developed out of the work of postsecondary writing teachers. This process is described in White et al.'s 1996 collection, *Assessment of Writing: Politics, Policies, Practices.* By the mid 90s, ways of measuring the construct of writing became more nuanced. Compositionists realized that writing is socially situated and began to publish research findings supporting this position. Understanding the people who designed and participated in the assessments and the multiple ways in which they were enacted across different institutional sites became a key component of writing assessment.

The increased accountability context within educational settings in North America resulted in innovative programmatic responses. The Council of Writing Program Administrators (CWPA) started to discuss and collaborate on whether "a pithy and effective list of objectives for writing [and] programs existed" (as cited in Harrington et al., 2003, p. xv). These conversations among members at all levels of expertise were enabled by many compositionists joining the then newly created WPA-L email listserv in the late 1990s. The members of this group recognized the multiple stakeholders who were invested in the outcomes of first-year composition. This exigence resulted in the development of the *WPA Outcomes Statement for First-Year Composition* (*WPA OS*) "a statement . . . plain enough to speak to those outside the discipline, yet rooted in disciplinary language enough to have status in the field" (Harrington et al., 2003, p. xvi). The WPA OS is a consensus document detailing the expectations for first-year writing common to most postsecondary institutions in North America (Harrington et al., 2001). Kathleen Yancey (2003) says that the *WPA OS* was intentionally written as outcomes and not standards for performance that needed to be achieved.

By framing and modeling curricular and assessment work as driven by faculty and local contexts, the collaborators of the *WPA OS* also began to formalize a new area of research. This new area of local programmatic response to assessment had several offshoots as contextually situated responses to assessment

and accountability mandates. In *Reclaiming Assessment: A Better Alternative to the Accountability Agenda,* Chris Gallagher (2007) describes his locally focused efforts with colleagues across K-12 programs in Nebraska. Gallagher and his collaborators argue that accreditation programs developed by teachers with students and learning in mind result in the best programs. Others like Christine Farris (2014) from Indiana University led the Writing and Reading Alignment Project which intended "to help teachers examine their current instructional practices and goals for student learning and develop new strategies to promote skills in critical reading, evidence-based writing and discussion as expected in college-level coursework." (Indiana University, 2014)

Wendy Sharer and her colleagues (2016) describe their efforts at Eastern Carolina University to reclaim accountability and assessment for postsecondary settings. In their edited collection, *Reclaiming Accountability: Improving Writing Programs through Accreditation and Large-Scale Assessments*, they provide models responding to the call of a 2007 WPA Executive Board letter that proclaimed "those who teach writing and those who administer writing programs need to be involved in defining the terms and setting the parameters of large-scale writing assessment so that any changes implemented in response to assessment are in keeping with what research and practice have demonstrated to be truly effective in helping student writers" (p. 3). In Behm's edited collection (2013), *The WPA Outcomes Statement: A Decade Later,* the effect of this situated research agenda is apparent. Topics in the book cover personal identity, its application to writing across the curriculum and disciplines, extensions into global settings, use with second language approaches, and impacts on technology. Much of the research on writing assessment explored its connections to instruction.

THE PROMISE AND PERILS OF TECHNOLOGY

Beginning with No Child Left Behind (NCLB) in the early 2000s, large scale assessment and accountability efforts moved testing and accountability to the center of educational policy and practice. As a result, the challenge of testing hundreds and thousands of students across institutions became a reality. Initially, technology was seen as the remedy to manage such a large-scale endeavor. But this effort was hampered by limitations within the technologies to reliably and validly evaluate a significant amount of writing as a socially situated, complex construct. Later these educational reform efforts morphed into the Common Core State Standards Initiative (CCSSI) further elevating the role for writing throughout the K-12 curriculum.

CCSSI marked a new period in the assessment landscape in which educational reformers, largely nonprofit and philanthropic organizations like the

Gates and Lumina foundations, had strong legislative support to reshape American education into one focused on the preparation of workers to advance the American economy. To measure their progress, these efforts partnered with testing companies like Pearson and ETS. With complex assessments being implemented on such a large scale, the possibilities of machine or computer scoring of writing ascended to the forefront. The challenges for assessing learning through writing remained and were amplified in these large-scale assessments.

These emerging curricular efforts recognized the importance of teaching writing as situated within disciplinary genres from the beginning of school. As accountability efforts moved from NCLB to the Smarter Balanced Assessment Consortium (SBAC) and Partnership for Assessment of Readiness for College and Careers (PARCC) national assessments, the initial response was to use technology, particularly the potentials for automated essay scoring, to support the integration of writing across the K-12 curriculum. These large-scale assessments have meant that writing assessment has taken a much more central role in accountability efforts. The challenge remains to develop computer-based scoring that represents the complexity of writing taught and assessed in the classroom.

PART ONE. TECHNICAL ISSUES IN THE ASSESSMENT OF WRITING: RELIABILITY AND VALIDITY

Part One of this collection focuses on technical issues in the assessment of writing, particularly reliability and validity as published in the *Journal of Writing Assessment*. Raters' approaches to texts are one of the most vexing issues in writing assessment. Controlling for individual raters' idiosyncrasies is one of the longest running issues in writing assessment reaching back to Paul Diederich's work at ETS in the 1960s. Writing assessment researchers' work on reliability and validity has taken many forms. In his piece, David H. Slomp, editor of *Assessing Writing*, contextualizes and responds to these changes in the technical constructs in writing assessment. Slomp's response focuses on the ways in which *JWA*'s legacy bridges the gap between educational measurement and writing studies. He explores the implications for research and practice that emerge from dialogues between these two fields. Slomp frames and responds to the following key articles from *JWA*.

Peggy O'Neill's "Reframing Reliability for Writing Assessment" (2011) shifts away from traditional discussions about inter-rater reliability as the ultimate goal–the single most important form of reliability within writing assessment at the time. She argues that both writing studies and psychometrics offer multiple forms of reliability that need to be attended to in the building of writing assessment systems and in research about writing assessment. Drawing on Lakoff's

(2002, 2004, 2006) work, O'Neill suggests that by moving discussions of reliability in writing assessment beyond inter-rater reliability, more nuanced, and more accurate, forms of writing assessment can be developed. These emerging forms of writing assessment might not only acknowledge but also account for—in a psychometrically rigorous way—variations across readers and variations across tests in the ways that Pamela Moss (1998) and Richard Haswell (1998) recognize as hermeneutic or rhetorical practices.

Diane Kelly-Riley's "Validity Inquiry of Race and Shared Evaluation Practices in a Large-Scale, University-wide Writing Portfolio Assessment" (2011) advances the field's understanding not only of the balance between reliability and validity but also brings into the conversation vital contextual elements involving race and racism. Her article takes on the question of race—and in more subtle ways racism—by looking at the implementation of a locally-developed, context-rich writing portfolio assessment system. Kelly-Riley's article is a precursor to the consideration of fairness and antiracist practices in writing assessment by providing an empirical study that looked at how raters understand and apply race in an assessment context. Writing assessment has struggled to develop an operational definition of race. Race is often defined by government agencies that collect data on race, but the experience in the writing classroom calls for more nuanced representations of race.

O'Neill and Kelly-Riley's work lead toward approaches outlined in Elliot et al.'s "Three Interpretative Frameworks: Assessment of English Language Arts-Writing in the Common Core State Standards Initiative" (2015). Elliot, Rupp, and Williamson examine how standards-based definitions of validity, reliability/precision, and fairness were integrated into the Smarter Balanced Assessment Consortium (SBAC) and Partnership for Assessment of Readiness for College and Career (PARCC) English Language Arts – writing assessments. They encourage stakeholders to be informed consumers when interpreting and using SBAC or PARCC scores about students' writing. Their work foreshadows a move within writing assessment research and practice encouraging stakeholders (WPAs, students, teachers, and parents) to not just accept the scores from large-scale state or national-level writing assessments at face value but to integrate how they will be used, to examine their meaning and their use value.

PART TWO. POLITICS AND PUBLIC POLICY OF LARGE-SCALE WRITING ASSESSMENT

Part Two explores the political dimensions of writing assessment. In her contextualization of this section, Carolyn Calhoon-Dillahunt, Yakima Valley College, Past Chair of the Conference on College Composition and Communication

and Past President of the Two-Year College Association, synthesizes these major educational reform movements and how they impact writing assessment scholarship and practices. This section highlights work by Edward M. White, Arthur N. Applebee, Hammond and Garcia, and Toth et al. All of these authors anticipate and wrestle with large scale writing assessment in terms of political and policy issues. Political changes across educational reform movements have both shaped and responded to assessment issues. The critiques of both placement in two-year colleges and of AES have centered around how students' writing must be considered and evaluated as contextual, rather than stripped of context for a placement decision afforded by the cost savings of having software evaluate a piece of writing.

Edward M. White's "The Misuse of Writing Assessment for Political Purposes" (2005) and Arthur N. Applebee's "Issues in Large-Scale Writing Assessment: Perspectives from the National Assessment of Educational Progress" (2007) set the stage for early political discussions around writing assessment. White argues that many large-scale writing assessments are motivated "by political rather than educational, administrative, [or] professional concerns." For White, No Child Left Behind and its reliance on testing "without the resources and leadership for students to achieve the skills they will be tested on" is a crucially flawed educational policy and a misuse of writing assessments based on politicians' misunderstanding of what educational testing can tell us. He considers a wide range of mandated, large-scale writing assessments ranging from required state-level testing of secondary students through placement exams for incoming college students to graduation requirements for college students. He suggests that the misuses of writing assessments "[are derived] from an exaggerated, even a credulous misunderstanding, of what particular kinds of assessments can accomplish." Such observations continue to underscore the misuse of assessments in educational settings.

In contrast to White's critique of assessment as gatekeeping, Applebee's "Issues in Large-Scale Writing Assessment: Perspectives from the National Assessment of Educational Progress" focuses on the contributions of the large-scale, national-level programmatic assessment conducted through the National Assessment of Educational Progress (NAEP). Applebee suggests the NAEP writing assessment is valuable because it is not tied to the assessment of individual students, but rather a way of looking at how students' writing is developing and comparing achievements in writing across states. White's and Applebee's works are both polemic, but research like J. W. Hammond and Merideth Garcia's show the legacy of informed and principled approaches documenting the effects of large-scale assessment on teachers.

Writing assessment, politics, and public policies in the first two decades of the twenty-first century requires that we address the effects of No Child Left

Behind (NCLB) and the Common Core State Standards (CCSS). Hammond and Garcia's "The Micropolitics of Pathways: Teacher Education, Writing Assessment, and the Common Core" (2017) and Toth et al.'s " Writing Assessment, Placement, and the Two-Year College" (2019) describe the impacts of these initiatives on teachers and students. Hammond and Garcia take the Common Core State Standards (CCSS) as their point of departure. They examine how teacher education programs frame the CCSS for their teachers-in-training. Their works suggest postsecondary faculty, teachers, and teachers-in-training "micropolitically interpret" the Common Core. In fact, Hammond and Garcia suggest that writing teachers and writing teachers-in-training foreground their own local writing assessments since teachers seem most focused on curriculum and instruction issues and secondarily on CCSS and pathway-related reforms to education. One of the key findings from Hammond and Garcia's work is the value of adopting a micropolitical perspective when considering writing curricula, instruction, and assessment.

The emphasis on learning pathways was championed in the educational reforms promoted through CCSS. As such, pathway-based reforms had a dramatic effect on community colleges, Toth et al.'s introduction to the *Journal of Writing Assessment's* Special Issue on Placement and Two-year Colleges takes up the overlapping issues of educational reform and how writing assessments have been used in placement decisions. In their ambitious and wide-ranging article, Toth, Nastal, Hassel, and Giordano review the history of two-year colleges within American higher education. They attend to the ways in which this history and pathway-based educational reform movement intersects with models for assessing and placing students within ESL, basic writing, or first-year composition courses. They then extend their discussion by turning to questions around the validity and the uses for writing assessment and placement systems. Toth et al.'s attention to sociocultural factors highlights the ways in which questions of writing assessment are being looked at at a systems level rather than only at the level of individual students.

PART THREE. IMPLICATIONS OF AUTOMATED SCORING OF WRITING

In Part Three, key pieces published in *JWA* explore possibilities and pitfalls with technology and writing assessment. Large-scale assessments became more commonplace as the accountability movement gained traction in public educational settings. During the late 1990s, No Child Left Behind was implemented across K-12 public school systems and the challenge of assessing each and every student became a reality. As accountability systems evolved, partnerships between

testing companies, educational reform nonprofit organizations with strong legislative support, and textbook publishing companies evolved into new initiatives connected to career and college readiness and capitalized upon the economic investment in public education. A focus emerged on secondary and post-secondary education to prepare students in economic terms. This resulted in more complex curricula—such as the Common Core State Standards Initiative—which emphasized students' readiness for workplace or college challenges. The backbone of this curriculum was writing—where it became embedded across multiple disciplines across grade levels. Writing was a primary means to demonstrate and assess student proficiency across disciplinary areas. To meet the challenge of assessing student performance across the country, test developers and researchers turned to automated scoring of student writing. Assessing the construct of writing when it is socially situated presents new challenges difficult for technology to address alone. This more robust, and socially situated construct of writing better represents what occurs in classroom settings, but it also requires the development of writing assessment systems that connect human readers and writing technologies.

In her introduction to Part Three, Laura Aull, Associate Professor and Writing Program Director at the University of Michigan, responds to the major developments in Automated Essay Scoring (AES) and contextualizes the issues in relation to key publications from *JWA* from the past twenty years. She discusses the major possibilities and limitations of these technologies, particularly as they relate to the ongoing implications of socially situating writing assessment. Her response highlights the intersections of Artificial Intelligence and AES in education and measurement, learning analytics, and user-centered design. Aull considers the impacts of these emerging writing assessment technologies on educational equity. As this book was going to press, ChatGPT had recently emerged and reignited the importance of this scholarship to the writing assessment community.

In "Validity of Automated Scoring: Prologue for a Continuing Discussion of Machine Scoring Student Writing", Michael Williamson (2003) lays out the tension between the field of writing studies and educational measurement as automated scoring of writing took hold. Williamson encourages writing studies scholars and practitioners to learn the language of educational measurement in order to weigh in on these evolving conversations which would inevitably bend toward a socially situated context because of the consideration of the use of test results and their impacts on test takers. Williamson notes that automated scoring of writing would likely become prevalent given the millions of pieces of writing that required evaluation in the large-scale assessment systems.

Next, Les Perelman's critique of AES software and its uses in assessing writing challenge the widespread adoption of this technology. His work points to

the important distinction that writing is a complex, socially situated activity and the necessary reductions that must occur to the construct of writing when it is assessed by computer software and algorithms. In "Critique of Mark D. Shermis & Ben Hamner, 'Contrasting State-of-the-Art Automated Scoring of Essays: Analysis'," Perelman (2011) notes that Shermis and Hamner reported high reliability between human and machine readings of student work, but Perelman argues that the samples are very short and were written in response to literary analysis or reading comprehension passages. Such writing samples are hardly representative of complex writing situated in context, genre, and social circumstance.

Finally, Jordan Canzonetta and Vani Kannan explore these writing assessment technologies in "Globalizing Plagiarism & Writing Assessment: A Case Study of Turnitin" (2016). This piece highlights the additional uses of automated essay scoring and its integration within learning management systems for plagiarism detection or online writing support. In order to be reliable, automated essay reading and scoring systems must operate with narrowly defined constructs of writing. Canzonetta and Kannan explore the implications of importing the US construct of writing into other cultures. In their view, the global reach of Turnitin privileges western academic writing and stigmatizes nonwestern writing. As a result, the plagiarism software reinforces western values about authorship not necessarily representative in other places. As Artificial Intelligence and systems like ChatGPT become more prevalent, it's important for writing assessment researchers and scholars to document and understand the possibilities and limitations within them.

REFERENCES

American Educational Research Association, American Psychological Association, & National Council of Measurement in Education. (2014). *Standards for educational and psychological testing.* American Educational Research Association.

Behm, N. N., Glau, G. R., Holdstein, D. H., Roen, D., & White, E. M. (Eds.). (2013). T*he WPA Outcomes Statement: A decade later*. Parlor Press.

Calfee, R. C., & Perfumo, P. (Eds.). (1997). *Writing portfolios in the classroom: Policy and practice, promise and peril*. Routledge.

Elbow, P. & Belanoff, P. (1997). Reflections on an explosion: Portfolios in the '90s and beyond. In K. B. Yancey & I. Weiser (Eds.), *Situating portfolios: Four perspectives* (pp. 21-33). Utah State University Press.

Elliot, N. (2005). *On a Scale: A social history of writing assessment in America*. Peter Lang.

Gallagher, C. (2007). *Reclaiming assessment: A better alternative to the accountability agenda*. Heinemann.

Harrington, S., Rhodes, K., Fischer, R. O., & Malenczyk, R. (2003). Introduction: Celebrating and complicating the Outcomes Statement. In S. Harrington, K. Rhodes, R. O. Fischer, & R. Malenczyk (Eds.), *The outcomes book: Debate and consensus after the WPA Outcomes Statement* (pp. xv-xix). Utah State University Press. https://doi.org/10.2307/j.ctt46nwgv

Harrington, S., Malenczyk, R., Peckham, I., Rhodes, K., & Yancey, K. B. (2001). WPA outcomes statement for first-year composition. *College English, 63*(3), 321-325.

Haswell, R. H. (Ed.). (2001). *Beyond outcomes: Assessment and instruction within a university writing program*. Ablex.

Haswell, R. H. (1998). Multiple inquiry in the validation of writing tests. *Assessing Writing, 5*(1), 89-110.

Haswell, R. H. (2004). Post-secondary entry writing placement: A brief synopsis, CompPile.org. https://wac.colostate.edu/docs/comppile/pd/writingplacementresearch.htm

Haswell, R. H. & Elliot, N. (2017). Innovation and the California State University and Colleges English Equivalency Examination, 1973-1981: An Organizational Perspective, *Journal of Writing Assessment*, *10*(1), https://escholarship.org/uc/item/7rt5v9p2

Hawisher, G. E., & Selfe, C. L. (1997). Wedding the technologies of writing portfolios and computers. In K. B. Yancey & I. Weiser (Eds.), *Situating portfolios: Four perspectives* (pp. 305-321). Utah State University Press.

Herman, J. L., Gearhart, M., & Baker, E. L. (1993). Assessing writing portfolios: Issues in the validity and meaning of scores. *Educational Assessment*, *1*(3), 201-224.

Herter, R. J. (1991). Writing portfolios: Alternatives to testing. *English Journal*, *80*(1), 90.

Indiana University. (2014). IU to collaborate with high school teachers on Writing and Reading Alignment Project. https://education.indiana.edu/news/2014/2014-06-11-01.html

Lakoff, G. (2002). *Moral politics: How liberals and conservatives think* (2nd ed.). University of Chicago Press.

Lakoff, G. (2004). *Don't think of an elephant! Know your values and frame the debate*. Chelsea Publishing.

Lakoff, G. (2006). Simple framing. *Rockridge Institute*. http://www.rockridgeinstitute.org/projects/strategic/simple_framing

Moore, C., O'Neill, P., & Crow, A. (2016). Assessing for learning in an age of comparability, In W. Sharer, T. A. Morse, M. F. Eble, & W. B. Banks (Eds.), *Reclaiming accountability: Improving writing programs through accreditation and large-scale assessments* (pp. 17-35). Utah State University Press.

Moss, P. (1998). Testing the test of the test: A response to "Multiple Inquiry in the Validation of Writing Tests" [by Richard H. Haswell]. *Assessing Writing 5(1)*, 111-122.

Sharer, W., Morse, T. A., Eble, M. F., & Banks, W. P. (2016). Introduction: Accreditation and assessment as opportunity. In W. Sharer, T. A., Morse, M. F. Eble, & W. B. Banks (Eds.), *Reclaiming accountability: Improving writing programs through accreditation and large-scale assessments* (pp. 3-13). Utah State University Press.

White, E. M., Lutz, W. D., & Kamusikiri, S. (1996). *Assessment of writing: Politics, policies, practice*s. Modern Language Association.

White, E. M. (1978). Mass testing of individual writing: The California model. *Journal of Basic Writing, 1*(4), 18-38. https://doi.org/10.37514/JBW-J.1978.1.4.03

Wright, E. A., Bordelon, S., & Halloran, M. (2020). "Available means of rhetorical instruction": "Broadening perspectives" on rhetorical education prior to 1900. In J. J. Murphy & C. Thaiss, (Eds.), *A short history of writing instruction: From ancient Greece to the modern United States* (pp. 244-271). Routledge.

Yancey, K. B. (1999). Looking back as we look forward: Historicizing writing assessment. *College Composition and Communication, 50*(3), 483-503.

Yancey, K. B. (2003). Standards, outcomes, and all that jazz. In S. Harrington, K. Rhodes, R. O. Fischer, & R. Malenczyk (Eds.), *Debate and consensus after the WPA Outcomes Statement* (pp. 18-23). Utah State University Press.

PART 1.

TECHNICAL ISSUES IN THE ASSESSMENT OF WRITING: RELIABILITY AND VALIDITY

RETROSPECTIVE.

FROM ISOLATION TO INTEGRATION: TECHNICAL ISSUES IN THE ASSESSMENT OF WRITING

David H. Slomp
University of Lethbridge

I approach this commentary as an editor of the other major writing assessment journal: *Assessing Writing*. *The Journal of Writing Assessment* and *Assessing Writing* share similar geneses having both been founded by Brian Huot and Kathi Yancey. In their first editorial of *JWA* Huot and Yancey (1993) explain the unfortunate circumstances that led to their founding a second, independent journal for the field (See their introduction to the first issue of JWA). Despite these circumstances, it has been fortunate for the field that we have two rigorous and highly respected journals dedicated specifically to the scholarship on writing assessment.

While both journals began with a primary focus on the assessment of writing in North America, under the editorship of Liz Hamp-Lyons, *Assessing Writing* developed a more international focus. With that internationalization came an increase in attention to the assessment of writing in second or additional language contexts. *The Journal of Writing Assessment,* however, maintained its emphasis on writing assessment in the North American context with a focus on program assessment, historical perspectives on assessment, assessment theory, and educational measurement. Both journals have recently celebrated milestone events: *Assessing Writing* commemorated its 25th anniversary in 2019, while the *Journal of Writing Assessment* marks its 20th anniversary in 2023. These milestone events signal a maturation of our field that in itself should be celebrated. At the same time, these events provide an opportunity for critical reflection on the programs of research that have emerged and developed in our field over the past two and half decades.

As an editor of one of these two journals, I have the privilege of a front row seat to the enormous diversity, complexity, and richness of our field. My experience in part informs the perspective that I bring to this commentary on the

DOI: https://doi.org/10.37514/PER-B.2024.2166.2.21

interplay between the scholarship on writing assessment from within and across the measurement and writing studies communities. In this commentary, I will focus on the ways in which *JWA*'s legacy bridges the gap between educational measurement and writing studies in three selected articles, and I will also explore the implications for research and practice that emerge from dialogues between these two fields. I begin, though, with an exploration of several tropes that have shaped our thinking about the interplay between educational measurement and writing studies communities.

FROM OPPOSITION TO COLLABORATION

When *Assessing Writing* was first published 25 years ago, the writing studies and educational measurement fields were constructed as being in conflict with one another. It was observed that the field of measurement approached the challenge of assessing writing with a different set of goals, perspectives, and values than that of composition and rhetoric. It was also suggested, that working in concert with political, policy, and educational leaders, measurement specialists imposed these values and goals on writing programs and educational systems with minimal concern for the consequences this was having on how writing was taught in schools. At the same time, compositionists and rhetoricians—writing studies specialists—were framed as those who were close to the consequences of these assessment systems, who saw their impact on students, colleagues, and the discipline as a whole and who worked to ensure assessment systems were designed to support student learning. In 2003, the first issue of the *Journal of Writing Assessment* carried this framing forward applying it to proxies for the measurement community—state departments of education—unfavorably contrasting externally mandated and imposed assessment programs against locally developed assessments (Huot & Yancey, 1993).

Ongoing research continued to reinforce this trope. Adler-Kassner & O'Neil's (2010) *Reframing Writing Assessment to Improve Teaching and Learning*, for example, argued that writing studies specialists needed to replace the measurement-based theoretical framing that has structured writing assessment research with more generative frames. In 2016, Broad argued that structured ethical blindness prevented measurement experts from understanding the harm their work is causing. He observed:

> [M]ass-marketers of standardized tests should not be blamed for failing to see the harms their products do, because the structures of human psychology, society, and economy prevent and prohibit such self-critical vision. This is the meaning

> of *structured ethical blindness*: not that people should be condemned for failing to see the damage they do, but rather that the rest of society must take on responsibility for handling those harms precisely because most good people meaning to do well cannot squarely face the harms they inadvertently bring about. (para. 23)

Broad's insight is in some manner also a critique of this early framing of the field.

Broad's work nods to a second trope that has been emerging over the past decade. Already in 2003, Huot's second *JWA* editorial highlighted the need for multidisciplinary framing and collaboration; he makes the point that "a writing assessment literature that is current and relevant to new issues and challenges while at the same time sophisticated in its treatment of theories and principles in both measurement and language education is a future goal and not a current reality" (p 82). Behizadeh and Engelhard's (2011) review of the integration of discourses from within and across the writing and measurement communities makes clear that this future goal remains a work in progress. My own more recent review of 25 years of scholarship published in *Assessing Writing* (2019) revealed that scholarship in the field remains rather siloed (by discipline, geographic location, and linguistic context). A similar analysis by Zheng and Yu (2019) showed that in *Assessing Writing*, this siloed nature extends to the theoretical frameworks that shape the papers published in the journal. Between 2000-2009, 67% of papers were framed through a writing studies lens, while 34% were framed through a measurement lens. This distribution shifts so that between 2010-2018, 58% of papers were framed with a writing studies lens, while 40% were framed with a measurement lens. While achieving the goal Huot envisioned remains a work in progress, it is fair to say that the disconnects of the past have lessened, creating bridges for new innovations that will shape writing assessment in the future.

In 2012, Elliot and Perelman's edited book, *Writing Assessment in the 21st Century: Essays in Honor of Edward M White,* called for the tensions of the past to give way to a spirit of multidisciplinary collaboration. Rather than casting the tension between these two founding disciplines as reason for division, Elliot and Perelman pointed to the generative potential this tension gives rise to. They identified four commonalities shared between educational measurement and writing assessment communities that can help drive forward a shared, collaborative, multidisciplinary research agenda. These include:

1. Developing theory and identifying the practical application of those theories to educational systems and settings;

2. Advancing the art and science of construct modeling;
3. Attending to assessment design that is principled, critical, and focused on the promotion of opportunity;
4. Identifying consequences of assessment design and use so that negative consequences can be mitigated and positive consequences can be promoted.

Underlying this vision is a third trope: The work we do as writing assessment scholars and professionals is inherently consequential. Given the ubiquity of writing assessments at all levels of educational systems in all corners of the world, millions of people are impacted every year by the inferences and decisions that are made about them, based on their performance on the writing assessments they have participated in.

Though founded on the first trope, the evolving story of *JWA* has been its contribution to the second and third: a generative focus on multidisciplinarity driven by an ethic of responsibility for the consequences of assessment design, implementation, and use. The three articles from the archives of *JWA* that are presented in this section demonstrate that evolution.

REFRAMING RELIABILITY AS A CATEGORY OF EVIDENCE

Huot (2002) highlights the technocentric foundations of writing assessment practices, grounded largely in the search for reliability. He observed that framed within a technocentric mindset, writing assessment focused on technical rather than humanistic solutions to the key challenges the discipline faced. Reliability, therefore was cast as a technical problem in search of technical solutions. O'Neill's (2011), "Reframing Reliability for Writing Assessment," calls for more diverse and integrated approaches to addressing issues of reliability and validity. Drawing on the work of Moss (1994) and Parkes (2007) she argues that rather than focusing on the statistical methods for operationalizing reliability, writing assessment developers and users need to focus on the values of accuracy, dependability, stability, consistency, and precision that these measures are meant to represent. By focusing on the values rather than on statistical measures that stand in as proxies for those values, O'Neill argues that we can develop new methods for creating valid and reliable assessments.

O'Neill's treatment of reliability is situated within the tension between foundational epistemologies associated with measurement and writing studies. She calls for a pragmatic approach to navigating these tensions. Writing Assessment researchers need to understand how core measurement principles are framed and operationalized within a psychometric tradition. At the same time, she suggests,

we need to attend to the values that underpin the field of writing studies. Her argument echoes Onwuegbuzie and Leech (2005) who call for training future researchers within a pragmatist tradition so that they are capable of navigating both positivist and interpretivist models of research, drawing on and adapting methods from within both traditions as the research warrants. O'Neill sums up her position:

> In determining reliability, many of us responsible for writing assessments should collaborate as equal partners with colleagues who have the statistical expertise. Writing assessment practitioners and scholars need to accept our responsibility to develop and maintain writing assessments that are informed by both language-based and psychometric theory and research. We need to develop new methods for assessment as well as for determining reliability and validity if current methods do not work adequately for our purposes, as Parkes (2007) argued. This may mean collaborating with others who have different kinds of experiences and expertise, learning more about psychometric theory and practices, and engaging in difficult discussions with colleagues about what we value and why it matters. (pp. 59-60)

She further argues that by focusing on our values, by continually bringing these into the conversations about assessment design, appraisal, and use, we can help to reframe reliability so that our pursuit of the values of accuracy, dependability, stability, consistency, and precision in writing assessment can be engineered to serve our students and our programs.

There is certainly evidence within the field of writing assessment to support her claims. In North America, for writing assessment at the post-secondary level, the response to this call has been evidenced in the uptake of communal writing assessment (Broad et al., 2009; Lindhardsen, 2020), community grading (Shumake & Shah, 2017), contract grading (Litterio, 2016), and comparative judgment (Sims et al., 2020) models of scoring: processes that rely on rigorous discussion and documentation to demonstrate commitment to accuracy, dependability, stability, consistency, and precision.

The broader value of O'Neill's article is that it continues a tradition of arguing for the role that composition studies can and should play in shaping the discourses and practices surrounding writing assessment. Writing in *Education Measurement: Issues and Practice*, Newton (2017), draws on the field of writing assessment—and indirectly on the scholarship in the field's two major journals—to make the point that the measurement community needs to engage

these voices, "treating assessment design as a process of negotiation between complementary, and sometimes contradictory, perspectives" (p 13). He warns the measurement community that an ongoing failure to engage communities such as ours will lead to the design of writing assessments that are "suboptimal for the systems within which they need to function, even when seemingly optimal from a measurement perspective" (p 13).

Within a measurement perspective, Brennan (2001) made similar observations about the limitations of reliability in writing assessment. He raised concerns about the often superficial treatment of reliability arguments, in how particular the move from the "more-or-less assumption-free procedures for estimating reliability (e.g., alternate forms) to assumption laden procedures" (p. 313) often fail to ensure that the procedures chosen to estimate reliability are in fact consistent with the claims being made about the assessment. Data related to internal consistency, for example, do not support claims related to consistency over time and multiple iterations of an assessment. He observes, there are "as many reliabilities as there are specifications of universe of generalization that one or more investigators is (are) willing to assert as meaningful for some purpose" (p 301).

He illustrates this concern, noting that facets related to tasks, rubrics, training procedures, and occasions are often not accounted for in constructing a reliability argument for performance assessments (such as timed, impromptu essay exams). He observes that a score received on a performance task is derived from two sources: the work produced by the examinee, and the score assigned by the rater. Inter-rater reliability, the facet most attended to in writing assessment design and use, only accounts for one of these two sources—consistency of raters—but not the other. Therefore, he notes, such scores only enable test users to make claims about raters, but not about examinees. He complicates the quality of even those claims, however, observing that it is typical for most performance assessments to use single rubrics and to train raters using only one training protocol. He points out that this limits test users' capacity to observe the impact of rubric design and training procedures as sources of variability in scoring.

Elliot et al.'s (2016) study of ePortfolio scoring published in *JWA*, illustrated Brennan's point. In their study the authors carefully explained what data they collected and what limitations it posed for interpretations of e-portfolio scores. They further observed:

> Levels of inter-rater consensus and consistency evidence presented . . . reveal that standard gauge reliability guidelines are of little use in interpreting ePortfolio scores. If scores from complex writing assessments are to be interpreted and infor-

> mation from them used, then researchers are best served by calling into question the 0.7 correlation coefficient established by writing tasks associated with standardized testing. . . . High rates of inter-rater reliability are of little value if the construct representation is, as Kane (2006) has written, a "very narrow slice of the target domain of literacy." (p. 102)

Further, linking the relationship between validity and reliability to the issue of fairness, they note that in their study, the low degree of reliability in female students' scores for writing processes, and in Hispanic students' scores on rhetorical knowledge, knowledge of conventions, and composing in electronic environments, inferences about writing ability for these populations, on these aspects of writing should not be made. Their study beautifully illustrates how the shift from a technocentric to humanistic orientation toward assessment design and use enables thoughtful consideration of how reliability concerns can productively shape and inform validity arguments.

SITUATING RELIABILITY EVIDENCE WITHIN AN INTEGRATED VALIDATION MODEL

O'Neill's call for a more contextual view of reliability, and a more integrated approach to reliability and validity was already being addressed within the measurement community and broader language assessment communities. In particular, new models of validation such as Kane's (2006, 2013) Interpretive and Use Argument (IUA) model and Bachman and Palmer's (2010) Assessment Use Argument (AUA) began to treat reliability, not as a separate consideration from validity, but rather as an embedded element of a broader validity argument. Within these models of validation, an assessment's validity argument consists of a series of inferences or claims that must be tested and sustained. Kane describes these as scoring, generalization, extrapolation, and decision inferences, while Bachman and Palmer describe these as claims regarding consequences, decisions, interpretations, and assessment records. Though differences in the structure of the validity argument can be found across these two models, both embed concern for reliability within these broader sets of claims and inferences. On the one hand, within this formulation, reliability maintains a place of primacy: sustaining the scoring inference (assessment records) requires evidence of consistency and dependability of scoring procedures. On the other hand, this formulation balances concern for reliability against concern for validity: the scoring inference cannot be sustained if scoring criteria suffer from either construct irrelevant variance or construct under-representation. Both the AUA and IUA validation

models make explicit the link between validity, reliability, and the consequences of assessment design and use.

Kelly-Riley's (2011) study "Validity inquiry of race and shared evaluation practices in a large-scale, university-wide writing portfolio assessment" demonstrates how within an argument-based framework concerns for validity, reliability, and ultimately fairness can be motivated by the revealed consequences of an assessment's use. She examines a portfolio-based assessment program that had been in use for 20 years at an American university. The purpose of the assessment was to identify students who needed additional support in their upper-divisional writing requirements. When an African American student questioned the pass rates of BIPOC students compared to those of white students, unexamined questions related to fairness, reliability, and validity were brought into focus. In response to the student's question, Kelly-Riley's study examined how the assessment program in question might be unwittingly disadvantaging students of color.

Drawing on Kane's (2006) model of validation, Kelly-Riley links concerns for consequences with questions of construct representation and issues of score stability across populations of test-takers. While rightly cast as a validity study, this paper examines the scoring inference—a reliability issue.

Her investigation revealed that this assessment program was in fact designating students of color as "needs work" more frequently than it did white students who were more likely to receive a "pass" score. Analyzing the influence on student scores of race, perceived demographic profiles of students, and scoring criteria, she found, "race did not contribute to faculty raters' functional definition of 'good writing' for any of the frameworks whether in the timed exam format or for course papers" (p. 80). Instead, she found that "coherence, focus, and correctness all contribute significantly to the functional definition of "good writing" (p. 83). Additionally, she found that "large percentages of the variance of writing quality are accounted for through the Demographic framework—primarily through the rater's perception of the writer's intelligence and comfort with writing" (p. 84). At the same time, however, there remain statistically significant differences in performance by race on this assessment.

Kelly-Riley's (2011) study demonstrates the value of localism in writing assessment. As an administrator and professor she is well positioned to see firsthand the impact of the assessment program she is investigating on the students who walk through her door. In fact, it is the very questions and concerns raised by students who were made different by the assessment, that prompted the focus of her research. The power of Kelly-Riley's study is that it leverages contemporary validation frameworks to address these concerns and to advance local values of equity and opportunity. She positions her work in response to O'Neill's

(2003) observation that validation research on writing assessment from within the composition community tended to lack rigor and structure. This state of affairs reduces the effectiveness of this body of work in promoting change and in demonstrating the value of innovations in assessment design and use that have emerged from the field of writing studies.

Published three years before the most recent *Standards for Educational and Psychological Testing* (AERA, APA & NCME, 2014), Kelly-Riley's study foreshadows the elevation of fairness—within the measurement community—to a position of primacy in the design and appraisal of assessment programs. Her study demonstrates how fairness is essentially the application of validity and reliability to the testing of inferences and decisions about key populations of examinees. This shift adds an important new category of evidence to the work of appraising assessment programs. As our classrooms become increasingly diverse, this category of evidence becomes increasingly important.

Kelly-Riley's study also points to the limitations of contemporary validity theory, especially with respect to issues of race and fairness. Kelly-Riley struggles to reconcile the finding that race did not contribute to functional definitions of good writing with the finding that there were statistically significant differences in performance on this assessment by different racialized groups. The findings seem incompatible. Traditionally, within the measurement community, such disparities, if evidenced, were explained with respect to opportunity to learn. Randall (2021) explains that historically Opportunity to Learn was used to hide or explain away the racism embedded in an assessment program. By pointing to factors outside of the assessment itself, disparities in performance by racialized populations can be explained away without requiring a deeper investigation into the assessment itself. Randall (2021) observed:

> Opportunity to Learn should be investigated after (and only after) the assessment itself has been thoroughly interrogated for white-supremacist content, and antiracist content moved into its place. (p. 6)

While Opportunity to Learn can certainly be a factor in explaining differences in performance, it should always be the last place assessment developers and users should look.

Randall, Slomp, Poe & Oliveri (2022) observe, that "when the ongoing realities of social oppression are not recognized, the use of validity arguments becomes another racist tool, reproducing- rather than disrupting-systems of oppression." They propose an anti-racist validation framework that instead places the issue of race at the center of assessment design and use. This process begins with a critical appraisal of the construct underpinning the assessment,

and its stability across racial contexts. While this disparity in the opportunity to learn may help explain the disparity in performance by BIPOC students in Kelly-Riley's study, deeper scrutiny of the assessment itself is likely necessary. In particular, the writing construct underpinning the assessment, the scoring criteria, and the operationalization of that criteria, likely needs to be more critically examined. Cushman (2016) more succinctly made this point in her critique of validity theory:

> Fairness can address content of particular questions, but it does little to adjust the overall ways in which validity measures themselves, from the start, are based on colonial difference that they help to create and maintain. . . . In this instance, constructs will always be unrelated to the knowledges and language practices of the peoples made different by the construct and validity measures in the first place.

Cushman's observations highlight the value of seeking out pluriversal understandings; of seeking out multiple and varied experiences and perspectives in trying to understand how an assessment is functioning.

IN SEARCH OF REASONABLE PLURALISM

The final chapter in this section is, Elliot, Rupp, and Williamson's (2015) paper, "Three Interpretative Frameworks: Assessment of English Language Arts-Writing in the Common Core State Standards Initiative." Their paper is a case study of the Smarter Balance Assessment Consortia's program of research and development for the Partnership for Assessment of Readiness for College and Careers (PARCC). This study demonstrates the value of seeking out pluriversal understandings in writing assessment research. Their research team itself, a collaboration between a compositionist and two psychometricians, illustrates how multidisciplinary perspectives can help to bring forward concerns for validity, reliability, and fairness in assessment design and use.

Elliot, Rupp and Williamson (2015) propose a heuristic-based model of stakeholder engagement, to foster dialogue, understanding, and design options that reflect diverse stakeholder perspectives. Rather than approaching writing assessment design and use through isolated frames of references particular to specific disciplines, they advocate for collaborative design processes grounded in common referential frames—well articulated construct models, principled design frameworks, and well defined standards/conceptualizations of validity, reliability, and fairness. Their call for pursuing a "reasonable pluralism" brings us full circle to a founding

motivation of the *Journal of Writing Assessment*: to promote multidisciplinary dialogue and understandings of writing assessment research.

Using the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014) as a structure, Elliot et al. (2015) create a heuristic for interrogating assessment programs from the standpoint of multiple stakeholder groups: Students and guardians, teachers and administrators, legislators, and workforce leaders. Similar to the impetus behind Kelly-Riley's (2011) study, these heuristics empower stakeholder groups by providing them with principled questions that can be used to ensure assessment programs are achieving just outcomes. Expounding on this innovation, they explain that heuristic-based argumentation can be used to bridge the gap between "the logic of the assessment developer and the logic of the assessment user" (p. 117).

Their paper also highlights the value of principled design frameworks for supporting and centering such collaborations. These frameworks help multidisciplinary design teams "develop common language, mental models, design artifacts, and best practices" (p. 105) combined with heuristic-based reasoning models. These frameworks can support the development of consensus among stakeholders to the assessment.

In 2021, the *Journal of Writing Analytics* published a special issue (Olivieri et al., 2021) that tells the story of a multidisciplinary collaboration focused on the design of a scenario-based digital formative assessment platform for teaching and assessing workplace English communication skills. The project brought together experts in assessment design, cognitive science, curriculum and instructional design, educational policy, human-computer interaction, information visualization, task design, psychometrics, score report design, and writing studies. Slomp, Oliveri, and Elliot (2021) in the Afterword to that Special Issue report that principled design frameworks were critical to the success of this collaboration, enabling the research team to identify key questions that drove the design work forward while also structuring which sets of expertise were required to address each question.

Multi-disciplinary collaborations, however, are difficult to manage. Competing ontological, axiological, and epistemological perspectives often underpin differences in our approaches to key issues in assessment. Coming to terms with these differences, and how they shape our thinking about writing assessment is a critically important part of this work. Cushman (2016) captures this challenge:

> [Y]ou don't have to be a person of difference to dwell in borders, to think of ways in which social equity and pluriversal understandings can be achieved in everyday knowledge work of assessment design and research on assessment. The

> important thing is to actively seek out pluriversal (rather than universal) understandings, multiple and varied (rather than singular and narrow) ways of expression, integrated (rather than siloed) exercises in validity and reliability, whole and active (rather than atomized and static) language uses in an effort to name and respect a range of ontological, axiological, and epistemological perspectives. (p. 102)

Cushman outlines a vision for our field that builds on Huot's (2003) vision of a more integrated discourse between the fields that inform writing assessment scholarship and practice.

LOOKING FORWARD: IMPLICATIONS FOR RESEARCH AND PRACTICE

An ethic I have always appreciated about the scholarship published in the *Journal of Writing Assessment* is its appreciation for the consequential nature of the work we do, of the importance of our scholarship for those impacted by our assessment practices. In part, the journal's enduring focus on localism helps to foster this ethic, as authors and researchers are often very close to the consequences of the assessments they are studying: they and their students often live with the consequences stemming from the use of the assessment programs they are designing and investigating.

While it may be true that scholars in fields of composition have drawn more on scholarship and concepts in measurement than measurement scholars have drawn on work in composition and rhetoric, the evolution of validity, reliability, and fairness within the measurement community has often reflected the criticism of this discipline offered by the writing studies community. The move toward more integrated conceptions of validity, reliability, and fairness is an important example of this evolution.

Writing Studies scholars who work in the field of writing assessment have been effective in leveraging advances in measurement theories and concepts to benefit their students, colleagues, classrooms, and institutions. We have harnessed these theories to our local and disciplinary values. A brief walk through the last 5 years of issues in *JWA* demonstrates this. In 2016, the Special Issue on Ethics and Writing Assessment (Kelly-Riley & Whithaus) offered critical appraisals of contemporary theories of validity and fairness to offer up an integrated framework for writing assessment design and appraisal positing a theory of ethics as a mechanism for foregrounding disciplinary concerns for fairness and justice in the application of those theories. In 2019, a Special issue on Writing

Placement in the Two-Year College (Kelly-Riley & Whithaus, 2019) applies the frameworks developed in the 2016 SI to the design and use of placement tests in the Two-year college. In 2018, Pruchnic et al. advanced mixed methods approaches to collecting validity and reliability evidence designed to address the concerns of both measurement specialists and writing studies professionals.

This work, however, remains uneven. Sprinkled through these same issues are articles that continue to approach validity, for example, using dated models and approaches: that speak of validating instruments rather than inferences and decisions. I noted the same unevenness in how this concept was being handled in articles published in *Assessing Writing* over the past decade:

> One trend of concern across several of the papers published in the past 10 years is the characterization of validation studies as attempts to "establish" the validity of the assessments in question. This language suggests a confirmation bias that was not noticeable in the earlier validation studies published in ASW. It is important to remember that we do not validate assessments. Rather, we examine categories of evidence and then use that evidence to form an interpretation and use argument that is always contingent. Too often this contingency is not expressed. (Slomp, 2019, p. 14)

As we draw on contemporary theories of validity, reliability, and fairness, to assess the design implementation and use of locally developed assessment programs, a critical reflexive mindset remains important.

Looking forward, it is also important to recognize that measurement is not a unified and monolithic discipline. Many scholars within this discipline, too, are struggling with its roots and with its history. Stephen Sireci (2021), in his presidential address to the National Council on Measurement in Education, for example, called out the discipline for losing the public's confidence in their work. He cites four reasons for this: psychometric hypocrisy, psychometric censorship, psychometric paralysis, and the discipline's support for an educational culture of distrust. Other measurement scholars I've cited in this paper—Newton, Mislevy, Randal, Rupp, Oliveri—are but a few examples of scholars who are working to take measurement in a more humanistic direction. Their work demonstrates how collaborations with measurement scholars who share concern for the impact of measurement both on diverse populations of students and educators, and on systems of education, can support the development of a new generation of writing assessment programs that focus first on the needs of students and educators (Oliveri et al., 2021) for an example of such collaboration). The three articles highlighted in this section offer a prescription for supporting this work.

Challenge Assumptions: We must always challenge assumptions. In particular, we must question assumptions about raters and the scores they produce. We need to continue challenging the assumption that agreement equals validity. Pursuing an ongoing program of research that examines the basis for raters' scoring decisions and the construct relevant and irrelevant factors that shape those decisions will help us to better understand both what their scores mean, and how confident we can be in making inferences and decisions based on these scores. Kelly-Riley's study reminds us that we need to challenge assumptions about our validity arguments too. In the past it may have seemed reasonable to justify differences in performances on a writing assessment by populations as a function of differences in their opportunity to learn. This is no longer true. As we grapple with systemic racism, and as our classrooms become increasingly diverse, it is increasingly important that we examine how race, culture, and difference shape the constructs we measure, the scores our assessments generate, and the decisions and inferences we draw from those scores.

Strive for Conceptual Clarity: O'Neill's study challenges us to always be pushing for conceptual clarity. As we challenge assumptions, we need to continue to think about how our ontological, axiological, and epistemological positions inform how we conceptualize the standards that shape our work. This search for conceptual clarity grounded in the values of our discipline will continue the innovation and evolution of writing assessment practices. Randall (2021) and Cushman's (2016) work point to the importance of questioning the very validity frameworks that we have used to guide the design and assessment of writing programs. One wonders, for example, how an anti-racist validation framework will open up possibilities for future innovations in writing assessment design.

Attend to Consequences: Kelly-Riley's study powerfully demonstrates the importance of attending to the consequences resulting from the design, implementation, and use of our assessment programs. In the absence of concern for consequences so much of the work we do can be dismissed as mere disciplinary and theoretical debate; work that only serves ourselves. The issues we explore matter precisely because they carry consequences for the millions of people every year who are subjected to writing assessments. Attending to those consequences will provide fruitful avenues for programs of research into both the theoretical and practical challenges our field is facing.

Pursue Purposeful Pluralism: Elliot, Rupp, & Williamson's study highlights the importance of purposeful pluralism. They draw our attention to the value of seeking out multiple, critical perspectives on the work that we are engaged in; of the importance of listening to those voices, carefully considering the hopes, concerns, and insights those voices infuse into our work; of the imperative that we respond to what we hear. Looking back to the earliest issues of both *Assessing*

Writing and *the Journal of Writing Assessment*, we see a clear ethic of critical scholarship, and openness to exploring possibility, of listening to critique, and of responding to it. Harnessing that ethic to a spirit of purposeful pluralism will serve our discipline well as it innovates for the future.

Together these principles position us to approach our work with a sense of humility and purpose, reminding us that the work of writing assessment research should be done in the service of others. As our fields continue to evolve purposeful, principled pluralism will be a key tool we can leverage to ensure that writing assessments programs serve all students, promote quality learning, and structure opportunity. In this spirit, educators—writing studies specialists—need to increasingly insist on having a seat at the table, and they need to come to that table equipped to engage with the measurement theories set before them, while not neglecting to add to the conversation the insights and concerns of our discipline, and in particular our enduring concern for the social consequences of our assessment programs on students, educators, and systems of education.

REFERENCES

American Educational Research Association, American Psychological Association, & National Council of Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Adler-Kassner, L., & O'Neill, P. (2010). *Reframing writing assessment to improve teaching and learning*. Utah State University Press.

Bachman, L. & Palmer, A. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world.* Oxford University Press.

Baker-Bell. A. (2010). Playing with the stakes: A consideration of an aspect of the social context of a gatekeeping writing assessment. *Assessing Writing*, *15*(3), 133-153.

Barkaoui, K. (2010). Variability in ESL essay rating processes: The role of the rating scale and rater experience. *Language Assessment Quarterly*, *7*(1), 54-74.

Barkaoui, K. (2011). Effects of marking method and rater experience on ESL essay scores and rater performance. *Assessment in Education: Principles, Policy & Practice*, *18*(3), 279-293.

Behizadeh, N., & Engelhard Jr., G. (2011). Historical view of the influences of measurement and writing theories on the practice of writing assessment in the United States. *Assessing Writing*, *16*(3), 189-211.

Brennan, R. L. (2001). An essay on the history and future of reliability from the perspective of replications. *Journal of Educational Measurement*, *38*(4), 295-317.

Broad, B. (2003). *What we really value: Rubrics in teaching and assessing writing*. Utah State University Press.

Broad, B. (2016). This is not only a test: Exploring structured ethical blindness in the testing industry. *Journal of Writing Assessment*, *9*(1). https://escholarship.org/uc/item/2bt3m3nf

Broad, B., Adler-Kassner, L., Alford, B., Detweiler, J., Estrem, H., Harrington, S., McBride, M., Stalions, E., & Weeden, S. (Eds.). (2009). *Organic writing assessment: Dynamic criteria mapping in action.* Utah State University Press.

Cushman, E. (2016). Decolonizing validity. *Journal of Writing Assessment*, *9*(2). https://escholarship.org/uc/item/0xh7v6fb

Dryer, D. B. (2012). At a mirror, darkly: The imagined undergraduate writers of ten novice composition instructors. *College Composition and Communication*, *63*(3), 420-452.

Elliot, N. (2005). *On a scale: A social history of writing assessment in America*. Peter Lang.

Elliot, N., & Perelman, L. (Eds.). (2012). *Writing assessment in the 21st century: Essays in honor of Edward M. White*. Hampton Press.

Elliot, N., Rudniy, A., Deess, P., Klobucar, A., Collins, R., & Sava, S. (2016). ePortfolios: Foundational measurement issues. *Journal of Writing Assessment*, *9*(2). https://escholarship.org/uc/item/7bm4t40t

Elliot, N., Rupp, A. A., & Williamson, D. M. (2015). Three interpretative frameworks: Assessment of English language arts-writing in the Common Core State Standards initiative. *Journal of Writing Assessment*, *8*(1). https://escholarship.org/uc/item/4zb222xg

Goodwin, S. (2016). A Many-Facet Rasch analysis comparing essay rater behavior on an academic English reading/writing test used for two purposes. *Assessing Writing*, *30*, 21-31. https://doi.org/10.1016/j.asw.2016.07.004

Huot, B. (2002). *Re-articulating writing assessment*. Utah State University Press.

Huot, B. (2003). Introduction. *Journal of Writing Assessment, 1*(2), 81-84. https://escholarship.org/uc/jwa/1/2

Huot, B., & Yancey, K. (2003). Introduction. *Journal of Writing Assessment, 1*(1), 1-4.

Jølle, L. (2014). Pair assessment of pupil writing: A dialogic approach for studying the development of rater competence. *Assessing Writing*, *20*, 37-52.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed.) (pp. 17–64). American Council on Education; Praeger.

Kane, M. T. (2013). Validating the interpretation and uses of test scores. *Journal of Educational Measurement, 50(1)*, 1–73.

Kelly-Riley, D. (2011). Validity inquiry of race and shared evaluation practices in a large-scale, university-wide writing portfolio assessment. *Journal of Writing Assessment*, *4*(1). https://escholarship.org/uc/item/7m18h956

Kelly Riley, D., & Whithaus, C. (2016). Introduction to a special issue on a theory of ethics for writing assessment. *Journal of Writing Assessment*, *9*(1). https://escholarship.org/uc/item/8nq5w3t0

Kelly-Riley, D., & Whithaus, C. (2019). Editors' introduction: Special issue on two-year college writing placement. *Journal of Writing Assessment*, *12*(1). https://escholarship.org/uc/item/7vg91466

Klein, J., & Taub, D. (2005). The effect of variations in handwriting and print on evaluation of student essays. *Assessing Writing*, *10*(2), 134-148.

Knoch, U. (2007). "Little coherence, considerable strain for reader": A comparison between two rating scales for the assessment of coherence. *Assessing Writing*, *12*(2), 108-128.

Lindhardsen, V. (2020). Co-equal participation and accuracy perceptions in communal writing assessment. *Journal of Writing Assessment*, *13*(1). https://escholarship.org/uc/item/20f7s465

Litterio, L. M. (2016). Contract grading in a technical writing classroom: A case study. *Journal of Writing Assessment, 9*(2). https://escholarship.org/uc/item/02q4g1gt

Marefat, F., & Heydari, M. (2016). Native and Iranian teachers' perceptions and evaluation of Iranian students' English essays. *Assessing Writing*, *27*, 24-36.

Mislevy, R. J., (2018). *Sociocognitive foundations of educational measurement*. Routledge.

Moss, P. A. (1994). Can there be validity without reliability? *Educational Researcher, 23*(4), 5-12.

Newton, P. E. (2017). There is more to educational measurement than measuring: The importance of embracing purpose pluralism. *Educational Measurement: Issues and Practice*, *36*(2), 5-15.

Oliveri, M. E., Slomp, D., Rupp, A., Mislevy, R., Vezzu, M., Tackitt, A., Phelps, J., Osborn, M. (2021). Introduction: Meeting the challenges of workplace English communication in the 21st Century. *Journal of Writing Analytics 5(1).* 1-33. https://wac.colostate.edu/docs/jwa/vol5/intro.pdf

O'Neill, P. (2011). Reframing reliability for writing assessment. *Journal of Writing Assessment*, *4*(1). https://escholarship.org/uc/item/6w87j2wp

Onwuegbuzie, A. J., & Leech, N. L. (2005). On becoming a pragmatic researcher: The importance of combining quantitative and qualitative research methodologies. *International journal of social research methodology*, *8*(5), 375-387.

Parkes, J. (2007). Reliability as argument. *Educational Measurement: Issues and Practice, 26*(4), 2-10.

Pruchnic, J., Susak, C., Grogan, J., Primeau, S., Torok, J., Trimble, T., Foster, t., & Barton, E. (2018). Slouching toward sustainability: Mixed methods in the direct assessment of student writing. *The Journal of Writing Assessment*, *11*(1), https://escholarship.org/uc/item/9z65k7wj

Pula, J. J., & Huot, B. A. (1993). A model of background influences on holistic raters. In M. M. Williamson & B. A. Huot (Eds.), *Validating holistic scoring for writing assessment: Theoretical and empirical foundations* (pp. 237-265). Hampton Press.

Raczynski, K. R., Cohen, A. S., Engelhard Jr, G., & Lu, Z. (2015). Comparing the effectiveness of self-paced and collaborative frame-of-reference training on rater accuracy in a large-scale writing assessment. *Journal of Educational Measurement*, *52*(3), 301-318.

Randall, J. (2021). "Color-neutral" is not a thing: Redefining construct definition and representation through a justice-oriented critical antiracist lens. *Educational Measurement: Issues and Practice, 40*(4), 82-90. https://doi.org/10.1111/emip.12429

Randall, J., Slomp, D., Poe, M., & Oliveri, M. E. (2022). Disrupting white supremacy in assessment: Toward a justice-oriented, antiracist validity framework. Educational Assessment, 27(2), 170-178. https://doi.org/10.1080/10627197.2022.2042682

Rosenblatt, L. (1938). *Literature as exploration*. D. Appleton-Century.

Shumake, J., & Shah, R. W. (2017). Reciprocity and power dynamics: Community members grading students. *Reflections*, *17*(2), 5-42.

Sims, M. E., Cox, T. L., Eckstein, G. T., Hartshorn, K. J., Wilcox, M. P., & Hart, J. M. (2020). Rubric rating with MFRM versus randomly distributed comparative judgment: A comparison of two approaches to second-language writing assessment. *Educational Measurement: Issues and Practice*, *39*(4), 30-40.

Sireci, S. G. (2021). NCME presidential address 2020: Valuing educational measurement. *Educational Measurement: Issues and Practice*, *40*(1), 7-16.

Slomp, D., Oliveri, M. E., Elliot, N. (2021). Afterword: Meeting the challenges of workplace English communication in the 21st century. *Journal of Writing Analytics 5(1),* 342-370. https://wac.colostate.edu/docs/jwa/vol5/afterword.pdf

Slomp, D. H., (2019). Complexity, consequence, and frames: A quarter century of research in assessing writing. *Assessing Writing, 42*(4), 1-17

Wang, J., Engelhard Jr, G., Raczynski, K., Song, T., & Wolfe, E. W. (2017). Evaluating rater accuracy and perception for integrated writing assessments using a mixed-methods approach. *Assessing Writing*, *33*, 36-47.

Wind, S. A., & Engelhard Jr, G. (2013). How invariant and accurate are domain ratings in writing assessment?. *Assessing Writing*, *18*(4), 278-299.

Winke, P., & Lim, H. (2015). ESL essay raters' cognitive processes in applying the Jacobs et al. rubric: An eye-movement study. *Assessing Writing*, *25*, 38-54.

Wiseman, C. S. (2012). Rater effects: Ego engagement in rater decision-making. *Assessing Writing*, *17*(3), 150-173.

Wolfe, E. M. (2005). Uncovering rater's cognitive processing and focus using think-aloud protocols. *Journal of Writing Assessment*, *2*(1), 37–56. https://escholarship.org/uc/item/83b618ww

Zheng, Y., & Yu, S. (2019). What has been assessed in writing and how? Empirical evidence from assessing writing (2000–2018). *Assessing Writing, 42*. https://doi.org/10.1016/j.asw.2019.100421

CHAPTER 1.

REFRAMING RELIABILITY FOR WRITING ASSESSMENT

Peggy O'Neill
Loyola University Maryland

This essay provides an overview of the research and scholarship on reliability in college writing assessment from the author's perspective as a composition and rhetoric scholar. It argues for reframing reliability by drawing on traditions from fields of college composition and educational measurement with the goal of developing a more productive discussion about reliability as we work toward a unified field of writing assessment. In making this argument, the author uses the concept of framing to argue that writing assessment scholars should develop a shared understanding of reliability. The shared understanding begins with the values—such as accuracy, consistency, fairness, responsibility, and meaningfulness—that we have in common with others, including psychometricians and measurement specialists, instead of focusing on the methods. Traditionally, reliability has been framed by statistical methods and calculations associated with positivist science although psychometric theory has moved beyond this perspective. Over time, the author argues, if we can shift the frame associated with reliability, we can develop methods to support assessments that lead to improvement of teaching and learning.

Writing an essay about reliability and writing assessment presents several challenges. One comes from determining what we mean by writing assessment because as a field it encompasses teachers and researchers in K-12 education as well as higher education. Some of these professionals are trained in educational measurement, but many others are trained primarily as literacy educators. The field also includes test developers employed by testing companies, some of whom may provide testing services for institutions, and government employees, typically in departments of education, who work on assessments such as NAEP or others. Another challenge concerns the very concept of reliability, which is deeply embedded in statistical theories and methods. Many educators who teach

DOI: https://doi.org/10.37514/PER-B.2024.2166.2.01

writing and work in college writing assessment have been educated primarily in humanities departments and are immersed in the subject of literacy education; they are not psychometricians and are not experts in statistical theories and methods, which seem to dominate approaches to reliability. Because of these challenges, college writing assessment practitioners often side-step reliability to some extent. They report instead, for example, a co-efficient about rater agreement or percentages of samples needed to be scored by three or more readers, but do not delve into the complexity of the issues associated with issues such as calculating coefficients. Yet, reliability is an important component of writing assessment that needs to be considered not just in its own right but also as part of the validation process because it addresses consistency and generalizability, among other values.

As writing assessment practitioners and scholars, we need to grapple with the challenges associated with reliability by examining how it has been used in writing assessment scholarship, especially within the college composition community, and how we can reframe it so that it both engages with what writing teachers value and contributes appropriately to validation efforts. With the interest in large-scale assessments (including writing assessment) and higher education increasing, college writing faculty will need to address several issues, many of which are related to reliability as well as validity (Adler-Kassner & O'Neill, 2010). One such issue is automated scoring, which is generally critiqued by college composition professionals (Herrington & Moran, 2001; Ericsson & Haswell, 2006) but which has found more support in the psychometric community (Williamson 2003). According to Williamson (2003):

> Two things are certain. One, automated scoring programs can replicate scores for a particular reading of student writing, and this technology is reliable, efficient, fast, and cheap. Two, automated scoring has been and will continue to be used in various large-scale assessments of student writing. (p. 256)

Carl Whithaus (2005) also acknowledged the role of automated scoring in large-scale testing and encouraged writing instructors not only to accept automated evaluation systems but also to integrate them (as well as other technologies) into their teaching (p. 13). Williamson, who doesn't go as far as Whithaus in supporting the use of automated evaluation, argued for a "productive alliance" between those in educational measurement and those invested in teaching writing (p. 101). To develop this kind of relationship, college writing instructors and program administrators need to "examine the research methodology or social sciences as it impinges on assessment" and to "explore the potential for collaborative research, not just within a social science or humanistic tradition"

(Williamson 2003, p. 101). Williamson (2003) identifies validity as a focal point of this research (p. 101), and I would add that we especially need to attend to reliability not only because of its contributions to validity but also because of the role it has played historically in writing assessment. If we can engage in these discussions, we may be able to begin reframing reliability by first developing a better understanding of reliability and then becoming full partners in the discussions—and development—of writing assessment that extend beyond our programs and institutions.

In what follows, I provide an argument for reframing reliability in writing assessment for those who come from the field of college composition as well as those whose approach is grounded in educational measurement. This analysis comes from my perspective as a composition and rhetoric scholar, but my goal is to begin a more productive discussion about reliability as we work toward a unified field of writing assessment (Huot, 2002).

THE CONCEPT OF FRAMING

In thinking about reframing reliability, I begin with the concept of framing in general. While there are many theories associated with framing, the basic idea is that we view ideas, experiences, and events through frames (akin to what Kenneth Burke called "terministic screens" and what Thomas Kuhn identified as "paradigms"). These frames usually operate at an unconscious level and are constructed by society. Members of a particular culture are conditioned to make certain connections and to understand new information through a particular frame or lens. Frames are stronger when they connect to stories shaped by the same frame (Hertog & McLeod, 2001). Communication and cultural theorists, such as Stuart Hall (1983), have explained that the media play a dominant role in this process through cultural conditioning, which is established by the boundaries media set around the stories that they cover. These boundaries, which determine what is and is not covered, create tacit connections and connotations for members of that culture. For example, current debates about education policy and funding are framed by the concept that education should prepare students for college or careers. This has become so ubiquitous in the media that it is hard to articulate other purposes of education, such as civic engagement. This perspective links to other stories about education, such as the often repeated story about US students lagging behind other students, which is reinforced by, among other things, the "Race to the Top" initiatives supported by the US Department of Education. In today's culture with non-stop access to news through 24-hour cable channels and the Internet, mass media is an especially powerful means of creating frames.

Cognitive linguists acknowledge the importance of culture and media in creating frames, but they take it deeper. Framing, according to cognitive linguists, is about how our minds work to make meaning from language and images. George Lakoff (2006) explained that framing is "a conceptual structure used in thinking" and that every word evokes a frame. In fact, words have no meaning outside of frames, which fit together to form systems. The frames are reinforced the more they are evoked: "Every frame is realized in the brain by neural circuitry. Every time a neural circuit is activated, it is strengthened" (Lakoff, 2006, n.p.). In other words, frames are connected to the way our brains are wired. Every word evokes a frame, and words defined within a frame, evoke the frame. The example Lakoff uses to explain this concept is "Sam picked up the peanut with his trunk," which evokes the frame of elephant because we understand trunk in this sentence within that frame. Even negating a frame evokes it as in Lakoff's example, "Don't think of an elephant!" which is impossible to carry out because as soon as elephant is mentioned, we think of it. Every time a frame is evoked—whether negatively or positively, whether directly or indirectly—it is strengthened. So in arguing against a frame, we are actually reinforcing it.

Frames are so powerful that they shape the way we understand facts. Facts, explained Lakoff (2002), are understood within our frames so that people with different worldviews understand and process the facts differently. In other words, as Lakoff (2002) argued, it isn't a matter of just getting the most accurate information into the debate, it is critical that that information is framed in ways that make sense to the audience. Other ways of reasoning, including framing and categorization, creates "huge variability in normal, everyday human reasoning" (Lakoff, 2002, p. 373). What one person sees as clear, rational commonsense can be understood in completely different ways by others depending on the individuals' frames, which makes communication more difficult. Many academics and researchers may experience this kind of communication disjunction when trying to discuss issues related to their scholarly work with a non-expert. Sometimes the difficulties are simply related to terminology, for example, the term "grammar" is often used by non-experts to discuss the teaching of writing to cover a wide range of issues to address in teaching writing from mechanics and punctuation to style, organization, use of evidence and a myriad of other aspects. In this type of situation, further discussion and probing can usually clear up the confusion. However, communication problems can also be rooted in different frames.

How individuals frame a concept such as "teaching writing" will depend on their own experiences, education, expertise and values as well as how it has been depicted in the culture. This understanding, furthermore, may or may not align with what a particular person means when she uses the phrase "teaching writing." An individual's frame will be reinforced every time it is evoked. So, for example,

when the Common Core State Standards Initiative identifies writing as one of the key areas for "college and career readiness," readers will understand this section through the frame they have already have about writing. Parents, teachers, policymakers and assessment experts may not all share the same framework and so they may understand the standards differently. The authors of the standards try to mitigate this situation by providing preliminary material that defines terms, explains situations, and even articulates what is not covered by the standards. However helpful that information is, it can also act as a away to reinforce what the reader already thinks and believes because many of the associations and assumptions work at unconscious levels. If we consider writing assessment, then, the same theory applies. What seems practical and rational to writing teachers may seem completely unreasonable or just wrong to policymakers or psychometricians, who may approach the activity through completely different frames—different values, experiences, assumptions, and world views. Specific technical terms associated with assessment, such as validity and reliability, will also be understood differently depending on the frame surrounding them.

If we want to change a concept or redefine a concept, we need to consider the frame that surrounds it and how that influences the way a term is understood. Trying to change the term without taking into account the bigger picture will not be successful, in Lakoff's view, because much of what is evoked happens automatically and unconsciously. Making visible the dominant associations and assumptions so we can see the frame that currently in place is the first step in trying to reframe writing assessment in general and reliability in particular.

TRADITIONAL FRAMING OF WRITING ASSESSMENT

For the last hundred years, reliability has been the dominant frame surrounding writing assessment, pre-occupying scholars and test developers (Huot & Neal, 2006; Williamson, 1993; Huot, 2002; Elliot, 2005; O'Neill, Moore, & Huot, 2009). Although reliability, as a psychometric concept, encompasses a broad range of concerns, in writing assessment this quest has focused primarily on scoring, specifically getting scorers to agree at an acceptable rate, which is referred to as interrater reliability. As Huot and Neal (2006) concluded in their techno-history of writing assessment: "Throughout the history of writing assessment and whether we refer to technologies like the indirect tests of grammar usage and mechanics, the use of rubrics and rater training, or the machine-scoring of student writing, we are basically referring to technological solutions to the problem of scoring consistency" (pp. 418-19). For example, the College Entrance Examination Board (CEEB) ostensibly abandoned essay exams in 1941 as part of the war effort to streamline student matriculation for potential armed forces

recruits. In truth, the CEEB had been piloting the SAT for scholarship students who needed to apply earlier and had found that the reliability and efficiency of the SATs to be much superior to that of the essay examination. The development of holistic scoring procedures in the 1960s, done by Educational Testing Service researchers Godschalk, Swineford, and Coffman (1966), revitalized essay testing because it provided a reliable way to score essays.

By the 1980s, holistically scored essays enjoyed widespread use for a variety of writing assessments across educational levels, but especially in college. Edward White (1993) claimed: '[W]hen a university or college opens discussion of the measurement of writing ability these days, the point of departure is usually a holistically scored essay test" (p. 89). The holistic scoring of essay exams depended upon standardization of procedures for the test administration, of the tasks and topics, and of the scoring. The holistic scoring sessions became, according to White (1993), not just a method for scoring but also a means of professional development as readers discussed anchor papers and practiced scoring samples to internalize the scoring rubric so they could apply it in a consistent way. These scoring sessions also required careful record keeping and checks for agreement between two independent raters.

While White focused on the benefits of holistic scoring both in terms of professional development and achieving acceptable reliability rates, Cherry and Meyer (1993) critiqued the way reliability has been handled in writing assessment. They explained that reliability "refers to how consistently a test measures whatever it measures" (p. 110). The consistency of a measurement, Cherry and Meyer (1993) explained, can come from the test design and administration, the students, or the scoring. For essay testing, particular sources of error may include the prompt—which may not produce reliable results—as well as the administration and scoring of the essays. After reviewing research in direct writing assessment from Starch and Elliot's 1912 article, "Reliability of the grading of high school work in English," through several pieces in the mid to late 1980s, Cherry and Meyer (1993) concluded that there have been four serious problems with reliability as reported in writing research and evaluation (p. 116).

First, according to Cherry and Meyer (1993), reliability discussions (with a few notable exceptions) have been limited to interrater reliability although there are many other aspects of reliability that need to be considered. For example, if students' performances are not accurate in terms of their writing abilities because of the prompt design, then results are not reliable no matter how consistently raters apply the rubric and how much they agree with each other (Hoetker, 1982).

Second, there has been confusion over reliability and validity in influential studies of writing assessment. Cherry and Meyer (1993) critiqued Godshalk,

Swineford, and Coffman (1966) because they identified differences in results across topics as a reliability issue when in fact these differences are about validity. Variation across topics/prompts, they explain, can be a validity issue because the underlying construct being tapped is different if the writing tasks are different.

Third, there has been a lack of agreement on appropriate statistical methods for determining interrater reliability. Cherry and Meyer (1993) reported that at least eight different methods had been used in computing and reporting interrater reliability statistics and that many studies never even explained how they calculated the reliability co-efficient (p.119). However, the variable ways for calculating the interrater reliability co-efficient can yield drastically different results. For example, using a straight percentage of agreement between raters, the Pearson correlation coefficient or Cronbach's alpha to calculate the interrater reliability will produce different statistics for the same data.

Finally, various procedures used in holistic scoring sessions directly affected the reliability statistics. Cherry and Meyer (1993) explained that sometimes reliability rates are based on "practice sessions," not live scoring, which can artificially inflate the interrater reliability statistic. Other problems come from the practice of "resolving" differences between two raters by using a third rater. In fact, Cherry and Meyer (1993) recommended discontinuing the practice of resolving differences all together (p. 122) because "interrater reliability formulas are quite sensitive to the manipulation of data" through these methods even when a low percentage of scores are affected (p. 123). Haertal (2006), in discussing reliability and ratings of products, echoed their concern: "It must be emphasized that when adjudication is used, assumptions for many statistical models are violated" (p. 102).

Stemler (2004) argues that interrater reliability needs to be unpacked. He contends "the widespread practice of describing interrater reliability as a single, unitary concept is at best imprecise, and at worst potentially misleading" (p. 2). He identifies three categories of interrater reliability—consensus estimates, consistency estimates, and measurement estimates—and details the assumptions, interpretations, advantages, and disadvantages of each (p. 2). The statistical methods for determining the different types of interrater reliability also vary, and Stemler (2004) reviews these as well. Although Stemler (2004) is not limiting his focus to writing and literacy assessments, he seems to agree with Cherry and Meyer (1993) that researchers do not address the nuances of interrater reliability enough.

While Cherry and Meyer (1993) articulated several problems with reliability reported in writing assessment research, Hayes and Hatch (1999) focused on problems with reliability in literacy research in general, including rating of student work whether for a testing or research purposes. Hayes and Hatch (1999)

also critiqued the method of calculating and reporting reliability found in the literature, especially on more recent studies. They argued that interrater reliability rates should be determined by statistical correlations and not the percentage of agreement between the two independent raters. Hayes and Hatch (1999) explained that reliability calculated using a statistical correlation formula takes into account the role of chance in the agreement rate while the percentage method doesn't. Depending on the scoring scale and the distribution of scores, chance can account for a significant portion of agreement. For example, the fewer score points on the rating scale, the greater the influence chance has on the agreement rate; or, the more scores tend to cluster around certain scores, the more influence chance has on the reliability measure.

Like Cherry and Meyer (1993), Hayes and Hatch (1999) noted that different methods for calculating reliability lead to different results, yet they also found many researchers did not report the method for calculating reliability correlations. Both Cherry and Meyer (1993) and Hayes and Hatch (1999) also agreed that when researchers do not fully disclose how they determined reliability estimates, it is difficult for readers to determine if the method is appropriate, to compare reliability across studies, and to avoid confusion. Hayes and Hatch (1999) concluded their essay with an acknowledgment that other methods exist for measuring reliability, including generalizability measures, than those they address although they don't discuss them.

Both Cherry and Meyer (1993) and Hayes and Hatch (1999) framed reliability in writing assessment using classical test theory. Shale (1996), however, advocated using generalizability theory instead, arguing that it is more appropriate for addressing the issues associated with reliability in writing assessment because it can address the multiple sources of error that can arise in a writing assessment. In most writing assessments, Shale (1996) contended, reliability is vague because it is only considered within the classical test theory, which was developed for multiple-choice testing: "Considerable ambiguity arises because the full sense of reliability as understood within the context of multiple-choice testing does not transfer well to the world of essay testing" (p. 77). Shale explained that the consequences of considering reliability only in terms of classical test theory has resulted in a "fixation on marker disagreement" which has led to a distortions and limitations in writing assessment practices (p. 78). Shale (1996) as with Cherry and Meyer (1993) and Hayes and Hatch (1999), also noted the paucity of rigorous inquiry into reliability in writing assessment scholarship. Reliability, how we should approach it, and what we mean by the term is still an issue in college writing assessment as I discuss in more detail later.

While concerns about reliability of essay exams preoccupied writing assessment scholars for a long time and, in effect framed writing assessment, the

validity of essay testing was not seriously challenged because essay testing required students to write instead of completing multiple-choice items about language conventions and grammar. White (1993) articulated the assumptions that supported holistic scoring of essay exams: "It is a direct measure of writing, measuring the real thing, and hence is more valid than indirect measures" such as fill in the bubbles multiple choice exams and editing tests (p. 90). By the 1990s, however, writing assessment scholars (as well as measurement theorists) began to turn their attention to validity arguing that a portfolio of writing was preferable to a single-sample, timed impromptu essay (Elbow and Belanoff, 1986).

The shift to validity began to take the focus away from reliability as a purely statistical concept and to frame it as part of a validity argument, which addresses both theoretical and quantitative, statistical evidence (Messick, 1989). Camp (1993) addressed the tension between classical test theory and emerging theories of writing and literacy. Camp (1993) argued: "Very likely we are seeing the signs of a growing incompatibility between our views of writing and the constraints necessary to satisfy the requirements of traditional psychometrics—in particular, of reliability and validity narrowly defined" (p. 52). Camp (1993) explored this tension, identifying some of the key factors that may need to be addressed to develop writing assessments that take into account what we know about writing as well as the principles of fairness, equity, and generalizability—concepts, she explained, that are associated with reliability. The challenge, according to Camp, has been to apply these principles in ways that lead far beyond the narrow focus on score reliability and constricted definitions of validity that have characterized earlier discussions of writing assessment (p. 68). At the time of Camp's essay, portfolios (like other performance assessments) were growing in popularity and Camp concluded with a brief discussion of some portfolio projects.

Since the early 1990s, the popularity of portfolios in college writing programs has continued to spread for both teaching and assessment (although essay testing also remained popular). Although writing portfolios seemed to be a substantive departure from impromptu essay testing, the discussion of reliability, however, did not changed very much. The focus was still narrowly on interrater reliability. As White (1993) looked to the future of portfolios, he identified reliability of portfolio scoring as the major issue to deal with, which in effect, continued to frame writing assessment in terms of reliability. At that time, he recommended adapting many of the same procedures for portfolios that were used for holistic scoring of essays: "At a minimum, each portfolio should receive two independent scores, and reliability data should be recorded. While reliability should not become the obsession for portfolio evaluation that it became for essay testing, portfolios cannot become a serious means of measurement without demonstrable reliability" (p. 105).

Compositionists in college writing programs, following Elbow and Belanoff (1986), developed an assortment of writing portfolio assessments for placement into first-year writing (Willard-Traub, Decker, Reed, & Johnson, 1999; Daiker, Sommers, & Stygall, 1996; Borrowman, 1999; Lowe & Huot, 1997; Hester et al., 2007) and proficiency (Roemer, Shultz, & Durst, 1991; Nelson, 1999; Haswell, 2001). Many of these portfolio assessments adapted holistic scoring methods used for essay exams to portfolios, reporting the interrater reliability and in many cases, doing so in ways that are problematic according to Cherry and Meyers (1993), Shale (1996), and Hayes and Hatch (1999). For example, Borrowman (1999), reporting the reliability of placement portfolio system at the University of Arizona, presented the reliability co-efficient for the program but did not explain how the figure was calculated. He did, however, devote three pages to discussing reliability and how the high interrater reliability is achieved: "the physical conditions in which the scoring of portfolios takes place and the generation of the scoring rubric" (12). Borrowman (1999) addressed the tension between reliability and validity but he only considered interrater reliability in his discussion, which is a very limited understanding of reliability.

RECONSIDERING THE TRADITIONAL FRAMEWORK

While White (1993) was correct that reliability is a critical issue to address, his assumption that the same methods associated with holistic scoring are the minimum requirements for portfolio assessment demonstrates how writing assessment practitioners and scholars often have a limited reliability as a theoretical construct. It also illustrates how the narrow psychometric frame continued to dominate many of the discussions of reliability in college composition. Yet, in spite of the focus on validity, the critique of traditional treatment of reliability in writing assessment, and discussions about scoring and reliability, many college writing assessment programs still failed to address the reliability issues that Cherry and Meyer (1993) identified in the literature associated with essay exams. While a few writing assessment scholars began pushing against reliability (Smith, 1992, 1993; Haswell & Wyche, 1996; Broad, 1994; Lowe & Huot, 1997), as a field we didn't grapple with it too directly. So when we encountered Moss's (1994) question, "Can we have validity without reliability?" we seemed to respond with an enthusiastic "Yes!" Reliability, however, is an important theoretical construct, and can't be dismissed or ignored. Mislevy (2004), as part of a special section of the *Journal of Educational and Behavioral Statistics*, responded to Moss's (1997) question as well as other commentaries on it, explaining that reliability in psychometrics encompasses a wide range of issues so that "a measure wholly unreliable in the more fundamental sense would consist only of error and could not support valid inferences" (p. 1).

We need to explore the concept more fully, considering it in light of what we know about writing and learning to write, as well as psychometric theory, because as Camp (1993) said, the principles that inform reliability are important.

Moss (1994), in fact, didn't reject reliability outright. Rather she encouraged assessment researchers and practitioners to explore it as a theoretical construct in light of validity. She explained that "less standardized forms of assessment . . . [such as portfolios] present serious problems for reliability, in terms of generalizability across readers and tasks as across other facets of measurement" (p. 6). Though carefully trained readers can achieve acceptable rates of reliability, Moss (1994), an educational measurement theorist, argued that with "portfolios, where tasks may vary substantially from student to student, and where multiple tasks may be evaluated simultaneously, inter-reader reliability may drop below acceptable levels for consequential decisions about individuals or programs" (p. 6). Moss concluded that "although growing attention to the consequences of assessment use in validity research provides theoretical support for the move toward less standardized assessment, continued reliance on reliability, defined as quantification of consistency among independent observations, requires a significant level of standardization," (p. 6). However, these less standardized forms of assessment are often preferable "because certain intellectual activities" cannot be documented through standardized assessments (p. 6).

Moss (1994) suggested that in educational assessment, we look beyond psychometric theories and practices in cases where acceptable reliability rates are difficult or impossible to achieve. She challenged the assessment community to consider its definitions of reliability—and here we in writing assessment need to remember that reliability is more than a quantification of consistency among independent observations. Moss recommended a hermeneutic approach because as a philosophical tradition, it values a "holistic and integrative approach to interpretation of human phenomena" (p. 7). After summarizing the key perspectives of hermeneutics, Moss explained how this methodology would work:

> A hermeneutic approach to assessment would involve holistic, integrative interpretations of collected performances that seek to understand the whole in light of its parts, that privilege readers who are most knowledgeable about the context in which the assessment occurs, and that ground those interpretations not only in textual and contextual evidence available, but also in a rational debate among the community of interpreters. (p. 7)

Critical features of this type of assessment include the recognition of disagreement or difference in interpretations as evaluators bring their expertise and experience to bear on the work. Positions of individual evaluators can change

as rational debate ensues, with the final decision coming out of consensus or compromise. In supporting this approach in specific situations, Moss (1994) reminded readers that reliability and objectivity are no guarantors of truth and that they can, in fact, work against "critical dialogue" and can lead "to procedures that attempt to exclude, to the extent possible, the values and contextualized knowledge of the reader and that foreclose[s] on dialogue among readers about specific performances being evaluated" (p. 9). Mislevy (2004), saw benefits in Moss's idea but also commented:

> In assessment, as in other fields, difficulties arise when novel problems appear and the usual heuristics fail. We now envisage assessments that target inferences more subtle than proficiency in a specified domain of tasks. . . . We must return to first principles to establish the credentials of this evidence . . . The hermeneutic tradition does offer insights into drawing inferences from disparate masses of evidence, and we can indeed learn much from dialectic between psychometrics and hermeneutics. (p. 2)

He advises, though, that a first step is to acquire "a deeper understanding of psychometric methods, an understanding of principles behind methods that will not be found in common wisdom, familiar testing practices, or standard textbook presentations" (p. 2).

Moss's comments about a hermeneutical approach to complex performance assessment echoed what writing assessment scholars praised about holistic scoring sessions and alternative methods for evaluating student writing (whether portfolios or essays). White (1993, 1994), who has been a stalwart supporter of holistic scoring of student writing has often expounded on the benefits associated with norming and scoring sessions. Scholars, reporting on portfolio assessments, made similar statements such as Hamp-Lyons and Condon (2000):

> Instead of focusing on scores, readers spend time bringing their reading processes into line with each other. They read and discuss samples with an eye toward developing and refining a shared sense of values and criteria for scoring. In other words, this method fosters a reading community in which reliability grows out of the readers' ability to communicate with each other, to grow closer in terms of the ways they approach samples. (p. 133)

Although Hamp-Lyons and Condon (2000) addressed reliability in this way, they still used more traditional reliability evidence to justify portfolio assessment:

"The reliability obstacle, in some local contexts, has been overcome. Miami University's reliability statistics, like Michigan's, are within the .8 range of holistic essay assessments . . ." (p.91). Their position echoed White's (1993) concerns about portfolios. However, Hamp-Lyons and Condon (2000) and others did not address the concerns about reliability articulated by Cherry and Meyers (1993), Shale (1996), and Hayes and Hatch (1999).

Other scholars pushed against the traditional holistic scoring approach designing methods that privileged those most knowledgeable about the context, that encouraged critical dialogue, and that used holistic and integrative judgments. Smith (1992, 1993) found that placement decisions for students entering college composition were more reliable with an expert reader system than when made via traditional holistic scoring procedures. In Smith's system, readers made decisions based on the most recent course they taught, either accepting or rejecting the student for the course or rejecting. Haswell and his colleagues at Washington State University (Haswell & Wyche, 1996; Haswell, 2001) developed a two-tiered expert reader system in which readers made the initial decision of whether or not a student should start in the regular composition course—the one most students take. A panel of expert readers made decisions for those students who did not fit neatly into this course. In making their decisions, the panel of readers could consult and discuss difficult cases instead of following the standardized, objective procedures associated with holistic scoring. Writing program administrators at the University of Cincinnati used a system of portfolio assessment to replace the first-year composition essay exit exam (Roemer, Shultz, & Durst, 1991; Durst, Roemer, & Shultz, 1994). The portfolio scoring system used large group "norming" sessions in conjunction with trios of writing teachers who worked independently to determine if students met the basic requirements to successfully exit the composition program. These alternative systems were still interested in reliability but not in achieving acceptable rates through the conventional approach to holistic scoring.

Others (e.g., Broad, 1994; Lowe & Huot, 1997; and Hester et al., 2007) challenged the traditional holistic scoring approach that characterized most portfolio assessments. Broad (1994) and White (1993, 1995) represented the concerns about reliability that circulated around the use of portfolios as a large-scale assessment method, but writing assessment scholars as a field still did not interrogate the concept of reliability. More recently, White (2005) noted the difficulty in reaching acceptable reliability rates that has plagued portfolio assessments and proposed a scoring method for portfolios "derived conceptually from portfolio theory, rather than essay-testing theory" (p. 583), overturning his earlier position that portfolios are basically just expanded essay tests (White, 1995). Although White (2005) seemed to be advocating a method of portfolio evaluation distinct from holistic

scoring, he describes his approach, which focuses on the reflective letter or self-assessment and clear statements of learning goals, this way:

> Now we can speak sensibly of scoring, even holistic scoring, of the reflective letter, which needs to meet certain quite specific criteria. We are back to a single document, the basic material for which holistic scoring was designed, and we can usually agree on the quality of that document, though we may disagree on the quality of the items in the portfolio that support that document. With some labor, we can come up with a scoring guide and sample portfolios at various score points, just as we can do with single essays. (p. 593)

In short, White's new method was closely aligned with the old one and was designed to streamline the portfolio scoring by focusing on a single text. Granted, he explained how the portfolio contents were used along with the writer's self-assessment, but he still framed of reliability in traditional conventional ways.

While Moss (1994) recognized that reliability standards, within the psychometric tradition, are grounded in fairness to stakeholders, she contends that from a hermeneutic perspective, reliability "can be criticized as arbitrarily authoritarian and counterproductive" (pp. 9-10). In the end, Moss did not argue for abandoning reliability but rather advocated that alternative approaches to assessment theory and practice be considered when appropriate (p.10). Her position is especially relevant for those charged with writing assessments because writing is a complex, multidimensional, contextually situated activity. Importing psychometric theory and practices, especially in terms of reliability, may undermine the very usefulness of a writing assessment's results. However, psychometric theory cannot be dismissed out of hand; instead, writing assessment scholars and practitioners need to draw on language, literacy and psychometric theories as well as other interpretive traditions to design assessments. Some scholars in college composition have done this (Smith, 1992, 1993; Haswell & Wyche, 1994; Broad, 1994, 2003; Lowe & Huot, 1997; Huot, 2002) there are still many assessment practitioners who conform to more narrow approaches, relying on an interrater reliability statistic to demonstrate reliability as we saw with Borrowman (1999).

REFRAMING RELIABILITY

Moss's (1994) argument to reconsider reliability through alternative research traditions appeals to those of us in writing assessment more comfortable with literacy studies, literary theory, and qualitative research methods. However, it

doesn't necessarily resolve some of the conflicts we experience in confronting the pre-occupation with reliability, narrowly conceived, that dominates large-scale assessments. College composition scholars Penrod (2005) and Lynne (2004) argued that psychometric concepts such as validity and reliability are not pertinent for college writing assessment because they are rooted in a positivist epistemology that is incompatible with the social constructivist approaches of writing and meaning-making that inform most of the field's work. Both Penrod (2005) and Lynne (2004) drew on qualitative research traditions. Lynne (2004), who used Guba and Lincoln among other theorists, suggested isolating college composition from educational measurement and developing our own key assessment terms. She offered "meaningfulness" and "ethics" for use instead of validity and reliability (p. 117). While both Penrod's and Lynne's critiques of validity and reliability (and psychometric practices in general) addressed some important concerns, if we attempt to reject reliability, or ignore it, we will make writing assessment more vulnerable to methods and interpretations of results that contradict what we know about literacy and writing, ultimately compromising validity. Like Lynne (2004), Huot (2002) advocated for assessments that are meaningful (p. 101) and acknowledged our responsibilities in writing assessment (p. 57-58), but he called on us to participate as full partners with educational measurement colleagues (p. 57). Psychometric theory is, after all, compatible with what writing teachers and scholars value even if these shared values are not always emphasized in practice (Huot, 2002; O'Neill, Moore, & Huot, 2009).

While Huot's (2002) discussion of validity and reliability have been acknowledged as making a significant contribution to the field of writing assessment, most writing administrators and writing faculty are not seriously engaged in theoretical discussions of assessment in general or reliability in particular. We are often too focused on practice—solving an immediate need, refining an existent assessment procedure—to engage in theoretical debates about assessment terms and principles (Gere, 1980; Faigley et al., 1986). Many people charged with college writing assessments are not composition scholars let alone writing assessment experts. We can't reject basic principles or terms, especially when a term is invested with so much cultural capital and power as reliability, without a better understanding of what reliability brings to the table and what it represents in the wider assessment community.

Since the early 1990s, we have seen an assortment of assessment models (e.g., Smith 1992, 1993; Haswell, 1994, 2001; Broad, 2003) that challenge conventional approaches to reliability; however, most of us are still confronted with demands for reliability narrowly framed or are ill-prepared for discussions about reliability. What happens, in many cases, is that those of us charged with writing assessment, who also identify as literacy teachers and researchers, have found

ourselves in discussions with testing specialists, whether in our institutions or from outside vendors, but unable to communicate clearly with them. We need to think carefully about what values reliability taps into and how they connect to the values we hold about teaching writing and learning to write. As Haertel (2006) concluded in his discussion of reliability—which is not specific to writing assessment—we need "further integration of notions of reliability with evolving conceptions of test validity" (p. 103). We need to understand, as Camp (1993) argued, the principles that reliability encompasses. We also need to think more strategically about identifying what we value and how to communicate that in ways that will be persuasive to others—policymakers, administrators, test developers. As Lakoff (2004) reminded us, language not only provides form for our values, ideas, experiences and thoughts, it helps shape them and how we understand the world around us. This often occurs unconsciously so we need to be intentional and thoughtful about how we use language to frame writing assessment and reliability or we may be undermining our own efforts.

If we think more strategically, as Lakoff (2004) recommends, about how we want to frame reliability—and writing assessment more generally—we need to consider how we use reliability and what it evokes with the educational assessment culture, especially in the field of writing assessment. As noted above, reliability has been a longstanding issue in educational measurement and in writing assessment. It is associated with quantification—measurement, scoring, statistics—and it also evokes validity.

In some sense, however, college writing assessment as a field of study seems ambivalent toward reliability. As Cherry and Meyer (1993) explained, writing assessment practitioners have not been consistent in the methods or presentation of interrater reliability although we keep using the term and providing a co-efficient. By continually referring to reliability and presenting a statistic, we have reinforced the traditional frame for evaluating an assessment. Yet, we have not established consistent methods in determining reliability or even in discussing how we are approaching it and why. As the Hamp-Lyons and Condon (2000) example above illustrates, we seem to try to have it both ways: we report a reliability statistic but what we find most valuable is the discussion and debate we have to develop the community of knowledgeable readers. Peckham (2009) illustrates some of the difficulties college composition as a field has with reliability. Writing about a pilot online placement system in the flagship journal of the Conference on College Composition and Communication, Peckham (2009) addressed scoring of placement essays and the results compared to the placement students received based on the ACT score. He acknowledged Huot's (1996) more recent critique of interrater reliability but then argued for it in terms of values (fairness), which he equated to validity. He wrote:

> Although Brian Huot argues interrater reliability is over valued ("Toward a New Theory," p. 560), I think there is some question about the fairness (and thus the validity) of an assessment if readers frequently disagree on the placements. (p. 521)

Later he explained, in part, the results of the essay scoring he conducted, noting the interrater reliability of the readers (the specific term interrater reliability isn't used). Peckham also acknowledged that the high rate of agreement could be a problem if the raters were consistently wrong but seems to dismiss this as unimportant. He wrote::

> Of course, the notions of "right" or "wrong" are highly suspect in any discussion of writing assessment. The only thing we can say with confidence is that we recommended reassignment for about 42 percent of the students. The percentages in the differences and directions of recommended reassignments over the three years suggest that our five readers, who remained generally the same from year to year, were at least ranking the essays consistently. *We picked the readers from among the best teachers in our program. . . . Unsurprisingly, our agreement rate was high—only 3 to 4 percent of the essays in the three years needed a third reading. Admittedly, reader agreement does not guarantee a valid assessment; my readers could be consistently wrong, but assessment is not a question of wrong or right*: it is about best choices, in this case to place students on the basis of their writing and a controlled scoring or on the basis of multiple-choice exam. (Peckham 2009, p. 535; emphasis added)

Peckham (2009) also addressed other aspects of reliability, such as the reliability of the test itself and connected that to scoring. Two aspects of this selection, excerpted below, are noteworthy: 1) After acknowledging the benefit of using two writing samples, he explained they use one because of "simplicity," which as a value seems to be prized more than reliability; and 2) He implied that he isn't confident in the abilities of the raters' scoring, which seems contradictory to what he said about the raters' agreement above:

> [W]e realize that for a more reliable assessment, we should require at least two essays for two reasons: first, two essays in different genres might increase "test reliability," that is, that given similar testing situations, students will achieve relatively

> similar scores on both tests (White, "Apologia," p. 41); and second, the second essay would allow us to assess the student's ability to respond to a writing task based on one of the major assignments in our second semester course. But *we decided to forgo the probable increase in test reliability for greater simplicity*. Our experience has shown us that it is difficult to train teachers to agree on the criteria and rankings of anchor papers in one genre. *When we are confident about teachers' abilities to score essays in one genre, then we will move to two essays in different genres*. We expect to expand our submissions into electronic portfolios, but that's down the road. (Peckham 2009, p. 526; emphasis added)

Peckham knows reliability is important, but he also seems to indicate that there may be some problems with it as it applies to a writing assessment. He wants the assessment to be fair, and "valid" (p 521) and he believes consistency in the scoring is important (pp. 526 and 535). But he favors simplicity over other concerns about the test. After reviewing research on the correlations between direct and indirect methods of writing assessment, Peckham (2009) concluded that "I would go with the writing simply because we are more nearly looking at what we think we are trying to assess (i.e., the direct method has more testing validity)" (p. 532). Peckham's article illustrates how as a field, there is some degree of uncertainty about how to handle all the nuances and technical components of reliability (and, by extension, validity).

The point in detailing Peckham's references to reliability is not to critique him per se but rather to illustrate the ambivalence we as a field have around reliability and the difficulty we have in addressing it. His article, after all, was published in *CCC* which "reflects the most current scholarship and theory in the field," according to its website, and uses blind peer-review. Because of its publication in *CCC*, Peckham's discussion of reliability also serves as a powerful example of our reluctance to address the concept of reliability more directly and in more theoretically informed ways. It demonstrates how a purely quantitative, statistical approach to reliability does not fit well with what we value. However, it also shows that we recognize the significance of reliability and that there are some positive, useful values that reliability supports, so we cannot dismiss it out of hand. This is what Lynne (2004) realized in her attempt to replace the terms *validity* and *reliability*. However, while we might need to consider the language—as Lynne (2004) suggested—we need to focus on what we value, what concepts are most important, and what ideas are involved when discussing reliability because frames are ultimately about values, ideas and concepts—the

language merely evokes and reinforces the frame. Therefore, we need to be more intentional and thoughtful about the language we use in discussing reliability and writing assessment.

Lakoff (2006) explains that language choice is "vital" because "language evokes frames—moral and conceptual frames" (p. 7). So far, we have allowed the psychometric practitioners (and I would also argue conservative policymakers and their constituencies) to frame reliability in ways that privilege their worldview and support their values. We need to consider ways to reframe reliability so that it evokes the values that literacy teachers hold and support in their research about teaching, learning, and language. Thinking about reliability as a concept, an issue, as well as the frame it evokes and how we can communicate more effectively about what we value, is a role that literacy educators are able to tackle because it shifts the debate away from statistical methods and technical expertise to the concept of reliability, the values it promotes, and the ways these values are communicated (Parkes, 2007).

While few writing teachers and theorists are psychometricians or experts in advanced statistics, many more are experts in language and literacy. We understand communication theory and language development. We know about teaching, learning, and students. We have strong values and beliefs—such as a belief that all children can learn, that all deserve access to quality education, that context is critical in effective writing, and that writing assessment should improve teaching and learning. Smith (1992, 1993) explored multiple aspects of reliability in a series of ongoing studies that were, in effect, a process of validating the locally-designed placement system he developed (O'Neill, 2003). His goal was to make sure that students in his program were placed in the most appropriate first-year writing course. Huot (2002) in arguing for a new theory of writing assessment that values context, local control, rhetorical principles, and accessibility considered reliability as part of the validation process.

Reframing a concept as ingrained and complex as reliability requires a commitment because frames are developed overtime, unconsciously in most cases, through repetition and reinforcement. Everyone has frames—and they are not always theoretically consistent or compatible—although people are usually not aware of them because they function at the unconscious level. To reframe an issue, Lakoff (2006) explained, we need to be strategic. With reliability, we can start by determining how it has been framed and then how we can reframe it in ways that support our beliefs about teaching and learning. One place to start is with the standard reference manuals in the field of psychometrics. To that end, below are excerpts of basic explanations of reliability from the most recent editions of two mainstream measurement reference manuals: *The Standards of Educational and Psychological Testing* (AERA, APA, & NCME, 1999) and

Educational Measurement, 4th edition (ACE, 2006). From the *Standards*, here's the opening paragraph on the section "Reliability and Errors of Measurement":

> A test, broadly defined, is a set of tasks designed to elicit or a scale to describe examinee behavior in a specified domain, or a system for collecting samples of individual's work in a particular area. Coupled with the devise is a scoring procedure that enables the examiner to quantify, evaluate, and interpret the behavior or work samples. *Reliability* refers to the consistency of such measurements when the testing procedure is repeated on a population of individuals or groups. (p. 25)

According to the glossary in the *Standards* (AERA, APA, & NMCE, 1999), reliability is "the degree to which test scores for a group of test takers are consistent over repeated applications of a measurement procedure and hence are inferred to be repeatable for an individual test taker" (p.180). It also includes the "degree to which scores are free of errors of measurement for a given group" (p. 180). Haertel (2006), in the fourth edition of *Educational Measurement*, opens the chapter on reliability this way:

> The concern of reliability is to quantify the precision of test scores and other measurements . . . Like test validity, test score reliability must be conceived relative to particular testing purposes and contexts. The definition, quantification, and reporting of reliability must each begin with considerations of intended test uses and interpretations. However, whereas validity is centrally concerned with the nature of the attributes tests measure, reliability is concerned solely with how the scores resulting from measurement procedure would be expected to vary across replications of that procedure. Thus reliability is conceived in more narrowly statistical terms than is validity. (p. 65)

Both of these explanations highlight the statistical, technical apparatus that typically frames reliability. In this frame, quantification and measurement are invoked. Measurement implies a finite amount of something. This epistemology is associated with objectivity that was the central to psychometrics in the early and mid-twentieth century (Williamson, 1993, 1994). However, in these excerpts, values of consistency and accuracy are also identified. Haertel (2006) even acknowledged context as a value when he notes that it "must each begin with considerations of intended test uses and interpretations" (p. 65) since these aspects of an assessment will define, in part, the particular situation. And in

fact, these are also values that were central to the development of psychometrics. Parkes (2007) argued that reliability as a concept has been conflated with its methodology and that what we need to do is remember that it is the values that are primary. Camp (1993) made a similar point. The methods to demonstrate reliability should not be more important that the values that reliability represents. Parkes (2007) explained it this way:

> The outcomes of the use of these tools—reliability coefficients, dependability coefficients, standard errors of measurement, information functions, agreement indices—serve as evidence of broader social and scientific values that are critically important in assessment. So a reliability coefficient is a piece of evidence that operationalizes the values of accuracy, dependability, stability, consistency, or precision. In practice and in rhetoric, however, the methodologies for evidence reliability are often conflated with the social and scientific values of reliability. (p. 2)

If the methods cannot produce the evidence needed to support reliability, then we need to develop better methods. Parkes (2007) contended that reliability, like validity, needs to be considered as an argument. According to Parkes (2007), a reliability argument has six components, the first and most critical of these is determining the social and scientific values clearly. He argued that in constructing a reliability argument, assessment developers need to

1. Determine the social and scientific values (dependability, consistency, etc.) that are most relevant and decide which ones are most important.
2. Articulate clear statements of the purpose and context of the assessment, which includes making explicit the reasons the information is needed and how it will be used.
3. Define "replication" in the particular context, specifically structural versus conceptual replication.
4. Determine the "tolerance" or level of reliability needed.
5. Collect the evidence from the assessment, which may include traditional reliability data but it might also include other information such as narrative evidence.
6. Pull all of the information together to make the judgment and explaining how the evidence supports the final judgment. (pp. 6-7)

Parkes also emphasized that at the start, it is "easy to think of methods . . . rather than values" first but that it is "critical to stay focused on the value itself" and to determine which value or values are more important than others (p. 6).

Is consistency, for example, more important that stability? Or is precision more important than consistency? At this point, Parkes (2007) explained, it is very important to think about the construct being assessed, which introduces validity into the process. In other words, while reliability is distinct from validity, an appropriate argument for a context-specific form of reliability should be part of any validity argument.

While Parkes did not use Lakoff's concept of reframing, his approach helps us to reframe reliability in ways consistent with Lakoff because Parkes (2007) focused on values, which is what Lakoff recommended in reframing, and both called for articulating values and then using (or developing) methods that support those values. Parkes' (2007) approach to reliability also highlights the significance of purpose and context, which are critical components in effective communication and in assessing writing (Huot, 2002; CCCC, 2006). Haertel (2006) emphasized this point as well: "It bears repeating that in describing score accuracy, the statistics used and the ways they are interpreted must be suitable to the context and purpose of the measurement" (p. 67).

In supporting his approach to reliability, Parkes (2007) used an extended example of a classroom-based assessment of collaboration, performed by a classroom science teacher, to explain how reliability can—and should—work in performance-based assessments of complex, multi-dimensional activities. Using Parkes' (2007) position to reframe reliability in writing assessment would change the focus of the discussion from interrater reliability statistics to issues of purpose, context, evidence, tolerance, and effectiveness without dismissing reliability as unimportant, irrelevant, or impossible. Instead of asking what the statistics are for rater agreement, one might consider other questions, as Smith (1992, 1993) did. Smith reframed the question about reliability of the placement test results. Instead of looking exclusively at the interrater reliability statistic for the group, which was typical, Smith thought about agreement of raters in a much more nuanced way, examining raters' agreement with him/herself as well as within pairs of raters. He also looked at raters' disagreements to see if they were consistent. Ultimately, Smith's focus on reliability was considered in terms of the adequacy of placement: Were students adequately placed into the composition sequence? This reframing put the scoring reliability in the service of the validity of the placement exam results and situated it in terms of the particular writing program and course. Instead of "scoring essays," Smith had teachers placing students into the courses. He still wanted to be sure that students were being placed reliably—would the same student be put in the same course if the essay was read by another reader? By another pair of readers?-but he developed different methods for achieving reliable and valid results (O'Neill, 2003).

While Smith worked with single sample impromptu essays in developing his system, Haswell used single sample impromptus and portfolios to develop a two-tiered expert reader system for a Junior Writing Portfolio assessment program (Haswell & Wyche, 1996). The systems developed by Smith and Haswell, which were implemented over fifteen years before Parkes' (2007) essay, demonstrate one of Parkes' (2007) main points—that the focus of reliability needs to be on the values (such as accuracy, consistency and fairness) associated with reliability within the context of the assessment's purpose and context. By emphasizing this approach, new methods can be developed that produce both reliable and valid results. Parkes' (2007) framework for reliability can also help us communicate more clearly about a writing assessment so that it is framed by our values, purposes and theories.

CONCLUSION

Writing assessment scholars and practitioners have had significant influence in promoting performance-based assessments as well as in developing methods for scoring them (Lane & Stone, 2006). However, these assessment experts have not always been experts in language and literacy but in psychometrics and educational measurement. In many ways, writing specialists have been content to assign reliability and reliability methods to psychometricians, distancing ourselves from it. Parkes' (2007) contention, that reliability (like validity) needs to be considered as an argument, demands language and literacy experts to participate in discussions of reliability because constructing the reliability argument requires knowledge of more than psychometric statistics and methods. Reframing reliability to emphasize our values about writing, teaching writing, and learning to write will emphasize finding methods to build an effective reliability argument instead of merely reporting reliability co-efficients, which scholars have demonstrated to be problematic in writing assessment practice (Cherry & Meyer, 1993; Hayes & Hatch, 1999).

In Parkes' (2007) approach to reliability, writing assessment administrators would need to explain how reliability is being determined, why this approach is appropriate in the particular context, how specifically reliability is being calculated, the threshold for acceptable reliability and a justification for it, the limitations of the reliability, and how reliability contributes to the overall validation of the assessment's results. In determining reliability, many of us responsible for writing assessments should collaborate as equal partners with colleagues who have the statistical expertise. Writing assessment practitioners and scholars need to accept our responsibility to develop and maintain writing assessments that are informed by both language-based and psychometric theory and research.

We need to develop new methods for assessment as well as for determining reliability and validity if current methods do not work adequately for our purposes, as Parkes (2007) argued. This may mean collaborating with others who have different kinds of experiences and expertise, learning more about psychometric theory and practices, and engaging in difficult discussions with colleagues about what we value and why it matters.

By emphasizing values, we can begin to not only reframe reliability but also build more collaborative relationships with the educational measurement community. In writing assessment, this reframing can help writing teachers and administrators discuss and negotiate appropriate writing assessments with institutional administrators and others in more nuanced and effective ways. We must remember that validity and reliability connect to values such as accuracy, consistency, fairness, responsibility, and meaningfulness that we share with others, including psychometricians and measurement specialists. Focusing on these values and working to develop methods for upholding them can lead to the development of writing assessment methods that not only support teaching and learning but also are supported by evidence-based and theoretically-informed arguments. Over time, we will be able to shift the frame associated with reliability away from statistical methods and calculations to values that these methods—as well as methods not yet developed—should be supporting.

I believe we can be successful in our efforts to reframe reliability; after all, we were instrumental in resisting the move away from essay exams made in the 1940s, insisting that student writing needed to be evaluated in writing assessment. This position led to the development of holistic scoring and other methods for evaluating performance assessments (Huot & Neal, 2006; Lane & Stone, 2006). As scholars, teachers, and assessment practitioners, we need to engage in thoughtful ways to reframe reliability so that our assessments serve students and programs as they enact what we know about language and literacy.

REFERENCES

Allen, M. S. (1995). Valuing differences: Portnet's first year. *Assessing Writing*, *2*(1), 67-89.

American Educational Research Association, American Psychological Association, & National Council of Measurement in Education. (1999). *Standards for educational and psychological testing*. American Educational Research Association.

Belanoff, P., & Dickson, M. (Eds.). (1991). *Portfolios: Process and product*. Boynton/Cook.

Black, L., Daiker, D., Sommers, J. & Stygall, G. (Eds.). (1994). *New directions in portfolio assessment: Reflective practice, critical theory and large-scale scoring*. Boynton/Cook.

Black, L., Helton, E., & Sommers, J. (1994). Connecting current research on authentic and performance assessment through portfolios. *Assessing Writing, 1(1)*, 247-266.

Borrowman, S. (1999). Trinity of portfolio placement: Validity, reliability, and curriculum reform. *Writing Program Administration*, *23(1-2)*, 7-27.

Broad, B. (2003). *What we really value: Beyond rubrics in teaching and assessing writing*. Utah State University Press.

Broad, R. L. (1994). "Portfolio scoring": A contradiction in terms. In L. Black, D. Daiker, J. & Stygall, G. (Eds.). *New directions in portfolio assessment: Reflective practice, critical theory and large-scale scoring* (pp. 263-277). Boynton/Cook.

Burke, K. (1966). *Language as symbolic action*. University of California Press.

Calfee, R. & Perfumo, P. (Eds.). (1996). *Writing portfolios in the classroom*. Lawrence Erlbaum.

Camp, R. (1993). Changing the model for the direct assessment of writing. In M. M. Williamson and B. A. Huot (Eds.), *Validating holistic scoring for writing assessment: Theoretical and empirical foundations* (pp. 45-78). Hampton.

Cherry, R. D., & Meyer, P. R. (1993). Reliability issues in holistic assessment. In M.M. Williamson and B. A. Huot (Eds.) *Validating holistic scoring for writing assessment: Theoretical and empirical foundations* (pp. 109-141). Hampton.

Common Core State Standards Initiative. n.d. National Governors Association and Council of Chief State School Officers. http://www.corestandards.org/

Conference on College Composition and Communication. (Nov. 2006). Writing assessment: A position statement (Rev. ed.). National Council of Teachers of English. http://www.ncte.org/cccc/resources/positions/123784.htm

Daiker, D. A., Sommers, J., & Stygall, G. (1996). Pedagogical implications of a college placement portfolio. In E. M. White, W. D. Lutz, & S. Kamusikiri (Eds.), *Assessment of writing: Politics, policies, practices* (pp. 257-270). Modern Language Association.

Diederich, P. B. (1974). *Measuring growth in English*. National Council of Teachers of English.

Durst, R. K., Roemer, M., & Schultz, L. (1994). Portfolio negotiations: Acts in speech. In L. Black, D. Daiker, J. Sommers & G. Stygall (Eds.) *Validating holistic scoring for writing assessment: Theoretical and empirical foundations* (pp. 286-300). Boynton/Cook.

Elbow, P. & Belanoff, P. (1986). Staffroom interchange: Portfolios as a substitute for proficiency examinations. *College Composition and Communication*, *37(3)*, 336-339.

Elliot, N. (2005). *On a scale: A social history of writing assessment in America*. Peter Lang.

Faigley, L, Cherry R. D., Jolliffe, D. A., & Skinner, A. M. (1985). *Assessing students' knowledge and processes of composing*. Ablex.

Gere, A. R. (1980). Written composition: Toward a theory of evaluation. *College English*, *42(1)*, 44-48, 53-58.

Godshalk, F. I., Swineford, F. & Coffman, W. E. (1966). *Measurement of writing ability*. (CEEB RM No. 6.) Educational Testing Service.

Haertel, E. H. (2006). Reliability. In R. L. Brennan (Ed.), *Educational measurement* (4th ed.) (pp. 65-109). ACE Praeger Series in Higher Education.

Hall, S. (1983). The narrative construction of reality. *Context*. https://www.academia.edu/34971009/Stuart_Hall_The_Narrative_Construction_of_Reality_1984_

Hamp-Lyons, L. & Condon, W. (2000). *Assessing the portfolio: Principles for practice, theory, and research*. Hampton Press.

Haswell, R. H. (Ed.). (2001). *Beyond outcomes: Assessment and instruction within a university writing program*. Ablex.

Haswell, R. H., & S. Wyche. (1996). A two-tiered rating procedure for placement essays. In T. W. Banta (Ed.), *Assessment in practice: Putting principles to work on college campuses* (pp. 204-207). Jossey-Bass.

Haswell, R. H., Johnson-Shull, L. & Wyche-Smith, S. (1994). Shooting Niagra: Making portfolio assessment serve instruction at a state university. *Writing Program Administration*, *18,* 44-54.

Hayes, J. R. & Hatch, J. (1999). Issues in measuring reliability. *Written Communication*, *16*, 354-367.

Hertog, J., & McLeod, D. 2001. A multiperspectival approach to framing analysis: A field guide. In S. D. Reese, O. H. Gandy, & A. E. Grant (Eds.), *Framing public life* (pp. 131-161). Lawrence Erlbaum.

Hester, V., O'Neill, P., Neal, M., Edgington, A., & Huot, B. (2007). Adding portfolios to the placement process. In P. O'Neill (Ed.), *Blurring boundaries: Developing writers, researchers, and teachers* (pp. 61-90). Hampton Press.

Hoetker, J. (1982). Essay examination topics and student writing. *College Composition and Communication*, *33(4),* 377-392.

Huot, B. (2002). *(Re)Articulating writing assessment for teaching and learning*. Utah State University Press.

Huot, B. & Neal, M. (2006). Writing assessment: A techno-history. In C. A. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (pp. 417-432). Guilford Press.

Kearns, E. (1993). On the running boards of the portfolio bandwagon. *Writing Program Administration*, *16(3)*, 50-59.

Kuhn, T. S. (1962). *Structure of scientific revolutions*. University of Chicago Press.

Lakoff G. (2002). *Moral politics: How liberals and conservatives think* (2nd ed). University of Chicago Press.

Lakoff, G. (2004). *Don't think of an elephant! Know your values and frame the debate*. Chelsea Publishing.

Lakoff, G. (2006) Simple framing. *Rockridge Institute*. http://www.rockridgeinstitute.org/projects/strategic/simple_framing

Lane, S. & Stone, C. A. (2006). Performance assessment. In R. L. Brennan (Ed.) *Educational measurement* (4th ed). (pp. 387-431). ACE Praeger Series in Higher Education.

Larson, R. L. (1996). Portfolios in the assessment of writing: A political perspective. In E. M. White, W. D. Lutz, & S. Kamusikiri (Eds.), *Assessment of writing: Politics, policies, practice*. (pp. 271-283). Modern Language Association.

LeMahieu, P. G., Eresh, J. T., & Wallace, R. C. (1992). Using student portfolios for a public accounting. *School Administrator*, *49*(11), 8-13.

LeMahieu, P. G., Gitomer, D., & Eresh, J. (1995). Portfolios in large scale assessment: Difficult but not impossible. *Educational Measurement: Issues and Practice*, *14(3*), 11-28.

Lowe, T. J., & Huot, B. (1997). Using KIRIS writing portfolios to place students in first-year composition at the University of Louisville. *Kentucky English Bulletin*, *46*, 46-64.

Lynne, P. (2004). *Coming to Terms: A theory of writing assessment*. Utah State University Press.

Messick, S. (1989). Meaning and value in test validation: The science and ethics of assessment. *Educational Researcher*, *18(2)*, 5-11.

Mislevy, R. J (2004). Can there be validity without "reliability?" *Journal of Educational and Behavioral Statistics*, *29(2)*, 241-245.

Moss, P. A. (1994). Can there be validity without reliability? *Educational Researcher*, *23*(4), 5-12.

Murphy, S. & Underwood, T. (2000). *Portfolio practices: Lessons from schools, districts and states*. Christopher Gordon.

Nelson, A. (1999). Views from the underside: Proficiency portfolios in first-year composition. *Teaching English in the Two Year College*, *26*, 243-253.

Nystrand, M., Cohen, A., & Dowling, N. (1993). Addressing reliability problems in the portfolio assessment of college writing. *Educational Assessment*, *1*(1), 53-70.

O'Neill, P. (2003). Moving beyond holistic scoring through validity inquiry. *Journal of Writing Assessment*, *1(1)*, 47-65. https://escholarship.org/uc/item/4qp611b4

Parkes, J. (2007). Reliability as argument. *Educational Measurement: Issues and Practice*, *26*(4), 2-10.

Peckham, I. (2009). Online placement in first-year writing. *College Composition and Communication*, *60(3)*, 517-540.

Penrod, Diane. (2005). *Composition in convergence: The Impact of new media on writing assessment*. Lawrence Erlbaum.

Roemer, M., Schultz, L. M., & Durst, R. K. (1991). Portfolios and the process of change. *College Composition and Communication*, *42(4)*, 445-469.

Shale, D. (1996). Essay reliability: Form and meaning. In E. M. White, W. D. Lutz, & S. Kamusikiri (Eds.), *Assessment of writing: Politics, policies, practices*. (pp. 76-96). Modern Language Association.

Smith, W. L. (1992). The importance of teacher knowledge in college composition placement testing. In J. R. Hayes (Ed.), *Reading empirical research studies: The rhetoric of research*. (pp. 289-316). Ablex.

Smith, W. L. (1993). Assessing the reliability and adequacy of using holistic scoring of essays as a college composition placement program technique. In M. M. Williamson and B. A. Huot (Eds.), *Validating holistic scoring for writing assessment: Theoretical and empirical foundations* (pp. 142-205). Hampton Press.

Sommers, J., Black, L., Daiker, D., & Stygall, G. (1993). Challenges of rating portfolios: What WPAs can expect. *Writing Program Administration*, *17(1-2)*, 7-29.

Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research & Evaluation*, *9*.

U.S. Department of Education. Race to the Top Fund. (n.d). http://www2.ed.gov/programs/racetothetop/index.html

Underwood, T., & Murphy, S. (1998). Interrater reliability in a California middle school English/Language Arts portfolio assessment program. *Assessing Writing*, *5*(4), 201-230.

White, E. M. (1993). Holistic scoring: Past triumphs and future challenges. In M. M. Williamson and B. A. Huot (Eds.), *Validating holistic scoring for writing assessment: Theoretical and empirical foundations* (pp. 79-108). Hampton Press.

White, E. M. (1994). *Teaching and assessing writing* (2nd ed). Jossey Bass.

White, E. M. (1995). Apologia for the timed impromptu essay test. *College Composition and Communication*, *46(1)*, 129-139.

White, E. M. (2005). The scoring of writing portfolios: Phase 2. *College Composition and Communication, 56(4),* 581-600.

Whithaus, C. (2005). *Teaching and evaluating writing in the age of computers and high-stakes testing*. Erlbaum.

Wiggins, G. (1993). Constant danger of sacrificing validity to reliability: Making writing assessment serve writers. *Assessing Writing*, *1(1)*, 129-139.

Willard-Traub, M., Decker, E., Reed, R., & Johnston, J. (1999). Development of large-scale portfolio placement at the University of Michigan 1992-1998. *Assessing Writing*, *6*(1), 41-84.

Williamson, M. M. (1993). Introduction to holistic scoring: The social, historical, and theoretical context for writing assessment. In M. M. Williamson and B. A. Huot (Eds.), *Validating holistic scoring for writing assessment: Theoretical and empirical foundations* (pp. 1-43). Hampton Press.

Williamson, M. M. (1994). Worship of efficiency: Untangling theoretical and practical considerations in writing assessment. *Assessing Writing*, *1,* 147-174.

Williamson, M. M. (2003). Validity of automated scoring: prologue for a continuing discussion of machine scoring student writing. *Journal of Writing Assessment*, *1(1)*, 85-104. https://escholarship.org/uc/item/8nv3w3w8

Yancey, K. B. (Ed.) (1992). *Portfolios in the writing classroom: An introduction*. National Council of Teachers of English.

Yancey, K. B. (1999). Looking back as we look forward: Historicizing writing assessment. *College Composition and Communication*, *50(3)*, 483-503.

Yancey, K. B., & Weiser, I. (1997). *Situating portfolios: Four perspectives*. Utah State University Press.

CHAPTER 2.

VALIDITY INQUIRY OF RACE AND SHARED EVALUATION PRACTICES IN A LARGE-SCALE, UNIVERSITY-WIDE WRITING PORTFOLIO ASSESSMENT

Diane Kelly-Riley
Washington State University

This article examines the intersections of students' race with the evaluation of their writing abilities in a locally-developed, context-rich, university-wide, junior-level writing portfolio assessment that relies on faculty articulation of standards and shared evaluation practices. This study employs sequential regression analysis to identify how faculty raters operationalize their definition of good writing within this university-wide writing portfolio assessment, and, in particular, whether students' race accounts for any of the variability in faculty's assessment of student writing. The findings suggest that there is a difference in student performance by race, but that student race does not contribute to faculty's assessment of students' writing in this setting. However, the findings also suggest that faculty employ a limited set of the criteria published by the writing assessment program, and faculty use non-programmatic criteria—including perceived demographic variables—in their operationalization of "good writing" in this writing portfolio assessment. This study provides a model for future validity inquiry of emerging context-rich writing assessment practices.

The best defense against inequitable assessment is openness. Openness about design, constructs, and scoring will bring out into the open the values, and biases of the test design process, offer and opportunity for debate about cultural and social influences, and open up the relationship between the assessor and the learner.

– C. Gipps

DOI: https://doi.org/10.37514/PER-B.2024.2166.2.02

An African American student came to the Writing Assessment Office at our western, land-grant public university and stated that she had heard that Black students failed our mid-career, university-wide Writing Portfolio at higher rates than other students. My office staff and I could not answer her because, since the program's inception in 1991, the Writing Assessment Office had never collected information regarding student race or ethnicity. The Writing Assessment Program fashioned itself as progressive: we administered a different kind of test than standardized ones so widely disparaged in writing circles. Our test was a portfolio that required students to turn in work produced for their regular coursework as well as complete an impromptu writing sample. A diagnostic evaluation was made by faculty from across the disciplines regarding the level of support needed for the student to successfully navigate the upper-division discipline-specific writing in the major courses required at our institution. Faculty raters used shared evaluation methodologies in which local context drives the articulation of assessment standards. As such, the connection between assessment, instruction, and curricular context was much stronger than standardized tests since much of the evaluation was based on coursework produced in undergraduate classroom settings, and the shared evaluation methodology relied on the expertise of classroom teachers in making these judgments. Students either passed the assessment or demonstrated a need for additional help, "Needs Work," mitigating the stakes for the test. The worst thing that happened to students was they were required to take structured instructional support as they navigated their upper-division writing requirements. The "Needs Work" rating did not follow the students: once they passed the additional coursework, the students' Writing Portfolio ratings were recorded as "Pass" on their university transcripts. In other words, students couldn't "fail" the Writing Portfolio. Program administrators tended to be satisfied with innovations developed for testing and contributions of the new shared evaluation rating procedures of our program, and adopted a stance consistent with other writing assessment scholars who claimed that "the advantages of portfolio assessment [had] overridden its problems, and as we [moved] into the twenty-first century portfolios achieved standing as the writing assessment method of choice" (White, 2005, p. 583). However, such a stance is detrimental to furthering an understanding of the complexity of shared evaluation practices in performance-based assessments and the effects they have on students. Schmidt and Camara (2004) confirmed the promise subscribed to performance assessments to

> reduce differences among groups because they provide students with hands-on opportunities to demonstrate their knowledge and understanding of how to solve problems

> rather than requiring students to simply recall facts. . . . Unfortunately, few large scale studies have examined differences among racial groups on performance assessments. (p. 193)

The one notable exception would be Breland et al.'s (2004) inquiry into the 'new' SAT which found "no significant prompt type effects for ethnic, gender or language groups, although there were significant differences in mean scores for ethnic and gender groups for all prompts" (p. 1). Cary-Lemon (2009) notes that "discourse about 'race' in [Composition Studies] reflects a fluctuating scholarly space" (W12), and argues for a self-critical look at the topics we have examined within our field related to race to see what has been included and excluded in our inquiries to examine these "reflections of racialized ideology over time" (W2).

While writing portfolio assessment tends to feel better to administrators and teachers, a limited number of quantitative or qualitative validational studies have been conducted through the revised framework of validity inquiry (AERA, APA, NCME, 1999; Kane, 2006). Such inquiries need to consider the interpretation and use of test scores as well as their consequences for students who take them. Kane (2006) asserts that validation "involves the development of evidence to support the proposed interpretations and uses [of test results] . . . to show that [such use] is justified . . . [and to assess] the extent to which the proposed interpretations and uses are plausible and appropriate" (p. 17). Perhaps owing to validity's psychometric roots, scholars in composition studies have had a general mistrust of validity research (Sharton, 1996; Lynne, 2004; Murphy, 2007). O'Neill (2003) documents how "validity has been—and continues to be—misconstrued in most of composition's assessment literature" (p. 49) highlighting the troubling "lack of rigorous composition research" (p. 51) into writing assessment methods regarding validity. Haswell (2005) also noted a general lack of replicable, aggregable, and data-driven scholarship in composition studies, characterizing the situation as an all-out war against this type of inquiry.

In spite of this, scholars have called for attentiveness to issues of validity in testing and assessment. Huot (1996) called for a "theory of writing assessment . . . [that recognizes] the importance of context, rhetoric, and other characteristics integral to a specific purpose and institution" (p. 552) and laid the groundwork for researchers to investigate composition-related issues of validity. The revised concept articulated in the Standards states that "validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by the proposed uses of tests" (AERA, 1999, p. 9). Validity inquiries should include examinations of the consequences to the individuals taking the tests, and are no longer just comprised of different and individual components of validity (construct, content, predictive). The process of validation involves accumulating evidence to

provide a sound scientific basis for the proposed score interpretations. It is the interpretations of test scores required by proposed uses that are evaluated, not the test itself. When test scores are used or interpreted in more than one way, each intended interpretation must be validated. (AERA, p. 9) Kane (2006) asserted that

> validation employs two kinds of argument. An *interpretive argument* specifies the proposed interpretations and uses of test results by laying out the network of inferences and assumptions leading from the observed performances to the conclusions and decisions based on the performances. The *validity argument* provides an evaluation of the interpretive argument. (p. 23)

The relevance of validity to writing assessment practitioners is apparent when validity is understood as an ongoing argument to be made rather than a static state to be achieved and justified. O'Neill (2003) contends that "validation arguments are rhetorical constructs that draw from all the available means of support" (p. 50). Huot and Schendel (1999) assert that validity and "assessment must be discussed in the context of ethics, for the consequences of assessment procedures are closely tied to the political and social contexts in which they take place" (p. 40). O'Neill (2003) argues that such lines of inquiry and research "[demonstrate] how systematic, ongoing validity research [function] to enhance a particular local test and contributes—both theoretically and practically to the scholarship of writing assessment" (p. 48). However, in spite of innovations and implementations of new contextually-based college writing assessment practices, systematic and rigorous validity inquiry into emerging college writing assessment practices have been limited.

O'Neill notes the reductive tendency in composition studies to simplify validity to mean "honesty . . . accuracy . . . and rightness" (2003, p. 49) that limits the complexity of the construct. There are many important theoretical calls for the discipline to wrestle with validity issues contextually or hermeneutically (Huot, 1996; Huot and Schendel, 1999; Moss, 1998a; Murphy, 2007; Inoue, 2007) and few forays of actual research and practice into validity inquiry in college writing assessment (Smith, 1993; Williamson and Huot, 1993; Haswell, 1998a and 2000; Broad, 2000; O'Neill, 2003; Hester, O'Neill, Neal, Edgington, & Huot, 2003; Elliot, Briller, & Joshi, 2007; Gere, Aull, Green and Porter, 2010). Researchers and scholars have neglected to conduct validity inquiries of locally developed writing assessment practices and so have not documented contributions or innovations these practices embody, and they fail to be attentive to students who take the exams. Kane (2006) says "there are, potentially, a large number of assumptions in any interpretive [validational] argument. We

take many of these assumptions for granted, at least until evidence to the contrary develops" (p. 23). To unearth some of these assumptions, previous scholars' criticism of standardized testing helps articulate where to begin: "what kind of proof do we have that students are wrong when they say, 'I don't belong in this dummy class'?" (Elbow, 1996, p. 93). While Elbow originally leveled this question at holistic or standardized tests, it is still relevant as a question for writing assessment programs that employ shared evaluation practices—locally developed, context-rich, practices that rely on faculty articulation values—whether via portfolios, direct-self placement, or other methods. Students who don't meet standards for writing tests face consequences that require completing additional coursework, spending additional time, spending additional money (perhaps), and dealing with the stigma of not passing the "test". Moss (1995) cites Cronbach and argues that "when the anticipated consequences [of assessment] 'impinge on the rights and life chances of individuals' (Cronbach, 1988, p. 6) . . . the investigation of consequences becomes particularly salient" (p. 11).

Rigorous validity inquiry allows for in-depth investigation of issues that we observe anecdotally—from student outrage at perceived unfair testing practices to patterns of course enrollment that may have more students of color populating the required writing support courses. Rigorous validity inquiry enables informed practice in a setting and directly addresses concerns of power highlighted by Huot and Williamson (1997) who note "assessment procedures [are] instruments of power and control, revealing so-called theoretical concerns as practical and political" (p. 44). They "fear that unless we make explicit the important power relationships in assessment, portfolios will fail to live up to their promise to create important connections between teaching, learning and assessing" (p. 44). Such a fear is applicable to any form of writing assessment that uses shared evaluation practices, particularly as these issues relate to test fairness. Camilli (2006) asserts while there are many aspects of fair assessment, it is generally agreed that tests should be thoughtfully developed and that the conditions of testing should be reasonable and equitable for all students . . . fairness issues are inevitably shaped by the particular social context in which they are embedded. (p. 221)

Certainly, as Schmidt and Camara (2004) observe, there have been "persistent score differences among racial groups" (p. 189) for a variety of standardized tests. Similar studies for performance-based assessments are still inconclusive but suggest that "subgroup gaps on traditional tests remain for [performance based] assessments" (p. 193). Most of this research, though, has occurred at the primary and secondary school level and not the college level.

Camilli (2006) states that "large differences are commonly encountered in test scores among groups of different races and ethnicities, and it is important to understand the extent to which these differences are artifacts of a test rather than

true proficiency" (p. 243). To address this, I conducted an empirically-based, descriptive validity inquiry into the large-scale writing portfolio assessment responsive to the African American student's question at my university. This inquiry begins by examining general performance trends by student race. It then conducts a sequential regression analysis into the construct of good writing as applied in the shared evaluation methodology used to assess the Writing Portfolio to identify the variables that raters actually use in the evaluation of students' writing, and to see if race is among them. This validity inquiry follows Moss' (2007) identification of

> productive directions for research in validity theory . . . [to develop] cases for validity research to both illustrate validity theory and to critique it . . . [including] cases as empirically based descriptions of the actual practices of working scientists, and . . . cases as critical analyses that locate our theories and practices in the sociohistorical-political contexts in which they are developed and used. (p. 96)

The question posed by the African American student regarding students of colors' performances on the Writing Portfolio opened up an avenue of research relevant for college writing assessment: Could the shared evaluation processes used by the university-wide Writing Portfolio assessment—and by other contextually defined writing assessment practices—be inadvertently complicit in perpetuating a system of discrimination? In other words, could teachers/evaluators unwittingly be disadvantaging students of color in a large-scale writing assessment program because of unstated biases related to race?

For this study, the operational definition of race is based upon the categories employed by my institution for collecting data related to race. These categories were based on an older definition of racial designations articulated by the federal Office of Management and Budget. These categories were not based on the most recent 1997 OMB revision to these designations articulated in Camilli (2006). The categories used in this study are American Indian or Alaska Native; Black or African American; Asian, Pacific Islander, Native Hawaiian; Hispanic or Latino; and White. This results in a less than nuanced view of race in this study, and, along with others, I recognize the limitations in such categorizations of race. Specifically, the American Anthropological Association (1997) asserted:

> Race and ethnicity both represent social or cultural constructs for categorizing people based on perceived differences in biology (physical appearance) and behavior. Although popular connotations of race tend to be associated with biology

> and those of ethnicity with culture, the two concepts are not clearly distinct from one another.

The APA Task Force on Diversity Issues at the Precollege and Undergraduate Levels of Education in Psychology (1998) argued that:

> "Race" has social meaning often accompanied by stereotyping; it suggests one's status within the social system and introduces power differences as people of different "races" interact with one another. 'Ethnicity,' on the other hand, connotes common culture and shared meaning. It includes feelings, thoughts, perceptions, expectations, and actions of a group resulting from shared historical experiences.

This study represents a starting point for this type of research, and hopefully future studies can include more complex representations of race and ethnicity.

VALIDITY INQUIRY AND WRITING PORTFOLIO INNOVATIONS

In the early 1990's, validation efforts for this program's Writing Portfolio focused on the Simple Pass methodology as this affected the largest number of students (about 60% of students who completed the Portfolio—roughly 2500 students out of 4200 who complete their Writing Portfolios each year), and at the time presented the most controversial and innovative contribution to the field of college writing assessment. The methodology of the writing assessment system represented a shift in writing assessment practices from holistic writing assessment—in which raters assigned an external numeric value to students' writing—to the expert-rater system (Haswell and Wyche-Smith, 1996; Haswell, 1998b; Haswell, 2001), a shared evaluation methodology which relies on context and teachers' judgments about students' abilities to manage the writing challenges of specific courses. The shared evaluation system used by our institution was based on the placement work of William Smith (1993) and was adapted to an upper-division context. Faculty raters review impromptu writing exams that sort writing obviously ready for upper-division writing intensive coursework from writing that was either very strong or very weak. Writing at either end of the spectrum—weak or strong—was sent on for further review and consultation by more experienced raters. The process assumed that additional focused rating time, information about the student's writing abilities through three additional course paper submissions, and faculty expertise would ensure the validity of the Writing Portfolio results. Virtually no validity attention was given to the results of "Needs Work," perhaps because such

widespread feeling existed among faculty about the poor quality of student writing. And, perhaps, the shift from holistically evaluating writing to relying on an innovative system of evaluation represented a significant enough move to not immediately surface new issues that embedded in the new methodology.

In response to the African American student's question, I investigated students of colors' performances on the Writing Portfolio for Academic Year 2004-05 according to the racial classifications collected by my institution. At that time, this institution reported the demographic profile of undergraduate students as 76 percent White; 1 percent American Indian Alaskan Native; 2 percent Black; 6 percent Asian Pacific Islander (API); 4 percent Hispanic; 3 percent non-resident aliens; and 8 percent unknown. Students' racial affiliation was obtained from this institution's Institutional Research Office by U.S. Census Bureau/ OMB categories, and then was combined with students' Writing Portfolio results. Tables 1.1 and 1.2 document the difference in performance percentages for the impromptu exam portion of the Writing Portfolio and the final review of the entire Writing Portfolio.

Tables 1.1 and 1.2 illustrate an unevenness in performance on the Writing Portfolio by race. Simply examining the percentages does not indicate whether these differences are significant. An analysis of variance was conducted on the performances of students on the timed writing portion of the Writing Portfolio by race. A random sample of 508 timed writing records were selected from the 5347 Writing Portfolio records recorded during AY 2004-2005. Students who spoke English as a second language were omitted from this analysis. An analysis of variance showed that the difference in performance by race on the timed writing portion of the Writing Portfolio was significant, $F(4, 503)=6.032$, $p=.000$. Post hoc analyses using Tukey's LSD for significance indicated that Black ($M=1.58$, $SD=.496$), API ($M=1.59$, $SD=.509$), and Hispanic ($M=1.7$, $SD=.462$) students' timed writing performances were significantly lower than White students ($M=1.84$, $SD=.550$).

Table 1.1. Comparison of Performance Rates on the Writing Portfolio Impromptu Exam by Race

Population	Pass	Distinction	Needs Work
Combined—all students	58.8%	8.6%	32.6%
American Indian	52.2%	8.7%	39.1%
API	47.3%	6.1%	46.6%
Black	48.6%	4.3%	47.1%
Hispanic	55.7%	5.1%	39.2%
White	60.3%	8.2%	31.5%

Note. Source: Writing Assessment Office, Database, (AY 2004-2005)

Table 1.2. Comparison of Performance Rates on the Final Writing Portfolio Review by Race

Population	Pass	Distinction	Needs Work
Combined—all students	78.1%	8.6%	13.3%
American Indian	82.6%	8.7%	8.7%
API	71.6%	5.3%	22.9%
Black	77.6%	2.9%	20.0%
Hispanic	82.3%	3.8%	13.9%
White	82.4%	6.9%	10.7%

Note. Source: Writing Assessment Office, Database, (AY 2004-2005)

Additionally, a second ANOVA was run to compare students' performances by race for the final Writing Portfolio review. A random sample of 749 final Writing Portfolio performances by race was selected from the 5378 available records for AY 2004-2005. Again, multi-lingual speakers were omitted from this analysis. The results indicated a significant difference in the performance on the final Writing Portfolio by race, $F(4, 744)=3.120$, $p=.015$. Post hoc analyses using Tukey's LSD for significance indicated that Black students ($M=-1.81$, $SD=.429$) performed significantly lower on the final Writing Portfolio review than all other students: American Indian ($M=2.00$, $SD=.434$), API ($M=1.97$, $SD=.412$), Hispanic ($M=1.93$, $SD=.411$), and White ($M=1.94$, $SD=.493$).

An analysis that ended here would purely speculate about the reasons underlying the differences in performance by race and wouldn't address "the extent to which these differences are artifacts of a test rather than true proficiency" (Camilli, p. 243) or whether they are result of a "construct-irrelevant variance [which] refers to the degree to which test scores are affected by processes that are extraneous to its intended construct" (AERA et al., 1999, p. 10). The *Standards* (1999) state:

> The idea that fairness requires overall passing rates to be comparable across groups is not generally accepted in the professional literature. Most testing professionals would probably agree that while group differences in testing outcomes should in many cases trigger heightened security for possible sources of test bias, outcome differences across groups do not in themselves indicate that a testing application is biased or unfair. (AERA, 1999, p. 75)

Breland et al.'s (2004) study approached testing difference by race from the perspective of reliability, but as Broad (2000) noted, writing assessment scholars

tend to feel a tension between what he characterized as 'validity and reliability debates' that occur between positivistic and hermeneutic traditions. Complicating this issue further, the Writing Portfolio uses non-parametric data for its system of measurement—ratings are recorded as Needs Work, Pass, or Pass with Distinction—and so have limited transferable numeric value resulting in equally limited statistical analyses. The question raised by the African American student was apt because it highlighted our own program's general tendency—as well as that of composition studies—to neglect to attend to students of colors' experiences in our writing assessment systems.

In composition studies, there has been a great deal of agenda setting and calls for research regarding potential biases against students of color in writing assessment practices, (Farr and Nardini, 1996; Lippi-Green, 1997; Mountford, 1999; Hamp-Lyons and Condon, 2000; Murphy, 2007) but no empirical or qualitative inquiry into students of color's actual experiences in college-level writing assessment systems. Farr and Nardini (1996) suggest that a dominant paradigm of writing instruction exists called "essayist literacy" in which "high value is placed on language, either oral or written, that is rational, decontextualized, explicit, and carefully ordered internally" (p. 108), "[and] . . . the social and cultural mindset that construes rationality, explicitness and order as fundamental values of literate text—namely, the (primarily white and male) Anglo-American analytic orientation" (p. 117).

Other researchers note possible deleterious effects of race applicable to context-rich assessment situations. Omi and Winant (1994) describe how racial formation occurs in everyday face-to-face experience in "the many ways in which, often unconsciously, we 'notice' race . . . One of the first things we notice about people when we meet them (along with their sex) is their race. We utilize clues about *who* a person is" (p. 59). They argue that "our ability to interpret racial meanings depends on preconceived notions of a racialized social structure. . . . We expect people to act out their apparent racial identities" (p. 59). In an assessment system predicated on faculty articulation of values, could Writing Portfolio raters have unstated expectations for student writing and who they think might "write" like students of color resulting in biased assessment of their writing? Moss and Shutz (2001) assert "even in the most intimate settings, issues of inequality, cultural and racial difference, gender, and class affect dialogues in subtle ways giving some voices more authority while silencing others" (p. 42). Ball (1997) concluded that holistic writing assessment procedures used in middle schools disadvantaged African American students because they did not share the same linguistic features as middle-class, Anglo American students and the middle class European teacher who evaluated their writing. In an assessment context, such findings are troublesome because these instances suggest an unfair educational

system and that the assessments may be perpetuating distressing consequences on particular groups of students who take these tests. Bond (1995) argued that

> performance assessments are, at least potentially, less biased and more fair to traditionally disadvantaged students because such tests, when properly used, can merge instruction and assessment rather than test abilities . . . that are only remotely connected to the everyday experience of these students (p. 21).

Bond concurs with writing assessment researchers who tout the value of portfolio assessment, but warns that performance assessments still have significant unresolved issues regarding bias and validity. The lack of straightforward validity evidence for portfolio assessment is corroborated by LeMahieu et al. (1995) and Griffee (2002). In particular, Bond cautions that examination of consequential aspects of validity should "not only [include] the elimination of elements in assessment that unduly *disadvantage minority persons* but also the elimination of construct-irrelevant elements that may subtly *advantage majority persons over others"* (p. 23) by asserting:

> People also hold purely prejudicial beliefs that can affect their objective assessment of others' ability . . . it would take an extraordinary effort on my part to give the same evaluation to two individuals who are identical in every way except that one has a high British accent, and the other a deep southern drawl! (pp. 23-24)

Taken in the context of a shared evaluation setting, Bond implies the potential for raters to privilege or diminish students' writing based on how the writing fits a pre-conceived notion of 'good writing' and that this definition of good writing may be susceptible to bias.

Moss (1998a) critiqued limitations in our program's early forays into validity inquiry of the junior-level Writing Portfolio assessment (Haswell 1998a) advocating that our program consider "to what extent . . . the writing program [is] complicit in simply reproducing a narrow model of academic writing (and the understanding of knowledge it entails) without providing opportunity for the values implicit in the model to be illuminated and self-consciously considered" (p. 120). Moss (1998b) argues and Huot and Schendel (1999) later reiterate that "we need to study the actual discourse and actions that occur around products and practices of testing" (p. 7). In a shared evaluation system, then, a primary validity focus should be on how faculty articulate and operationalize the standards of good writing for the particular context. For this study, students' readiness for upper-division, disciplinary-specific writing in the major work represents that

context. The Writing Portfolio assessment requires the articulation of commonalities of student readiness requisite for their entry into diverse, disciplinary-specific discourse communities. Given the differences in the general Writing Portfolio performance data between different racial groups, consideration of race is key in how faculty operationalize standards for good writing.

METHODS

This validity inquiry examines how raters functionally define good writing through sequential regression analysis techniques that examine actual student products submitted for the Writing Portfolio. These analyses are conducted through three frameworks: the Writing Portfolio assessment criteria, an Alternate set of writing criteria, and Demographic criteria. Each of these frameworks are applied to the two distinct writing tasks—impromptu writing and coursework written for regular undergraduate courses across the disciplines—selected for inclusion in the Writing Portfolio as representative of the student's best writing.

This inquiry allows for more sophisticated statistical analysis of the factors that account for the variability in the writing quality scores of the Writing Portfolio using a finer grained instrument, the Writing Portfolio Differential Scale for Writing and Demographic Information (see Appendix A). The Writing Portfolio Differential Scale was developed by this researcher to interpret raters' evaluation behaviors and determine the criteria they seemed to actually use to evaluate writing; the criteria that seemed to carry more weight in their evaluation process; and whether demographic features perceived about writers accounted for any part of the evaluation results. Guiding questions for this inquiry include:

1. What is the definition of "good writing" that faculty raters apply when evaluating the Writing Portfolio?
2. Do faculty raters make demographic assumptions about students based on their writing that effect the results?
3. Does this evaluation privilege forms of writing according to race?

Sequential regression analysis is used to assess the relationship between a dependent variable (like writing quality) and several independent variables (like criteria that comprise quality—focus, organization, or use of Standard American English and so on) by entering variables in a specific order into regression equations to identify which variables account for the variability in—or the criteria that comprise—the overall score (Tabachnick & Fidell, 2006). The methodology for this study was piloted in an earlier project by the researcher in which

the Writing Portfolio Differential Scale was tested and the order of the criteria variables were established for the regression analysis (Kelly-Riley, 2006).

The Writing Portfolio Differential Scale for Writing and Demographic Information was adapted from the work of Piche, Rubin, Turner, and Michlin (1978) and Osgood (1957). Piche et al. used the work of Osgood to examine whether teachers evaluated Black elementary students' writing differently from their White counterparts. Osgood created semantic differential scales that "relate to the functioning of representational processes in language behavior and hence may serve as an index of these processes" (p. 9). Osgood's work developed out of experimental psychology to establish pairs that exist in what he called semantic space, "which are assumed to represent a straight line function that passes through the origin of this space, and a sample of such scale then represents a multidimension space" (p. 25). His work attempts to quantify the complexity inherent in measuring a construct like writing. Piche et al. (1978) developed their scale items based on the research of Osgood (1957) and the application of these scales by Williams, Whitehead, and Miller (1971) who examined relationships of attitudes and children's speech. Piche et al. examined teachers' responses to their scale items by presenting teachers with different samples of writing. Some of the samples contained inserted types of speech the researchers characterized as African American Vernacular English (AAVE). Actual samples of Black students' writing were not used for their study. Instead, they used a piece of writing, and added features identified as AAVE into the text.

In addition, Rubin and Williams-James (1997) examined the ways teachers responded to international students' writing using similar scales. These researchers created a text and inserted types of speech that appeared to be consistent with writers from different nationalities. They did not use actual student products for their evaluation. The instrumentations of these differential scales, however, set a precedent to examine instructors' impressions of students' writing.

For this study, the Writing Portfolio Differential Scale contains three separate criteria frameworks: Writing Portfolio Criteria, the programmatic areas articulated, published and evaluated by the Writing Program (Comprehension of the Task, Focus, Organization, Support, and Proofreading) and two other frameworks of criteria—Alternate Writing Criteria and Demographic Criteria—which were previously used by Piche et al. and Rubin and Williams-James. The Alternate Writing Criteria include Coherence, Use of Standard American English, Logic, Grammar, Creativity, Level of Language Passivity, and Quality of Writing; and the Demographic Criteria include the rater's perception of the writer in many areas: Strength of Writer, Intelligence, Socio-Economic Status, Level of Cultural Advantage, Confidence, and Comfort as a Writer. This study examined actual samples of student writing composed by actual students for

undergraduate courses across the disciplines subsequently submitted for their university required Writing Portfolios.

A group of faculty Writing Portfolio raters were trained to apply the different variables of the Writing Portfolio Differential Scale to individual components of students' Writing Portfolio submissions. Two rating sessions were held in the fall of 2006, and consisted of thirty-three raters. Four raters (12%) were tenure-line faculty; thirteen (39%) were adjunct faculty; fourteen (42%) were graduate teaching assistants with extensive teaching experience; and the other two raters (6%) were other position classifications. Of this group, 24 were white; 4 multiracial; 1 Hispanic; 3 Asian Pacific Islander; and 1 African American. Eighteen percent were multi-lingual and 82% were native speakers of English. Seventy-six percent of the raters were female and 24% of the raters were male.

Two hundred and fifty writing portfolios were selected for this study—fifty from each racial classification used by the researcher's institution (American Indian Alaska Native, Asian Pacific Islander, Hispanic, Black, and White). The selected Writing Portfolios had been submitted between 2003-2006. Each Writing Portfolio contained an impromptu exam and three course papers. Overall, one thousand samples of writing were evaluated for this study.

The analysis examined a random sample of 150 impromptu exams and 300 individual course paper submissions. The samples were selected by using the SPSS option to create a randomized list for data analysis. In order to instill confidence in the results, sample sizes for the analyses followed Tabachnick and Fidell's (2005) "simple rules of thumb [for sample size for regression analyses]: N>50+8m (where *m* is the number of [independent variables] for testing the multiple correlation" (p. 123). Shavelson's (1996) rule of thumb encouraged at least fifty subjects, and ten times as many cases as independent variables, which would be at least 120 for each analysis. Again, separate regression analyses were conducted for the three different frames for the impromptu exams and for the three different frames for the course paper submissions. Each analysis conducted a sequential regression analysis on each of the scale frameworks: Writing Portfolio criteria, Alternate Writing criteria, and Demographic criteria. In other words, a total of six regression equations were calculated: three sequential regression equations were established for each criteria framework for each type of writing resulting in equations that account for the variability of writing quality scores.

Sequential regression analyses were conducted on a random sample of 150 Writing Portfolio impromptu exams to account for the variability in the quality of the writing through the Writing Portfolio criteria, Alternate Writing criteria, and Demographic criteria. Each variable was rated on a scale from 1 to 6.

Students' racial identities were converted to a nominal scale (American Indian=1; African American=2; Asian Pacific Islander=3; Hispanic=4; White=5) and were entered first into each regression equation. A sequential regression analysis was conducted and the entry order of the variables was based upon the stepwise regression analysis results from Kelly-Riley (2006). Table 1.3 details the order of the criteria variables were entered into the sequential equation as well Cronbach's alpha.

Table 1.3. Variable Entry Order into the Sequential Regression Analysis and Reliability Data for the Timed Writing analysis

Criteria Framework	Variable Order of Entry	Cronbach's Alpha
Writing Portfolio	Race, Focus, Proofreading, Support, Comprehension of task, and Organization	.8946
Alternate Writing	Race, Coherence, Logic, Creativity, Grammar, use of Standard American English, Language passivity	.8902
Demographic	Race, and raters' perceptions of writers' Confidence, Intelligence, Comfort with writing, Socio-economic status, and Cultural advantage	.7902

Note: N=150

Table 1.4 details the order of the criteria variables were entered into the sequential equation as well Cronbach's alpha. The entry order of the variables are slightly different based on the results from the stepwise analysis conducted by Kelly-Riley (2006).

Table 1.4. Variable Entry Order into the Sequential Regression Analysis and Reliability Data for the Course Paper analysis

Criteria Framework	Order of Entry	Cronbach's Alpha
Writing Portfolio	Race, Focus, Proofreading, Support, Organization and Comprehension of task	.8512
Alternate Writing	Race, Coherence, Logic, Grammar, Creativity, Use of Standard American English, Language passivity	.8647
Demographic	Race, and raters' perceptions of writers' Comfort with writing, Intelligence, Confidence, Socio-economic status, and Cultural advantage.	.8560

Note: N=300

RESULTS

The results from the sequential regression analyses for both types of writing—impromptu and course paper submissions—suggest that the definition of good writing is based more on variables of coherence, correctness, and confidence as applied by faculty raters in this large scale Writing Portfolio assessment. Race did not contribute significantly to faculty raters' functional definition of "good writing" for any of the frameworks whether in the timed exam format or for the course papers. Surprisingly, faculty raters operationalize their assessment of "good writing" based on criteria accounted more through non-programmatic evaluation criteria of the Alternate Writing framework variables. In addition, a high percentage of the writing scores—for impromptu writing as well as papers written for courses—included demographic considerations of the writer. Higher percentages of writing quality scores were accounted for through coherence and grammar, part of the Alternate Writing framework. These variables overlap with focus and mechanics, which account for writing quality through the Writing Portfolio framework, although the Writing Portfolio variables account for a slightly lesser percentage of the writing quality score. A surprisingly high percentage—nearly two thirds of the score—of writing quality is also accounted for through raters' perceptions of student writers' intelligence and comfort with writing. For each type of writing—impromptu exams and course paper submissions—race did not contribute significantly to writing quality score through any of the frameworks.

Table 1.5 details the separate regression equations that account for the variability in the impromptu exam analysis. The Alternate Writing Criteria accounted for the most variability in the impromptu writing score.

Findings for the Impromptu Exam Analysis

Table 1.5. Raw and Standardized Regression Coefficients and Percent of Variance in Timed Writing Quality explained by (A) Writing Portfolio Criteria, (B) Alternate Writing Criteria and (C) Demographic Criteria

Significant Regression Equations	B	b	% of variance explained
(A) Focus + Mechanics +-Support	.198	.224**	61.6
(B) Coherence+ Creativity + Grammar	.223	.236**	68.0
(C) Intelligence+ Comfort	.546	.506**	60.9

*Note. Each frame represents a separate regression equation with the variables in the order in which the regression analysis specified. N=150 *p<.05. **p<.01.*

Tables 1.6, 1.7, and 1.8 provide detailed analysis of the significant regression equations and the differences in the percentages of the variance explained for each of the three separate criteria frameworks.

Table 1.6. Raw and Standardized Regression Coefficients and Percent of Variance in Timed Writing Quality explained by Writing Portfolio Criteria

Predictor Variables/Regression Equations	B	b	% of Variance Explained
Focus	.214	.235**	38.1
Focus + Mechanics	.467	.492**	59.5
Focus + Mechanics+ Support	.198	.224**	61.6

*Note. N=150 *p<.05. **p<.01.*

Table 1.7. Raw and Standardized Regression Coefficients and Percent of Variance in Timed Writing Quality explained by Alternate Writing Criteria

Predictor Variables/Regression Equations	B	b	% of Variance Explained
Coherence	.518	.512**	62.0
Coherence +Creativity	.218	.208**	65.5
Coherence +Creativity +Grammar	.223	.236**	68.0

*Note. N=150 *p<.05. **p<.01.*

Table 1.8. Raw and Standardized Regression Coefficients and Percent of Variance in Timed Writing Quality explained by Demographic Criteria Variables

Predictor Variables/Regression Equations	B	b	% of Variance Explained
Intelligence	.533	.396**	42.0
Intelligence +Comfort	.546	.506**	60.9

*Note. N=150 *p<.05. **p<.01.*

Findings for the Course Paper Analyses

More of the variance in the writing quality score was accounted for in the assessment of the course papers. Table 1.9 details the separate regression equations that account for the variability in the course papers. The Alternate Writing criteria accounted for the most variability in the course paper review.

Table 1.9. Raw and Standardized Regression Coefficients and Percent of Variance in Course Paper Writing Quality explained by (D) Writing Portfolio Criteria, (E) Alternate Writing Criteria and (F) Demographic Criteria

Significant Regression Equations	B	b	% of Variance Explained
(D) Focus + Mechanics + Organization	.198	.188**	72.0
(E) Coherence + Logic + Grammar	.420	.452**	77.0
(F) Comfort + Intelligence + Confidence	.187	.158**	64.1

*Note. Each frame represents a separate regression equation with the variables in the order in which the regression analysis specified. N=300 *p<.05. **p<.01.*

Tables 1.10, 1.11 and 1.12 provide detailed analysis of the significant regression equations and the differences in the percentages of the variance explained for each of the three separate criteria frameworks for the review of the course papers.

Table 1.10. Raw and Standardized Regression Coefficients and Percent of Variance in Course Paper Writing Quality explained by Writing Portfolio Criteria

Predictor Variables/Regression Equations	B	b	% of Variance Explained
Focus	.253	.222**	38.2
Focus + Mechanics	.545	.592**	70.8
Focus + Mechanics Organization	.198	.188**	72.0

*Note. N=300 *p<.05. **p<.01.*

Table 1.11. Raw and Standardized Regression Coefficients and Percent of Variance in Course Paper Writing Quality explained by Alternate Writing Criteria

Predictor Variables/Regression Equations	B	b	% of Variance Explained
Coherence	.421	.360**	64.2
Coherence+ Logic	.192	.162**	68.2
Coherence+ Logic+ Grammar	.420	.452**	77.0

*Note. N=300 *p<.05. **p<.01.*

Table 1.12. Raw and Standardized Regression Coefficients and Percent of Variance in Course Paper Writing Quality explained by Demographic Criteria Variables

Predictor Variables/Regression Equations	B	b	% of Variance Explained
Comfort	.465	.426**	55.6
Comfort + Intelligence	.369	.318**	63.5
Comfort + Intelligence +Confidence	.187	.158**	64.1

*Note. N=300 *p<.05. **p<.01.*

DISCUSSION AND IMPLICATIONS

The first research question focused on the definition of good writing used by faculty raters and the results from the six separate regression analyses show the surprising ways that faculty operationalize this construct. First, a comparison of the two Writing frameworks (Writing Portfolio criteria and Alternate Writing criteria) show that coherence, focus, and correctness all contribute significantly to the functional definition of "good writing" applied by faculty raters in the context of this mid-career diagnostic assessment. All of these variables contribute significantly to writing quality in both the impromptu writing situation and for the course paper evaluation (in which students theoretically would have time to plan, draft, and revise). More variance in writing quality is accounted for in the evaluation of the course papers than the impromptu exams. Nearly a third of the impromptu writing quality score is unaccounted for while less than a quarter is unaccounted for in the course papers.

Secondly, raters seem to apply more non-programmatic variables not overtly articulated by the Writing Program. More of the variance in the writing quality score—for both impromptu writing and course papers—is accounted for by the non-programmatic Alternate Writing criteria. The variable of Coherence accounts for the largest percentage of the variance in which Coherence, by itself, accounts for 62% of timed writing quality and 64.2% of course paper writing quality. On the other hand, Focus, as a standalone variable, accounts for only 38.1% of the variance of timed writing quality and 38.2% of course paper quality. For impromptu writing, Creativity is a variable considered by raters whereas the Writing Portfolio criteria include Support. These two variables are dissimilar. However, there is some overlap between the variables of the two frameworks of writing criteria as Mechanics and Grammar are included in all four regression equations, and logic and organization are similar variables included in the course papers frameworks. In spite of the published and articulated Writing Portfolio criteria, raters

seem to apply idiosyncratic criteria that fall outside of the intended assessment. Perhaps this disconnect can be explained by the explicit instructions in the rating sessions for raters to reference their classroom writing experiences and expectations and to be guided by the Writing Portfolio criteria in the assessment. They are asked to operationalize the criteria as relevant to their disciplinary realities.

Similarly surprising, raters evaluate impromptu writing with slightly different expectations than the writing done within courses. While Focus and Mechanics are included both Writing Portfolio criteria, Support is used by raters to assess impromptu writing quality while Organization replaces it in the evaluation of the course papers. Likewise, this trend is observed in the Alternate Writing criteria. Coherence and Grammar account for the variance in writing quality for impromptu writing and course papers, but Creativity is important in impromptu writing whereas Logic replaces it in the course paper writing. These results suggest that faculty have different expectations for the two different writing tasks included in the same Writing Portfolio. These results do not differentiate between one set of criteria being better than the other; they only indicate that faculty seem to view these tasks differently. Certainly, this interesting result deserves further study.

The second research question examines whether the operationalized definition of good writing included demographic information. The findings suggest that large percentages of the variance of writing quality are accounted for through the Demographic framework—primarily through the rater's perception of the writer's intelligence and comfort with writing. The variables of race, perceived economic status, and perceived cultural advantage did not contribute significantly to the writing quality score. While the two writing frameworks have more obvious overlap, the demographic criteria seem to overlap with writing issues too. The demographic criteria that faculty use to account for writing quality are based on variables that would be reasonable to identify a writer as needing help: the student's comfort level with writing, the student's confidence with writing, and the teacher's perception of the student's intelligence. The variables are not related to demographic features that are irrelevant to the classroom.

The third question examined whether the assessment process privileged forms of writing according to race. The findings from this study suggest that race is not a significant contributor to the faculty's assessment of students' writing for either the impromptu writing or papers written for courses. The results of the sequential regression analyses suggest that race does not significantly account for the variance in good writing. However, students' performances by race on the Writing Portfolio are significantly different like the studies conducted by Schmidt and Camara (2004) and Breland et al. (2004), but the rating processes used by faculty raters do not seem to the cause for these differences.

Such concern about the relationship between the rater and the writer is warranted. Ball (1997) documented potential bias by readers for writers based on dissimilar cultural backgrounds. Smitherman's extensive research (highlighted in Smitherman and Villanueva, 2003) has documented different linguistic structures of African American students and their implications in educational settings. This study, though, found somewhat different results. The rating corps used for this study represented a linguistically and culturally diverse set of faculty—who were also representative of the regular Writing Portfolio rating corps—attempting to address some of the concerns raised by Ball. Admittedly, this study did not intend to address the specific relationship between rater and writer.

Furthermore, the extent to which Mechanics contributes to writing quality is interesting in the light of Smitherman's research, but this study included more racial categories than Smitherman's studies, which focused primarily on African Americans. Such distinct differences between raters and writers with a multitude of different backgrounds may not be as detectable as comparisons that look at only two racial groups. Even though race was not a variable that accounted for any writing quality in this study, some of Smitherman's findings that connect race and linguistic structure might seem supported by this study. Specifically, Mechanics accounts for a great deal of the Writing Portfolio quality score. Mechanics accounts for 32.6% to the variance in impromptu writing and 21.4% of the course papers. Overall, though, the Writing Portfolio criteria account for less of the writing quality than the Alternate Writing criteria. In the Alternate criteria, Grammar, while a significant contributor to quality, did not account for as much as Mechanics, with 8.8% of the variance explained for impromptu writing and 2.5% in the course papers. Issues of Coherence that go beyond Grammar seem to be more important in raters' assessment of students' writing.

While the findings from this study suggest that race does not contribute significantly to raters' operationalization of good writing, it is disconcerting that there are statistically significant differences in performances by race on the Writing Portfolio. While the reason may not be in how the raters evaluate student writing, the subject requires further investigation. Schmidt and Camara (2004) summarize the prevailing theories used to explain the gap in standardized test performances by race: inequitable educational preparation, poverty, discrimination, poor educational opportunities, and lack of access to educational resources. Studies such as these would be useful in large-scale performance-based assessment programs, and studies examining the effectiveness of the structured support required by these programs would be the next logical step in validity research.

Validity, again, refers to the use and interpretation of test scores in a particular setting. What do these results mean for the use and interpretation of the test

scores in the university-wide Portfolio context? The purpose of the Writing Portfolio is to assess students' readiness for the upper-division Writing in the Major courses. In the rating sessions, faculty are overtly asked to draw on their classroom experiences and expectations for the assessment situation. Perhaps this request for raters to draw on pedagogical reference points helps explain the large role that the Alternate Writing criteria play in accounting for writing quality. These findings are consistent with Broad's qualitative study that document the frustration faculty felt in rubric-based assessments in first year writing programs. Since the Writing Portfolio relies on the multitude of disciplinary definitions of good writing, it's important to have the starting point of common language articulated in the Writing Portfolio criteria and to begin to fully acknowledge the additional role that other non-programmatic criteria play. Additionally, in this Writing Portfolio system, frustration levels are mitigated in that faculty don't have to agree about a static definition of writing; faculty simply place student writers into three broad categories of placement: Pass, Pass with Distinction, and Needs Work. The broadly defined functional placements mask the complex process behind the rating behaviors. These rating behaviors need to be routinely examined.

These findings suggest that writing assessment program administrators need to play more of active role in looking at published program criteria, standards for writing, and faculty enactment of these standards. The focused time allotted to the evaluation of writing at the ends of the spectrum (weak or strong) in the shared evaluation, expert-rater system does not translate into a systematic application of the criteria of good writing as articulated through the programmatic rubric. Locally developed writing assessment programs—whether portfolios or directed self-placement or other mechanisms which rely on faculty articulation of standards—need to compare the published criteria used by their programs to the criteria used functionally through the rating process. This point is the "significant site of power and knowledge" (O'Neill, 2003, p. 62) so often ignored by compositionists.

A tension exists between the criteria the Writing Program articulates and publishes, and the actual multi-dimensional criteria enacted by faculty raters. The ways in which programmatic criteria and disciplinary expectations intersect must be examined further because they most certainly inform and reform each other in a system that intends to be responsive to validity and reliability concerns. The absence or limited contribution of the some of the programmatic Writing Portfolio criteria—comprehension of the task, organization, and support—to the writing quality score points to a disjuncture. The findings suggest that these criteria—as faculty use them—either contribute minimally to the quality score or not at all for both the impromptu and course paper evaluations. This omission raises the

question as to whether raters—who are hired for these positions based on their extensive teaching expertise—don't value or don't know how to evaluate for these criteria areas. While the program advertises and publishes specific criteria for evaluation of Writing Portfolios, more than half of these criteria areas are not utilized by raters in the evaluation setting. This omission questions the extent to which these criteria are employed in classroom settings. The program administrators must be aware of the tendency of raters to draw on personal pedagogical expectation, and to move the raters toward the programmatic criteria particularly for decisions that fall on the ends of the spectrum. Improved rater training and overt conversations about this tendency in norming sessions might be a way to begin to further identify and address these issues.

Finally, research that includes more nuanced considerations of race and ethnicity into these large scale writing assessment practices need to be more commonplace. Educational research already has a robust agenda of research in standardized tests related to race and ethnicity, but most times, the standardized tests are separated from the instructional or local context. Composition studies needs to embrace a similar research agenda which considers the hermeneutically-oriented assessment approaches that are rooted in local context. While this study only examines the construct of good writing as applied by raters, there are many other angles of necessary research and validity inquiry for students of colors' experiences in context-rich, locally-developed writing assessment programs. Given the more mainstream position that college writing assessment methodologies have garnered of late, such inquiry is important, timely, and vital—not only to examine the quality of the practices, but to ensure that such methodologies are not intentionally or unintentionally leveling consequences for students—particularly those represented by small populations who may be easily overlooked.

REFERENCES

American Educational Research Association, American Psychological Association, & National Council of Measurement in Education. (1999). *Standards for Educational and Psychological Testing*. American Educational Research Association.

American Anthropological Association (1997). Response to OMB directive 15. *Race and ethnic standards for federal statistics and administrative reporting*.

American Psychological Association Task Force on Diversity Issues at the Precollege and Undergraduate Levels of Education in Psychology (1998). Enriching the focus on ethnicity and race. *Monitor, 29*(3).

Ball, A. (1997). Expanding the dialogue on culture as a critical component when assessing writing. *Assessing Writing, 4*(2), 169-202.

Bond, L. (1995). Unintended consequences of performance assessment: Issues of bias and fairness. *Educational Measurement: Issues and Practice*, *14(*4), 21-24.

Breland, H., Kubota, M., Nickerson, K., Trapani, C., Walker, M. (2004). *New SAT writing prompt study: Analyses of group impact and reliability* (Report No. 2004-1). College Entrance Examination Board.

Broad, B. (2000). Pulling your hair out: Crises of standardization in communal writing assessment. *Research in the Teaching of English, 35*(2), 213-260.

Camilli, G. (2006). Test fairness. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed) (pp. 221-256). American Council on Education/Oryx Press Series on Higher Education.

Clary-Lemon, J. (2009). The racialization of composition studies: Scholarly rhetoric of race since 1990. *College Composition and Communication, 61*(2), W1-W17.

Cronbach, L. (1988). Five perspectives on validity argument. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 3-17). Erlbaum.

Elbow, P. (1996). Writing assessment in the 21st century: A utopian view. In L. Z. Bloom, D. A. Daiker, & E. M. White (Eds.), *Composition in the twenty-first century: Crisis and change* (pp. 83-100). Southern Illinois University Press.

Elliot, N., Briller, V., & Joshi, K. (2007). Portfolio assessment: Quantification and community. *Journal of Writing Assessment, 3*(1), 5-30. https://escholarship.org/uc/item/8nm1m6xc

Farr, M. & Nardini, G. (1996). Essayist literacy and sociolinguistic difference. In E. M. White, W. D. Lutz & S. Kamusikiri (Eds.), *Assessment of writing: Politics, policies, practices* (pp. 108-119). Modern Language Association.

Gere, A. R., Aull, L., Green, T., and Porter, A. (2010). Assessing the validity of directed self-placement at a large university. *Assessing Writing, 15(3)*, 154-176.

Griffee, D. (2002). Portfolio assessment: Increasing reliability and validity. *The Learning Assistance Review: The Journal of the Midwest College Learning Center Association, 7*(2), 5-17.

Hamp-Lyons, L. and W. Condon. (2000). *Assessing the portfolio: Principles for practice, theory and research*. Hampton Press.

Haswell, R. (1998a). Multiple inquiry in the validation of writing tests. *Assessing Writing*, *5*(1), 89-109.

Haswell, R. (1998b). Rubrics, prototypes and exemplars: Categorization and systems of writing placement. *Assessing Writing*, *5*(2), 231-268.

Haswell, R. (2000). Documenting improvement in college writing: A longitudinal approach. *Written Communication*, *17(3)*, 220-236.

Haswell, R. (Ed.). (2001). *Beyond outcomes: Assessment and instruction within a university writing program*. Ablex.

Haswell, R. (2005). NCTE/CCCC's recent war on scholarship. *Written Communication*, *22*(2), 198-223.

Haswell, R, & S. Wyche. (1996). A two-tiered rating procedure for placement essays. In T. W. Banta (Ed.), *Assessment in practice: Putting principles to work on college campuses* (pp. 204-207). Jossey-Bass.

Hester, V., O'Neill, P., Neal, M., Edgington, A., & Huot, B. (2007). Adding portfolios to the placement process. In P. O'Neill (Ed.), *Blurring boundaries: Developing writers, researchers, and teachers* (pp. 61-90). Hampton Press.

Huot, B. (1996). Toward a new theory of writing assessment. *College Composition and Communication*, *47*(4), 549-566.

Huot, B. & Schendel, E. (1999). Reflecting on assessment: Validity inquiry as ethical inquiry. *Journal of Teaching Writing*, *17*(1-2), 37-55.

Huot, B. & Williamson, M. M. (1997). Rethinking portfolios for evaluating writing: Issues of assessment and power. In K. B. Yancey and I. Weiser (Eds.), *Situating portfolios: Four perspectives* (pp. 43-56). Utah State University Press.

Inoue, A. (2007). Articulating Sophistic rhetoric as a validity heuristic for writing assessment. *Journal of Writing Assessment*, *3*(1), 31-54. https://escholarship.org/uc/item/64n8z5mz

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.) *Educational measurement* (4th ed.) (pp. 17-64). American Council on Education/Oryx Press Series on Higher Education.

Kelly-Riley, D. (2006). A validity inquiry into minority students' performances in a large-scale writing portfolio assessment. (Doctoral Dissertation, Washington State University).

LeMahieu, P. G., Gitomer, D. H., & Eresh, J. T. (1995). Portfolios in large-scale assessment: Difficult but not impossible. *Educational Measurement: Issues and practice*, *14*(3), 11-28. https://doi.org/10.1111/j.1745-3992.1995.tb00863.x

Lippi-Green, R. (1997). *English with an accent: Language, ideology, and discrimination in the United States*. Routledge.

Lynne, P. (2004). *Coming to terms: A theory of writing assessment*. Utah University Press.

Moss, P. A. (1995). Themes and variations in validity theory. *Educational Measurement: Issues and Practice*, *14*(2), 5-13.

Moss, P. A. (1998a). Testing the test of the test: A response to "Multiple inquiry in the validation of writing tests." *Assessing Writing*, *5*(1), 111-122.

Moss, P. A. (1998b). The role of consequences in validity theory. *Educational Measurement: Issues and Practice*, *17*(2), 6-12. https://doi.org/10.1111/j.1745-3992.1998.tb00826.x

Moss, P. A. (2007). Joining the dialogue on validity theory in educational research. In P. O'Neill (Ed.), *Blurring boundaries: Developing writers, researchers, and teachers* (pp. 91-100). Hampton Press.

Moss, P. A & Schutz, A. (2001). Educational standards, assessment and the search for consensus. *American Educational Research Journal*, *38*(1), 37-70.

Mountford, R. (1999). Let them experiment: Accommodating diverse discourse practices in large-scale writing assessment. In C. R. Cooper & L. Odell (Eds.), *Evaluating writing: The role of teachers' knowledge about text, learning, and culture* (pp. 366-396). National Council of Teachers of English.

Murphy, S. (2007). Culture and consequences: The canaries in the coal mine. *Research in the Teaching of English*, *42*(2), 228-244.

Omi, M. & Winant, H. (1994). *Racial formation in the United States: From the 1960's to the 1990's*. Routledge.

O'Neill, P. (2003). Moving beyond holistic scoring through validity inquiry. *Journal of Writing Assessment*, *1*(1), 47-65. https://escholarship.org/uc/item/4qp611b4

Osgood, C. E., Suci, G. J., & Tannenbaum, P. (1957). *The Measurement of meaning*. University of Illinois Press.

Piche, G. L., Rubin, D. L., Turner, L. J. & Michlin, M. L. (1978). Teachers' subjective evaluations of standard and Black nonstandard English compositions: A study of written language and attitudes. *Research in the Teaching of English*, *12(2)*, 107-118.

Rubin, D. L., & Williams-James, M. (1997). The impact of writer nationality on mainstream teachers' judgments of composition quality. *Journal of Second Language Writing*, *6*(2), 139-154.

Scharton, M. (1996). The politics of validity. In E. M. White, W. D. Lutz & S. Kamusikiri (Eds.), *Assessment of writing: Politics, policies, practices* (pp. 53-75). Modern Language Association.

Schmidt, A. E. & Camara, W. J. (2004). Group differences in standardized test scores and other educational indicators. In R. Zwick (Ed.), *Rethinking the SAT: The future of standardized testing in university admissions* (pp. 189-201). Routledge Falmer.

Shavelson, R. J. (1996). *Statistical reasoning for the behavioral sciences* (3rd ed.). Allyn and Bacon.

Smith, W. L. (1993). Assessing the reliability and adequacy of using holistic scoring of essays as a college composition placement technique. In M. M. Williamson and B. A. Huot (Eds.), *Validating holistic scoring for writing assessment: Theoretical and empirical foundations* (pp. 142-205). Hampton Press.

Smitherman, G. & Villanueva, V. (Eds.). (2003). *Language diversity in the classroom: from intention to practice*. Southern Illinois University Press.

Tabachnick, B. & Fidell, L. S. (2006). *Using multivariate statistics* (5th ed.). Allyn and Bacon.

White, E. M. (2005). The scoring of writing portfolios: Phase 2. *College Composition and Communication*, *56*(4), 581-600.

Williams, F., Whitehead, J. L. & Miller, L. M. (1971). *Attitudinal correlates of children's speech characteristics* (USOE Project No. 0-0336). Center for Communication Research.

Williamson, M. M. & Huot, B. A. (1993). *Validating holistic scoring for writing assessment: Theoretical and empirical foundations*. Hampton Press.

APPENDIX: WRITING PORTFOLIO DIFFERENTIAL SCALE FOR WRITING AND DEMOGRAPHIC INFORMATION

Paper code_____

Circle the appropriate number to indicate your evaluation of the writing.

1. Conception of topic

Unclear	1	2	3	4	5	6	Clear

2. Focus

Unclear	1	2	3	4	5	6	Clear

3. Organization

Disorganized	1	2	3	4	5	6	Organized

4. Support

Not provided	1	2	3	4	5	6	Provided

5. Mechanics

Not Effective	1	2	3	4	5	6	Effective

The writing seems to be

6. Incoherent	1	2	3	4	5	6	Coherent
7. Non-Standard American English	1	2	3	4	5	6	Standard American English
8. Illogical	1	2	3	4	5	6	Logical
9. Ungrammatical	1	2	3	4	5	6	Grammatical
10.Unimaginative	1	2	3	4	5	6	Imaginative
11. Passive	1	2	3	4	5	6	Active
12. Poorly written	1	2	3	4	5	6	Well written

The Student Writer is

13. Weak Writer	1	2	3	4	5	6	Strong Writer
14. Unintelligent	1	2	3	4	5	6	Intelligent
15. Low socio-economic class	1	2	3	4	5	6	High socio-economic class
16. Culturally disadvantaged	1	2	3	4	5	6	Culturally advantaged
17. Unsure	1	2	3	4	5	6	Confident
18. Uncomfortable as a writer	1	2	3	4	5	6	Comfortable as a writer

CHAPTER 3.

THREE INTERPRETATIVE FRAMEWORKS: ASSESSMENT OF ENGLISH LANGUAGE ARTS-WRITING IN THE COMMON CORE STATE STANDARDS INITIATIVE

Norbert Elliot
New Jersey Institute of Technology

Andre A. Rupp
Educational Testing Service

David M. Williamson
Educational Testing Service

We present three interpretative frameworks by which stakeholders can analyze curricular and assessment decisions related to the Common Core State Standards Initiative in English Language Arts-Writing (CCSSI ELA-W). We pay special attention to the assessment efforts of the Smarter Balanced Assessment Consortium (Smarter Balanced) and the Partnership for Assessment of Readiness for College and Careers (PARCC). Informed by recent work in educational measurement and writing assessment communities, the first framework is a multidisciplinary conceptual analysis of the targeted constructs in the CCSSI ELA-W and their potential measurement. The second framework is provided by the Standards for Educational and Psychological Testing (2014) with a primary focus on foundational principles of validity, reliability/precision, and fairness. The third framework is evidence-centered design (ECD), a principled design approach that supports coherent evidentiary assessment arguments. We first illustrate how Standards-based validity arguments and ECD practices have been integrated into assessment work for the CCSSI ELA-W using Smarter Balanced and PARCC assessment reports. We then demonstrate how

DOI: https://doi.org/10.37514/PER-B.2024.2166.2.03

all three frameworks provide complementary perspectives that can help stakeholders ask principled questions of score interpretation and use.

By the end of the nineteenth century in the United States, demand for universal public education had become equated with assurance of participatory democracy. In 1869-1870, 7.48 million students enrolled in kindergarten and grades one through eight. By 1899-1900, that number had risen to 14.98 million. This increase was accompanied by a dramatic rise in high school enrollment as advanced education became necessary for better paying jobs. In 1869-1870, 80,000 students were enrolled in grades nine through twelve. In 1899-1900, that number had risen to 519,000 (Snyder, 1993, p. 34, Table 8).

Accompanying this new influx of students were those who believed they knew best how to shape the curriculum. Archetypal responses—the humanism of Charles W. Eliot (1892), the developmentalism of G. Stanley Hall (1883), the social efficiency of Joseph Mayer Rice (1893), and the social meliorism of Lester Frank Ward (1883)—were to continue throughout the twentieth century (Kliebard, 2004). Today, one may identify these enduring themes in the calls for equity by Diane Ravitch (2010), the cognitive modeling of Howard Gardner (2006), the emphasis on effective teaching by Bill and Melinda Gates (2015), and the progressivist agenda of Arne Duncan (2015).

With enrollment projections for the school year 2015-2016 estimated at 49.8 million public elementary and secondary school students (Snyder & Dillow, 2015, p. 86, Table 203.10), these and other voices emerge to give council on how best to spend a projected education budget of no less than $669 billion (Snyder & Dillow, 2015, p. 58, Table 106.10). There is a loud roar of voices accompanying initiatives associated with the term "educational reform," which has become nearly deafening as the national debate has turned to the Common Core State Standards Initiative (CCSSI) and associated state-led curricular guidelines for a national school curriculum assessed by two consortia: the Smarter Balanced Assessment Consortium (Smarter Balanced) and the Partnership for Assessment of Readiness for College and Career*s* (PARCC).

As the most comprehensive effort in American history to leverage uniform goal-based instruction, the CCSSI is designed to ensure that high school graduates are prepared to take credit-bearing courses in two- or four-year college programs or enter the workforce. At the present writing, forty-two states, the District of Columbia, four territories, and the Department of Defense Education Activity have adopted the CCSSI. Assessments in English language arts and mathematics have taken place in the 2014-2015 school year, and preliminary results are being released at the time of this writing.

The development of the CCSSI and its assessment has been accompanied by three categories of criticism: warnings of the dangers of neoliberalism; concerns over the constraint of the writing construct; and fears that the achievement of equity continues to elude educational reform. From their creation (in order to enhance global competitiveness and workplace success) to their solicitation (in order to encourage proposals for next generation assessment systems), the CCSSI have been informed by "a form of cultural politics and a set of economic principles, policies, and practices devoted to handing over as much of social life as possible to private interests" (Gallagher, 2011, p. 453).

Referencing this depiction of neoliberalism, Wilson has been critical of the ways that such framing has diminished teacher agency (Shannon, Whitney, & Wilson, 2014). In interacting with students and teachers, she argued, "you see what matters, and you realize that these grand plans that Bill Gates has for how it is that we're going to improve education just don't make any sense" (p. 299). In similar fashion, Addison and McGee (2015) warned that the role of the Gates Foundation compromises local efforts such as those sponsored by the National Writing Project, "to gain compliance" with the CCSSI (p. 215). Concentrating on the limits of construct representation following from the neoliberal policy climate, Kristine Johnson (2015) found curricula based on the CCSSI "would focus almost exclusively on expository/informational and fact-based argumentative writing, with some narrative descriptive writing"—a "narrowing effect" that diminishes coverage of the writing construct (p. 520). Applebee (2013) has also identified this narrowing effect in his identification of four areas—separate emphasis on foundational skills, grade-by-grade standards, absence of a developmental writing model, and implementation issues—with "equal potential to distort curriculum" (p. 28).

While public debate swirls around societal impact, often absent are voices of stakeholder groups directly involved with students: parents and guardians; teachers and administrators; legislators; and workforce leaders. It is our aim in this paper to suggest directions of inquiry for those stakeholder groups. Specifically, we seek to empower these stakeholder groups by discussing how a deeper understanding of the traditions, terminologies, and best practices of educational measurement and writing assessment provide an excellent way to ask critical questions about new curriculum and assessment initiatives.

Such strategies are needed to navigate a maze of complex debates in which everything and its opposite both appear to be true. As researchers in writing assessment (Elliot), cognitively-grounded diagnostic measurement (Rupp), as well as automated scoring and modern psychometrics (Williamson), we are positioned to enter the controversial roar in a very precise way.

While we acknowledge and honor the ontological and axiological force of voices interested in the social dimension of assessment, we focus in this paper

on structuring discussions around significant technical issues in assessment design and use for the CCSSI in English Language Arts-Writing (CCSSI ELA-W). These issues are discussed through the lens of three interpretative frameworks that provide complementary perspectives and ways of thinking about key issues for stakeholders: multidisciplinary research on writing (e.g., Elliot & Perelman, 2012), the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014), and evidence-centered design (ECD) (e.g., Mislevy, Steinberg, Almond, & Lukas, 2006).

While we are certainly encouraging readers to consider the different perspectives we present through these frameworks, our discourse modes are primarily expositive, descriptive, and narrative. That is, we do not seek to criticize the CCSSI or the work of Smarter Balanced and PARCC in any absolute or relative terms; rather, we want to illustrate how the three interpretative frameworks provide conceptual scaffolds for asking critical questions that lead to enriched discussions among stakeholders. We believe that such discussions—and the associated heightened awareness of the complexities of many curricular and assessment design decisions—can help the diverse communities affected by the CCSSI gain a stronger appreciation for the relative strengths and weaknesses of various political, instructional, and assessment efforts.

INTERPRETATIVE FRAMEWORK 1: MULTIDISCIPLINARY RESEARCH ON WRITING

Part of the discipline of education, the field of educational measurement finds its origin in 1892 with the founding of the American Psychological Association (Fernberger, 1932) and the subsequent 1945 designation of Division 5, Evaluation and Measurement (Benjamin, 1997). Part of the discipline of English language and literature, the field of writing assessment finds its origin with the founding of the National Council of Teachers of English in 1911 (Lindemann, 2010) and the 2010 designation of Rhetoric and Composition/Writing Studies as its own specialized field (Phelps & Ackerman, 2010).

Recent multidisciplinary research between educational measurement and writing assessment has addressed the present landscape of writing assessment, as well as methodology, consequence, and future directions for the field (Elliot & Perelman, 2012). Clearly, the two fields have begun to influence each other; the acknowledgment of mutually beneficial research agendas, for instance, has resulted in recommendations for next-generation assessments to focus on social and rhetorical knowledge, domain knowledge and conceptual strategies, writing processes, and knowledge of conventions (Sparks, Song, Brantley, & Liu, 2014). Such a multidisciplinary perspective provides a way to frame the CCSSI

assessment of ELA-W in terms of reflective attention to definitions and measurement of the writing construct.

CONSTRUCT DEFINITION

A construct such as writing, which is the core focus of the definition and empirical representation of models of student competence for CCSSI ELA-W assessment, is generally defined rather broadly. Its description, however, should be as concrete, comprehensive, and systemic as possible to be useful for instructional guidance and assessment development. The operationalization of the way the construct is measured through assessment tasks and their associated scoring rules is a great leverage point for obtaining clarity about the boundaries of the construct definition as targeted in an assessment.

Beginning with the protocol analyses of Flower and Hayes (1981), writing has been understood as a complex process in which readers and writers construct meaning through detailed, often internal, cognitive iterations concerning variables such as discourse conventions, social context, language, purpose, and knowledge. In negotiating meaning, writers create "webs of intention, carrying out complex, individual, and socially bounded purposes, shaped by attitudes and feelings, and other people" (Flower, 1994, p. 54). In recent iterations of the model, attention has been drawn to the importance of source-based investigation, the design of visual content, and management of attention and motivation (Hayes, 2012; Leijten, Van Waes, Schriver, & Hayes, 2014). As evidence of their enduring presence, Beringer (2012) has documented the origin, traditions, and future directions of cognitive perspectives on writing research. Based on construct models derived from these perspectives, Deane and his colleagues (2015) have recently developed a key practice framework linking ECD, scenario-based assessment, and cognitively-based assessment in order to create English Language Arts task sequences that support both instruction and assessment. Social cognitive models are understood to yield high quality, specific information about both the writing construct and its boundaries.

Informed by models of social cognition, the CCSSI ELA-W is designed to specify performance-level objectives—knowledge descriptions that can be mapped to grade levels. By these strategies, the CCSSI ELA-W models writing from kindergarten through grade 12. That is, in the CCSSI ELA-W, the construct is defined in actionable terms: "Students should demonstrate increasing sophistication in all aspects of language use, from vocabulary and syntax to the development and organization of ideas, and they should address increasingly demanding content and sources" (CCSSI, 2015c). By extension, writing is also viewed as part of the broader construct of ELA:

> The Common Core asks students to read stories and literature, as well as more complex texts that provide facts and background knowledge in areas such as science and social studies. Students will be challenged and asked questions that push them to refer back to what they've read. This stresses critical-thinking, problem-solving, and analytical skills that are required for success in college, career, and life. (CCSSI, 2015a)

As a blend of both reading and writing, this definition advances a conception of language arts that envisions writing and reading as integrated constructs.

In turn, this blended, integrated construct is then rendered specific within grade levels across kindergarten through grade 12. For example, the standards for grades 11 and 12 are further defined in terms of the following conceptual anchors: text types and purposes (to "write arguments to support claims in an analysis of substantive topics or texts, using valid reasoning and relevant and sufficient evidence"); production and distribution of writing (to "produce clear and coherent writing in which the development, organization, and style are appropriate to task, purpose, and audience"); research to build on present knowledge (to "conduct short as well as more sustained research projects to answer a question [including a self-generated question] or solve a problem; narrow or broaden the inquiry when appropriate; synthesize multiple sources on the subject, demonstrating understanding of the subject under investigation"); and range of writing (to "write routinely over extended time frames [time for research, reflection, and revision] and shorter time frames [a single sitting or a day or two] for a range of tasks, purposes, and audiences") (CCSSI, 2015b).

CONSTRUCT MEASUREMENT

While the CCSSI ELA-W is research-based, it is important to understand that the conceptual model—the way the elements of writing are understood in their relationship to each other within the given construct—was based on consensus opinion. Distinct from construct definitions based on evidence from reflective latent variable models (Graham, McKeown, Kiuhara, & Harris, 2012; Graham & Perin, 2007; Hillocks, 1986; Rogers & Graham, 2008), this consensus definition is, in reality, a "stew" of elements that might or might not be empirically related to each other (National Research Council, 2012). Put differently, as a consensus model, the development and instantiation of the CCSSI has, so far, been a state-led effort based on adoption, not on data collection. The means of assessing students and the information resulting from that assessment are left to

the discretion of the states as an activity distinct from the CCSSI—a very complex task for individual states and collections of states.

The era of modern assessment has arguably been characterized by a focus on creating writing tasks that are closely aligned with modern views of writing from expert communities. In fact, without this involvement of the writing community it would be difficult to imagine how this new generation of assessment would be different than the print-born bubble and booklet tests of the past. This involvement has led to the use of digitally-delivered stand-alone writing tasks and the embedding of writing activities in domain or profession-specific complex performance tasks (Tucker, 2009). Designed to capture blended constructs, integrated tasks incorporating content from source materials offer benefits such as providing realistic, challenging activities, engaging students in writing responsible to specific content, obviating practice effects associated with conventional item types, evaluating language abilities consistent with integrated models of literacy, and offering diagnostic value for instruction or self-assessment.

Challenges nevertheless remain. Cumming (2013) has noted that integrated writing tasks have associated risks. These include confounding measurement of writing ability with abilities to comprehend source materials, merging assessment and diagnostic information together in ineffective ways, and invoking genres that are emerging and therefore difficult to score. As we discuss below, navigating the complex system of tradeoffs when designing individual assessments and systems of assessments over time for CCSSI ELA-W can be substantially facilitated, integrated, and scrutinized using the *Standards* and ECD frameworks as guidance.

INTERPRETATIVE FRAMEWORK 2: STANDARDS FOR EDUCATIONAL AND PSYCHOLOGICAL TESTING

Recently revised, the *Standards* and their adaptations by testing companies (e.g., Educational Testing Service, 2014) can be seen as cohesive interpretative frameworks that lend focus to assessment design. Use of standards-based reasoning results in logical approaches to evidence in light of desired arguments about individual test-takers, test-taker groups, and the assessments themselves.

A consensus statement of its own, the *Standards* (2014) are intended "to provide criteria for the development and evaluation of tests and testing practices and to provide guidelines for assessing the validity of interpretations of test scores for the intended test uses" (p. 1). A consensus statement of its own, the *Standards* (2014) are intended "to provide criteria for the development and evaluation of tests and testing practices and to provide guidelines for assessing the validity of

interpretations of test scores for the intended test uses" (p.1). However, while *Standards* are designed for raising awareness and guiding decision-making about assessment systems. However, while *Standards* are designed for raising awareness and guiding decision-making about assessment systems at a high conceptual level, the document is not designed to be step-by-step instructions of how to do the necessary work on a day-to-day basis. That role falls to principled assessment design frameworks like ECD, which we discuss in the next section.

Calls for increased assessment literacy such as those found in the *Standards* (pp. 192-193) are not incidental to our purpose in this paper. Any fixed set of curricular approaches or assessment methods yields particular kind of interpretation and any such methodological exclusivity is inappropriate when dealing with complex assessments such as the CCSSI-ELA-W. In fact, assessment of the CCSSI-ELA-W is designed to generate the kinds of evidence needed to validate multiple proposed interpretations and uses.

While the present version of the *Standards* is our concern here, the 4th revision (1999) was the common referential point for both the Smarter Balanced and PARCC consortia. Indeed, the five sources of validity evidence identified by Sireci (2012) in his report of the Smarter Balanced research agenda—a report to which we will turn later in order to establish the informed view of validity used to support score interpretation and use (Kane, 2013, 2015) in the design of the CCSSI ELA-W assessment—are taken directly from the 1999 version. The *Standards* have played, and will continue to play, a significant role in the development of assessments related to the CCSSI.

In their present form, the *Standards* are divided into three sections: foundations, operations, and applications. By far, the foundations section is the most significant in terms of assessment of the CCSSI ELA-W. It is here we find extended discussion of the three overarching principles of validity, reliability/precision, and fairness. Because these foundational concepts deeply inform Smarter Balanced and PARCC assessment designs, a brief definition and discussion of each is warranted. Nevertheless, the concepts are not intended to be separated; rather, validity, reliability/precision, and fairness are intended to be used in support of proposed interpretation and use of scores associated with the CCSSI ELA-W assessment.

VALIDITY

In the *Standards*, validity is defined as the "degree to which accumulated evidence and theory support a specific interpretation of test scores for a given use of a test. If multiple interpretations of a test score for different uses are intended, validity evidence of each interpretation is needed" (p. 225). Although still considered by many as an "up-or-down vote" or a simple "stamp of approval,"

the 2014 edition is clear on the imprecision of such summary judgment: "Statements about validity should refer to particular interpretations and consequent uses. It is incorrect to use the unqualified phrase 'the validity of the test'" (p. 23).

While the origin of this characterization of validity may be found in the 1985 edition of the *Standards*, it is important to reflect on just how enduring the work of Messick (1989) has become in his characterization of validity as "an integrated evaluative judgement of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of *inferences* and actions based on test scores or other modes of assessment" (p. 13, emphasis in original). Equally important is the work of Kane (2013) and his call for evidence-based interpretation and use arguments: "To validate an interpretation or use of test scores is to evaluate the plausibility of the claims based on the test scores" (p. 1).

Validation therefore requires a clear statement of the claims inherent in the proposed interpretations and uses of the test scores. "Public claims require public justification" (Kane, 2013, p. 1). Influential in the development of the *Standards* and their manifestation in the assessment of the CCSSI ELA-W, Kane (2015) has offered a two-step approach to validation:

> First, the interpretation and use is specified as an interpretation/use argument, which specifies the network of inferences and assumptions leading from test performances to conclusions and decisions based on the test scores. Second, the interpretation/use argument is critically evaluated by a *validity argument*. (p. 4, emphasis in original). As a result of this orientation, validity becomes a property of score interpretations—not as a property of the assessment: "Once we adopt an interpretation, it can make sense to talk about 'the validity of a test', but the 'validity' is relative to that interpretation" (Kane, 2015, p. 2).

This "flexible framework for validation," as Kane terms it, is important in that it allows for—indeed, encourages—multiple interpretations that may arise from multiple groups. As Kane concludes, "[T]o restrict our conception of validity to one kind of interpretation seems unnecessary and would greatly limit our ability to respond to the varied applications of test scores" (2015, p. 3).

RELIABILITY/PRECISION

Reliability/precision is defined as:

> The degree to which test scores of a group of test takers are consistent over repeated applications of a measurement proce-

> dure and hence are inferred to be dependable and consistent for an individual test taker; the degree to which scores are free of random errors of measurement for a given group. (AERA, APA, & NCME, 2014, pp. 222-223)

In other words, the empirical quantification of reliability requires the existence of replication of assessment across conditions that are comparable (e.g., test forms, administration conditions, subsets of items, and sets of raters).

Once seen strictly as quantifiable by the familiar statistical coefficient of classical test theory, reliability was re-conceptualized by Lord (1980) through a more complex mathematical model for the relationships among test item performance, item characteristics, and test taker proficiency with respect to the construct(s) under examination. This framework is known in the educational measurement literature as item response theory (IRT) (e.g., de Ayala, 2009; de Boeck & Wilson, 2004) and is the most commonly applied framework for large-scale assessment apart from classical test theory. IRT can accommodate reporting on single and multiple dimensions, the existence of nested data structures (e.g., students nested in schools nested in districts), and the inclusion of variables to explain performance differences for test-takers and tasks. It can be effectively used to create large banks of tasks that can be used for adaptive assessment systems and the efficient delivery of comparable assessments with varying composition for international, national, and state-wide survey purposes.

As is the case with validity, misunderstanding about reliability abounds. For example, still considered by many as the equivalent of the railroad standard gauge, the value of 0.7 for a single reliability coefficient such as internal consistency, inter-reader agreement, or cross-administration score correlation often appears to be the sole level of attainment in the hearts and minds of many. However, with frameworks like IRT the notion of precision of measurement can be assessed more finely at different points of the reporting scale, which is important for optimizing pass-fail decisions or test assembly in high-volume testing contexts.

Consequently, the authors of the *Standards* do their best to dispel such reductionism and offer general guidelines that allow for the proper use of modern measurement approaches for capturing evidence about reliability/precision, validity, and fairness. To this end, the authors of the *Standards* also underscore that reliability and validity must be considered in conjunction with fairness considerations. For example, while the need for precision at some points of the scale increases as the consequence of score use increase, the authors acknowledge that the sacrifices in reliability/precision that may result from using performance-based writing tasks instead of multiple choice items may, in fact, be acceptable. Despite being more costly to score, these tasks may reduce construct-irrelevant variance (difference in

scores attributable to elements extraneous to the test) and/or diminish construct underrepresentation (failure to tap significant aspects of the construct that the assessment is designed to measure), which lessen the validity of the intended interpretation/use argument and its critical evaluation by the validity argument.

FAIRNESS

In the *Standards* fairness is defined as:

> The validity of test score interpretations for intended use(s) for individuals from all relevant subgroups. A test that is fair minimizes construct-irrelevant variance associated with individual characteristics and testing contexts that otherwise would compromise the validity of scores for some individuals. (p. 219)

This section of the *Standards* has been expanded substantially over previous revisions, with emphasis given to fairness for all examinees. Again, we see the presence of Messick (1989) who linked forms of validity with consequences related to score use—an emphasis that has been maintained by Kane (2006, 2013).

Significantly, special attention is given in the *Standards* to the opportunity to learn—"the extent to which individuals have had exposure to instruction or knowledge that affords them the opportunity to learn the content and skills targeted by the test" (p. 56). In an analysis consistent with this emphasis on exposure, Pullin (2008) has highlighted connections among assessment, equity, and opportunity to learn, as both a reflection of the learning environment and a concept demanding articulated connections between the assessment and the instructional environment. Such characterizations afford identification and removal of barriers to valid score interpretation for the widest possible range of individuals and subgroups, interpretative validity for examined populations, and the development of suitable testing accommodations and safeguards to protect fair score usage.

Equally associated with fairness—and of special interest in terms of equity to all stakeholders—is adherence to the principles of universal design. An approach to assessment that strives to minimize construct distortion and maximize fairness through uniform access for all intended examinees, universal design has been identified in the *Standards* (2014) as a way to leverage fairness for all examinees (p. 63). As Ketterlin-Geller (2008) has established, when student characteristics are considered during the conceptualization, design, and implementation phase of test development under principles of universal design (e.g., specifying content and cognitive complexity in the test blueprint, as well as information about the target and access skills), test performance of students with special needs is more likely to reflect their construct knowledge. Furthermore, Mislevy et al. (2013)

has demonstrated that a combination of ECD and universal design results in an increased sense of fairness as construct-irrelevant barriers to student success are proactively removed in comprehensive efforts to provide all students with an opportunity to perform at their best during assessment episodes.

OPERATIONS AND APPLICATIONS

As the authors of the *Standards* wrote, test design "begins with considerations of expected interpretations for intended uses of the scores to be generated by the test" and therefore "test design and development procedures must support the validity of the interpretations of test scores for their intended uses" (p. 75). The influence of Kane is again palpable. Issues related to validity, reliability/precision, and fairness are thus interwoven into the development process from the creation of test specifications to the copyright responsibilities of test users; as we will see, this perspective is embodied by the ECD framework that we discuss in the next section.

While the foundations discussed in the first three sections of the *Standards* are essential for understanding and navigating the complex decision-making space surrounding assessments, additional guidance is needed to put these articulated principles into practice. In the assessment operations section of the *Standards*, chapters are devoted to test design and development processes that lead to reported scores, scales, and norms as well as processes for score linking (processes used to facilitate score comparisons) and cut score setting (processes used to divide scores in order to act upon them). The authors also included chapters on test administration; scoring, reporting, and interpretation; supporting documentation for tests; the rights and responsibilities of test takers; and the rights and responsibilities of test users. The final section of the *Standards* is devoted to testing applications. Attention is given to psychological, workplace, and educational assessment, as well as the role of tests in program evaluation, policy studies, and accountability.

INTERPRETATIVE FRAMEWORK 3: EVIDENCE-CENTERED DESIGN (ECD)

As the discussions in the previous *Standards* section have made abundantly clear, to build an evidentiary argument for assessment scores so that intended interpretations and decisions comply with the *Standards* is a complex process. This complex process is exemplified in the CCSSI assessment aim as it is identified by Smarter Balanced: "The assessment system being developed by the Consortium is designed to provide comprehensive information about student achievement

that can be used to improve instruction and provide extensive professional development for teachers" (Sireci, 2012, p. 4). As such, "the assessment system focuses on the need to strongly align curriculum, instruction, and assessment, in a way that provides valuable information to support educational accountability initiatives" (p. 4).To help facilitate the construction of arguments supporting such aims and to imbue the assessment ecosystem with appropriate characteristics that support intended interpretations and decisions, a principled design framework for practice such as ECD is needed. Proposed to make explicit the evidentiary reasoning process of assessment interpretation and decision-making, ECD helps organize assessment practices in ways that yield cohesive integrated thinking about assessment aims, delivery capability, and justification of score use. As such, ECD can be viewed as providing the "evidentiary grammar" for evidence-based assessment arguments.

At its best, ECD is a powerful professional development tool that can help interdisciplinary teams of experts (e.g., assessment developers, statisticians, information technology specialists, policy-makers, and other stakeholders) develop common language, mental models, design artifacts, and best practices. In addition, it can help such teams utilize these capacities to develop targeted artifacts that move the assessment process forward in ways that best capture the connected thinking underlying the design process. These goals are always laudable and important, of course, but become especially important as the assessments become more performance-oriented, more reliant on models of social cognition, more responsive to correlates such as engagement or motivation, and more situated within community practices. In short, ECD is highly relevant for task-based CCSSI assessments of ELA-W.

Mislevy, Sternberg, and Almond (2003) identified five core structural/conceptual elements for ECD and arrange them in what they term the conceptual assessment framework: student models that characterize knowledge and skill; task models that provide constructed response test items to elicit student knowledge and skills; evidence models that provide a chain of inferential reasoning from student test performance to knowledge and skill, with emphasis on scores and their measurement; assembly models that specify how individual tasks are combined to produce the final assessment; and presentation models that specify how individual tasks are administered to students. In practice, spelling out these different models means creating artifacts such as databases, spreadsheets, and text files to document the key decisions that underlie the reasoning process.

Thus, a second layer in the day-to-day practice of assessment development is putting the decisions captured in these artifacts into practice by setting up a delivery, scoring, and reporting architecture, which Mislevy, Sternberg, and Almond described as a four-process model of activity selection (the process of

selecting and sequencing assessment tasks), presentation (the process of presenting the assessment task to the student), response processing (the process that evaluates the essential features of the student response to the task), and summary scoring (the process that produces inferences about student ability based on evidence accumulated across the task). Each of these processes emanates from an understanding of the domain that inferences are tied to and the processes of analyzing and modeling the domain tasks for assessment development purposes (Almond, Steinberg, & Mislevy, 2002).

As noted above, ECD is a framework or mechanism for making explicit the evidence-based reasoning practices of interdisciplinary teams charged with assessment design, delivery, scoring, and reporting. At a fine-grained technical level the decomposition of the argumentation is based on Toulmin's argument schema (1958/2003), which is well known to the writing assessment community (White, Elliot, & Peckham, 2015, Figure 3.5) and the educational measurement community (Mislevy, 2007, Figure 1). Moreover, Bachman (2005) extended the Toulmin diagram/argument from assessment interpretations to assessment decisions. Recent scholarship has elaborated on the Toulmin model as a way to formalize three credentials of an evidential datum—relevance, credibility, and inferential force—that must be established in analyzing its relationship to a hypothesis (Anderson, Schum, & Twining, 2005). As the Toulmin model reveals, evidence, warrants, claims, and qualifications are important in establishing the two aspects of overarching validation arguments proposed by Kane (2006, 2013, 2015) noted above: an interpretive argument, which documents the network of inferences and assumptions leading from the performance to the conclusions and decisions on use; and the validity argument, which serves as a check on the interpretative argument by evaluating its plausibility. As Mislevy (2007) has observed, the Toulmin model serves an important function, which is to render the validity argument "public, sharable, and reusable" (p. 437).

For CCSSI ELA-W assessments, the validity argument is used as a vehicle to articulate the characteristics and boundaries of a designated construct. In the next section we describe how the *Standards* and the ECD framework have been instrumental in the development of curricular and assessment efforts surrounding the CCSSI ELA-W.

STANDARDS-BASED VALIDITY ARGUMENTS AND ECD PRACTICES: INTEGRATION INTO CCSSI ELA-W ASSESSMENT

In this section we use three Smarter Balanced and PARCC assessment reports to illustrate how *Standards*-based validity arguments and ECD practices have been integrated into assessment work for the CCSSI ELA-W.

Consider first the report entitled "Smarter Balanced Assessment Consortium: Comprehensive Research Agenda" (Sireci, 2012). The author's detailed validity argument is intended to "put potential misperceptions to rest" that the Consortium has adopted a research agenda that has unfortunately resulted in fragmentation (p. 63). To counterbalance these claims, Sireci advanced seven principles, or claims, of the assessments: that they are grounded in a standards-based curriculum and are part of an integrated system; that they produce evidence of student performance; that they are part of a state-led effort with a transparent and inclusive governance structure; that they are structured to continuously improve teaching and learning; that they provide useful information on multiple measures educative for all stakeholders; that their implementation strategies adhere to established professional standards; and that teachers have been integrally involved in the development and scoring of the assessments.

The claims are then followed by two tables: one providing the details of 55 studies proposed by the Consortium; and the other providing a way to map the studies to the five sources of evidence—validity based on test content, internal structure, response processes, relationships to other variables, and consequence—identified by the consortium. Explicitly and by name, the report utilizes ECD as a way to evaluate the degree to which the assessment specifications represent the CCSSI and the degree to which the constructed response items themselves capture the assessment specifications (p. 25).

Second, consider the "Memorandum on Instructional Sensitivity Considerations for the PARCC Assessments" (Way, 2014). The author uses a validity argument to map a research agenda of the instructional sensitivity of the assessments, defined as the extent to which a test item is sensitive to instruction. Rather than viewing instructional sensitivity as an isolated concept, Way proposed that it is "tied up with related concepts governing what is supposed to be taught in the classroom, what is actually taught in the classroom, and how well tests and items align with what is taught" (p. 3).

Way noted that while the PARCC assessments are designed to measure integrated stills (such as those that require evaluation, synthesis, analysis, reflective thought, and research), this particular type of integration might not be taught in a given school year. As such, the assessments could possibly become tests in search of a curriculum. To address this dilemma, Way proposes the use of IRT plots as predictors based on ability level, as well as classroom observations and teacher reports of classroom content. Framing a research agenda in anticipation of validity argument used to establish assessment and curricular connections suggests the centrality of evidentiary reasoning throughout the CCSSI design process.

Finally, consider the PARCC report "Evidence and Design Implications Required to Support Comparability Claims" (Luecht & Camara, 2011). In it, the

authors have paid close attention to score use—to the ways to compare student performance across schools, districts and states, to measure growth across grade levels, and to evaluate year-to-year changes. Because of the importance of such comparisons and goal setting, the authors emphasized the need for "well-articulated, cognitively-based constructs" based on the CCSSI, which should be developed in order to establish the ordered claims and evidence requirements by grade level.

Luecht and Camara noted that the ECD approach "may offer some advantages over conventional item design and test specifications because such new design approaches prioritize more explicit connections between items from task models which are directly derived from evidence" (p. 15). Task models resulting from ECD, as the report acknowledges, allow designers to control for content through an emphasis on cognitive demand and yield greater efficiency in development of the assessment over time.

As these three examples demonstrate, strategic use of *Standards*-based and ECD frameworks at the planning stage yields a validity agenda and evidentiary processes. In the next section, we provide some guiding questions for stakeholder networks that can help to raise awareness about what it means to translate the different concepts in the *Standards* and ECD into thoughtful assessment practice that supports meaningful interpretations and decisions.

GUIDING QUESTIONS FOR STAKEHOLDERS

In this section we turn to four key stakeholder groups—students and guardians, teachers and administrators, legislators, and workforce leaders—and provide questions intended to empower each to grapple with the decisions that must be made as a result of information issuing from the three interpretative frameworks discussed above. It is our belief that these stakeholders would be well served by raising a series of such very specific questions that can lead to informed judgments regarding score use stemming from the assessment of the CCSSI ELA-W by Smarter Balanced and PARCC. Made on a state-by-state basis this judgment will, we argue, be best made if informed by the perspective gained when key stakeholders think along the same lines.

More broadly, the perspective offered by these questions is commensurate with comprehensive validation arguments and coherent evidentiary reasoning practices embodied in the *Standards* and ECD, respectively. It is therefore appropriate to think of the questions raised in Tables 3.1 and 3.2 as applicable to any large-scale assessment of ELA-W that has been created under the contemporary evidentiary reasoning practices presented in this paper. As evidence of the force of multidisciplinary research, we note that our perspective is congruent with the

emphasis on networks and their logic proposed by Gallagher (2011); that is, the questions we provide are intended to provide "analytic tools for understanding how actors exercise power by virtue of their *locations* and *relations*" (p. 466, emphasis in original).

HEURISTICS AND BIAS

We have informed our questions by the heuristics and biases research of Amos Tversky and Daniel Kahneman. Together, these scholars in the field of decision-science advanced a program of research since the early 1970s that revolutionized our understanding of human judgment (Kahneman, 1973; Tversky & Kahneman, 1973). Their system is too complex for discussion save its core concept: attention to the heuristics that we use to ask questions and the cognitive biases that result in tangled reasoning. Defined as "a simple procedure that helps find adequate, though often imperfect, answers to difficult questions," Kahneman (2011, p. 98) had found that heuristics are a consequence of intuition (termed System 1 thinking) and strategy (the corrective System 2). While we think associatively, metaphorically, and causally with some ease and accuracy as a result of intuition, he noted, even the most educated have trouble thinking about more abstract concepts like probabilities and uncertainties to make appropriate strategic inferences.

Complexities that arise from the overestimation of what we know and the underestimation of chance are potentially important for two reasons in educational assessment and measurement. First, as we have demonstrated in our three interpretative frameworks, modern assessment requires that we embrace evaluative techniques as complex as the humans we seek to learn about. In this process, meaningful and informed questions are of paramount importance lest we underestimate the demands of assessment. Just below the surface, foundational concepts are associated with probabilities, and the nuanced nature of the evidence produced from modern assessment systems requires acknowledgment of contingency. Second, while we are experientially familiar with the forms of logic that assessment designers use in test design, we know less about the forms of logic that the stakeholders use to make interpretations and decisions based on assessment scores. The more we can learn about the logic of stakeholder networks, the better we will be able to communicate our evidentiary processes.

In the absence of such information, the questions in Table 3.1 and Table 3.2 are intended to help networks of non-specialists structure conversations that may, in turn, help specialists learn more about the cares and concerns of all stakeholders. The guiding questions are designed to help uncover implicit assumptions, potential biases in reasoning, and connections between various

design decisions within the teaching and assessment ecosystem. We deeply believe that it is of value to connect the logic of educational measurement and writing studies research with the logic of heuristics and biases research, if only to remind everyone that complex ventures obligate us to think in complex ways.

In each table, we have used the *Standards* to generate a series of broad foundational and operational questions that, in turn, are made specific by focusing on specific facets of measurement. Because our focus is on an educational assessment, we have integrated that application into the foundational and operational question and, hence, no additional table is provided for that section of the *Standards.*

Table 3.1. Foundational Questions for Stakeholder Groups in English Language Arts-Writing

Standard	Students and Guardians	Teachers and Administrators	Legislators	Workforce Leaders
Validity: "Clear articulation of each intended test score should be set forth, and appropriate validity evidence in support of each intended interpretation should be provided" (p. 23).	How will scores be used? • Will scores be used to draw conclusions about an individual student's present writing ability? • Will scores be used to make decisions about an individual student's ability to perform in subsequent courses?	Has validity evidence been provided that will allow interpretation of test scores for a specified use? • Has the sample of test takers been defined from which scores have been drawn? • How does this sample represent the population of interest in terms of socio-demographic or developmental characteristics?	What evidence has been provided that the assessment has positive consequences for stakeholders? • If unintended consequences have occurred, have investigations been made of both categories of validity evidence and factors external to the assessment?	What evidence has been provided that the assessment captures a construct that is relevant in the workplace? • If the scores are to be used for credentialing, how will they be distributed and what interpretative materials will be provided?
Reliability/Precision: "Appropriate evidence of reliability/precision should be provided for the interpretation for each intended score use" (p. 42).	Have estimates of reliability/precision of scores been provided so that scores use can be justified? • Have estimates of reliability/precision been provided for each relevant student subgroup so that comparisons can be made between individual and group performance?	How do the methods for estimating subscores contribute to the interpretation and justification of score use? • In the case of automated scoring of essay items, have descriptions of the scoring algorithms and scores associated with those algorithms been made available?	What evidence has been provided that administrative conditions of the assessment have remained stable? • What evidence has been provided of reliability/precision to justify score interpretation and use?	When compared to a meaningful workplace criterion variable, what evidence has been provided that the assessment reliably predicts workplace performance? • Is workplace performance reliably predicted for subgroups of employees?

Standard	Students and Guardians	Teachers and Administrators	Legislators	Workforce Leaders
Fairness: "All steps in the testing process, including test design, validation, development, administration, and scoring procedures, should be designed in such a manner as to minimize construct-irrelevant variance and to promote valid score interpretations for the indented used of all examinees in the intended population" (p. 63).	What evidence has been provided that scores contribute to equality of opportunity and opportunity to learn for individual students? • Has each student been provided with the opportunity to learn the construct as it is being assessed?	What evidence has been provided that principles of universal design have been followed in creating the assessment? • Have barriers been identified and mitigated that impede access to the construct as it is being assessed?	Have safeguards been developed to discourage inappropriate score interpretations and score use? • If value added methods have been considered in determining school or teacher performance based on test scores, does evidence justify a fixed weight in decision-making?	What evidence is available that the scores have the same meaning for all individuals? • If meanings differ for different individuals or groups, how will evidence be provided to justify score interpretation and use?

Table 3.2. Operational Questions for Stakeholder Groups in English Language Arts-Writing

Standard	Students and Guardians	Teachers and Administrators	Legislators	Workforce Leaders
Test Design and Development: "Tests and testing programs should be designed and developed in a way that supports the validity of interpretations of the test scores for their intended uses. Test developers and publishers should document steps taken during the design and development process to provide evidence of fairness, reliability, and validity for intended uses for individuals in the intended examinee population" (p. 85).	What is the relationship among the following: the curriculum at the individual student's school, the curricular goals, and the assessment? How have the steps of the assessment processes been documented and communicated by those responsible for developing the assessment?	How have assessment specifications been provided regarding the construct under examination, the examinee populations, and the proposed interpretations of scores and their use? How have the assessment developers communicated the standards for item review, the administration and scoring procedures, and the basis for revision of the assessment?	How have the assessment developers demonstrated that they have designed their assessments in ways to support the validity, reliability/precision, and fairness associated with their intended use? What processes have been established, and what funds have been designated, to revise the assessment based on new information resulting from the present administration?	How have the assessment developers demonstrated that their test development and design process have taken into consideration important workplace needs associated with construct competency? How have rationales been developed that justify linkages between test design and development processes and workplace needs for credentialing, selection, placement, and promotion?

Standard	Students and Guardians	Teachers and Administrators	Legislators	Workforce Leaders
Scores, Scales, Norms, Score Linking, and Cut Scores: "Test scores should be derived in a way that supports the interpretation of test scores for the proposed uses of tests. Test developers and users should document evidence of fairness, reliability, and validity of test scores for their proposed use" (p. 102).	If decisions regarding placement and progression are to be made from the assessment, have cut scores been established for categories of student performance? If cut scores have been established, has the procedure been documented and communicated in terms of both technical specifications and policy decisions?	If cut scores have been established, are these scores to be used for descriptive or decision-making purposes? • How have assurances been made that the establishment of cut scores does not undermine the validity of score interpretations?	How have the assessment developers demonstrated that scores have been normed with student populations similar to those found at individual schools or school districts? • How have differentiated norms been established for different gender, race/ethnicity, language, disability, economically disadvantages, grade, and age groups?	How have the assessment developers demonstrated that the norms and cut scores established are congruent with workforce populations and employment needs? • How have interpretations been established to help employers interpret and use the established norms and cut scores?
Test Administration, Scoring, Reporting, and Interpretation: "To support useful interpretations of score results, assessment instruments should have established procedures for test administration, scoring, reporting, and interpretation. Those responsible for administering, scoring, reporting, and interpreting should have sufficient training and supports to help them follow the established procedures. Adherence to the established procedures should be monitored, and any material errors should be documented and, if possible, corrected" (p. 114).	How have the assessment developers designed the digital administration so that technical disruptions do not contribute to construct-irrelevant variance? Have distinctions been made between accommodations for test takers based on need and accommodations based on misalignment between the digitally-based assessment and the print-based curriculum?	Because different stakeholder groups may administer, score, report, and interpret the assessment, how have procedures been established to ensure that score interpretation and use are not compromised by failure of standardization? How have assessment developers demonstrated that standardization will ensure that students have the same ability to demonstrate their competency?	How have resources been leveraged to ensure that the diverse stakeholder groups needed to administer, score, report, and interpret the assessment have the competency and resources necessary to ensure standardization? In cases of students with disabilities or different language backgrounds, how have nonstandard models been established that will allow these students to demonstrate competence?	How have test administration, scoring, reporting, and interpretation processes been designed so that scores can be used to establish connections with workplace needs? How have standardization processes resulted in the anticipation and removal of construct-irrelevant variance so that scores from the assessment can be used on a long-time basis?

Standard	Students and Guardians	Teachers and Administrators	Legislators	Workforce Leaders
Supporting Documentation for Tests: "Information relating to tests should be clearly documented so that those who use tests can make informed decisions regarding which test to use for a specific purpose, how to administer the chosen test, and how to interpret test scores" (p. 125).	When scores are released, how have interpretations appropriate for both students and their guardian been communicated? When technical information on development and scoring is released to students and guardians, has this information been adequately explained so that score interpretation is informed?	How have documents been prepared so that teachers and administrators can understand and communicate to students and their guardians the development process, administration and scoring, and appropriate use of scores associated with the assessment? What milestones have been established so that these supporting documents are made available to teachers and administrators in a timely manner?	How have resources been allocated so that supporting documentation has been examined for its intended audiences? Based on knowledge about aim, genre, and discourse communities, have supporting documents been prepared so that they will discourage score misuse and contribute to justified score interpretation?	How has supporting documentation been prepared so that workplace users of the assessment will be able to receive additional interpretative support when summaries of technical information are needed to interpret scores? In cases where the workplace is international in nature, have supporting materials been prepared in digital form and translated into languages users will need to interpret assessment scores?
Rights and Responsibilities of Test Takers: "Test takers have the right to adequate information to help them prepare for a test so that the test results accurately reflect their standing on the construct being assessed and lead to fair and accurate score interpretations. They also have the right to protection of their personally identified score results from unauthorized access, use, or disclosure. Further, test takers have the responsibility to present themselves accurately in the testing process and to respect copyright in test materials" (p. 133).	How has the student been provided with accurate, free information about the assessment? • As a means of reducing construct-irrelevant variance, how has the student been provided with practice access to the digital environment in which the test will be administered?	How has the instructor provided students with information about the assessment, intended score use, scoring criteria, administrative policy, available of accommodations, and confidentiality? • How have the students been informed of their rights and the rights of their parents to access assessment results and be protected from unauthorized use of results?	In order to protect students from potentially adverse consequences, how has the legislative process been used to delay justified score use? • If the legislative process has been used to delay score use, how have specific determinations been made regarding a range of decisions and a timeline for justified score use?	If assessment scores are to be used to determine workplace competency, how have assurances be established to assure that students have information about how employers are using scores? • If assessment scores are to be transferred to employers, how have the data systems be designed to assure confidentiality?

Standard	Students and Guardians	Teachers and Administrators	Legislators	Workforce Leaders
Rights and Responsibilities of Test Users: "Test users are responsible for knowing the validity evidence in support of the intended interpretations of scores on tests that they use, from test selection through the use of scores, as well as common positive and negative consequences of test use. Test users also have a legal and ethical responsibility to protect the security of test content and the privacy of test takers and should provide pertinent and timely information to test takers and other test users with whom they share test scores" (p. 142).	What assurances exist that those who use assessment scores have the training and credentials necessary for responsible score interpretation and use? • How have those individuals been prepared to deliver consistent and timely interpretations of scores and their use?	How has a clear and distinct role been established for instructors in the communication of assessment results? • If teachers and administrators disagree with justified interpretation and use, have processes been designed to allow warranted disagreement while maintaining a stance that will not compromise student motivation or parental interest?	In order to protect students from potential misinterpretations of scores, how have legislators minimized these foreseeable misrepresentations? • What processes have legislators put in place to prevent score misrepresentations?	How have workplace leaders been educated to interpret and use scores in ways leading to the advancement of equity and opportunity to learn? • How have workplace leaders been educated about anticipating negative consequences of score use?

A TOWN HALL THOUGHT EXPERIMENT

To envision how the questions in Table 3.1 and Table 3.2 might be used together, we propose a thought experiment: a series of town hall meetings in which local stakeholders are brought together to address assessment issues associated with the CCSSI ELA-W. If frequently asked questions arising from these tables were prepared and distributed in advance, fact finding could occur before the meeting and the participants could then focus on establishing common ground.

Imagine that town hall meeting were to occur in the beginning of the 2015 school year, a time at which many questions of proper score interpretation and use remain unanswered. Using questions from Table 3.2 in order to establish the relationships among validity, reliability/precision, and the operational obligations of assessment developers, curriculum developers, and teachers, students and their guardians might justifiably ask how scores have been established for categories of student performance and if those scores will, in turn, lead to decisions regarding promotion and placement.

During the imagined town meeting, attention might be drawn to the Smarter Balanced Consortium (2014b) document entitled "Interpretation and Use of Scores and Achievement Levels" that we discussed in the previous section. Recall that scale scores and achievement level descriptors are identified in alignment with the *Standards* in the document. Using the validity questions from Table 3.1, teachers and administrators might focus on discussing the relationship between test results and the curriculum in their classrooms, schools, and districts. Choices in test design, administration, and reporting become critical as questions are raised regarding the constructive alignment—the integrated instructional and assessment systems and efforts used to map learning activities to outcomes (Biggs & Tang, 2011)—that must be established among the individual student's school, the CCSSI ELA-W, and Smarter Balanced and PARCC assessments. Critically discussing the implications of various decisions based on questions around constructive alignment would help establish a common understanding of the extent to which the scores are faithful demonstrations of individual student ability.

Similarly, in using the questions to investigate sources of evidence related to reliability, teachers and administrators would benefit by paying attention to the concept of measurement precision and not just an overly simplistic single descriptive statistic (Sireci, 2012). Estimates of score reliability (internal consistency) and those based on examining students more than once (parallel forms) thus become important sources of information to consider when determining appropriate and less appropriate interpretations of scores.

For students, guardians, teachers, and administrators, questions of what constitutes appropriate score interpretation and use would be especially relevant in light of the disaggregated information about student performance obtained from the Smarter Balanced field test that was administered between March and June 2014 (Smarter Balanced, 2014a). The test revealed clear performance differences among key student subgroups that allow for a critical discussion of how these differences are related to potential differences in opportunities to learn.

Specifically, at the Grade 11 level, 40.9 percent of total students examined (n = 31,018) met the cut score of Level 3 (or above) in achievement levels ranging from Level 1 (novice) to Level 4 (advanced). Among American Indian/Alaskan Native students (n = 777), 26.6 percent passed; Asian students (n = 2,334) passed at 54.1 percent; Black/African American students (n = 2,552) passed at 21.2 percent; Hispanic/Latino students (n = 10,041) passed at 32.4 percent; Native Hawaiian/Other Pacific Islander students (n = 195) passed at 32.8 percent; White/Caucasian students (n = 16,020) passed at 46.2 percent; Multi-ethnic/Multi-racial students (n = 889) passed at 45.1 percent. Among those enrolled in an Individualized Education Program (n = 2,084), 9.0 percent passed; among those classified as Limited English Proficient/English language learners (n =

1,767), 5.7 percent passed; among those classified under special program enrollment preventing discrimination based on disability (n = 366), 36.1 percent passed; among those classified as Economically Disadvantaged students (n = 13,962), 32.6 percent passed (Smarter Balanced, 2014a, p. 12).

The literature associated with opportunity to learn is a particularly rich framework for advancing instructional equity among student groups (Moss, Pullin, Gee, Haertel, & Young, 2008). In terms of the fairness questions raised in Table 3.1, using scores as a way to promote opportunity to learn can help in identification of barriers to success and creation opportunities to foster educational advancement. Making *Standards*-based conceptual and empirical connections among issues around validity, reliability/precision, and fairness through the lens of opportunity to learn is, we believe, an especially powerful logic that can be used to guide discussion of assessment results.

Because the continuum among school, college, and workplace writing appears to exhibit more disjuncture than congruence (Burstein, Elliot, & Molloy, in press; Melzer, 2014), Table 3.2 might be used to call attention to the especially difficult generalization inference between academic and workplace writing established by the CCSSI ELA-W. Because the CCSSI specifically identifies both academic and workplace readiness, it is reasonable for post-secondary academic and workplace leaders to ask questions that allow them to obtain more clarity on critical assessment design, delivery, and scoring decision. Moreover, it is important that the ensuing discussions are used to elucidate any remaining ambiguities around how performance certification decisions should be informed by scores from CCSSI ELA-W assessments. In terms of the report "Interpretation and Use of Scores and Achievement Levels" that we discussed in the previous section, questions of score use become especially important in light of the fact that parallel operational definitions and frameworks are still under development for career readiness (Smarter Balanced Consortium, 2014b, p. 2). Present at the imagined town meeting, academic and workplace leaders could certainly highlight issues regarding the learning continuum.

Legislators will want to attend to both the intended and unintended consequence of the CCSS ELA-W in terms of validity evidence and factors external to the assessment. Determination of score use is especially important in the case of value-added methods used to make inferences about teacher performance, especially when current research reveals that the scores resulting from such procedures may be systematically biased in favor of some instructors and against others (Haertel, 2013). In anticipating legal issues associate with CCSSI ELA-W assessment, stakeholders will find the empirical techniques associated with quantifying disparate impact equally useful (Poe, Elliot, Cogan, & Nurudeen, 2014) so that they can meaningfully help to advance opportunities to learn.

CONCLUSION

As these examples from our town hall thought experiment illustrate, while the questions in Table 3.1 and Table 3.2 are not meant to be exhaustive, they might prove useful for three reasons. First, because their phrasing is informed by the program of research begun by Tversky and Kahneman (2011), it is possible that such questions might act as a bridge between the kinds of evidence-based, argumentative logic that assessment designers employ in ECD (Mislevy, Steinberg, & Almond, 2003) and the availability, representativeness, and adjustment involved in heuristic reasoning that other assessment stakeholders may use in decision-making (Gilovich & Griffin, 2002). Bridging the logic of the assessment developer and the logic of the assessment user is a worthy goal that might be served by attention to decision-making under uncertainty. Tables 3.1 and 3.2 contribute to our desire to help stakeholders ask principled questions about assessment design, score use, and consequences. Second, attention to diverse reasoning processes is inherent in the social cognitive view of writing that informs the CCSSI ELW-W and its assessment. As Gilovich & Griffin (2002) have observed, the heuristic reasoning program fits well with our present understanding of how the mind works. Third, the imagined town meeting as the forum for deliberative discussion suggests the need for the development of what Rawls (2001) has referred to as overlapping consensus. The aim of reasonable pluralism is a worthy goal that may be achieved if common referential frames are established of the kinds we have suggested here.

The concepts we have presented in this paper are complex, and the challenges we have identified are real and must be addressed. We believe that our collective logic can be guided by interpretative frameworks such as the three presented here that speak to core issues associated with advancement of opportunity to learn. As present curricular and assessment innovations merge to produce information about student performance, many questions nevertheless remain. Especially notable are questions regarding the relationship between assessment and opportunity structure. Future work must turn to questions left unanswered here.

REFERENCES

Addison, J., & McGee, S. J. (2015). To the core: College composition classrooms in the age of accountability, standardized testing, and Common Core State Standards. *Rhetoric Review*, *34(2)*, 200-218.

Almond, R. G., Steinberg, L. S., & Mislevy, R. G. (2002). A four process architecture for assessment delivery, with connections to assessment design. Educational Testing Service. https://www.education.umd.edu/EDMS/mislevy/papers/ProcessDesign.pdf

American Educational Research Association, American Psychological Association, & National Council of Measurement in Education. (1999). *Standards for educational and psychological testing*. American Educational Research Association.

American Educational Research Association, American Psychological Association, & National Council of Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Anderson, T., Scum, D., & Twining, W. (2005). *Analysis of evidence* (2nd ed.). Cambridge University Press.

Applebee, A. N. (2013). Common Core State Standards: The promise and the peril in a national palimpsest. *English Journal*, *103(1)*, 25-33.

Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly, 2*, 1-34.

Benjamin, L. T. (1997). The origin of psychological species. *American Psychologist*, *52(7)*, 725-732.

Berninger, V. W. (Ed.). (2012). *Past, present, and future contributions of cognitive writing research to cognitive psychology*. Taylor and Frances.

Biggs, J., & Tang, C. (2011). *Teaching for quality learning at university* (4th ed.). McGraw-Hill.

Burstein, J., Elliot, N., & Molloy, H. (2016). Informing automated writing evaluation using the lens of genre: Two studies. *Calico Journal*, *33*(1), 117-141.

Common Core State Standards Initiative. (2015a). English language arts standards. http://www.corestandards.org/ELA-Literacy/

Common Core State Standards Initiative. (2015b). English language arts standards, writing, grade 11-12. http://www.corestandards.org/ELA-Literacy/W/11-12/

Common Core State Standards Initiative. (2015c). English language arts standards, writing, introduction, 6-12. http://www.corestandards.org/ELA-Literacy/W/introduction-for-6-12/

Cumming, A. (2013). Assessing integrated writing tasks for academic purposes: Promises and perils. *Language Assessment Quarterly*, *10*, 1-8.

de Ayala, R. J. (2008). *The theory and practice of item response theory*. Guilford Press.

De Boeck, P., & Wilson, M. (Eds.). (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. Springer.

Duncan, A. (2015). Remarks by U.S. Secretary of Education Arne Duncan on the 50th anniversary of Congress passing the Elementary and Secondary Education Act. http://www.ed.gov/news/speeches/remarks-us-secretary-education-arne-duncan-50th-anniversary-congress-passing-elementary-and-secondary-education-act

Educational Testing Service (2014). *ETS standards for quality and fairness, 2104*. Educational Testing Service. https://www.ets.org/s/about/pdf/standards.pdf

Eliot, C. W. (1892). Wherein popular education has failed. *The Forum*, *14*, 411-428.

Elliot, N., & Perelman, L. (Eds.). (2012). *Writing assessment in the 21st century: Essays in honor of Edward M. White*. Hampton Press.

Fernberger, S. W. (1932). The American Psychological Association: A historical summary, 1892-1930. *Psychological Bulletin*, *29(1)*, 1-89.

Flower, L. (1994). *The construction of negotiated meaning: A social cognitive theory of writing*. Southern Illinois University Press.

Flower, L., & Hayes, J. R. (1981). A cognitive process theory of writing. *College Composition and Communication*, *32(4)*, 365–387.

Gallagher, C. W. (2011). Being there: (Re)making the assessment scene. *College Composition and Communication*, *63(3)*, 450-476.

Gardner, H. (2006). *Five minds for the future: Leadership for the common good*. Harvard Business School Press.

Gates, B., & Gates, M. (2015). College-ready education. http://www.gatesfoundation.org/What-We-Do/US-Program/College-Ready-Education

Gilovich T. & Griffin, D. (2002). Introduction—Heuristics and biases: Then and now. In T. Gilovich, D. Griffin, & D, Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 1-18). Cambridge University Press.

Graham, S. (2006). Writing. In P. Alexander & P. Winne (Eds.), *Handbook of educational psychology* (pp. 457-478). Erlbaum.

Graham, S., McKeown, D., Kiuhara, S. A., Harris, K. R. (2012). A meta-analysis of writing instruction for students in the elementary grades. *Journal of Educational Psychology*, *104*, 879-896.

Graham, S., & Perin, D. (2007). A meta-analysis of writing instruction for adolescent students. *Journal of Educational Psychology, 99(3),* 445-476.

Hall, G. S. (1883). The contents of children's minds. *Princeton Review*, *11*, 249-272.

Haertel, E. H. (2013). Reliability and validity of inferences about teachers based on student test scores. [William H. Angoff 14th memorial lecture]. National Press Club, Washington, DC. https://www.ets.org/Media/Research/pdf/PICANG14.pdf

Hayes, J. R. (2012). Modeling and remodeling writing. *Written Communication*, *29(3)*, 369-388.

Hillocks, G. (1986). *Research on written composition: New directions for teaching*. National Council of Teachers of English.

Johnson, K. (2013). Beyond standards: Disciplinary and national perspectives on habits of mind, *College Composition and Communication*, *64(3)*, 517-541.

Kahneman, D. (1973). *Attention and effort*. Prentice-Hall.

Kahneman, D. (2011). *Thinking fast and slow*. Farrar, Straus and Giroux.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed.). (pp. 17-64). American Council on Education and Praeger.

Kane, M. T. (2013). Validating the interpretation and uses of test scores. *Journal of Educational Measurement*, *50(1)*, 1-73.

Kane, M. T. (2015). Explicating validity. *Assessment in Education: Principles, Policy & Practice*, *23(2)*, 198-211.

Ketterlin-Geller, L. R. (2008). Testing student with special needs: A model for understanding the interaction between assessment and student characteristics in a universally designed environment. *Educational Measurement: Issues and Practice*, *27(3)*, 3-16.

Kliebard, H. M. (2004). *The struggle for the American curriculum, 1893-1958* (3rd ed.). Routledge.

Leijten, M., Van Waes L., Schriver, K., & Hayes, J. R. (2014). Writing in the workplace: Constructing documents using multiple digital sources. *Journal of Writing Research*, *5(3)*, 285-337.

Lindemann, E. (Ed.). (2010). *Reading the past, writing the future: A century of American literacy education and the National Council of Teachers of English*. National Council of Teachers of English.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Erlbaum.

Luecht, R. M., & Camara, W. J. (2011). Evidence and design implications required to support compatibility claims. http://www.parcconline.org/sites/parcc/files/PARCCWhitePaperRLuechtWCamara.pdf

Melzer, D. (2014). *Assignments across the curriculum: A national study of college writing*. Utah State University Press.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.) (pp. 13-103). American Council on Education and Macmillan.

Mislevy, R. J. (2007). Validity by design. *Educational Researcher, 36(8)*, 463-469.

Mislevy, R. J. (2008). How cognitive science challenges the educational measurement tradition. http://umdperg.pbworks.com/f/CommentaryHaig_Mislevy.pdf

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2004). A brief introduction to evidence-centered design (ETS Research Report-03-16). Educational Testing Service. https://www.ets.org/Media/Research/pdf/RR-03-16.pdf

Mislevy, R. J., Haertel, G., Cheng, B., Ructtinger, L., DeBarger, A., Murray, E., Rose, D., Gravel, J., Colker, A. M., Rustein, D., & Vendlinski. T. (2013). A "conditional" sense of fairness in assessment. *Educational Research and Evaluation: An International Journal on Theory and Practice*, *19(2-3)*, 121-140.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessment (with discussion). *Measurement: Interdisciplinary Research and Perspective*, *1*, 3-62.

Moss, P. A., Pullin, D. C., Gee, J. P., Haertel, E. H. & Young, L. J. (Eds.), (2008). *Assessment, equity, and opportunity to learn.* Cambridge University Press.

National Research Council. (2012). *Education for life and work: Developing transferable knowledge and skills in the 21st century*. Committee on Defining Deeper Learning and 21st Century Skills, J. W. Pellegrino, & M. L. Hilton (Eds.). Board on Testing and Assessment and Board on Science Education, Division of Behavioral and Social Sciences and Education. The National Academies Press.

Phelps, L. W., & Ackerman, J. M. (2010). Making the case for disciplinarity in rhetoric, composition, and writing studies: The visibility project. *College Composition and Communication*, *18(1)*, 180-215.

Poe, M., Elliot, N., Cogan, J. A., & Nurudeen, T. G. (2014). The legal and the local: Using disparate impact analysis to understand the consequences of writing assessment. *College Composition and Communication*, *65(4)*, 588-611.

Pullin, D. C. (2008). Assessment, equity, and opportunity to learn. In P. A. Moss, D. C. Pullin, J. P. Gee, E. H. Haertel, & L. J. Young (Eds.), *Assessment, equity, and opportunity to learn* (pp. 333-351). Cambridge University Press.

Ravitch, D. (2010). *The death and life of the great American school system: How testing and choice are undermining education*. Basic Books.

Rawls, J. (2001). *Justice as fairness: A restatement*. Harvard University Press.

Rice, J. M. (1893). *The public-school system of the United States*. Century.
Rogers, L., & Graham, S. (2008). A meta-analysis of single subject design writing intervention research. *Journal of Educational Psychology, 100(4)*, 879-906.
Shannon, P., Whitney, A. E., & Wilson, M. (2014). The framing of the Common Core state standards. *Language Arts*, *91*, 295-302.
Sireci, S. G. (2012). Smarter balanced assessment consortium: Comprehensive research agenda. *Report of recommendations*, 1-90.
Smarter Balanced Assessment Consortium. (2014a). Disaggregated data from the Smarter Balanced field test.
Smarter Balanced Consortium (2014b). Interpretation and use of scores and achievement levels.
Snyder, T. D. (1993). *120 years of American education: A statistical portrait*. National Center for Education Statistics. http://nces.ed.gov/pubs93/93442.pdf
Snyder, T. D., & Dillow, S. A. (2015). *Digest of education statistics 2013* (NCES 2015-011). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education.
Sparks, J. R., Song, Y., Brantley, W., & and Liu, O. L. (2014). Assessing written communication in higher education: Review and recommendations for next-generation assessment (ETS Research Report No. RR-14-37). Educational Testing Service.
Toulmin, S. E. (1958). *The uses of argument*. Cambridge University Press.
Toulmin, S. E. (2003). *The uses of argument*. (2nd ed.). Cambridge University Press.
Tucker, B. (2009). The next generation of testing. *Educational Leadership*, *67*, 48-53.
Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, *5*, 202-232.
Ward, L. F. (1883). *Dynamic sociology, or applied social science as based upon statistical sociology and the less complex sciences*. Appleton.
Way, W. (2014). Memorandum on instructional sensitivity considerations for the PARCC assessments.
White, E. M., Elliot, N., & Peckham, I. (2015). *Very like a whale: The assessment of writing programs*. Utah State University Press.

PART 2.

POLITICS AND PUBLIC POLICY OF LARGE-SCALE WRITING ASSESSMENT

RETROSPECTIVE.

THE POLITICS AND PUBLIC POLICY OF LARGE-SCALE WRITING ASSESSMENT

Carolyn Calhoon-Dillahunt
Yakima Valley College

For more than 150 years, standardized testing has been a part of the U.S. education system. Almost from the outset, standardized testing was inextricably linked to writing assessment and, thus, to writing instruction and, ultimately, to writing as a discipline. Early concerns about the "problem" of student writing revealed by standardized assessments resulted in increased attention to writing and writing instruction for teachers, for schools, and, eventually, for policymakers. As a result, for good and bad, writing (granted, often defined and assessed in reductive ways) holds a position of primacy in assessment and in educational policy, a position that garners attention and resources, but also scrutiny and intrusion.

In this section introduction, I briefly trace the history of large-scale writing assessment and how it has been entwined with politics and policymaking, situating the specific essays featured in Part Two of this collection in the "reform and accountability era" of large-scale standardized testing. From there, I discuss core themes around which these distinct articles coalesce: the policy intentions for and resulting uses and misuses of large-scale writing assessment in the 2000s; the consequences of mandated writing standards and high stakes writing assessments on curriculum, teachers and teaching, and students; and the possibilities enabled through some large-scale writing assessments.

A BRIEF HISTORY OF THE RELATIONSHIP BETWEEN LARGE-SCALE WRITING ASSESSMENT AND POLICY

Although the purposes of standardized assessments have shifted over the past century and a half, gatekeeping and tracking have been primary among them. The earliest standardized tests focused on achievement of basic skills, such as language and literacy skills. Such tests were quickly taken up by selective colleges to determine admissions (National Education Association [NEA], 2020) and

DOI: https://doi.org/10.37514/PER-B.2024.2166.2.22

placement into "remedial" writing coursework, starting with Ivy League schools in the late 1800s (Haswell, 2004). These early forays into writing assessment as gatekeeping planted the seeds of both basic writing and near universal first-year writing requirements in postsecondary study.

By the turn of the century, the founding of the College Entrance Examination Board meant that admissions testing and thus writing assessment became "outsourced," and assessment became a professional, prolific, and profitable industry separate from the institutions that relied on their results (Huot, O'Neill, & Moore, 2010). In *Before Shaughnessey,* Ritter (2009) observes that accessibility of higher education, increasingly available to the masses after WWI and even more so with the GI Bill post-WWII, shifted the focus of writing assessment. Writing assessment became preoccupied with surface-level correctness, and remediation was prescribed to resolve students' perceived lack of preparation for college-level writing. Over the course of the 20th century, writing placement also became increasingly disconnected from writing curriculum, as many institutions, especially open-admissions institutions, shifted from locally scored timed writing exams to externally scored standardized indirect writing assessments (Haswell, 2004).

In the 20th century, standardized testing expanded to assess proficiency, aptitude, intelligence, and more. However, according to Rosales and Walker (2021), "since their inception almost a century ago, the tests have been instruments of racism and a biased system," founded on the pseudo-science, eugenics, and grounded in *white racial habitus* (Inoue, 2015). Nowhere is this racism more apparent than in standardized writing assessments. The purposes for such testing grew beyond simple gatekeeping for university admissions to diagnosing deficits, measuring skill sets, and predicting future performance. As a result, standardized testing was increasingly tied to educational decision-making (NEA, 2020), with the results of a single measure–generally an indirect measure embedded in White language and culture supremacy–being used to classify, rank, track, and exclude students. These approaches disproportionately affected historically underserved students, particularly students of color. Political support of large-scale testing as an important educational tool was sealed with the passage of the 1965 Elementary and Secondary Education Act. The first national assessment, the National Assessment of Academic Progress (NAEP), addressed in Applebee's article in this section, was administered in 1969.

In the later 20th century, alarming reports of an impending literacy crisis, a crisis of "mediocracy," and its implications for the U.S. economy, such as *Newsweek*'s "Why Johnny Can't Write" (Sheils, 1975), *A Nation at Risk* (National Commission on Excellence in Education, 1983), and *Time for Results* (National Governors Association, 1985), led to calls for reform and accountability. These

calls for action resulted in a range of state-level policy solutions. One common action was increased implementation of statewide standards and assessment of students, from elementary to secondary-level, often in the form of direct assessments of writing and other basic skills. These standards and assessments were designed to impact curriculum and instruction and frequently were developed in response to employer demands, but with the influence of disciplinary experts. For instance, Sandra Murphy's (2003) *Journal of Writing Assessment* article, "That Was Then, This Is Now: The Impact of Changing Assessment Policies on Teachers and the Teaching of Writing in California," describes the California Assessment Program. This program developed in the early-1980s and was regarded as cutting edge for its focus on direct writing assessment. Murphy (2003) notes that half the states also were conducting direct writing assessments by the mid-1980s.

The essays in this section were published during a new era of large-scale assessment focused on educational "accountability." These approaches assumed test scores and high stakes could be used to raise standards. Literacy and writing remained key areas of concern and focus. By the late 1990s, many legislatures were moving toward holding schools and teachers accountable for improving students' performance on state-delineated standards, such as California's 1999 Public Schools Accountability Act (Murphy, 2003); however, the passage of the No Child Left Behind Act (NCLB) of 2001 mandated regular state-wide standardized testing coupled with financial performance-based penalties and rewards to push educational reform.

Although problems with one-dimensional accountability–and accountability resting entirely on the test scores of "hapless students" (White, 2005, p. 148)—were evident early on in K-12 education, this high stakes, testing-centered approach to educational accountability quickly "trickled up" to higher education. The 2006 Spellings Commission Report, which called for improving "accessibility, affordability, and accountability" in higher education, resulted in the 2008 Higher Education Opportunity Act, ushering in a wave of new accountability measures, increased federal regulation and data reporting requirements and a greater federal oversight role in institutional accreditation (Eaton, 2008).

Under the Obama administration, the accountability movement accelerated and increasingly gravitated toward the neoliberal economic policies of "paying for performance," what Toth, Sullivan, and Calhoon-Dillahunt (2016) describe as "a dubious method of improving educational outcomes through financial penalties and rewards already well-tested (and failing) in K-12 reform efforts" (p. 392). In elementary and secondary education, Race to the Top competitive grants, funded through the 2009 American Recovery and Reinvestment Act, helped propel states toward adopting the newly-minted Common Core

State Standards (CCSS), which Hammond and Garcia (2017) studied in their piece in this section. The English Language Arts and Mathematics Common Core, initiated by the Council of Chief State School Officers (CCSSO) and the National Governors Association (NGA) with the support of Achieve, Inc., were taken up by nearly every state (CCSS Initiative, 2022), often alongside the PARCC or Smarter Balanced online tests designed to measure these standards.

According to Adler-Kassner (2017), this accountability age has been driven by increased external influence on educational standards and outcomes by lawmakers, influential corporations, and many groups and actors that make up the reform-minded Educational Industrial Complex (EIC), who tell the story of "The Problem with American Education and How to Fix It" (p. 320). Toth et al. (2019) note, "over the last few decades, calls among both state and federal policymakers to improve student retention and degree completion have increasingly been framed as a matter of institutional 'accountability'" (p. 2). According to Calhoon-Dillahunt (2018), the EIC's solutions "privilege proficiency and efficiency (aka 'success' and 'completion') over learning and development" and their view of 'accountability' is market-oriented, with 'value' measured almost exclusively in economic terms" (p. 281). As a result, developmental and first-year writing are primary targets in "the EIC's quest to streamline and economize higher education" (Calhoon-Dillahunt, 2018, p. 281). In the past decade, some states–Florida and Connecticut, for instance–have intruded into policies that were once institutionally determined, such as placement and developmental education, and most states have enacted performance-based funding policies in an attempt to drive reform.

ACCOUNTABILITY CONSEQUENCES AT STATE AND NATIONAL LEVELS

The four chapters in this section are situated directly in the reform and accountability era. While the scale of "large-scale" and the policy implications—local, state, or national—vary with each assessment studied, the chapters together examine the intentions, politics, and misperceptions behind externally imposed writing standards and high stakes writing assessments and the resulting material and policy ramifications of these reform and accountability efforts.

In "The Misuse of Writing Assessment for Political Purposes," Edward M. White (2005) identifies three focal areas of writing assessment that have been shaped by politics and public policy: high school proficiency testing, college placement, and mid-career assessments in colleges. The latter, "junior" writing assessments, which are addressed only in White's piece, are comparable to high school proficiency testing in many ways. The remainder of the collection of

articles focus primarily on one of two significant and long-standing types of assessments White describes: secondary-level writing proficiency assessments and college writing placement testing.

In addition to White, Arthur N. Applebee and co-authors J. W. Hammond and Meredith Garcia all address K-12 writing proficiency testing and standards at the state and national level. Applebee's "Issues in Large Scale Writing Assessment: Perspectives from the National Assessment of Educational Progress," and Hammond's and Garcia's "The Micropolitics of Pathways: Teacher Education, Writing Assessment, and the Common Core" detail national writing standards and writing assessments and their consequences broadly. Applebee (2007) discusses the National Assessment of Educational Progress (NAEP), a congressionally mandated assessment across multiple subject areas, including a writing assessment, given to a representative sample of elementary and secondary students across the country. Applebee documents issues with large-scale writing assessments and the ways disciplinary expertise has been leveraged to improve the test and its utility. Hammond and Garcia (2017), on the other hand, focus on the Common Core State Standards (CCSS). Rather than analyzing the large-scale assessments associated with CCSS, PARCC, or Smarter Balanced (SBAC), they study how teachers navigate these common national standards in their own local contexts.

Along with White, co-authors Christie Toth, Jessica Nastal, Holly Hassel, and Joanne Giordano interrogate college writing placement in the age of high stakes. In "Introduction: Writing Placement, Assessment, and the Two-Year College," which is part of a *JWA* special issue on two-year college writing placement, Toth et al. (2019) outline how two-year college writing placement has become a particular target for educational reformers, which has resulted in a reconsideration of the role of placement and common placement practices.

Collectively, these four chapters coalesce around three core themes:

- The intentions behind and (mis)use of mandated writing standards and assessments for accountability purposes.
- The consequences of large-scale, high stakes writing assessments on curriculum, teachers, and students.
- Positive outcomes and spaces for possibility among some large-scale writing assessments and the policy implications.

INTENTIONS, ASSUMPTIONS, AND (MIS)USES

Educational reforms and policies are often well-intended, but how they are enacted and enforced is often troubling and troublesome, especially in the

"accountability era." In their articles in this section, the authors share that intentions behind common writing standards and standardized assessments often seem reasonable and even laudable. For instance, White (2005) asserts that it is entirely logical to expect high school students to demonstrate a certain level of reading and writing skill upon graduating. High school writing standards and accompanying writing proficiency tests are promoted as a way to prepare students for postsecondary writing. Hammond and Garcia (2017) describe how definitions of "preparedness" became codified in the Common Core State Standards, enabling measurement of this elusive idea of "college and career readiness." According to the Common Core State Standards Initiative (2021) website, a consistent, nationwide set of standards can be used to articulate and measure student progress and to ensure students have acquired the necessary skills and knowledge to achieve success in postsecondary education and the workforce. Standardized testing, then, is viewed by policymakers and others involved in education reform as a way of raising standards and monitoring progress. According to the 2004 National Commission on NAEP report, a "high school diploma was no longer the culminating degree for most students" (Applebee, 2007, p. 86). Applebee also observes that about half of high school students who continued on to college were placed into developmental education, suggesting that many students were graduating from high school underprepared to do the sort of writing required in higher education. Thus, assessing 12th graders' readiness for college, military, and career seems essential.

According to White (2005) and Toth et al. (2019), in some ways, placement testing aligns with intentions for high school writing proficiency testing, ensuring students are "ready" to do college work. The theory behind placement assessments is to match students to appropriate coursework, which allows college writing programs to maintain high standards in first year writing while providing support for underprepared students before or as part of their first-year writing coursework (White, 2005). In their article, Toth et al. (2019) share Willingham's 1974 algorithm for understanding the role of placement assessments, a logic still pervasive in placement and developmental writing today. This logic suggests that, by identifying students with poor writing skills and matching those students to coursework designed to improve those skills, student learning and retention in writing courses will be improved.

Holding institutions accountable for student learning and achievement is also reasonable, according to White (2005): "it is wholly appropriate for politicians and citizens to inquire into whether the schools are accomplishing established goals" (p. 25). After all, states and local taxpayers, in particular, invest heavily in education, and they should expect students to graduate with the knowledge and skills needed for postsecondary pursuits. However, as Toth, Sullivan, and

Calhoon-Dillahunt (2016) have observed, accountability measures often fail to acknowledge that "the academic playing field is not level. An institution's record of 'success' is largely shaped by its student demographics and resources" (p. 401). Moreover, high stakes measures offer limited information about student achievement and potential, yet are used, often singularly, to make consequential educational decisions.

In the "accountability era" of education, reforms are enforced through high stakes assessments. Problematically, accountability for these educational reforms is one-dimensional and one-directional, with consequences for schools (and thus students), regardless of their capacity and resources. White (2005) questions this one-way accountability that holds teachers and schools responsible for students' performance on a single assessment without consideration of other influential factors, including school environment, quality and experience of teachers and administrators, learning support for students and teachers, among others, and, importantly, without consideration for policymakers' own responsibility to ensure equal access to education and to appropriately support and fund basic education as well as their ambitious new educational initiatives.

Regardless of how well-intended, education reform in the "accountability era" is too often driven by oversimplified perceptions and a lack of understanding of what motivates, creates, and indicates change. Hammond and Garcia (2017) observe that education reform typically tries to "manage educational pathways," using standards and assessments to regulate how students move through the educational system and "in the process, managing student advancement, opportunity, and attainment." However, they note, "educational complexity is not so easily tamed," and, ultimately, "[r]eform initiatives can only standardize so much" (Hammond & Garcia, 2017, p. 2). High stakes assessments enter the equation under the assumption that financial penalties and rewards will inspire desired reforms and create desired results. Linda Darling-Hammond (2007) asserts that accountability-oriented policies like NCLB misidentify the problems in education, assuming that "what schools need is more carrots and sticks rather than fundamental changes." In an *NPR* interview, NCLB cheerleader turned outspoken critic Diane Ravitch adds that "measure and punish" is not an effective way to prompt change: "incentives and sanctions may be right for business organizations, where the bottom line–profit–is the highest priority, but they are not right for schools" and, in fact, have led to manipulation, dishonesty, and even cheating as schools compete for or try to preserve scarce resources (Inskeep, 2010).

Not only is educational reform founded on misperceptions about how to implement change, but the writing assessments used to measure intended changes are based on fundamental misunderstandings about writing and how students

learn to write. Linda Adler-Kassner (2017) argues that "this lament, this story that students 'can't write,' works from the premise that writing is 'just writing.' It's a thing that writers bang out. It is constituted of words that are clear, that mean the same thing to everyone, that are easily accessible and need only to be plugged into forms" (p. 317). Toth et al. (2019) describe the foundational logic of traditional writing placement in much the same way; it's built on the notion that such writing skills are attainable, measurable, and relevant to subsequent college-level writing coursework and that assessing these generic skills–and placing students accordingly–will lead to improved writing. In describing the development of the revised framework for the 2011 NAEP writing assessment, Applebee (2007) references a range of scholars who have challenged the "traditional emphasis on writing as a generic skill, taught primarily in English language arts or composition classes, and assessable through generic writing tasks detached from particular disciplinary or socially constituted contexts" (p. 163), yet the myths that "writing is just writing" and that "good writing" can be measured by a single test and without regard to context persist.

Raising the stakes on writing assessments and at the same time basing such assessments on fundamental misunderstandings about writing, assessment, and accountability has led to misuse rather than reform. For instance, the perception of writing as a generic skill has led to assessment tools that are often built to prioritize ease of measurement rather than achievement of higher order skills, resulting in assessments that focus on editing skills or formulaic writing tasks (Applebee, 2007). According to Toth et al. (2019), "The widespread reliance on commercially produced [writing placement] tests that measure a very limited construct of writing has prioritized knowledge of Edited American English conventions at the expense of any other outcome, primarily because these are the skills that can be easily measured" (p. 219). Thus, the tools that determine whether consequences will be meted out do not capture the lofty goals of the reform movement, and they are also biased against historically marginalized and minoritized students *by design*, essentially ensuring that the schools that serve such students will be penalized. These misuses are costly, in all senses.

In some cases, the high stakes assessments work against the very reforms they are trying to institute, case in point, high school writing proficiency testing. As several authors in this section articulate, the intentions behind large-scale high school writing assessments are to raise standards and increase student proficiency in writing for their postsecondary pursuits, as writing is a perceived "problem" despite the fact that high school graduation rates are over 85 percent and about two-thirds of those students enroll in postsecondary education after high school (U.S. Department of Education, 2021). To "inspire" students and teachers to take these standard-raising writing assessments seriously, many states tie earning

diplomas to passing state-mandated tests. Inevitably, implementing policies to solve one perceived problem, students' lack of preparedness for postsecondary pursuits, created many others. Policymakers were unprepared to admit that large portions of graduating seniors didn't demonstrate proficiency (White, 2005), although, given the frequency of testing students in K-12, they had fair warning about the likely results. Paradoxically, with the proliferation of dual enrollment programs in high school (NACEP, 2019), it's entirely possible for a student to simultaneously succeed in postsecondary coursework–and even earn a postsecondary degree–in high school, while simultaneously failing single-measure assessments designed to certify a student's "college-readiness."

As a result, grade 12 assessments are now given earlier in students' academic career, to allow more time for remediation and retakes. Assessments have been simplified to increase pass rates; instead of raising the bar, the assessments now represent the minimum level of competence required, and, even then, some students may not be able to pass them, so, according to White (2005), "exemptions, exceptions, and fraud enter the assessment system" (p. 146). Ultimately, these assessments create a Catch-22: students are deemed "unprepared" for postsecondary writing, although there is little consensus about what "college and career-ready" writing means (Applebee, 2007), by high school proficiency tests and writing placement tests, assessments largely disconnected from the writing curriculum. The number of "unprepared," as defined by student performance on these same high stakes assessments, leads policymakers to demand greater accountability, using high stakes assessments as the measure and mechanism for change.

CONSEQUENCES

Attaching penalties and rewards to student performance on single assessment measures in order to drive educational reform and accountability policies has had far-reaching repercussions. The authors in this section address the negative consequences that have resulted from the use of mandated standards and high stakes writing assessments in three particular areas: curriculum, teachers and teaching, and students.

Impact on Curriculum

One of the most well-studied consequences of high stakes standardized testing is its impact on curriculum. Sandra Murphy (2003) notes high stakes assessments do not just measure achievement; they define it. Several authors in this section observed the ways that such assessments narrow, constrain, and distort writing

curriculum. Applebee (2007) argues that attempts to shape curriculum and assessment around abstract notions of "career and college readiness" have generally resulted in "a system of curriculum and assessment that focused on basic skills or on generic workplace tasks (e.g., business letter format) that easily degenerated into formulas with little real-world relevance" (p. 167).

The curricular impact of high stakes assessments can also be seen in postsecondary writing placement. According to Toth et al. (2019), "[i]n the nation's open-admissions two-year colleges, where students enter from a wide range of academic trajectories and often have not taken any kind of admissions exam, placement assessment is nearly universal" (p. 215), and the use of commercial placement products predominates. One of the results of this sort of placement mechanism is that most two-year colleges offer multiple levels of pre-college writing courses, which may be similarly disconnected from first-year writing curriculum, focused instead on the "basic skills" developmental writers seemingly lack, and which sometimes prohibit students from accessing other college-level courses outside of English. On the other end of the spectrum, some colleges may exempt high performing students from the first-year writing requirement altogether, which suggests that first-year writing curriculum is not about introducing students to a discipline, but, instead, teaching generic "writing" skills.

Writing assessments that are disconnected from a college's first-year writing curriculum provide limited utility for authentic placement, but they send powerful messages about how the institution views and values writing. Toth et al. (2019) recognize that writing placement "is not a neutral action" (p. 218); it communicates particular values and ideologies that affect how students, local high schools, and others perceive writing, and as a result, it can impact both high school curriculum and perceptions about the role of developmental and first-year writing on college campuses. Simultaneously, commercial placement tests also fail to communicate anything particular about a writing program, the theory that underlies its curriculum, and the practices it values; such assessment instead perpetuate the narrow conceptions of writing many students bring with them from high school and the commonly held notion that first-year writing is a course they need to "get out of the way." Additionally, writing curriculum is impacted, negatively and positively, by current reform movements that seek to limit and accelerate developmental writing offerings (Toth et al., 2019).

Impact on Teachers and Teaching

Externally mandated standards and high stakes writing assessments also have a profound impact on teachers and the teaching of writing. Murphy (2003) argues standardized testing has deprofessionalized teachers, constraining their

opportunities for professional growth, undermining their autonomy and professional authority, devaluing their expertise, and blaming them for poor student performance on tests. White (2005) asserts that high stakes assessments are politically motivated and are used in disrespectful and manipulative ways toward teachers. He notes that many teachers are wary of large-scale assessments "because it almost inevitably narrows and often reduces what they do to simple numbers that will be used against their students and them" (p. 144).

Postsecondary writing instructors do not face the same types of blame and control as their secondary-level colleagues, but the reliance on placement tests, particularly at two-year colleges and other open-admissions institutions, have contributed to the notion that developmental writing and even first-year writing courses do not require professionalized writing teachers. The use of standardized placement tools that deem many–even a majority of students "unprepared" for college-level writing has led to a proliferation of basic writing courses. These courses are often viewed and even taught as "basic skills" courses, as courses designed to "re-teach" what students should have already learned in high school and, thus, not worthy of much investment. Toth et al. (2019) argue that the disconnect between theory and practice in writing placement assessment also detracts from the professional status of faculty who teach developmental and first-year writing.

Impact on Students

While the studies included in this set of articles don't address the impact of high stakes testing on students directly, the implications are clear: students bear the brunt of the consequences of standardized writing assessments. There is a long history of using writing assessments to gatekeep and rank students, and the consequences are even greater for students, especially historically underserved students, when assessments are tied to diplomas for college-level access. White (2005) argues that "Each of these assessments [high school proficiency exams, placement tests, mid-career writing assessments] represents a gate through which students must pass if they are to gain access to the privileges and enhanced salaries of college graduates, and so they carry a particular social weight along with their academic importance" (p. 145). The negative impacts of accountability policies and high stakes assessments previously described, from penalizing already under-resourced schools to narrowing the curriculum and reducing teacher agency and professionalization, also affect the quality of education students receive.

Toth et al. (2019) discuss most directly the impact standardized assessments have had on students in the context of placement. The authors cite Haswell's work

on the lack of predictability of writing placement tests; this lack of predictability of success has been corroborated by many others to reveal the ramifications for under-placement, which can extend student costs and time to degree, and for over-placement, which can cause failure, which is also costly, time-consuming, and can result in academic penalties. The consequences for misplacement disproportionately affect historically underserved student populations (Toth et al., 2019). Additionally, performance-based funding policies can penalize open-admissions institutions for student performance, which may incentivize those institutions to limit or refuse entry to students who, based on their placement scores, seem unlikely to succeed and, thus, threaten the college's funding. This disparately impacts minoritized and marginalized students (Toth et al., 2016).

POTENTIAL AND POSSIBILITIES

This section makes clear that high stakes standardized writing assessments have often been detrimental to teaching and learning, to public perceptions about writers and writing, and to educational policy decision-making. However, the enterprise of large-scale writing assessment has not been without utility and even, at times, positive effects. Several of the articles in this chapter provide examples of well-designed standardized writing assessments that, when used as intended and without adding penalties and rewards that subvert their aims, serve a productive educational purpose and have contributed to our understanding of writing and writing assessment. Hammond and Garcia's study shows that teacher involvement in developing and mediating standards and assessments creates conditions for assessments to be used in ways that inform and improve curriculum and instruction, which are precisely the goals of these educational policy reforms.

Applebee's review of the framework for 2011 National Assessment of Educational Progress reveals that constructing and revising large-scale assessments, especially in consultation with teachers and disciplinary experts, enables writing assessment to reflect and shape research and scholarship in writing studies. Applebee (2007) reports on significant questions the National Assessment Governing Board (NAGB) considered about how to assess student writing in a valid, fair, and purposeful way as it revised its writing assessments for 2010 and beyond. According to Applebee (2007), in preparation for the revised 2011 NAEP writing assessment, the NAGB addressed questions about everything from the types of writing to be assessed, the prompts to use to generate writing, and the aspects of writing achievement to be measured to computer-mediated writing, test-taking accommodations, and time allotments. Such thoughtful consideration of assessment content and design leads to more informed and informative

assessments, especially when the stakes for such assessments remain relatively low for students, teachers, and institutions.

Because, as Applebee (2007) indicates, NAEP also served as a model for many state-developed assessments, NAEP's conscientiously designed and theoretically grounded assessment in writing had reverberating and likely positive effects on other large-scale writing assessments. Granted, the NAEP assessment, which appears to have been largely replaced at the high school-level by the CCSS-connected Smarter Balanced and PARCC assessments, still struggles with its intended goal of assessing student writing in ways that inform "preparedness for postsecondary endeavors," likely impossible to measure within a single, standardized assessment. However, the results have provided a fertile ground for study, on a large scale, which has enabled the field of writing studies to evolve.

Of course, mandated writing standards and large-scale writing assessments largely remain externally directed and developed. However, Hammond and Garcia (2017) remind us that policies have to be put into practice: "Standards . . . are never as autonomous or agentive as sometimes imagined; they are largely contingent on interpretation and implementation by the very actors they are intended to coordinate and perhaps constrain" (p. 184); indeed, they continue, "reforms put in place are seldom as stable and standardized as intended" (Hammond & Garcia, 2017, p. 186). The fact that policy is not determinative, is "not so easily tamed," means that policy requires support and buy-in to be fully enacted. Policy implementation is also negotiated and navigated within particular contexts: "Homogenizing educational projects like the CCSS are always alloyed with heterogeneous local perspectives, assumptions, and aims. While perhaps obscured by standardizing efforts, local differences are not erased by them." (Hammond and Garcia, 2017, p. 185). These mediated spaces are places of possibility, enabling the tools of policy implementation to be productively adapted and providing agency for those involved in their implementation.

In their study of student teachers, mentor teachers, and field instructors at three midwestern high schools, Hammond and Garcia (2017) observed that, while all teachers involved in their study utilized CCSS in some way in their curriculum development, they used and assessed the standards in different ways and for their own purposes, tied to their own local contexts. Study participants tended to curate and even "retrofit" the standards, rather than adopt them outright, which enabled the participants to select and prioritize the outcomes that fit their curriculum and goals and their students' needs as well as to use low stakes, classroom-based assessment practices to determine mastery. The study revealed that, instead of finding CCSS restrictive, the participating teachers tended to use the standards as a rhetorical tool, "as a medium for managing communication with stakeholders and—by extension—signaling professional participation in

the collective enterprise of American education" (p. 4). Some found the CCSS provided a common language for teachers, students, and parents to facilitate teaching and learning in the discipline, and others found using this "professional lingua franca" validated their work to external audiences, whether administrators, community members, or policymakers. Hammond and Garcia's study reveals that, while policymakers may devise standards, teachers are the ones who enact them; the possibilities of educational reforms are tied to teacher buy-in and teacher agency to implement such reforms in context.

Further, their study suggests that teacher agency in determining and designing curriculum and assessment in context facilitates "professional accountability," as described by Linda Darling-Hammond. According to Darling-Hammond (1989), "[p]rofessional accountability" requires that teachers are knowledgeable and engaged practitioners, who participate collectively in all aspects of teaching and learning, including assessment and local decision-making. Professional accountability has much more potential to drive positive and lasting change than the "carrot and stick" approaches associated with "accountability era" reforms. Hammond and Garcia's (2017) work reveals that when teachers have agency in curricular decisions and when they are not threatened with punitive consequences, teachers often view imposed standards and large-scale assessments favorably.

CONCLUSION

Education reform's "accountability" turn has often been framed in terms of "value added," with value defined–and "accountability" enforced–through neoliberal economic ideologies. Ravitch argues this competitive, market-based approach is wrong for public schools, which should function collaboratively and should share what works with others (Inskeep, 2010). In the "reform and accountability" era, large-scale writing assessments have often enabled these competitive and punitive policies. However, Rose (2012) asserts that "our philosophy of education—our guiding rationale for creating schools—has to include the intellectual, social, civic, moral, and aesthetic motives as well. If these further motives are not articulated, they fade from public policy, from institutional mission, from curriculum development" (p. 185). Because it's connected to policy, mission, and curriculum–and, in fact, should emerge from these areas, writing assessment is foundational to how we articulate and ascertain "value" in education, and the future direction of writing assessment should consider "value-added" from the broader perspective Rose identifies.

To this end, the chapters in this section suggest a range of possibilities for future research. White's, Applebee's, and Hammond and Garcia's work all

recognize the critical role of teachers in education reform and reveal the importance of teacher engagement with standards and assessments and of assessments emerging from and shaping curriculum. As teachers are enactors of reform policies, more attention should be directed toward understanding the impact of education policies and large-scale assessments on their practice and the role professionalization and "professional accountability" plays in facilitating educational reform. Such research may reveal that investing in the changemakers, teachers, rather than investing in large-scale assessment tools may yield better results. Additionally, few studies talk to *students* about the ways in which they are experiencing "accountability" reforms, particularly how such policies and high stakes assessments affect their development and self-perceptions as writers and their conceptions of writing.

Toth, Nastal, Hassel, and Giordano's work highlights the importance of assessing assessment tools. The work of researchers that questioned the validity, reliability, and predictability of commonly used commercial placement tests has resulted in many institutions abandoning such tests in favor of local alternatives or reducing the stakes by using such tests as one consideration, among others, for placement. These studies also led to revisions in commercial products themselves, often including a direct assessment of writing, albeit computer-scored. Not only is it important to assess validity and reliability in large-scale writing assessments, Toth et al. remind us of the importance of assessing the fairness of writing assessment tools and methodologies, especially in large-scale and high stakes assessments. Given that standardized writing assessments are rooted in White Language Supremacy and ableism, studying the consequences of writing assessments, in particular the disparate impacts of such assessments, can provide direction for how to redesign and even reimagine writing assessment tools that attend to local contexts and value diverse students. Toth et al. argue–and I agree–that two-year colleges are important spaces in which to conduct this research, as two-year colleges serve diverse students and communities and, with their open admissions policies, often serve as the primary access point for postsecondary education for the least advantaged students.

Finally, writing assessment research is one key way to change the public narrative around writing and to help policymakers develop informed solutions to the educational problems they are trying to solve. Writing researchers and scholars can contribute by asking different questions that counter the predominant failure-driven narrative. For instance, how can writing assessments provide evidence that student writing isn't a "problem" and instead highlight the rich and rhetorically conscious ways students language and compose in classrooms with professionalized teachers developing curriculum appropriate to local contexts and students' needs? How can large-scale writing assessments account for the

varied ways students demonstrate proficiency and success, for instance, in considering multiple measures instead of single assessments? How do lower stakes assessments provide more meaningful information and yield more positive results? How can large-scale writing assessments provide evidence of "college and career-readiness" by centering rhetorical dexterity and situated language practices instead of facility with Edited American English?

In addition to researching in ways that change the dominant discourse around writing, writing researchers and scholars can also practice their own rhetorical dexterity by sharing writing research in accessible ways with public audiences and policymakers. In other words, it is incumbent upon writing researchers to "[find] ways to communicate our expertise to those outside of our discipline and [seek] opportunities to participate in public conversations about literacy education" (Calhoon-Dillahunt, 2015). Future writing researchers can also take a page from two-year college teacher-scholar-activists who view engagement in educational policy as a professional responsibility, which requires "undertak[ing] the public work of defending educational access, teaching for democratic participation, and advocating for practices and policies grounded in disciplinary knowledges" (Toth, Sullivan, & Calhoon-Dillahunt, 2019).

REFERENCES

Adler-Kassner, L. (2017). Because writing is never just writing. *College Composition and Communication*, *69*(2), 317-340.

Applebee, A. (2007). Issues in large-scale writing assessment: Perspectives from the National Assessment of Educational Progress. *Journal of Writing Assessment, 3*(2), 81-98. https://escholarship.org/uc/item/0zx1m9fg

Calhoon-Dillahunt, C. (2015). Finding our public voice. In D. Cambridge & P. Lambert Stock (Eds.). *Structural kindness: Essays on literacy education in honor of Kent D. Williamson* (pp. 163-169). National Council of Teachers of English.

Calhoon-Dillahunt, C. (2018). Returning to our roots: Creating the conditions and capacity for change. *College Composition and Communication*, *70(*2), 273-293.

Common Core State Standard Initiative. (2022). *Frequently asked questions.* http://www.corestandards.org/about-the-standards/frequently-asked-questions/

Darling-Hammond, L. (2007). *Evaluating 'No Child Left Behind'.* SCOPE: Stanford Center for Opportunity Policy in Education. https://edpolicy.stanford.edu/library/blog/873

Eaton, J. S. (2008). The Higher Education Opportunity Act of 2008: What does it mean and what does it do? *Inside Accreditation*. https://www.chea.org/higher-education-opportunity-act-2008-what-does-it-mean-and-what-does-it-do

Hammond, J. W. & Garcia, M. (2017). The micropolitics of pathways: Teacher education, writing assessment, and the Common Core. *Journal of Writing Assessment, 10*(1). https://escholarship.org/uc/item/392847v0

Haswell, R. (2004). *Post-secondary entry writing placement: A brief synopsis of research.* CompPile. http://www.comppile.org/profresources/writingplacementresearch.htm

Huot, B., O'Neill, P., & Moore, C. (2010). A usable past for writing assessment. *College English, 73*(5), 496-517.

Inoue, A.B. (2015). *Antiracist writing assessment ecologies: Teaching and assessing for a socially just future.* The WAC Clearinghouse; Parlor Press. https://doi.org/10.37514/PER-B.2015.0698

Inskeep, S. (2010). *Former 'No Child Left Behind' advocate turns critic.* NPR. https://www.npr.org/templates/story/story.php?storyId=124209100

Murphy, S. (2003). That was then, this is now: The impact of changing assessment policies on teachers and the teaching of writing in California. *Journal of Writing Assessment, 1(*1). https://escholarship.org/uc/item/1fg1531r

NACEP. (2019). *Fast Facts.* National Alliance of Concurrent Enrollment Partnerships. https://www.nacep.org/resource-center/?_sft_resourcetypes=fast-facts

National Commission on Excellence in Education. (1983). *A nation at risk: The imperative for educational reform.* https://edreform.com/wp-content/uploads/2013/02/A_Nation_At_Risk_1983.pdf

National Education Association. (2020). *History of standardized testing in the United States.* https://www.nea.org/professional-excellence/student-engagement/tools-tips/history-standardized-testing-united-states

National Governors' Association. (1986). *Time for results: The governors' 1991 report on education* (ED279603). ERIC. https://eric.ed.gov/?id=ED279603

Ritter, K. (2009). *Before Shaughnessy.* Southern Illinois University Press.

Rosales, J. & Walker, T. (2021). *The racist beginnings of standardized testing.* NEA. https://www.nea.org/advocating-for-change/new-from-nea/racist-beginnings-standardized-testing

Rose, M. (2012). *Back to school: Why everyone deserves a second chance at an education.* The New Press.

Sheils, M. (1975, December 8). *Why Johnny can't write.* Newsweek. https://www.leetorda.com/uploads/2/3/2/5/23256940/why_johnny_cant_write__newsweek_1975___1_.pdf

Toth, C., Nastal, J., Hassel, H. & Giordano, J. (2019). Introduction: Writing assessment, placement, and the two-year college. *Journal of Writing Assessment, 12(*1). https://escholarship.org/uc/item/8393560s

Toth, C., Sullivan, P. & Calhoon-Dillahunt, C. (2016). A dubious method of improving educational outcomes: Accountability and the two-year college. *Teaching English in the Two-Year College, 43*(4), 391-410.

Toth, C., Sullivan, P. & Calhoon-Dillahunt, C. (2019). Two-year college teacher-scholar-activism: Reconstructing the disciplinary matrix of writing studies. *College Composition and Communication*, *71*(1), 86-116.

U.S. Department of Education. (2021). *Public high school graduation rates.* Institute of Education Sciences, National Center for Education Statistics. https://nces.ed.gov/programs/coe/indicator/coi

White, E. M. (2005). The misuse of writing assessment for political purposes. *Journal of Writing Assessment, 2(*1). https://escholarship.org/uc/item/9hg796d1

CHAPTER 4.

THE MISUSE OF WRITING ASSESSMENT FOR POLITICAL PURPOSES

Edward M. White
University of Arizona

This chapter focuses on the political dimensions of writing assessment, outlining how various uses of writing assessment have been motivated by political rather than educational, administrative, and professional concerns. Focusing on major purposes for writing assessment, this article examines state-mandated writing assessments for high school students, placement testing for incoming college students, and upper class college writing assessments such as rising junior tests and other exit measures that are supposed to determine whether students can write well enough to be granted a college degree. Each of these assessments represents a gate through which students must pass if they are to gain access to the privileges and enhanced salaries of college graduates, and so they carry a particular social weight along with their academic importance. In other words, each of these tests carry significant consequences or high stakes. According to the most recent and informed articulations of validity, each of the cases examined in this article require increased attention to the decisions being made and the consequences for students, teachers, and educational institutions. In each case, this article addresses the political reasons why these assessments are set in motion and point to the inner contradictions that make it quite impossible for them ever to accomplish their vaguely stated purposes.

As I detail in a *College English* article, I first became involved with writing assessment as a result of political interference with the teaching of first-year composition (White, 2001a). In that article, I point out how I stumbled into the field of assessment more than 30 years ago as one of several English department chairs trying to protect our first-year composition programs from being defined by a demeaning test that the Cal State system chancellor wanted us to use to

DOI: https://doi.org/10.37514/PER-B.2024.2166.2.04

further his political career. Every year since, I have been involved in one way or another with the political dimension of assessment, a perspective that is usually oppressive, insensitive, disrespectful, and manipulative to teachers and students. I look back on three decades of struggling to live with such misuse of writing assessment, even as I have stressed in my scholarship over the last three decades the importance of teacher involvement and understanding of assessment as a professional responsibility, indeed one with undoubted political ramifications. Political figures love assessment because it allows them to posture about education and pretend to themselves and to others that they are improving education by measuring a simplified version of it. Teachers generally dislike and distrust assessment, because it almost inevitably narrows and often reduces what they do to simple numbers that will be used against their students and them. Meanwhile, those of us actually teaching writing use assessment of one sort or another all of the time in our classrooms (Huot, 2002; White, 2006). How else, for example, can we teach self-assessment and revision? Regardless of the centrality of assessment to the teaching of writing, we are forever fending off the efforts of politicians and testing companies to use assessment improperly, to prove that our students are not learning, and that we are at fault. Although I agree that teachers and writing program administrators (WPAs) are responsible for assessing those programs, the current assessment climate often makes teachers, students, and WPAs accountable to ill-conceived, poorly constructed, and misused assessments. No wonder that the very mention of assessment is enough to send many teachers racing from the room, even if it sends them back to their offices—to continue responding to this week's set of papers.

In this article, I focus on writing assessment in its political definition, not as the form of professionalism that allows us to do our jobs with our students. This is an important distinction because the mandated assessments from those ignorant of what we do have little or nothing to do with our teaching or our students. One canny reviewer of the MLA book I edited with two others entitled *Assessment of Writing: Politics, Policies, Practices* (White, Lutz, & Kamusikiri, 1996) wrote that it should really have been titled *Assessment of Writing: Politics, Politics, Politics*. So I am going to follow his advice here, attending solely to the politics of writing assessment, an aspect of the field that is, unfortunately, its most prominent and unexamined face. We do need to assess our students' work to help them improve and to assess our programs to see if they are doing what we expect them to do. But we also must dispute the view that testing, particularly testing using nationally normed tests, can determine if we are teaching well and responsibly.

I intend to look at three places where writing assessment is most prominently misused: the high school writing assessments, now afflicting students seeking

their diplomas in all but two states; placement testing, the usual sorting of first-year students into those supposedly ready for regular college work and those who are not; and, finally, mid-career assessments, required of college students as they move from the sophomore year to the junior year in an attempt to ensure that such students will have a certain level of ability at reading and writing, at least enough to placate their major professors in college and their employers after graduation. Each of these assessments represents a gate through which students must pass if they are to gain access to the privileges and enhanced salaries of college graduates, and so they carry a particular social weight along with their academic importance. In other words, each of these tests carry significant consequences or high stakes. In each case, I examine the political reasons why these assessments are set in motion and point to the inner contradictions that make it quite impossible for them ever to accomplish their vaguely stated purposes—which leads to a certain amount of thrashing about to identify the problems and possible solutions. Ultimately, I believe we need to reconstruct the stage for writing assessment, and I hope my discussion can begin this important work. We could thus cast this discussion as a study of violations of test validity, using modern definitions of validity that extend beyond score correlations into the entire context of a testing program, including consequences for test takers and anything else that affects the decisions made on behalf of a measure. But in a short article focusing on political issues, I focus specifically on the inherent problems and contradictions these programs represent and allude to some effective ways to approach the political goals in a responsible way. It bears mentioning that if test users and developers adhered to current conceptions of validity summarized in the most recent *Standards for Educational and Psychological Testing* (1999) most of the problems I explore in this article would not exist.

HIGH SCHOOL PROFICIENCY TESTS

What could be more logical than to require students seeking their high school diplomas to demonstrate on a test that they can read and write at the level we like to imagine we did at their age? And so, state after state has convened committees, task forces, and consultants to prepare the tests that will determine if teachers and students have done their jobs well—with some even withholding diplomas from students who do not pass and reassigning school principals from "failing schools." The tests are almost without exception unmitigated disasters, constantly being delayed in final implementation and forever being revised so that most of the students can pass them, but that has not deterred state agencies and our most prominent politicians from making such tests the keystones of their political campaigns. It hardly seems to matter if the tests are

multiple-choice and detrimental to learning, as they are in Arkansas, or teacher-devised and supportive of learning as they were at one time in California, before the religious right determined that asking students to write was an invasion of the privacy of the home and so subjective that it might lead to children asking uncomfortable questions of their parents. (The California advertising campaign that led the then governor to declare that the proficiency test in writing must be an "off-the-shelf" multiple-choice test featured a charming 13-year-old girl declaring that any test without clear right or wrong answers was unfair.) When the tests become high-stakes assessments, as many—but not all—of them are, with important implications for the budgets of schools and the futures of students, the writing proficiency measures become strange artifacts with little connection to reality. I remember fussing, for instance, when my local school district in California defined coherence (who can object to testing for coherence?) as a paragraph containing three sentences—any kind of sentences at all. Well, argued the district consultant, we can't fail more than half of the students, can we?

The tests are supposed to measure student abilities at the point of graduation. But when they are given to high school seniors, invariably a very large proportion of them fail. Supposedly, that is the point, but no state is prepared to say that 60% or more of its seniors cannot read and write well enough to graduate. So the tests must be given to students in time for them to buckle down and pass them after taking test preparation courses. This means in practice that the tests are actually given to eighth graders, so that the students who fail can work all the way through high school to pass their "proficiencies," as they call them. The senior-level test has now become an eighth-grade test, but, alas, some students still reach graduation without passing them. Nonetheless, political considerations demand that they must be gotten through somehow, so exemptions, exceptions, and fraud enter the assessment system. The courts sometimes get involved, particularly as it becomes clear that racial minorities and the children of the poor fail at an especially high rate (Lutz, 1996). One Florida court forced the state to postpone implementing such a test until the school system could demonstrate that African-American children were actually being taught to read and write, a matter in considerable dispute that had somehow escaped the attention of the politicians pressing onward with the testing.

Meanwhile, the tests have an unfortunate effect on the high school curriculum, generally turning it from instruction in reading and writing to instruction in how to pass multiple-choice tests or how to write formulaic prose. Two essays in this journal's first issue gave convincing argument and evidence for this devolution in learning: Sandra Murphy's (2003) "That Was Then, This is Now: The Impact of Changing Assessment Policies on Teachers and the Teaching of Writing in California" and George Hillocks' (2003) "How State Assessments Lead

to Vacuous Thinking and Writing." The Murphy study compares the effects of a careful test in 1988, designed largely by teachers, with a commercial standardized test given in 2001; the results of the later test showed a clear "narrowing and fragmentation of the curriculum" (p. 40). The Hillocks study looks closely at statewide tests in Texas and Illinois, concluding that they "work against the goal of learning how to think critically and argue persuasively" (p. 20).

In addition to the scholarly evidence for the unfortunate effects of these politically directed tests on students, teachers, and learning, I can add a personal experience, from my graduate course in writing research in California, one of the states where the SAT-9 was a high-stakes test, determining budgets and "success" for high schools. One of my students, a fine high school teacher, told me of her confrontation with the school principal, at a teachers' meeting. He had distributed the SAT-9 scores, which were down, and then informed the teachers that everything they did in class must be directed to improving those scores. My student, emboldened by my course, spoke out: "I'm an English teacher. Are you saying that I can't teach reading and writing because they're not on the test?" She spoke mournfully of his reply: "He pointed his finger at me and told me very forcefully that I was not to waste class time on reading and writing or I'd be fired!"

To the obvious contradiction of a senior-level high school test undermining the curriculum so that it can be passed by eighth graders, we need to add the further problem of college entrance. Shouldn't such a test serve for college placement? Well, logically yes. But in practice, almost half of the graduating high school seniors are not heading for college, so why should their high school diplomas depend on a college entrance measure? Besides, the test is in fact designed for eighth graders. Furthermore, it is quite possible that the best high school classes in both English and math are more demanding and set higher standards than the usual first-year college courses in those subjects, so we have no clear definition of what college-level proficiency means beyond particular college practice. National tests, one might imagine, pose a kind of definition; but these range from the relatively strict standards of the Advanced Placement Program to the most minimal multiple choice scores embodied by the General Examinations of the College-Level Examination Program, both administered by the Educational Testing Service, serving consumers at all levels; test criteria and standards move lower still as we look at the products of less professional testing firms. Because we have no reference point for the definition of "college-level" performance from such varied test criteria, we cannot take solace from national tests without national curricula, which nobody really wants. Thus, the stage is set for a continuing muddle, with the writing assessment asked to solve unsolvable problems and to assure everyone that all can be made well if only teachers

worked harder and the administration cracked down on the worst slackers and we tested students often enough.

To be sure, the issue of school accountability is neither trivial nor superficial. It is wholly appropriate for politicians and citizens to inquire into whether the schools are accomplishing established goals. But if they were serious about the matter, this accountability would not rest entirely on the hapless students taking more or less relevant tests. Genuine questions about school accountability would ask about the school environment (does it support learning and is it a supportive, well-maintained, and pleasant place?), teachers and administrators (are they well trained and well paid, the kind of people who should be entrusted with students?), and parents (are they respected as partners in student learning, do they participate?), as well as student test data; but these matters refer to political responsibility for schools in ways that do not allow the politicians to point fingers at others in nice sound bytes. So only the students are assessed, on the cheap and irresponsibly, and these student tests are assumed to represent the status of schools.

But in fact, nobody really pays much attention to the entire operation, aside from the politicians, pointing with pride to their efforts to raise standards, and the students, forced by punishments or induced by free doughnuts or some other bribe to take meaningless tests. The colleges and universities universally ignore the high school tests, preferring to use tests designed for college admission, and usually, sensibly, preferring their own placement procedures, tailored to their own students. (But that is probably going to change; see the following section of this article.) And high school graduates seem to read and write about as well or as badly as they did before all of these tests were instituted, despite test scores rigged to show improvement, because those actual proficiencies depend on the parents, teachers, and the school environment, the key ingredients in any education. It is not hard to imagine more constructive uses for the vast sums now being spent on testing, to very little purpose, in this sad pretense at school accountability.

COLLEGE AND UNIVERSITY WRITING PLACEMENT TESTS

The testing of entering college and university students in order to place them in an appropriate college, or pre-college, writing course has, for more than 100 years, seemed reasonable, responsible, and a nice compromise between high standards for the first-year course and social awareness of the needs of those with weak preparation for study. However, the actual practice of placement testing has never quite lived up to this theory, and many questions have been raised about the way in which college placement takes place, emerging from both the academic left (objecting to invalid testing, institutional tracking, negative

labeling, and retrograde employment practices) and the popular right (objecting to the use of university resources for those defined as not ready for university work). When we think systematically about placement into the first-year writing course, we encounter a tangle of academic, professional, political, and social issues that makes it difficult to decide on an appropriate course of action in general or at our own institutions. Again, as with high school proficiency tests, we find that political motives and naïveté about assessment normally lead to meaningless or destructive tests, useful primarily for political posturing and jockeying for funding.

The least satisfactory method of placement—and the most common in American colleges—is by means of some multiple-choice testing of editing skills, a quick impromptu writing sample, or some combination of both. The problems with this kind of assessment have become obvious. The multiple-choice test of editing skills does not require the production of text and so measures skills not directly related to the first-year writing course. Edgington, Ware, Tucker, and Huot (2005) report that more than 250 students placed in remedial courses through the COMPASS test (an untimed editing exercise on computers) were also placed by a writing sample into the regular first-year writing course, and all these students chose the higher placement. More than 70% of these students received an A or B in the course, and more than 90% of these students received at least a C. The indirect relation of such tests to writing is in much dispute and seems particularly weak for students from homes that do not speak the school dialect. Although a written impromptu placement test is certainly a better option than tests that do not contain any writing at all, we already have several examples of portfolio placement programs that are accurate, reliable, and affordable (Hamp-Lyons & Condon, 2000; Hester, Neal, O'Neill, & Huot, 2005; Willard-Traub et al., 1999; http://www.muohio.edu/portfolio/). On the other hand, as recently as a decade ago, at least half of all respondents to a national survey on placement indicated that they were using something other than student writing to make placement decisions (Huot, 1994). With the validity of these placement decisions so questionable, one must ask why they dominate American higher education. There are numbers of answers, of course, but political considerations are certainly behind most of them. I became convinced of this, a few years ago, when I tried to convince the writing directors of the California State University system to replace their outdated English Placement Test (EPT; whose development and implementation I administered in 1975-1977) with a more modern and more valid portfolio requirement. "Keep your hands off our EPT," they said, unified for once. "All of our financing depends on those scores."

I may be surprising some readers, because I have, for some decades been a strong advocate of placement testing, based on the theoretical arguments

supporting a targeted writing curriculum for entering college students, according to their abilities. There is compelling evidence that entering college students, defined by their institutions as having weak writing skills, will persevere and succeed in college at about the same rate as those with stronger writing ability, if the weaker students receive the extra help basic writing programs can provide; without such help, fewer than 15% of those less prepared students, several studies have shown, will still be in college after 2 years (Phipps, 1998; White, 2001b). There is also a commonsense argument that regular college composition courses have higher standards when the weakest students receive extra help before or during those courses. But, although I remain committed to providing opportunities for success to all admitted students by means of different levels of college composition instruction, I have at long last lost confidence in placement *testing* as an appropriate method for determining who should enroll at these different levels. That is, placement into an appropriate curriculum is both responsible and valuable; but placement testing as now generally practiced has shown itself to be a political rather than an academic activity.

You can tell an assessment is political and not serious academically when discussion starts with testing rather than learning and teaching. Placement is meaningless without considering what we are placing students into, a question inevitably ignored by every national placement device and many local ones. In other words, before we can argue about the validity of placement decisions, we must have data that confirm the educational benefits of each placement option. Everyone knows that some students are better prepared than others for college writing and that those others need some extra help. But that is as far as agreement goes. It is hard to find two colleges that define that extra help in the same way or that have the same descriptors for students needing help. Many open-enrollment schools will have several layers of basic writing; some colleges have none at all. Even the same institution, with little program supervision or coordination, might have requirements for some basic writing sections that are more demanding, in practice, than other courses, nominally for their best students. I have seen an institution mistakenly place some "remedial" students (according to its own criteria) into "honors" sections, where they performed perfectly well. Unpublished studies at the University of Arizona and the University of Louisville (Edgington et al., in press) have shown that many students placed into "remedial" courses by the COMPASS examination, or by a single impromptu essay, can succeed perfectly well in regular composition courses. Sometimes, basic writing courses mean a great deal of technology and drill; sometimes they mean small classes intent on confidence-building through approval of personal writing; sometimes they mean an extended time frame for the same work as regular classes; and sometimes they mean exile to a desert of grammar from which

the only escape is to leave college altogether. And meanwhile, everyone knows that such untested matters as social class, finances, motivation, self-confidence, reading experience, and family responsibilities play a large role in student success in every writing class. In other words, large-scale placement tests, which tend to measure editing skills on other people's prose or impromptu fluency on a writing topic about which there is little time to think, do not allow for the same kind of decision making into every college's writing program. They measure only a small component of what is needed for student success, and they cannot be responsive to the program into which they are placing students. They tend to be a social-sorting mechanism, useful for political posturing, but of limited use for students, teachers, or institutions.

So, how can we place students into a well-designed series of college writing classes, including a variety of basic writing instruction, that will lead to student and teacher satisfaction and to as much student success as possible? Clearly, the first step is for each college or university to design well-defined writing courses that are appropriate for its own student body, including some clear sense of what a student should be able to demonstrate in order to profit from a particular course. This is a crucial activity that large-scale placement testing, with its built-in illusion that all college programs are the same, has allowed most colleges to avoid. For them, it is cheaper and easier to let the tests place students, to staff the writing courses with part-time help whose voices on curricular matters will not be heard, to hope that whatever such teachers do in class will be minimally respectable, and (in too many cases) to wish that the students in need of extra help will blame themselves for their weak preparation and just go away quietly, after surrendering their tuition dollars. Regardless of what placement procedures an institution uses, there must be a systematic, rigorous program of validity inquiry in which placement decisions are studied from a variety of perspectives including but not limited to student success in the course and teacher and student satisfaction with the placement procedures.

One interesting and important innovation in placement shifts the proposed solution from assessing students' writing, editing, or grammar or vocabulary knowledge to an enhanced form of counseling. Part of the attractiveness of Directed Self-Placement (DSP) is that it proposes a way through this tangle, one that might keep the advantages of placement yet avoid the disadvantages of placement testing. The idea is deceptively simple. In place of testing students, the institution puts its efforts into informing students about the demands and expectations of the composition courses available to them and how they can meet the writing requirement. Then the student makes an informed choice, and takes full responsibility for that choice, instead of more or less grudgingly accepting test results and institutional placement. DSP assumes that students

will be mature enough to choose the course that is right for them, if they have enough information and pressure to choose wisely. DSP also assumes that there may be many reasons besides test performance for students to choose more or less demanding writing courses in their first year of college. And—perhaps the most perilous assumption of all—DSP depends on the institution clearly defining the requirements and proposed outcomes of its different writing courses, maintaining consistency in those definitions, and then communicating them to entering students. For DSP to be effective, the institution must develop some means of making that information meaningful to young students, generally bemused by the mass of lectures, warnings, greetings, and exhortations offered in the weeks before the opening of classes (Royer & Gilles, 2003).

Of course, DSP is no panacea, although its promise is encouraging. Like many other solutions to educational problems, DSP offers new problems in place of old. Yet, the new problems are those that postsecondary education should be meeting anyway: helping students take responsibility for their own learning, replacing reductive placement testing with sound counseling, developing clear curricular guidelines and outcomes, and becoming less paternal and more, shall we say, avuncular. At heart, DSP, like the concept of placement itself, is a conservative proposal, one that maintains the first-year writing requirement as an essential introduction to college-level writing, thinking, and problem solving. DSP is an answer to those unwisely calling for an end to college writing requirements as unnecessary in modern times of technological and vocational revolution. At the same time, DSP proposes a radical solution to the persistent problems of over testing, negative labeling, and student alienation from required coursework.

Will it work? That is, will it be able to convince those inside and outside of academe that it is meeting the political goals of assessment when it avoids assessment entirely? At this point, nobody really knows. Maybe entering college students are not really able to make wise course decisions; perhaps communicating with entering students about their choices is too difficult; maybe the curriculum is in too much disarray to become transparent. Many institutions will need to revamp their counseling procedures for new students to make DSP possible and such change is exceedingly difficult. All kinds of unforeseen problems lurk behind the implementation of DSP, perhaps most pointedly a shift in perception of who should be responsible for academic decisions. The critiques of DSP are appearing along with the encomiums, even in the Royer and Gilles book. But the concept is promising enough for widespread trials—now under way everywhere one looks—and we need to gather information about what happens, as concept becomes procedure at real institutions.

But, as we may expect, a simple and crude political solution to the issue of placement stands ready to replace existing local placement experiments and

abort the promise of DSP. Both of the major American college aptitude testing institutions, the College Board, and the American College Testing Service, have added short impromptu writing tests to their admissions testing programs in 2005. Because most students bound for 4-year colleges and universities take one of these tests, almost every admissions office will now have ready-made placement information at hand, paid for by the student rather than the college, and buttressed by an imposing set of comparative statistics. It will not matter that on many, perhaps most campuses, the information will be useless or worse; it will be politically difficult, if not impossible, to resist using it to place students. So we can anticipate that local placement procedures and the high promise of DSP will fade away in short order.

What is wrong with using national scores on a short piece of impromptu writing to place students in college writing courses? Think for a moment of devising a writing topic appropriate for the privileged students applying to Dartmouth and for the struggling residents of inner-city blighted neighborhoods; consider attempting to score such an examination—or, worse still, attempting to program a computer to score such an examination—with some regard for the diversity of its examinees; consider trying to understand the results when comparing students who grew up in homes using the school dialect to those for whom other dialects or even other languages were used at home. Locally administered placement tests, locally scored, have been able to deal with these problems in various ways, but all those accommodations will probably now be swept away with one universal score, based on national norms. Perhaps most damaging will be the effects of the new tests on the college composition curriculum (oh yes, that), now more or less tailored to the students who wind up sitting in actual classrooms. If we think of the essential purpose of placement, to match particular students to a particular curriculum at a particular campus, it becomes preposterous to even imagine that a single common test score can be used to make accurate, consequential decisions for more than 2 million students entering a variety of institutions. And because tests inevitably define their subjects, think of the high school students for whom writing will increasingly become narrow test preparation.

An additional cruel twist still awaits. Although the commercial firms devising and scoring these written tests are busy recruiting battalions of human readers to score them, does anyone doubt that those humans will shortly be replaced by computers, now moving rapidly into the scoring of writing? A grim satire looms: student computers writing out prose to be read by scoring computers, in turn placing the students into composition sections increasingly taught in computer centers by computer-based instruction. The economy and efficiency is stunning: Neither students nor teachers will need to write or read, or even show up on

campus. Of course, I exaggerate here for effect, and I'm not dismissing the very real and important role computer technology can play in the teaching of writing. On the other hand, my exaggeration has its point. We can emphasize technology at the expense of creating suitable environments for teaching and learning.

Leaving the futuristic satire for the present, we must agree that it will be a bold institution indeed willing to budget its own placement procedures, for its own students, in the face of the scores that will be arriving at no additional cost to the college. Where will we find the political will to fight such a battle? We can expect an impressive marketing campaign, arguing that the vexatious problem of coping with individual students and a broad writing curriculum has now been solved. We must hope that institutions and faculty will resist such false solutions and the mechanistic future they preshadow. As this essay goes to press it is heartening to observe that several members of the WPA listserv report some resistance to using the new ACT or SAT writing tests for placement purposes.

MID-CAREER WRITING ASSESSMENTS

What could be more efficient and reasonable than a mid-career writing assessment, particularly for universities enrolling large numbers of transfer students from community colleges? Such an assessment not only ensures that these students will meet the standards of the receiving institution, but also assures professors throughout the university that student writing issues have been taken care of by the test and so they need not assign or respond to student writing in their own classes. These are great virtues indeed for such a test, but when we look closely at this assessment, and its aftermath, we come to realize that most students are right to see it as an empty hurdle, doing more harm than good. Once again, a test is asked to do much more than it can, and its principal value is political, not academic.

These tests have various names and a long history. The "rising junior" examination at the State University of Georgia was the first large-scale mid-career writing assessment, more than two decades ago, and the California State University followed with its Graduation Writing Assessment Requirement, which took effect in 1981 on its then 19 (now 23) campuses. About the same time, the University of Arizona called it the Undergraduate Writing Proficiency Examination (UDWPE), and other vaguely comic acronyms followed across the land. My favorite is the relatively common "Written English Proficiency Test" (WEPT), which suggests many students' responses after receiving their scores. Many other universities and university systems followed, all with the best of intentions. But, as with the high school proficiency exams, the results have been much less positive than anticipated, while the unintended consequences have been unfortunate to some and devastating to others.

The problems with the rising junior exams are not as severe as they are with the high school tests, at least on the surface. A faculty can often agree on what a student should be able to demonstrate in order to succeed in upper division courses: the ability to read texts of moderate difficulty and write about them clearly enough to show that understanding; the ability to assert some kind of idea and develop it coherently for a few pages; the ability to use source material to support an assertion rather than to substitute for one; and ability to edit written work so it is reasonably free from distracting or embarrassing errors. Sounds easy. But as various departments begin to consider their special needs, more criteria start to appear: the ability to write about scientific or technical matters so a nontechnical reader can understand; the ability to use technology to write and revise; the ability to integrate data and charts into an argument; and so on.

Thus, the creation of a responsible test becomes either so complicated and wide ranging as to be very expensive and time-consuming, or so simple that it loses all credibility. As always, the national testing firms are prominent in the market with their multiple-choice tests, which few faculty respect, if they can even be cajoled into evaluating the instruments. Usually, the English department is told to manage the thing somehow and the rest of the faculty wash their hands of the matter. Meanwhile, about half of the students (those who can be forced or cajoled into taking the test) fail it, no matter what it is. They have been counseled to get first-year writing courses "out of the way," and have written little or nothing in their other lower division courses, so they struggle to remember how to do whatever is called for.

If the creation of the rising junior test is difficult and expensive, the scoring of it is more so. Large institutions wind up with hundreds, sometimes thousands of tests to grade and little money for paying graders. More than one such test has been abandoned for lack of money to pay readers (the University of Arizona's UDWPE, for example) and on some campuses absurd multiple-choice tests have been used as a way to keep the shell of the requirement in effect on the cheap (as one Texas university does). But even when the scoring is supported, by student fees or otherwise, the standards for scoring become a vexatious issue. Can we really expect the students in math or agriculture or physical education to come up to the same standards we might expect of English or history majors? To what degree should we tailor the writing topics and test standards as well as the criteria for scoring to the student's major? It is difficult to harmonize such matters as the preference for brevity and clarity in the sciences with the taste for complexity, metaphor, and wit in the humanities, especially when English faculty end up being responsible for constructing and scoring the tests. Even more vexing for scoring is the ambiguity behind the assessment's purpose: Is the test really a minimum proficiency exam, designed to catch only students whose writing is

so bad that it will be a public embarrassment to the university, or is it an exam defining the critical thinking and sophistication we actually wish our graduates would demonstrate? A minimum proficiency test satisfies the political needs of employers and the public, but the low standards of such an assessment diminish its credibility and participation among the faculty, eventually generating the same concerns from future employers and the public that motivated the tests in the first place. A genuine examination of advanced writing skills, however, will yield many failing scores, even from students with high grade point averages. Are such standards simply unrealistic and unfair?

But once the test is devised and, somehow, scored, the problems are just beginning. No matter how those issues are resolved, the institution is left with a group of students who have failed the test (otherwise, why give it?). Like the high schools trying to cope with the students who have failed their diploma proficiencies, the college must offer something besides sheer despair to such students. Constant repeats of the test are a version of despair, particularly when those whose first language is not English repeatedly and inevitably fail a timed impromptu brief writing sample; I have observed such a test at a California campus, where some students were taking the test for the 13th or 15th time after completing all other requirements for the degree. Surely, every campus in such a situation is obligated to provide some kind of institutional support for those who have met every requirement for graduation except the writing proficiency examination.

This leads to that particular abomination, the upper division remedial writing course, designed to get students, somehow, through the test. It is hard to tell whether the course is despised more by the students taking it or the teachers teaching it. Where the requirement can only be met by passing the test, the course may or may not be useful for actual writing or thinking; what really matters is test preparation. If passing the course is enough, without retaking the test, then the course bears a huge responsibility for enforcing university minimum standards, which are rarely defined with clarity. Both the test and the course are asked to carry the responsibility for writing that must, if it is to be meaningful, be carried by the faculty as a whole.

The best solution to this vexatious tangle of irresponsibility is the one set out by Rich Haswell and others in *Beyond Outcomes* (2001), which recounts the innovative program at Washington State University (WSU). Although based on a special version of portfolio assessment, it includes various other kinds of assessments, including an impromptu essay scored holistically and a certification sign-off option that has involved more than 1,000 WSU faculty members. More appropriate still, the assessment emerges directly from the curriculum, rather than being imposed on it from outside. WSU has invested substantial funds in

this assessment, a rare example of a happy confluence of political and academic goals working together.

Of course, such an elaborate system is not the only way for a university to enforce the reasonable demand that its graduates be demonstrably literate. Some institutions simply require a genuine upper division writing course, connected to a writing-across-the-curriculum program, with some common assessment options, if the political situation requires one. Other colleges require capstone courses in the major, with substantial writing part of the curriculum. And still others have established such a campus culture of writing that a student completing any major can be certified as sufficiently literate. But where those conditions do not exist, the university has to choose among ignoring the political demand for certification of writing beyond the curriculum, meeting that demand with an ineffective and empty assessment program with no real effect on students, or a major investment in a serious curricular and assessment effort as WSU has done. We should not be surprised that WSU stands almost alone at this time.

CONCLUSION

I want to be explicit here that I am not making a case against writing assessment. We will be better teachers of writing if we know how to assess our students' work responsibly, and our students will learn how to revise their work if they learn from us how to assess their own work. Furthermore, careful and responsible assessment of writing beyond the classroom is professionally important, as we have learned from much experience; if we do not meet the academic and political demand for writing assessment at various levels, others will happily take on that task, whether they know anything about the matter or not. Keith Rhodes (in conversation) has named my little proverb on this matter "White's first law of assessodynamics": *Assess thyself or assessment will be done unto thee*. Indeed, in some ways, the misuses of writing assessment I have been discussing are symptoms of our own failures to accept this responsibility. I am, in short, a strong supporter of the responsible uses of writing assessment.

But what I have been dealing with in this article is the misuse of writing assessment. In some ways, this misuse derives from an exaggerated, even a credulous misunderstanding, of what particular kinds of assessments can accomplish. In other ways, it merely reflects an all-too-American view that competition is a positive value and that it is good for society to have a few winners and many losers. In still other cases, it embodies a devious way to avoid difficult problems by substituting a test score—any old score from any old test—as a pseudo answer to such hard social problems as the meaning of a high school diploma or a college degree, or even for whom the doors of opportunity should swing open or

shut. This fast, easy and mis-use of assessment is an important part of the Bush Administration's No Child Left Behind Legislation that emphasizes an elaborate testing, standards and accountability program without the resources and leadership for students to achieve the skills they will be tested on. We must guard against the misuse of assessment while at the same time we promote a climate of responsibility in writing instruction and writing program administration. Just as administrators, politicians and the private sector urge us to be more accountable, we must also hold these people and the testing companies to the very principles of validity that should drive all test use.

REFERENCES

American Educational Research Association. (1999). *Standards for educational and psychological testing*. American Educational Research Association.

Edgington, A., Ware, K., Tucker, M., & Huot, B. (2005). The road to mainstreaming: One program's successful but cautionary tale. In C. Handa & S. McGee (Eds.), *Discord and direction: The postmodern WPA*. Utah State University Press.

Hamp-Lyons, L., & Condon, W. (2000). *Assessing the portfolio: Principles for practice, theory and research*. Hampton Press.

Haswell, R. (Ed.). (2001). *Beyond outcomes: Assessment and instruction within a university writing program*. Ablex.

Hester, V., Neal, M., O'Neill, P., & Huot, B. (2005). Adding portfolios to the placement process: A longitudinal study. In P. O'Neill (Ed.), *Blurring boundaries: Research and teaching beyond a discipline*. Hampton Press.

Hillocks, G. (2003). How state assessments lead to vacuous thinking and writing. *Journal of Writing Assessment, 1*(1). https://escholarship.org/uc/item/33k2v0w5

Huot, B. (1994). A survey of college and university writing placement practices. *WPA: Writing Program Administration, 17(3)*, 49-67.

Huot, B. (2002). Toward a new discourse of assessment for the college writing classroom. *College English, 65*(2), 163-180.

Lutz, W. (1996). Legal issues in the practice and politics in the assessment of writing. In E. M. White, W. Lutz, & S. Kamusikiri (Eds.), *Assessment of writing: Politics, policies, practices* (pp.33-34). Modern Language Association.

McNenny, G. (Ed.). (2001). *Mainstreaming basic writers: Politics and pedagogies of access*. Erlbaum.

Murphy, S. (2003). That was then, this is now: the impact of changing assessment policies on teachers and the teaching of writing in California. *Journal of Writing Assessment, 1*(1). https://escholarship.org/uc/item/1fg1531r

Phipps, R. (Ed.). (1998). *College remediation: What it is, what it costs, what's at stake.* The Institute for Higher Education Policy.

Royer, D., & Gilles, R. (Eds.). (2003). *Directed self-placement: Principles and practices.* Hampton Press.

White, E. M. (2001a). The opening of the modern era of writing assessment: A narrative. *College English, 63*(3), 306-320.

White, E. M. (2001b). Reconsidering the importance of placement and basic studies: Helping students succeed under the new elitism. In G. McNenny (Ed.), *Mainstreaming basic writers: Politics and Pedagogies of access* (pp. 19-28). Erlbaum.

White, E. M. (2006). *Assigning, responding, evaluating: A writing teacher's guide* (4th ed.). St. Martin's.

White, E. M., Lutz, W., & Kamusikiri, S. (Eds.). (1996). *Assessment of writing: Politics, policies, practices*. Modern Language Association.

Willard-Traub, M. (1999). The development of large-scale portfolio placement assessment at the University of Michigan. *Assessing Writing, 6*(1), 41-84.

CHAPTER 5.

ISSUES IN LARGE-SCALE WRITING ASSESSMENT: PERSPECTIVES FROM THE NATIONAL ASSESSMENT OF EDUCATIONAL PROGRESS

Arthur N. Applebee
University at Albany, SUNY

This chapter reviews the development of the framework for the 2011 National Assessment of Educational Progress in writing. An issue paper commissioned by the National Assessment Governing Board is used to consider a number of continuing issues in large-scale assessment of writing, including the definition of the domain of writing tasks, which tasks should actually be assessed at which grade levels, the relationship of the assessment to postsecondary demands, the role of commonly available tools such as word processing software in the construct of writing achievement, the specification and measurement of achievement, the development of appropriate topics for writing, the issue of time for writing, and accommodations for English learners, students with disabilities, and low achievers.

During 2006-2007, committees broadly representative of K-12 teachers, school administrators, state departments of education, university specialists in the teaching and assessment of writing, parents, the general public, and the business community worked to develop a new framework for the 2010-2011 writing assessment of the National Assessment of Educational Progress (NAEP). The previous NAEP writing framework dated to 1989-1990 with revisions, primarily to test specifications, in 1995-1996 (National Assessment Governing Board [NAGB], 2002). Much of the substance of the framework went back even further, to the objectives for the 1983-84 assessment (NAEP, 1982. See the appendix for a summary of the NAEP objectives from 1969-2011.)

DOI: https://doi.org/10.37514/PER-B.2024.2166.2.05

As part of the development process, the NAGB commissioned a background paper to frame issues and debates in writing assessment that could be constructive-ly addressed by the framework committees (Applebee, 2005). The article that follows incorporates that issues paper, adds the recommendations that emerged during the framework development process (NAGB, 2007), and extends the arguments beyond NAEP to concerns that relate more generally to large-scale assessment of writing.

The perspective that frames these issues is a personal and practical one, drawing on recent research and scholarship, changes in policy and practice over the past 20years, and the collective experience of myself and many others in the development, analysis, and reporting of NAEP assessments.

Underlying all of the specific issues that follow is a larger one: What information about how students write should NAEP and other large-scale assessments provide to interested members of the general public, policymakers, and educators? Although it is a seemingly simple question, buried within it are a variety of difficult issues on which there is currently little consensus, including how to describe the domain of writing tasks; the relationships among component skills, content knowledge, and generalized writing "fluency"; and the relevance of computer-based applications to definitions of writing achievement as well as to assessment techniques.

NAEP itself has a number of constraints and opportunities that set it apart from most other assessments of writing. The opportunities derive from the fact that NAEP does not report scores on an individual level. This allows it to use a matrix sampling design in which different students complete different tasks. It also allows NAEP to use a single rater in evaluating each writing sample (with appropriate checks for interrater reliability). As a result, NAEP assessments are able to include more than 20 writing tasks at a given grade or age level, many more than typically can be administered or scored in other writing assessments. Many states, for example, use a single task, as do the SAT and ACT college entrance examinations.

The constraints on NAEP assessments derive directly from the opportunities: In order to relate student performance across tasks and contexts, NAEP uses a complex balanced-incomplete block design (BIB spiraling) in which all tasks at a given grade level are paired with one another in overlapping subsamples of students. In order to do this, the assessment is organized in blocks of items that take equal time to complete for writing and other subjects being assessed. The result at present is that NAEP writing items are constrained to a maximum of 25 to 30 minutes of testing time. State writing assessments, in contrast, often offer considerably more time for writing and revision.

THE ISSUE: WHAT TYPES OF WRITING SHOULD BE ASSESSED, AND HOW ARE THEY RELATED TO ONE ANOTHER?

Recent research in writing has tended to emphasize the extent to which writing genres are socially situated and context-specific. This is true whether one begins with Miller's (1984) emphasis on genre as social action, or the systemic linguistics approach of the Australian genre theorists (Cope & Kalantzis, 1993; Halliday & Martin, 1993). These perspectives pose a challenge to the traditional emphasis on writing as a generic skill, taught primarily in English language arts or composition classes, and assessable through generic writing tasks detached from particular disciplinary or socially constituted contexts. They suggest that what counts as effective argument and persuasive evidence varies greatly in moving from one context to another, so that what counts as "good writing" is itself socially constructed and context-specific. As Halliday and Martin (1993) demonstrated, for example, science writing has many features such as reliance on technical vocabulary, use of the passive voice, and nominalization (use of verbs and adjectives as nouns) that English teachers would ordinarily find objectionable—although these features have evolved in science writing to serve particular communicative needs.

The current NAEP framework, which will remain in place through the analysis and reporting of results from the 2006-2007 writing assessment, derives from the work of Kinneavy (1980), Britton and colleagues (1975), and Moffett (1968) during the 1960s and 1970s, in interaction with perceptions of typical practice and school-based terminology for discussion of writing instruction. The domain of NAEP writing tasks is divided into three broad purposes for writing—informative, persuasive, and narrative. This framework encourages writing within each of these purposes involving a "variety of tasks" and "many different" audiences, triggered by a "variety" of stimulus materials (NAGB, 2002). There is no consensus in theory or practice, however, about the proper way to partition the domain of writing tasks, and there has always been a perception of overlap among the categories: Doesn't an author of an "informative" text implicitly intend to persuade a reader of the truth or accuracy of what is being said? Isn't narrative an important technique for both informing and persuading? ("Narrative" has itself evolved out of concerns in earlier versions of the assessment with "personal," "imaginative," or "expressive" writing, in an attempt to capture the genres of literature as well as of personal reflection.)

The problems in terminology extend to state writing assessments, which have often turned to NAEP as a starting point in designing their own assessments.

Texas, for example, requires writing for "various audiences and purposes," in a variety of forms, including "business, personal, literary, and persuasive texts." California instead treats these generalized purposes as part of "writing strategies," and specifies a variety of specific genres to be assessed (e.g., at Grade 11, fictional, autobiographical, or biographical narrative; responses to literature; reflective com-positions; historical investigation reports; and job applications and resumes.)

There are other alternatives. College entrance exams from the College Board and ACT both assume that good writing is a generic skill, at least in academic contexts; the College Board, for example, advises that high scores will go to "essays that insightfully develop a point of view with appropriate reasons and examples and use language skillfully" (College Board, 2008).

From the Australian genre-theory perspective, Martin and Rothery point in another direction, with a list of schooled nonfiction genres: recount, report, procedure, explanation, persuasion, and discussion. Their listing, like others from the Australian group, introduces terminology unfamiliar to American readers, and also collapses their original insights about the situated nature of genre knowledge into a generic set of "school" genres that are not all that distant from Britton et al.'s (1975) and Moffett's (1968) subcategories of informational or expository writing.

THE OUTCOME

Lacking a widely accepted way to resolve these problems in definition and categorization, the committees developing the 2011 NAEP framework (NAGB, 2007) proposed organizing the assessment around three broad purposes for writing that are closely related to the distinctions made in earlier assessments:

1. to persuade, in order to change the reader's point of view or affect the reader's action;
2. to explain, in order to expand the reader's understanding; and
3. to convey experience, real or imagined.

These represent an attempt to clarify and elaborate the categories of persuade, inform, and narrate in the previous assessment. The framework also attempts to separate purposes from the ways they are carried out, noting that there are a wide variety of strategies for thinking and writing that writers may use in addressing these purposes, including the traditional modes of narration and description, as well as processes such as analyzing and interpreting, and organizational strategies such as compare and contrast. Taking this notion of choices available to writers even further, the 2011 framework recommends that students in Grades 8 and

12 be allowed to choose the particular genre or form in which they will respond (e.g., letter, essay, brochure), rather than having the form of the response dictated by the writing prompt.

The purposes embodied in this proposal, despite the changes in terminology, will provide an easy transition for other assessments that look to NAEP for guidance. The proposal also acknowledges that there is at present no widely accepted alter-native in either theory or practice.

THE ISSUE: WHAT WRITING TASKS/TYPES SHOULD BE ASSESSED AT EACH GRADE LEVEL?

Tangled with the problem of specifying the domain of writing tasks is the distribution of tasks across grade levels. The framework in place through 2006-2007 assumes that each of the broad purposes for writing is appropriate even for primary grade writers, with development taking the form of the ability to complete ever-more sophisticated or specialized tasks within those purposes. Although informative writing tasks have been relatively uncontroversial across the grades, arguments have been raised against assessment of persuasive writing at the fourth grade level, and narrative (particularly story) writing at grade 12. At the fourth-grade level, the arguments have been that persuasive writing is

1. too difficult,
2. developmentally inappropriate, or
3. out of step with the curriculum.

At grade 12, the arguments have been that story writing is

1. too easy,
2. no longer relevant to the curriculum of most students, or
3. not consistent with the types of writing expected in college and the workplace.

The current framework addresses this issue by placing more emphasis on persuasive writing in Grade 12, and more on narrative writing in Grade 4.

NAEP itself offers some evidence on these arguments, in that achievement has been somewhat higher on narrative tasks and somewhat lower on persuasive ones. There has been a narrowing of the range of task difficulty over time, however, early assessments showed much greater between-task variation than is presently evident. This is the result of pilot-testing and task-selection procedures that have eliminated tasks that were very easy or very hard at a given grade level. In fact, the current framework cautions against items that are either too hard or too difficult (NAGB, 2002). One result of this has been

that it is no longer possible to comment on tasks that lower-achieving students can complete successfully, because these tasks are no longer included in the assessment.

THE OUTCOME

Most large-scale assessments have too few separate writing items to have a wide range of task difficulty. The NAEP framework for 2011 similarly recommends a focus on tasks that will encourage all students to write at some length, rather than including some unusually easy or unusually difficult tasks.

The NAEP framework for 2011 emphasizes the importance of writing for a wide range of purposes at all grade levels, including some attention to each of three broad purposes included in the assessment framework. In recognition of the shifting demands of the curriculum, however, the framework places somewhat more emphasis on writing to explain and to persuade in the upper grades, and correspondingly less emphasis on writing to convey experience (primarily storytelling and personal experience essays). For all three purposes, the framework recommends increasingly abstract content and more distant audiences in the upper grades.

The specific types of writing to be emphasized at different grades warrants careful consideration in any large-scale assessment. Curriculum has a tendency to narrow around the types that are assessed, often coupled with unintended effects on what counts as writing well (Hillocks, 2002). Assessments that have to rely on a limited number of tasks at a given grade level might do well to consider designs that sample from a larger range of possible tasks at each grade level assessed, rather than focusing on one or two types.

THE ISSUE: HOW CAN THE 12TH-GRADE ASSESSMENT BE STRUCTURED TO MEASURE PREPAREDNESS FOR POSTSECONDARY ENDEAVORS, INCLUDING COLLEGE, WORKPLACE TRAINING, AND ENTRANCE INTO THE MILITARY?

In 2003 the NAGB established the National Commission on NAEP 12th Grade Assessment and Reporting to review the 12th grade NAEP assessment and to recommend improvements to NAGB. The Commission's report (2004) noted that the high school diploma is no longer a culminating degree for most students; 88% of eighth graders report wanting to continue into higher education, and 70% of high school graduates actually do so within 2 years of graduation.

At the same time, 45%-55% of entering freshmen are unprepared for college work, as reflected in placements in remedial coursework during their first year in college.

Lacking any other national standard for measuring preparedness, the Commission recommended that new NAEP frameworks for the 12th grade be oriented toward assessing preparedness for the challenges of college, workplace training, and the military. At the same time, the Commission noted that there is little consensus on what "preparedness" means, and that validating measures of preparedness is likely to require extensive follow-up studies exploring how students at various achievement levels do in various post-high school contexts. The NAGB's Assessment Development Committee has endorsed this emphasis on 12th-grade preparedness, while noting that the issue is complex and the message that NAEP will send in this regard is very important.

The history of attempts to shape curriculum and assessment around preparedness for future life or work is not a happy one (Applebee, 1974). Past attempts to inventory necessary skills have tended to converge on simple skills that are easy to itemize (spelling, punctuation) rather than higher-level skills (e.g., thoughtful argument and use of evidence) that virtually everyone cites as essential goals of education. The result was usually a system of curriculum and assessment that focused on basic skills or on generic workplace tasks (e.g., business letter format) that easily degenerated into formulas with little real-world relevance.

The most extensive recent effort to relate high school achievement to preparedness both for college study and for the workplace is the American Diploma Project (2004). Drawing on studies of the skills needed in high-performance, high-growth jobs, as well as the requirements for college-level tasks, the American Diploma Project report emphasizes higher-level skills such as expressing ideas clearly and persuasively, and producing high quality writing resulting from careful planning, drafting, and meaningful revision. The report also includes extensive benchmarks meant to indicate the level of achievement appropriate for high school graduation. The 10 benchmarks for writing cover a wide range, from planning, drafting, and revising; to selecting language appropriate for purpose, audience, and context; to writing well-structured academic essays and work-related texts; to using appropriate software programs. Benchmarks under other headings also refer to writing tasks, however, including benchmarks labeled as research, logic, informational text, media, and literature. Although the overall emphasis remains on higher-level accomplishments, the benchmarks show some of the problems of earlier attempts, with appropriate citation of print or electronic sources emerging as a benchmark at the same level of importance as writing an academic essay.

THE OUTCOME

The issue of how current performance relates to the demands of future contexts is an important consideration in the development of any assessment. The new NAEP framework addresses this issue by stressing the continuity of skills that will be needed in postsecondary contexts, rather than by emphasizing particular postsecondary types of writing: Good writing at all levels entails appropriate development of ideas, logical organization, language facility, and use of conventions—all shaped by purpose and audience. Postsecondary contexts also emphasize effective analysis, interpretation, and problem-solving, which is reflected in the 2011 framework in a gradual increase in writing to explain and to persuade at Grades 8 and 12.

THE ISSUE: SHOULD THE WRITING ASSESSMENT BE COMPUTERIZED?

Computer use is becoming widespread in American schools, and by the 2011 assessment it should be even more so. In 2003, for example, virtually all schools reported having computers with Internet access, with no differences among schools serving demographically different populations. Student access to such computers for instructional use has also been increasing rapidly; there was one computer with internet access for every 4.4 students in 2003, compared with one computer for every 12.3 students in 1998 (U.S. Department of Education, 2005).

For writing instruction, the most important computer-based tool has been the word processor. Like the calculator in mathematics, word processing transforms the writing task, simplifying editing and revision and providing embedded tools for spelling and grammar checking. Although most assessments are still paper-and pencil, computer-based assessment that allows the use of word processing is becoming more widespread. When the Test of English as a Foreign Language (TOEFL) exam was recently revised, for example, is was moved to an Internet based format that assesses reading, writing, and spoken language skills; and the Canadian province of Alberta has, for a number of years, made provision for optional use of word processors for Diploma exams in English and other subjects (Alberta Ministry of Education, 2008; Russell & Plati, 2002).

Computer-based writing assessment nonetheless raises some difficult issues of equity and access. Writing produced on a computer tends to be longer than writing produced by hand, and longer writing tends to be more highly evaluated than shorter selections, perhaps because of the inclusion of more evidence or elaboration (Bereiter & Scardamalia, 1987). The bias arguments run in both directions: Not having access to a computer may penalize those who are used to

writing on a computer in school and at home; on the other hand, those who are not used to writing on a computer will either be handicapped by poor keyboarding skills, or if they compose by hand by the greater length of essays produced by their computer-using peers.

The research base on the effects of word processors on assessment results is slim and not particularly convincing; arguments that paper-and-pencil tests underestimate achievement of students who are used to writing on word processors treat writing as though it were being evaluated against an external, fixed standard (e.g., Russell & Plati, 2002), when in fact writing rubrics ordinarily reflect the circumstances of production. Rather than an overall increase in performance, a switch to a computerized assessment including word processing software is more likely to lead to changes in the benchmarks at each level in the scoring rubric to reflect the advantages accrued from the new format.

The most extensive study of the effects of computerizing a writing assessment is NAEP's 2002 study of writing online (Sandene et al., 2005). This special study compared performance on two NAEP writing tasks (one informative and one persuasive) at the eighth-grade level, when given as part of the regular paper-and-pencil assessment or given in a special Web- or laptop-based format that also included simple word processing tools. The detailed results show a number of topic-specific differences in performance across formats, but are generally encouraging. There were no equity-related differences in essay quality, although there was a 1% higher response rate for the paper-and-pencil version of one task. Males also wrote significantly longer responses on computer than on the paper-and-pencil version of one task, but their essays were not rated significantly higher.

Students with more hands-on computer skill (as measured by typing speed, error rate, and ability to use word processing tools) did better on both of the computer based writing tasks; the correlation between their overall writing score and the measure of computer skill was .42; even after adjusting for paper-and-pencil writing achievement, computer skill still accounted for about 11% of the variation in computer-based measures of writing achievement. The "hands-on" computer familiarity measure, however, had a significant literacy component that may account for much of this relationship. Other measures of computer experience, including frequency of completing various kinds of writing assignments on a computer, were unrelated to computer-based writing achievement.

Overall, the authors of the NAEP writing online study conclude that aggregated scores from online assessment do not differ significantly from paper-and-pencil results, although results for individual students may do so.

Although school-level data have recently suggested that equity issues in computer access have been reduced, at the student level issues of access have not been completely resolved. In 2003, for example, there were fewer computers

with Internet access available in schools serving high proportions of minority students than in schools with the lowest proportion of minority students (5.1 students per computer vs. 4.1 per computer). Data from the 2002 writing assessment suggest an even larger divide: Some 29% of White students reported using a computer for writing "a lot," compared with only 19% of Black and 18% of Hispanic students (NAEP Data Explorer, 2002 Writing Assessment).

THE OUTCOME

In designing the 2011 NAEP framework, the committees decided that writing in the 21st century will be computer-based. This is already how most students write, and it is certainly an expectation for writing in the workplace and in postsecondary education. Thus, the 2011 writing framework calls for assessing the ability to write using word processing software at Grades 8 and 12. The framework calls for students to write using "commonly available tools," including the various writing and editing tools widely available in commercial word processing programs. The framework also calls for a computer-based assessment to be phased in at Grade 4 over the life of the framework, as access to and experience with word processing becomes more widespread in the elementary grades.

Computer-based assessment seems almost an inevitable response to the frequency and scale of mandated assessments in all areas of the curriculum. For writing assessment, developers will need to consider how advances in computer use and availability are impacting writing instruction, and what this means for definitions of what it means to write well. If equity issues can be resolved, a computer-based assessment has a number of advantages in measuring writing achievement and in providing accommodations to students who need them (see section on accommodations). Equity issues, however, are much more acute for assessments that report individual scores than they are for NAEP, whose results can serve policy development by highlighting issues of access without penalizing individuals.

THE ISSUE: WHAT ASPECTS OF WRITING ACHIEVEMENT SHOULD BE MEASURED?

Just as there is no widely agreed on definition of the domain of writing tasks, there are many competing approaches to measuring the various interrelated components of writing achievement. Over time, the primary rubric used to measure writing achievement in NAEP has evolved from a holistic rating to a prompt-specific primary trait rating (Lloyd-Jones, 1977) to the current set of purpose-related rubrics (one for each of the three purposes for writing) that can be seen as either generalized primary trait or focused holistic. Although NAEP

reports have been organized around separate sections discussing informative, persuasive, and narrative writing, reporting has either remained at the level of individual writing prompts, or has been aggregated to a total writing score. There have been no separate subscales for types of writing in published reports or in the data available online (NAEP Data Explorer).

Other scoring systems have attempted to provide separate ratings for different features of a writing sample. The most widely used today is probably the 6-trait (or 6+1 trait) system disseminated by Northwest Regional Laboratory. This provides separate scores for ideas, organization, voice, word choice, sentence fluency, conventions, and (optionally) presentation. This system emerged out of the work of Paul Diederich and his colleagues at Education Testing Service (Diederich, French, & Carleton, 1961), and can be a useful tool in reminding teachers and students of the many dimensions of effective writing. As a measurement tool, however, it is not clear that the profiles that result yield psychometrically useful information (Hill, 2001). Diederich's (1974) suggestion was to use the traits for socializing raters to a common standard, and then to drop the traits and focus on total scores.

But there have been many attempts to measure other aspects of writing achievement, including syntactic complexity, ability to edit and revise, mastery of writing conventions (punctuation, capitalization, usage, spelling), organizational ability, and vocabulary level. Such features are arguably of interest in understanding writing achievement, but they have usually required time-consuming scoring procedures and been complicated by the fact that the results are task- and content-specific. Syntactic complexity is usually greater for an analytic or persuasive task than for a narrative task, for example, reflecting the typically embedded nature of clauses in argumentative discourse. Error rates in writing conventions similarly vary with task—with errors tending to increase as tasks become more difficult, presumably as the result of the deflection of cognitive and linguistic resources from one aspect of the task to another.

Many measures of interest that are tedious to derive by hand are very easy to derive by computer. There are now a range of text-analytic software programs available that will report features such as number of words, variety in word choice, syntactic complexity, vocabulary level, and error rates. Many also calculate an overall quality score. If the 2011 writing assessment is computer based, it would allow the assessment of aspects of writing development that can currently only be examined in special studies on limited subsamples of papers.

There is of course another psychometrically efficient option to obtain measures of some of these features. Knowledge of written language conventions and vocabulary level, for example, can be tested quite efficiently in multiple-choice formats. Such measures are highly reliable and have good predictive validity

(Breland, Camp, Jones, Morris, & Rock, 1987; Godshalk, Swineford, & Coffman, 1966); however, they have long been resisted by the community of writing educators because of their impact on curriculum and instruction. Such short-answer formats divert the focus of instruction away from student experiences with writing extended text.

Thus, another benefit of computer-based analyses of features of writing is the ability to derive these measures from samples of extended writing rather than from short answer or multiple-choice formats. This could provide a richer portrait of writing achievement without sacrificing the emphasis on the creation of complete texts.

THE OUTCOME

For the 2011 NAEP, the framework development committees have recommended a focused holistic scoring system with components tailored to the three purposes to be assessed (to explain, to persuade, and to convey experience). Raters will be trained to attend to the development of ideas, to organization, and to language facility and use of conventions, all as appropriate and relevant to the purpose and audience of each task.

The new framework also envisions a "Profile of Student Writing" that would examine in more detail each of these three components. The profile will rely to the extent possible on measures that can be computed automatically from the word processed writing samples, but will also include analytic scoring of a subsample of student writing for features that cannot be derived from computerized text analyses.

THE ISSUE: WHAT SHOULD STUDENTS WRITE ABOUT?

The current framework for the NAEP writing assessment emphasizes writing prompts that are accessible to all students. In practice, this results in an emphasis on common life experience, generic academic content (e.g., favorite books or music, fictional or historical figures, the value of space travel), and on writing that reflects public discourse in a democratic society (e.g., persuasive tasks about community or school issues). If content is provided, it is typically more illustrative than substantive—a brief "story starter," a picture stimulus, or a brief framing of sides of a "controversial" issue. (Real controversies that have political volatility do not make it through the item-review process.) When reading and language difficulties of English-language learners and low-achieving students are taken into account, the push in item development is toward simple and "clean" writing prompts with a low vocabulary load.

At the same time, writing plays a role in virtually all of the other subject area assessments in NAEP. Both short and extended constructed responses comprise major sections of the current assessments in science, history, geography, civics, and reading, as well as the frameworks for new assessments in economics and foreign languages. Rubrics in these assessments bear little similarity to the rubrics in the writing assessment, however, often emphasizing listing of specific content rather than the construction of an argument or explanation. This creates an artificial separation of writing from content knowledge. As Hillocks (2002) pointed out in his critique of state writing assessments, one of the biggest problems in many assessments is the lack of a substantive content base on which to base the writing. Without a content base, much of the writing that results is formulaic and shallow .

THE OUTCOME

For the 2011 NAEP, the framework committees have recommended a continued focus on generic, easily accessible content, including short reading passages, visual stimuli, or graphics. The one major change in the content of the writing tasks is the recommendation that students at Grades 8 and 12 be allowed to choose the genre or form they consider most appropriate to the audience and purpose specified in the prompt. The framework recommends pilot-testing items in a variety of formats (with form specified, without form specified, and with a choice of forms specified) in order to better understand the interaction between purpose, choice of genre or form, and student performance in an assessment context.

Other large-scale assessments vary in the degree to which they rely on generic, easily accessible content. Although many use items very similar to those in NAEP, others, such as New York State, base writing on extended reading passages, or include at least some classroom-based writing as part of the assessment (Kentucky, Vermont). A more general issue for assessment developers is whether it would be useful to increase the content load of student writing prompts, and if so, how this could be done within current assessment frameworks or through extensions of them. One possibility, particularly if writing and other assessments become computerized, would be through the adoption of some common metrics for assessing quality of writing across assessments in different content areas.

THE ISSUE: HOW SHOULD THE FRAMEWORK ADDRESS THE QUESTION OF TIME?

Time to write has been an issue for successive NAEP writing framework committees, and has led both to changes in time allotments and to special studies. From 1970 to 1979, NAEP writing assessments had items of variable length, from a few

minutes for completing forms to nearly 30 minutes on some essay tasks. The move to BIB spiraling in the 1984 assessment reduced the maxi-mum time to 15 minutes. Beginning with the 1992 assessment, this was increased to 25 minutes (with a subset of 50-minute writing tasks that was eliminated in the 2002 assessment).

Two issues usually dominate discussions of writing time: Do the results misrepresent overall writing achievement because students have too little time to write? And does the limited time allowed penalize some groups of students, particularly those whose classrooms have emphasized an extended process of writing and revision? (Conversely, will extended time frustrate lower achieving students and exacerbate achievement gaps?)

The issue of time has been driven by a tension between the constraints of assessment and the conventional wisdom on instruction. One of the accomplishments of the writing process movement in instruction was to remind teachers and students that writing takes place over time—that there are identifiable strategies for generating ideas, drafting, revising, editing, and sharing that shape and reshape a final written text. During the past 30 years of writing assessment, the proportion of teachers claiming to emphasize process-oriented approaches to writing instruction has risen sharply; by 1998 it was central to the instruction of 70% of fourth- grade teachers surveyed, and used to supplement instruction by another 28% (Applebee & Langer, 2006). Comparable figures were reported by Grade 8 and 12 students in the 2002 assessment. (Background questions and grade levels at which they are asked vary from assessment to assessment so there is no single set of data on which to draw.) Given the constraints of large-scale assessment, NAEP has always emphasized that the writing assessment focuses on first-draft writing (as do the College Board and ACT in their college entrance examinations).

Given the overall design of the assessment, when NAEP has included 50-minute tasks the trade-off has been these tasks have not been scalable. (With a 50-minute prompt, each student completes only one task, so interrelationships among tasks cannot be determined.) In 1998,the results of these longer tasks do not seem to have even been reported. Previous NAEP studies of the impact of additional time have yielded mixed results. One special study compared 11th-graders' performance on a persuasive writing task given in 16- or 50-minute time blocks but mixed together for scoring with identical rubrics. As common sense might suggest, the students who had more time for writing scored higher—although the gain was less than might have been expected: 45.4% produced adequate or better responses in 50 minutes, compared with 33.8% in the 16-minute format. The benefits of extra time were not equally distributed among students, however; the extra time made little difference to the weaker writers, increasing the performance gap between the two groups(Applebee, Langer, & Mullis, 1989). The 1992 assessment reported results for 50-minute as

well as 25-minute prompts, with achievement noticeably higher on a 50-minute informative writing task than on the other, 25-minute informative tasks .But comparable increases in achievement did not occur on 50-minute narrative and persuasive tasks included in the assessment; the report concluded that the differences were likely to be topic-related rather than a function of the increased response time (Applebee, Langer, Mullis, Latham, & Gentile, 1994).

These results do not mean that time is not an important factor in quality of writing; simply that the effects of time within the constraints of NAEP writing prompts as they are currently designed are not as large as might be thought, and may be topic-specific. It may be that for meaningful effects of time to emerge, the nature of the tasks would need to be radically reconstrued to incorporate, for example, significant content to be examined or reviewed, or significant feedback to be provided after an initial draft.

Related to the issue of time is whether to provide any special supports for students as they write, particularly supports related to how students use the time available to them. The current NAEP assessment format, for example, includes a blank space that students are encouraged to use to plan their writing. Students also receive a booklet, "Ideas for Planning and Reviewing Your Writing," that suggests planning and revision strategies.

THE OUTCOME

For the moment at least, the NAEP writing assessment remains constrained by a 25- or 30-minute format. Other large-scale assessments have the option to explore formats that go well beyond this, however, and to investigate the effects of variations in time and administrative procedures on student performance. New York State, for example, provides substantive material for students to read and write about, using extended, 3-hour time blocks. Kentucky pairs classroom-based writing with an assigned task, and also insists that some of the classroom-based writing come from subject areas other than English. Hillocks (2002) commented favorably on both of these assessments in his critical look at the quality of writing elicited by various approaches to writing assessment at the state level.

THE ISSUE: WHAT ACCOMMODATIONS SHOULD BE MADE FOR ENGLISH-LANGUAGE LEARNERS, STUDENTS WITH DISABILITIES, AND LOW-ACHIEVING STUDENTS?

NAEP policy is to include as many students as possible in all of its assessments, without altering the construct being measured. In practice, this is accomplished

by careful item-development procedures and, where necessary, by providing accommodations to students with disabilities and English-language learners. Typical NAEP accommodations include more testing time, small-group testing, and other appropriate accommodations depending on the NAEP subject being tested.

As noted earlier, recent writing framework committees have been concerned with making all writing prompts as accessible as possible to all students. This has usually meant a lightening of the vocabulary load and of content provided through the prompt, so that these students would not be put off by problems in understanding before even beginning with their own writing.

The inevitable consequence of this accommodation in the current booklet-based testing format has been that there has been little room to experiment with alternative formats that have the possibility of providing a more substantive context for at least some of the writing tasks.

A computer-based assessment in 2011 would open up a variety of new possibilities, particularly if paired with writing analysis software that could make rapid initial judgments about writing proficiency of individual students. A simple "rangefinder" task, for example, might be used to place students in alternative formats adjusted to their general literacy levels. (New Zealand, for example, uses a very simple and quick initial task in its reading assessment; see NEMP, 2000). Or the response level on the first task administered to each student could be used (with computerized scoring) to select a second task of appropriate difficulty. This could serve to provide accommodations for students who need them, and also to provide greater challenges for higher-ability students.

THE OUTCOME

The 2011 NAEP writing framework recommends typical accommodations such as large-print booklets, extended time, or one-on-one testing when needed. It also emphasizes item-development procedures that will ensure that every item is presented in a simple and clear format accessible to all students.

A move to a computer-based administration for NAEP and other large-scale assessments opens up the possibility of more tailored accommodations in the future, however. By taking advantage of the computer platform, future assessments might be able to individualize such factors as reading load and vocabulary level in ways that are not possible with paper-and-pencil assessment booklets. Assessment developers need to continue to give serious consideration to the effects of accommodations for poor readers and English-language learners on the overall content of the assessment, and look for alternatives that might provide a richer array of assessment options for all students.

CONCLUSION

The framework for the NAEP writing assessment has evolved significantly over the years, in the nature of the writing prompts, in the time available for each task, and in its emphasis on rhetorical features such as audience and purpose. The issues considered in developing a new framework for the 2011 writing assessment have no easy answers, but the changes recommended for 2011 represent an updating that reflects recent changes in scholarship and practice, and that will also return NAEP to its position as a leader in assessment practice and assessment technology. The most significant change involves the movement from paper-and-pencil booklets to a computer-based assessment, which carries with it potential changes in many different aspects of the assessment: in the underlying construct that is being assessed, in possibilities for analyzing the writing samples and reporting on student performance, and in adaptive testing. The challenges will be large, but the opportunity for improving our understanding of student performance is equally large.

States and other groups developing writing assessments will have to confront similar issues in the design of their own assessments, though the particular answers they reach will vary in response to their varying purposes and constraints.

REFERENCES

Applebee, A. N. (1974). *Tradition and reform in the teaching of English: A history*. National Council of Teachers of English.

Applebee, A. N. (2005). *NAEP 211 writing assessment: Issues in developing a framework and specifications*. National Assessment Governing Board.

Applebee, A. N., & Langer, J. A. (2006). *The state of writing instruction in America's schools: What existing data tell us* (A report to the National Writing Project and the College Board). Center on English Learning and Achievement.

Applebee, A. N., Langer, J. A., & Mullis, I. V. S. (1989). *Understanding direct writing assessments: Reflections on a South Carolina writing study*. Educational Testing Service.

Applebee, A. N., Langer, J. A., Mullis, I. V. S., Latham, A. S., & Gentile, C. A. (1994). NAEP *1992 writing report card*. U.S. Government Printing Office for the National Center for Education Statistics, Office of Educational Research and Improvement, U.S. Department of Education.

Bereiter, C., & Scardamalia, M. (1987). *The psychology of written composition*. Erlbaum.

Breland, H. M., Camp, R., Jones, R. J., Morris, M. M., & Rock, D. (1987). *Assessing writing skill*. College Entrance Examination Board.

Britton, J., Burgess, T., Martin, N., McLeod, A., & Rosen, H. (1975). *The development of writing abilities.* Macmillan Education.

College Board. (2008). S*trategies for success on the SAT essay*. http://www.collegeboard.com/student/testing/sat/pre_one/essay/pracTips.html

Cope, B., & Kalantzis, M. (Eds.). (1993). *The powers of literacy: A genre approach to teaching writing.* University of Pittsburgh Press.

Diederich, P. B. (1974). *Measuring growth in English*. National Council of Teachers of English.
Diederich, P. B., French, J. W., & Carlton, S. T. (1961). *Factors in judgments of writing ability. Research Bulletin RB-61-15*. Educational Testing Service.
Godshalk, F. I., Swineford, F., & Coffman, W. E. (1966). *The measurement of writing ability*. College Entrance Examination Board.
Halliday, M. A. K., & Martin, J. R. (1993). *Writing science: Literacy and discursive power*. University of Pittsburgh Press.
Hillocks, G., Jr. (2002). *The testing trap: How state writing assessments control learning*. Teachers College Press.
Kinneavy, J. (1980). *A theory of discourse: The aims of discourse*. W. W. Norton.
Martin, J. R., & Rothery, J. J. (1980). *Writing Project Report No. 1*. University of Sydney, Department of Linguistics.
Miller, C. R. (1984). Genre as social action. *Quarterly Journal of Speech*, *70(2)*, 151-167.
Moffett, J. (1968). *Teaching the universe of discourse*. Houghton Mifflin.
National Assessment of Educational Progress (NAEP). (1982). *Writing objectives 1983-84 assessment*. National Assessment of Educational Progress, Educational Testing Service.
National Assessment Governing Board. (2002). *Writing framework and specifications for the 1998 National Assessment of Educational Progress*. U.S. Department of Education.
National Assessment Governing Board. (2007). *Writing framework for the 2011 National Assessment of Educational Progress* (Prepublication edition). National Assessment Governing Board.
National Commission on NAEP 12th Grade Assessment and Reporting. (2004). *12th grade achievement in America: A new vision for NAEP. A report to the National Assessment Governing Board*. National Assessment Governing Board.
New Zealand National Education Monitoring Project (NEMP). (2000). *Reading and speaking assessment results 2000*. http://nemp.otago.ac.nz/read_speak/2000/
Sandene, B., Horkay, N., Bennett, R. E., Allen N., Braswell, J., Kaplan, B., & Oranje, A. (2005). *Online assessment in mathematics and writing: Reports from the NAEP technology-based assessment project*. National Center for Education Statistics.
U.S. Department of Education. (2005). Internet access in U.S. public schools and class-rooms: 1994-2003. *National Center for Education Statistics*. http://nces.ed.gov/surveys/frss/publications/2005015/2.asp

APPENDIX: THE EVOLUTION OF THE NAEP WRITING FRAMEWORK

Cross-Sectional Writing Assessments

1969-1970 Assessment

1. Write to communicate adequately in a social situation.
2. Write to communicate adequately in a business or vocational situation.
3. Write to communicate adequately in a scholastic situation.
4. Appreciate the value of writing.

1973-1974 and 1978-1979 Assessments

1. Demonstrates the ability in writing to reveal personal feelings and ideas(through free expression and through the use of conventional modes of discourse. [For 1978-1979, reinterpreted as "ability to engage in writing for expressive purposes."])
2. Demonstrates the ability to write a response to a wide range of societal demands and obligations. Ability is defined to include correctness in usage, punctuation, spelling, and form or convention as appropriate to particular writing tasks (social, business/ vocational, scholastic). [For 1978-1979, interpreted as explanatory or persuasive writing done for a particular audience.]
3. Indicates the importance attached to writing skills (recognizes the necessity of writing for a variety of needs, writes to fulfill those needs, and gets satisfaction, even enjoyment, from having written something well).

1983-1984 and 1987-1988 Assessments

1. Students use writing as a way of thinking and learning (for subject knowledge and self-knowledge).
2. Students use writing to accomplish a variety of purposes (informative, persuasive, and literary). [Literary was variously interpreted as "imaginative" and as "personal /imaginative narrative" in reports on these assessments.]
3. Students manage the writing process (generate, draft, revise, edit).
4. Students control the forms of written language (organization and elaboration, conventions).
5. Students appreciate the value of writing (for interpersonal communication, for society, and for self).

1991-1992, 1997-1998, 2001-2002, and 2006-2007 Assessments

1. Students should write for a variety of purposes: narrative, informative, and persuasive.
2. Students should write on a variety of tasks and for many different audiences.
3. Students should write from a variety of stimulus materials, and within various time constraints.
4. Students should generate, draft, revise, and edit ideas and forms of expression in their writing.
5. Students should display effective choices in the organization of their writing. They should include detail to illustrate and elaborate their ideas, and use appropriate conventions of written English.
6. Students should value writing as a communicative activity.

2010-2011 Assessment and Beyond

The 2011 NAEP writing assessment will assess the ability
1. to persuade, in order to change the reader's point of view or affect the reader's action;
2. to explain, in order to expand the reader's understanding;
3. to convey experience, real or imagined.
Beginning in 2010-2011, the assessment will be administered using commonly available word processing tools at Grades 8 and 12, with a similar assessment being phased-in at Grade 4 by 2018-2019.

Long-Term Trend Assessments

1969-1979 through 1983-1984

Writing prompts developed using the 1969-1970 framework were re-administered to study long-term trends through 1983-1984, although trend reports have reinterpreted prompts in light of the writing objectives in place at the time of reporting. Trends were analyzed at the item level rather than using scaled scores.

1983-1984 through 1995-1996

Writing prompts developed using the 1983-1984 framework were re-administered to study long-term trends through 1996, again with reinterpretation of prompts in light of later revisions to the writing framework. Two assessments (1993-1994 and 1995-1996) were limited to long-term trends. The last writing long-term trend assessment administered and reported was for 1995-1996. Although writing long-term trend data were collected in 1999, results were not reported due to instability of the score scale. NCES and NAGB determined that the writing long-term trend assessment should be discontinued because too few prompts were administered to enable reporting of viable trend results.

2010-2011 and Beyond

Writing prompts and procedures developed for the 2011 assessment will be used to establish a new trend line.

FRAMEWORK REFERENCES

Norris. E. L. (Ed.). (1969). *Writing objectives*. Committee on Assessing the Progress of Education.

National Assessment of Educational Progress. (1972). *Writing objectives: Second assessment*. Education Commission of the States.

National Assessment of Educational Progress. (n.d.). *Supplement to the 1973-74 writing objectives*. Unpaginated insert to *Writing objectives; Second assessment*.

National Assessment of Educational Progress. (1982). *Writing objectives: 1983-84 assessment*. Educational Testing Service.

National Assessment Governing Board. (2002). *Writing framework and specifications for the 1998 National Assessment of Educational Progress*. U.S. Government Printing Office.

CHAPTER 6.

THE MICROPOLITICS OF PATHWAYS: TEACHER EDUCATION, WRITING ASSESSMENT, AND THE COMMON CORE

J. W. Hammond
University of Michigan

Merideth Garcia
University of Michigan

Within writing assessment scholarship, disciplinary discussions about the politics of pathways regularly question how reforms mediate education and affect education actors. This article complements and complicates these conversations by attending to the micropolitics of pathways: how local education actors mediate reform-related standards, and, in the process, pave what they believe to be locally-meaningful pathways. Taking the Common Core State Standards (CCSS) as our point of departure, our study centers on one important site for micropolitical work that has, to date, gone unstudied in CCSS-focused writing assessment research: teacher education, which involves coordination between secondary and postsecondary actors who might differently interpret and engage with externally-imposed reforms. Our findings suggest that while standards may be politically intended to mediate education and standardize pathways, teachers micropolitically interpret and repurpose those standards—strategically drawing on them as a means to communicate about local writing instruction and assessment. For this reason, we argue conversations about pathway-related reforms can benefit from adopting a micropolitical perspective, sensitive to the participation of teachers in locally constructing and maintaining educational pathways.

Education reform often focuses on redesigning and managing educational pathways. Whether by introducing standards, assessments, or curricula, these reforms seek to regulate the flow of students across grade levels and school sites—in the

DOI: https://doi.org/10.37514/PER-B.2024.2166.2.06

process, managing student advancement, opportunity, and attainment. Whether we look to past struggles over the American curriculum (Kliebard, 2004), or to present-day resistance to large-scale testing-related reform (Stein, 2016) and systematic over-testing (Lazarín, 2014), it seems the politics of pathways are never fully settled, and are never far from our classrooms.

Writing assessment scholars are no strangers to these politics. They have written extensively, and often critically, about high-stakes, standardized testing-related reforms (e.g., Gallagher, 2011; Hillocks, 2002; Poe, 2008)—including the Common Core State Standards (CCSS) and its attendant large-scale assessments (e.g., Addison, 2015; Jacobson, 2015). In both pushing for curricular alignment and introducing assessments that purport to measure "college readiness," the CCSS participates in paving the pathways students navigate in and between courses—including secondary-postsecondary pathways (Addison, 2015; Bailey, Jaggars, & Jenkins, 2015, pp. 139-141). The CCSS is intended to articulate education institutions, classrooms, and actors:

> High standards that are consistent across states provide teachers, parents, and students with a set of clear expectations to ensure that all students have the skills and knowledge necessary to succeed in college, career, and life upon graduation from high school, regardless of where they live. ("Frequently Asked Questions", n.d., p. 1)

In this manner, "The new standards . . . provide a way for teachers to measure student progress throughout the school year and ensure that students are on the *pathway* [emphasis added] to success in their academic careers" ("What Parents Should Know"., n.d., para. 3).

To date, writing assessment scholarship has raised significant concerns about the CCSS (e.g., Addison, 2015; Clark-Oates, Rankins-Robertson, Ivy, Behm, & Roen, 2015; Ruecker, Chamcharatsri, & Saengngoen, 2015). The pathways it promises are too rigidly or narrowly constructed; however, for all their supposed pathway-defining power, these standards are neither self-interpreting nor self-implementing. "Policy directives—at whatever level of education—do not execute themselves," Gallagher (2011) told us (p. 463). Here is the tension at the core of the CCSS, and of pathway-defining standards, generally: Standards like the CCSS are never as autonomous or agentive as sometimes imagined; they are largely contingent on interpretation and implementation by the very actors they are intended to coordinate and perhaps constrain. In the words of Bridges-Rhoads and Van Cleave (2016), "we (and all teachers) *create* the meaning of the Standards in every instructional moment" (p. 271, emphasis in original).

Our article dwells on this tension. Turning to the CCSS, we explore the *micropolitics of pathways*, by which we mean the ways education actors negotiate and

mediate pathway-related reforms. That is to say, we consider how the CCSS's impacts on articulations, assessments, and curricula are micropolitically shaped by teachers. To borrow Gallagher's (2011) turn of phrase, "being there matters" (p. 468). Even as the CCSS affords teachers a common vocabulary, its local meanings and effects remain reliant on local education actors—each of whom might have a different interpretation of the CCSS and its value. Homogenizing educational projects like the CCSS are always alloyed with heterogeneous local perspectives, assumptions, and aims. While perhaps obscured by standardizing efforts, local differences are not erased by them. Our work seeks to restore the active and strategic participation of teachers in the micropolitics of pathways.

To this end, our research centers on an aspect of education that, while gestured to (e.g., Ruecker et al., 2015), remains unstudied in CCSS-oriented writing assessment scholarship: teacher education work. English Language Arts (ELA) teacher education is a professional space that articulates K-12 and postsecondary actors who might have different beliefs about writing assessment, goals for writing education, and interpretations of writing standards. As such, this space is a useful one for writing assessment scholars interested in how different educators interact with and through pathway-related standards and assessments, like those the CCSS advances. Teacher education helpfully highlights micro-level engagements with the politics of pathways, drawing our attention to the local meanings of standards and the limits of pathway-standardizing efforts. The process of teacher education requires pre-service ("student") teachers to navigate and negotiate novel organizational and professional expectations. As such, micropolitics are notably visible in teacher education and induction work (Blase, 2005; Kelchtermans & Ballet, 2002): Student teachers learn to engage with pathway-related reforms, while—at the same time—experienced educators explicitly guide them through this process. Our article draws on qualitative analysis of interviews conducted with nine educators engaged in secondary ELA teacher education: three field instructors, three mentor teachers, and three student teachers coordinated through a teacher education program at a large Midwestern university (henceforth, Midwestern University). These actors give voice to the micropolitical pathway work teachers routinely do when engaging with standards like the CCSS—work that existing writing assessment scholarship has remained largely silent on.

REFORMS, PATHWAYS, AND THE MESSINESS OF MICROPOLITICS

The ascendancy of national standards-and-assessment reform initiatives (like the CCSS) is only a recent entry in a saga that stretches back over a century (Addison & McGee, 2015)—the story of complex pathways, diverse teacher practices,

and how reformers have sought to manage them. In the past century, new assessment technologies, including writing scales, rubrics, normed holistic scoring, and automated essay scoring, have emerged in response to the perceived problem of heterogeneity (i.e., unreliability) across teacher assessments of student writing (Elliot, 2005). New pathway-related reforms have likewise proliferated, promising increased consistency, commonness, and standardization.

Still, educational complexity is not so easily tamed; the pathways that reforms put in place are seldom as stable and standardized as intended. This is as true for postsecondary reforms as it is for those primarily targeting K-12 education. To give one recent, community college-focused example, Bailey, Jaggars, and Jenkins (2015) suggested student outcomes can be raised through adoption of what they call "guided pathways," which provide students with directive guidance and a more focused curriculum—using faculty and advisors to coordinate (or guide) students "instead of letting students find their own paths through college" (p. 16). We might think of the guided pathways approach as something of a spiritual successor to the CCSS—at least to the extent that both reforms propose to manage the complexity of the curricular paths students take. Finding much promise in the guided pathways idea, Rose (2016) nevertheless reminded us of "how messy and unpredictable the process of reform can be" (para. 12), noting that reforms relying on articulation between faculty members can run into particular challenges: "faculty can have quite different beliefs about concepts like 'improving students' lives.' And some of these differing beliefs can present resilient barriers to change" (para. 18). Reform initiatives can only standardize so much; where their pathways lead is always partly contingent on the assumptions and aims of the teachers who maintain them.

Rose (2016) underscores that the politics of pathways—our overt contestation over the paths structured for students—can be complicated or confounded by the ways educators interpret and engage with reform initiatives, something Blase (2005) has called the *micropolitics of educational change*. The term "micropolitics" has been used in education research to account for the heterogeneity, dissensus, and complexity at the core of education work. In Achinstein's (2002) words, "Micropolitical theories . . . spotlight individual differences, goal diversity, conflict, uses of informal power, and the negotiated and interpretive nature of organizations" (p. 423). Adopting a micropolitical perspective sensitizes us to the idea that educator behavior is not fully shaped and determined by the structures educators participate in; instead, educators partly shape those structures through "the use of formal and informal power . . . to achieve their goals" (Blase, 1991, p. 11; see also Achinstein, 2002; Blase, 2005). This interpretive influence of teachers touches virtually every aspect of educational practice. The complex process of socializing new teachers is micropolitical (Kelchtermans & Ballet, 2002), as is the messy act of collaboration

(Achinstein, 2002; Adamson & Walker, 2011) and the often-overlooked strategic work of interpreting standards and reforms (Blase, 2005; März & Kelchtermans, 2013; see also Dover, Henning, & Agarwal-Rangnath, 2016).

Yet despite disciplinary understandings that teachers are necessary mediators of educational change (Blase, 2005; Gallagher, 2011), and that teachers' beliefs and perceptions affect their teaching (Hillocks, 1999), teacher interpretation and negotiation of the CCSS remains understudied. Work of this kind is essential, for "there are more scholars theorizing about the CCSS than those who are actually collecting and analyzing data from teachers who are responsible for implementing the standards" (Ajayi, 2016, p. 3). To date, larger-scale research on teacher perceptions of the CCSS suggests teachers hold broadly positive views of the CCSS (Matlock et al., 2016), and of the writing and language standards specifically (Troia & Graham, 2016). Even several years into CCSS adoption and implementation, teachers report widespread unfamiliarity with the ELA CCSS-related assessments (Troia & Graham, 2016; also Ajayi, 2016); they also hold conflicted views that those assessments are "more rigorous than their prior state writing tests" but "fail to address important aspects of writing development and do not accommodate the needs of students with diverse writing abilities" (Troia & Graham, 2016, p. 1740; Murphy & Haller, 2015; Ruecker et al., 2015). Perhaps understandably, in trying to take the general measure of emerging teacher engagements with the CCSS, this existing scholarship has focused on broad patterns in teacher perceptions of the CCSS, seldom digging deeper into the messiness of these perceptions—or how teachers micropolitically engage with and locally instantiate the CCSS.

METHODS

Participants

Participants worked together at three different high school sites in professional triads composed of field instructors, mentor teachers, and student teachers at each site. Participants and sites associated with them were assigned pseudonyms beginning with the same letter, chosen to alliteratively signal and clarify relationships. Sites and participants associated with Triad A all begin with "A"—Amanda, Anne, and Alicia at Allendale High; Triad B—Barbara, Brenda, and Brandon, at Bardstown High; and Triad C—Caleb, Cathy, and Cal at Clayville High (Appendix, Table 6.1).

Field Instructors. Recruitment began at Midwestern University by emailing field instructors in its ELA teacher education program. Three instructors expressed interest in participating—Amanda, Barbara, and Caleb (Appendix A,

Table 2). All had previously been secondary ELA teachers. We asked them to recommend participants from student and mentor teacher pairs in their cohorts. Field instructors facilitated communication between the university and mentor teachers, supported student teachers in a weekly course relevant to placement experiences, conducted at least three classroom observations of student teachers, and attended beginning- and end-of-semester meetings with student and mentor teachers. They completed evaluations required for teacher certification, and frequently wrote recommendation letters for students' applications to teaching jobs and graduate programs.

Mentor Teachers. This study includes three mentor teachers—Anne, Brenda, and Cathy (Appendix A, Table 6.3)—from among those recommended by our field instructors. Mentor teachers opened their classrooms to student teachers and field instructors, providing student teachers the opportunity to observe instruction daily and, for part of the year, to take responsibility for two or more classes. They guided student teachers in preparing lessons according to school and state requirements, and helped student teachers apply abstract content and procedural knowledges to real workplaces. Mentor teachers completed two formal evaluations of student teacher performance for inclusion in the student teacher's certification application.

Student Teachers. We draw on data from three student teachers—two (Alicia and Brandon) enrolled in the undergraduate teacher certification program, and one (Cal) in a Master's level certification program (Appendix A, Table 6.4). The Master's program placed students for the entire school year, while the undergraduate program placed students for one semester. Student teachers in both programs observed mentor teachers daily, coordinating with them to plan and enact instructional units (usually spanning four to six weeks) in at least two classes. They submitted unit plans to their field instructors for feedback and evaluation, and scheduled their field instructors' observations to showcase developing instructional skills.

School Sites

Secondary school sites were located in the same state as Midwestern University, a public Research I university whose teacher education program is accredited by the Teacher Education Accreditation Council. While a CCSS adoptee throughout data collection and the writing of this article, this state articulated its standards to and through a standardized test other than the Partnership for Assessment of Readiness for College and Careers (PARCC) and Smarter Balanced Assessment Consortium (SBAC) assessments. During the data collection period, Midwestern University hosted 22 secondary-level student teachers and

partnered with a number of secondary school sites, including the three represented in our study.

Allendale High. Alicia described Allendale High as having a "relatively homogeneous" student population, and Anne explained "we are about 1200 students, 9 through 12. We serve primarily suburban, upper-middle class or affluent families." She added, "primarily we're a school full of white students, but we do pull from a lot of other populations," and that "Allendale tends to pull from students whose parents have found a way to land in the neighboring city and get themselves into the district." State information indicates only 9.7% of the testing population scored not-proficient on the statewide standardized assessment test given to the 268 11th-graders enrolled in Allendale during the 2014-2015 school year.

Bardstown High. Brenda said that Bardstown High has "about 1900 students there, so it's large . . . and it's pretty homogeneous," serving a "mostly white" and "middle to middle-upper class" student population. She explained that parents selected Bardstown because "the scores are very high here . . . last year we had the number one AP scores in the state." Brandon concurred that, "It's one of the best schools in the state, and it's probably, I would say, probably considered one of the best schools in the Midwest for public schools." State information indicates 12.1% of the testing population scored not-proficient on the statewide standardized assessment test given to the 464 11th-graders enrolled in Bardstown during the 2014-2015 school year.

Clayville High. Cathy described Clayville High as "a small alternative education setting with at-risk students in an urban setting. We have about 235 students total that range in age from 14 to 25." In Cal's account, Clayville primarily served students who "have been kicked out or for other disruptive reasons have left their high school, and they are now here. It's really homogeneous. 99%, just about, African American. All are high needs, high trauma." State information indicates that 80.6% of the testing population scored not-proficient on the statewide standardized assessment test given to the 44 11th-graders enrolled in Clayville during the 2014-2015 school year.

DATA COLLECTION AND ANALYSIS

We conducted one-on-one semi-structured interviews (Appendix C) ranging from 30 minutes to an hour long. (One participant, Alicia, submitted responses in written form.) Filler words (e.g., "um," "uh") were excised during transcription. We began by independently coding the data, attending to how participants interpreted and mediated the CCSS through their classrooms, paying particular attention to writing instruction and assessment. We returned to the

data iteratively and collaboratively to tease out nuanced differences within and across participant responses. This analytic approach was supplemented with memos and notes shared between and reviewed by both researchers. At all analysis stages, we sought to document and learn from the diversity evident in participant accounts, rather than evaluate their comparative merits and omissions. Consequently, our work does not account for the myriad effects the CCSS might, in reality, have had—on pathways, curricula, and assessments—beyond those participants raised. Evaluating teacher perspectives and casting our analytic focus beyond them are crucially important projects, but they are not ours here.

FINDINGS

Sensitive to the intended pathway-consolidating function of the CCSS, participants described the CCSS as having the potential to put teachers and students across the country (in Barbara's words) "on the same page"—a phrase used also on the CCSS's official webpage: "With students, parents, and teachers all *on the same page* [emphasis added] and working together toward shared goals, we can ensure that students make progress each year and graduate from high school prepared to succeed in college, career, and life" (Common Core State Standards Initiative, "Read the Standards" n.p.). Yet while each of our participants reported using the CCSS in some way in their curricular planning, none held identical perceptions of the CCSS, and none described using it (or locally assessing it) in quite the same way. Importantly, none of our participants reported that the CCSS fully determined the educational pathways their own students traveled down. Instead, participants reported micropolitically interpreting and repurposing the CCSS—drawing strategically on the standards to supplement and support the local pathways they already had in mind for students.

Cathy, for instance, asserted that standards themselves are—without local curation, negotiation, and interpretation—improper guides for the educational pathways traveled by students. "I think they're [the CCSS] too restrictive," she told us, adding:

> I think standards in general are too restrictive. The needs of students change based on the environment the students live in and the environment that they're going to be going into. If it's a college prep school, standards might be, you know, a little bit, there should be higher expectations. Students that are just going to go out into the world, they just want to find jobs, and they just want their high school diploma so that they can have

> that, they're [the standards] not as important. And sometimes life lessons are more important than the school standards.

Here, Cathy's claim was not just that education must be calibrated to the needs of students, but also, more specifically, that pathways precede standards, not proceed from them—and that teachers appraise the uses and usefulness of standards against the backdrop of the pathways they already imagine for students. Cathy reported that as 11th- or 12th-graders entered her Clayville High classes, "they come to me sometimes and all they need is one English class to graduate, but they're only reading on a third or fourth grade level." Her solution was not to abandon externally-developed standards entirely, but to curate or retrofit them to serve local needs and preexisting pathways. Cathy confided that rather than covering the whole of the grade-level standards, she preferred to "take one or two standards and teach the crap out of them" because she "would rather have them [students] master a few than half-master all of them." Specifically, she focused on "the writing standards," judging these to be in closest alignment with student needs.

The CCSS as a Micropolitical Medium

Our participants described the CCSS as a medium for managing communication with stakeholders and—by extension—signaling professional participation in the collective enterprise of American education. In this way, they framed the CCSS less as a reform that imposes pathways in (and between) schools than as a kind of rhetorical instrument teachers could use when describing the local instructional pathways they constructed. The CCSS afforded teachers in our study a common vocabulary for making local education pathways externally legible. In other words, teachers engaged with the CCSS micropolitically, leveraging its vocabulary to satisfy the complex professional requirement to make instruction-and-assessment decisions intelligible and palatable to an audience made up of multiple stakeholders. In the work of teacher education, this professional requirement involved (at minimum) communication between student teachers, mentor teachers, and field instructors. However, as many participants indicated, the common language found in the CCSS also had a communicative reach that extended beyond the professional triads we interviewed for our study. Standards may be media, but they are media teachers can strategically use.

Curricular Curation and Communication. Alicia noted that adoption of the CCSS and its terminology was *not* the same as adopting a new set of practices or goals. Instead, she appeared to regard the CCSS as a kind of institutional prosthesis for teachers, helpfully facilitating professional conversation

by providing teachers with the terms necessary to voice what they were already doing. She told us, "The CCSS seems to allow quality teachers access to the language that describes the skills that they were likely already teaching in their classroom all along." Underpinning this perception was the idea that the standards represented little more than what good teachers are always already doing in their classrooms—albeit, perhaps, without the vocabulary to make their positive practices known to stakeholders. "My classmates and I, generally, agreed that it [the CCSS] is a document that only suggests skills and lessons that 'good teaching' should have anyway," she wrote. This sense was shared by field instructors Barbara and Caleb, the former of whom reported that, among teachers, "nobody's bothered by it [the CCSS]"; nodding to the communicative uses of the standards, she asked, "what's the big deal? It's nice that everybody's on the same page."

As a general touchstone for talking about "good teaching," the common language afforded by the CCSS was micropolitically useful to our participants, who drew on it to warrant their instructional decisions. Brandon, for his part, described the CCSS as micropolitically "beneficial" as a kind of professional lingua franca, enabling him to enter disciplinary conversations about professional expectations and practices. He confided:

> When I got to student teaching, when my mentor teacher started talking about Common Core, I wasn't like, deer in the headlights, or anything like that. I could discuss it with them. I talked to the principal a lot, and we had departmental meetings and things like that, and I wasn't just sitting there completely with a blank stare on my face. I could contribute to the conversation . . . .

The communicative uses of the CCSS were evident also in Brandon's lesson planning. When he and his mentor teacher developed a unit plan for *Huckleberry Finn*, they started with themes that they wanted to emphasize—"such as friendship, love, and trust, empathy. Those are the things that again, I just think they're so important for kids to learn about"—then moved to the final assessments they would locally implement: a group presentation on banned books, a multiple-choice test on *Huckleberry Finn*, and a portfolio "compiled of a bunch of things" students had composed throughout the unit. After the text, themes, and summative assessment were settled, Brandon and his mentor "went through all the Common Core Standards" and matched them to lessons where "they'll fit in." In other words, Brandon strategically curated the standards, selecting those that best matched his goals and assessments to signal compliance. Cal, too, discussed the CCSS in terms of its communicative uses. He relied on the CCSS

not to define the curricular path he paved for students, but instead to signal to outsiders that this path was an appropriate, standards-approved one.

Cal's sense, though, was less that the CCSS opened professional doors than that it closed opportunities for professional censure. Speaking of the CCSS, Cal described the negative assumptions that might accompany failure to draw on the vocabulary of the CCSS: "if you don't necessarily have it embedded into, like, you know your lesson plans and everything you're doing, then you're not necessarily an effective teacher." Like Brandon and Alicia, Cal's facility with the CCSS provided him a way to signal professional growth to the field instructor who supervised his lesson planning. As Cal tells us, though, this lesson planning was a subtle rhetorical task: "while he wants to make sure that I know them [the standards], he also wants to make sure that I don't know them, if that makes like any sense. It's like, 'Use them, but don't necessarily be pigeon-holed by them.'" Instead of letting the CCSS narrow his curriculum, Cal mined the CCSS for pieces that were relevant, "tak[ing] like bits and like two or three of them, varying on the grade level, and apply[ing] them for my unit plan on a weekly basis." What his description touched on was a pattern in the way our participants engaged with the CCSS: They rhetorically used the standards, and actively resisted being used by them.

Cal's mentor teacher, Cathy, also described curating the standards. She said of the CCSS, "After reading them, they're pretty straightforward, and they can be kind of twisted however you need to use them." Cathy went so far as to suggest that rhetorical engagement with the CCSS—strategically selecting, interpreting, and negotiating them—ought to be "a mandatory class," because such training would facilitate pedagogical self-awareness and develop a capacity to communicate about the educational pathways locally maintained in the classroom. Cathy argued teachers do not need to demonstrate equal fidelity to all educational standards, but she noted:

> It's important to know what you don't like. And it is important to be able to explain why. . . . I think every teacher needs to be educated to the point of being experts on these [the standards] because that's the only way we can get around them if we need to.

No classroom is an island; educational pathways are the shared jurisdiction of multiple stakeholders and, as such, must be negotiated. Using the standards as a shared local language secured for our participants a kind of self-determination that could only come with persuasively communicating with outside stakeholders.

A Common Language for Improving Pedagogy. Indeed, like many educators (including one of the present writers), Cathy is both a teacher and a parent.

As a parent, she appreciated that the CCSS provided a "user-friendly" means of communicating with her daughter's teachers about the educational pathways they locally supported. In this way, too, the CCSS provided a common terrain for communicating about—and rhetorically contesting—the local paths teachers paved. Anne (Alicia's mentor teacher) described herself as taking comfort in the communicative affordances of the CCSS when responding to parents who were "more demanding about the sort of tasks that their kids do." Anne gestured to the CCSS as a means to allay parents' concerns that their students were not on an appropriate educational path: "I find it easier to say, 'Look at all the things, the Common Core things, we're doing!'" For Anne, making the connection between her lessons and the standards was a documentary process that demonstrated the validity of what she was already doing. Amanda—working as a field instructor with Anne and Alicia—expressed a similar, if stronger, conviction about the demonstrative potential of the CCSS. In her estimation, one major problem confronting teachers was the need to "prove" the validity of local instructional decisions to external stakeholders—a communicative requirement she had met through recourse to a "big old curriculum binder" in her own (pre-CCSS) instructional days. For this reason, she taught her student teachers to employ the CCSS as a warrant for the decisions they made, insisting, "what they teach should always be able to be proven—I use that word—you know, with the Common Core. . . ." In this insistence, Amanda seemed to echo Cal's commentary on the uses of the CCSS: within a professional community that has adopted the CCSS, failure to speak the language of the standards was to court sanction—to be left without a persuasive micropolitical means to prove oneself to skeptical outsiders.

Brenda (Brandon's mentor teacher) stressed a "helpful thing [about the CCSS] is that it provides a common language," replacing what might be thought of as the normal, Babel-like diversity of teacher vocabularies for describing the same practices. As an example, she noted:

> What we used to call an "assertion"—people know it as a "thesis"—but it is so clear in the Common Core that you have to have a "claim." . . . It goes back to that language thing. It's very helpful, I mean, all teachers talk about a "claim" now, you can read online, and everyone uses the word "claim," and it's logical. You make a claim, you have to support it.

In this account, the terminological mess of teaching was brought under greater control by adoption of the CCSS's community-articulating vocabulary. As a National Writing Project-trained participant in district-wide writing-specific workgroups, Brenda was perhaps particularly sensitive to the diversity of ways

different groups of teachers related the same essential practices and processes. Importantly, Brenda's description of the CCSS here positioned it less as a reform of practice than of what we call that practice—with teachers adopting a new, shared convention for existing staples of their local work.

This terminological shift alone was one that Brenda believed beneficial. She held that educational pathways already in place would be more easily navigable by students, who would "have a language that transfers from teacher A to teacher B." This affordance of the CCSS was one voiced regularly in the interviews we conducted, with participants noting that a shared vocabulary provided educational pathways at least superficially an increased kind of intelligibility and coherence for students. When Brenda discussed movements between the classrooms of teachers A and B, she was thinking specifically of vertical course alignment, with students advancing from "English 11A" to the next course ("English 11B") in a sequence. However, other participants (like Anne and Caleb) noted the benefits of this shared vocabulary for students who, in Anne's words, "have to be mobile." As Caleb put it, an easily-navigable lateral movement from school to school is essential "so you can live in a country where people move a lot."

The Micropolitics of CCSS Assessment

When asked how they assessed whether students were mastering the CCSS, no participants offered large-scale standardized testing as a possibility. Instead of using the large-scale, seemingly high-stakes testing closely associated with the CCSS to guide their classrooms, participants consistently described developing local, often low-stakes writing assessments to appraise CCSS mastery. Participants expressed a range of perspectives concerning the standardized large-scale tests associated with the CCSS, but interestingly, none reported their classrooms as fully being captive to them. Rather than follow a narrow curricular pathway determined by the CCSS, the teachers in this study curated the CCSS, strategically determining which standards to emphasize and how best to assess student mastery of them.

Mentor Teachers. All mentor teachers had some practical knowledge of prior and emerging state-mandated standardized tests: Anne and Brenda had seen the SBAC test in professional development and practice-test contexts, and Cathy served as the test-coordinator for her campus. Anne considered adaptive testing-facilitated SBAC performance tasks a "big change. Instead of 25 multiple choice questions about grammar, we're looking at higher-order thinking skills." This increased rigor and complexity was not, by her account, an unproblematic good: "I'm thinking, 'Gosh! This is really, really cumbersome in its task, not only just physically, but also mentally.' To me, it involves a lot of technology

skill that I don't know if our students have." Rather than calibrating her classes to CCSS-aligned large-scale tests, though, Anne reported another (more local) way the CCSS was instantiated: formative classroom assessment. "I don't really prioritize essay grading. Instead, I try to prioritize whatever small chunks I can do and give them [students] feedback on immediately," Anne admitted. Such a trade-off was consistent with one of her personally-held "goals as a teacher[, which] is to get feedback to [her] kids in a more meaningful and timely way." She took CCSS-adoption as an occasion to develop a local system for appraising student writing—in small chunks, rather than essays—that reflected her own aims and beliefs as an educator.

Having observed an SBAC practice implementation, Brenda thought the test "was fabulous. I thought it was amazing." Praising the test for its explicit alignment, Brenda regretted the state's decision to pursue a different—potentially less-explicitly aligned—testing system, stating, "I was really sad that we didn't go to it." Even voicing this support, Brenda described her classroom not as caught in the thrall of large-scale standardized test preparation, but instead as backwards planned (Wiggins & McTighe, 2005) against meaningful, locally-developed summative goals. Brenda's school developed and implemented local CCSS-aligned "common final exams in the core classes" like English. In Brenda's own classes, essays were assessed against a rubric targeting "three focus correction areas per paper"—areas that "come from the Writing Core." Brenda's writing rubric was reframed, not narrowed, to reflect the CCSS's commonly-shared vocabulary: "to be honest, I don't think it's [the rubric] anything extraordinarily different [from past, pre-CCSS rubrics], but the wording and the language is going to match the wording and the language on the Common Core. In fact, it might even list the strand." In this way, the CCSS was micropolitically leveraged to signal (rather than determine) local writing assessment priorities—marking, in new terminology, the pathway Brenda had already set.

Cathy, too, reported that state-mandated large-scale tests had "gotten much more difficult" in response to the CCSS, but faulted the CCSS for misalignment to her students' needs: "I really think it focuses too much . . . towards the testing. I think that impacts students a lot, because our kids, . . . they need to be ready for the real world, and Common Core does not always address those needs." Departing from other mentor teachers, Cathy said of the CCSS, "Yes, it's raising the bar to a higher level, but sometimes students need more than that." Here, "raising the bar" was equated to something decidedly *less* than what students needed. Cathy identified the "need to write an argumentative essay" as something that, perhaps, "isn't really important for their [her student's] life needs. Can they write a resumé? That's more important. Do they know how to look up a job application and fill that out? That's more important for these kids." In

this rendition, the CCSS pulled too far in the direction of what were perceived to be postsecondary writing needs at the expense of more immediately valuable (professional) writing-related skills.

"I don't look at the standards as standards, I look at them as suggestions," Cathy claimed; "They're a good place to start, but they can go either way. You can advance them, take it to the next step, or you can cut things out." In a way, our study's mentor teachers all practiced what Cathy preached here—micropolitically mediating the CCSS in the service of preexisting commitments and beliefs regarding assessment. They used the CCSS as license to increase and explore their preferred formative assessment strategies (Anne), leveraged the CCSS's vocabulary to validate assessment practices they believed effective (Brenda), and strategically determined which standards were emphasized, how, and to what ends (Cathy).

Student Teachers. Echoing their mentor teachers, student teachers discussed negotiating and interpreting the CCSS by means of local assessment. Asked if the CCSS dictated or shaped classroom evaluations of student performance, Alicia stressed the multiple ways (beyond standardized testing) assessment could locally instantiate the CCSS:

> it is all about the type of interpretation you take toward the CCSS. I believe that our assessments and classroom evaluations of student performance should be loosely based off of the skills that are offered by the CCSS. However, I do not think that this means that teachers need to merely provide a standardized test assessing these skills. Quite the opposite, in fact. I believe that teachers should provide final assessments that ask students to use a wide range of tasks the CCSS focuses on (e.g., using evidence to support a claim, determining the central idea of a text, etc.).

Consistent with this line of thinking, Brandon speaks favorably about state-mandated standardized testing and its ability to help "comparison throughout the US education to be a little more accurate," but was careful to claim that such testing ought not drive curriculum, because "if you master the Common Core Standards, you're going to do well on standardized testing." Instead, when he discussed the local meaning the CCSS had for guiding assessment, Brandon thought not of state-mandated large-scale tests, but instead—like Brenda—talked of teachers in a department sharing common tests "based on the Common Core Standards." Appropriate assessment was a local matter; as a baseline for describing "good teaching," the CCSS provided local actors the vocabulary necessary to collaboratively develop (and discuss) shared local tests.

In stark contrast to Alicia and Brandon, Cal believed the CCSS was "dumbing down education," and worried formative and individualized assessments might be crowded out because "people could be kind of pigeon-holed to only have a couple of assessments that actually show the students are following the Common Core standard." Crucially, Cal's concerns related less to his classroom than those of other teachers more "willing to kind of take it [the CCSS] by law. . . ." By contrast, Cal's classes prioritized individualized writing assessment ("different benchmarks" indexed to individual students), privileging feedback-rich writing instruction grounded in the "notion of writing as rewriting"—all while carving out space for preparing students to create resumés (a local need identified by Cathy). Even under the CCSS, essay writing and assessment could serve, for Cal, as tools for teaching deeper life skills. "'Listen,'" he exhorted an imagined audience of his students:

> "Your writing can be something that can be extremely superb, but it is something . . . that you have to be willing to work on, meaning that you need a work ethic for [it], and it has to be something that you have to realize that you have to accept criticism for and seek it out for that." And, hopefully, again, [students will] kind of retain that to their life [sic].

Moving between the field instructors at Midwestern University and the mentor teachers at their high school placements, student teachers micropolitically negotiated the CCSS in the process of developing assessments they believed best supported student learning. In response to the CCSS's standardizing potential, they insisted on the need for multiple measures of student progress (Alicia); imagined that collaboratively-developed, standards-aligned assessments would naturally prepare students for success on large-scale standardized tests (Brandon); and advocated individualized, rather than standardized), assessment, tailored to student needs and preexisting pathways (Cal).

Field Instructors. Field instructors had limited direct knowledge of the CCSS-related state-mandated large-scale standardized tests and instead reported on what they gleaned secondhand in placement sites. Interestingly, though, all three field instructors expressed some form of support for the CCSS's large-scale standardized tests—their perspectives underpinned by a more general sense that the CCSS represented little more than (in Caleb's words) "a nice minimum" good teachers always already meet: "I remember the first time I read through them [the CCSS], . . . my feeling was, 'Well if you're not doing these things, what the heck are you doing in English class?' These are just the things that you should be doing." Broadly speaking, the field instructors thought of the CCSS as aligning with their own micropolitical sense of what was normal or appropriate

for "good" teaching. With this understanding in mind, field instructors expected the effects of CCSS-aligned standardized testing to be positive or unobtrusive—altering local practices only in cases of curricular negligence or ineptitude.

Within this context, Caleb referenced the idea that tests drive curriculum—a familiar concern in writing assessment scholarship (e.g., Hillocks, 2002)—but regarded this possibility as a feature rather than a flaw. Comparing the sample CCSS-aligned large-scale tests with previous state tests he had observed as a high school teacher, he told us:

> This [CCSS-related assessment] seems more difficult, so [much] more rigorous, more focused on critical thinking and synthesis of information, and I know that a lot of times, tests drive curriculum, so, I mean, I think the hope is that curriculum will become more rigorous than they were—than it was under [previous] state standards.

Importantly, though, when Caleb spoke about tests driving curriculum, he was not envisioning rote test preparation; his sense was, to the contrary, that quality instruction was already aligned to (and preparatory for) CCSS-aligned large-scale tests. Considering himself "ideologically aligned" with what he considered the CCSS's emphasis on teaching with "big questions" in mind, Caleb claimed "teaching those types of lessons well ensures that kids are going to do fine on the assessment." Caleb attributed the controversy over the CCSS large-scale tests to "misconceptions" in the wake of a weak introduction: "the way it [the CCSS] was sort of rolled out and implemented didn't really promote a lot of clarity, and I think some parents are refusing to let their kids test, and some states are opting out."

For their parts, Barbara and Amanda—neither of whom had seen a CCSS-aligned sample test—expressed regret that politics had complicated CCSS-aligned testing. Barbara's central complaint about the new testing regime seemed to be that state-level political forces had been unwilling to commit to standards and tests long enough for schools to gauge requirements and prepare adequately—a kind of politics of pathways that exchanged student futures for political pride. The state, she argued,

> was reluctant at first to go with the Core, and it's like, 'Everybody else is on board, why . . .?' The legislature again, feeling like, 'We have to be autonomous. We don't need to have the Core. We can have our own guidelines.' It's like, 'Why?' The same thing with the testing . . . .

Amanda, by contrast, regretted that for all the CCSS's promised commonness, its official large-scale tests remained plural, a promise of commonness

fragmented into the SBAC, PARCC, and other state-determined CCSS-related tests. "I thought . . . we would all have the same standardized test—which I hate—but if you've got to have 'em, I'd just as soon it's the same thing, you know, across the board," she maintained.

Whereas Barbara framed the problem of political equivocation in terms of local ability to adapt to new tests, Amanda's concern seemed more that pluralization of CCSS-related large-scale tests contradicted the spirit of the CCSS: "I feel like we're kind of wavering now. Like with the [state-specific CCSS-aligned test], why are we doing . . . ? Why don't we, h[ave]—I thought this was going to be like a national test, where everybody took the same test." Potential articulations and pathways proliferated, Amanda worried, complicating the standardizing promise of the CCSS through tests that mediated and stabilized its meaning in different ways. Indeed, while none of the field instructors regarded the CCSS's large-scale standardized tests as having an undue, constraining effect on classroom instruction or assessment, all of them identified these tests as plagued by overtly political problems. The problems identified with these assessments were less a matter of local, micropolitical engagement than of macropolitical controversy and chaos—with local interpretation and navigation complicated by a confusing rollout (Caleb), state-level indecision (Barbara), and a national inability to adopt a single, standardized assessment (Amanda).

DISCUSSION

When this study's participants communicated with one another about instruction and assessment, they invoked the CCSS as an articulating document. However, beneath the veneer of unity provided by the CCSS, we found substantive disagreement both within and across teacher groups, indicating that pathway-related reforms and the consensus they seek to impose are always fraught with local, micropolitical dissensus. For example, one might expect that field instructors would share an orientation toward the CCSS, using it to evaluate student teachers' lesson plans in a state that had adopted the CCSS. Yet among our field instructors, Amanda insisted that every lesson plan build on a specific standard, Barbara encouraged student teachers to plan around skills and match standards to them retroactively, and Caleb reported that students attended more closely to selecting texts than standards when planning. Consider, too, micropolitical dissensus within Caleb's Clayville triad: Where Caleb viewed the CCSS as introducing more rigor to the curriculum, Cal saw it as "dumbing down education," and Cathy explained that students needed more than what the standards offered. More generally, participants reported developing locally-meaningful lessons and assessments, and strategically curating the "common core" of standards to match

uncommon local needs. We might say, playing on the terminology favored by Bailey et al. (2015), that within the putatively "guiding" structure offered by the CCSS, these participants took a "cafeteria-style" approach.

Professional Secondary-Postsecondary Partnerships

When suggesting productive responses to the CCSS, writing assessment scholarship often recommends attention to teacher training, centering teacher perspectives, and some form of closer, more meaningful articulation between writing studies specialists and K-12 teachers (e.g., Addison, 2015; Clark-Oates et al., 2015). Consider, as one example, Ruecker et al.'s (2015) suggestions from the 2015 special issue of the *Journal of Writing Assessment* dedicated to the CCSS—released just months after completion of our own data collection for this paper. Ruecker et al. (2015) argued "teacher perceptions provide readers a situated perspective of the implementation of the CCSS that is often lost as politicians, test makers, and other individuals fight over the value of the CCSS and the continued push for high-stakes standardized assessment" (para. 3). What is lost, we might say, is the micropolitical perspective teachers bring with them. Suggesting "co-constructed workshops" and "co-teaching in both high school and college classes" as possible paths forward (para. 54), Ruecker et al. reminded readers that close collaboration between secondary and postsecondary educators (including those involved in teacher education programs) has potential "to improve writing instruction for all students"—noting that "it is important to ensure that these relationships are collaborative and not top-down" (para. 53).

We agree with this recommendation, and believe that an explicitly micropolitical perspective is helpful for more fully thinking it through. For instance, we might be led to ask: What does it mean to productively and non-hierarchically navigate around (or through) differences in perception, where standards and assessments are concerned? After all, what it means to "improve writing instruction for all students" is a matter subject to micropolitical negotiation. Because teacher education work is a space where secondary and postsecondary actors already collaborate closely—negotiating the meanings of standards, assessments, and the pathways they participate in—additional teacher education-centric research might aid writing assessment scholars in better understanding the micropolitics of teacher collaboration and conflict (Achinstein, 2002). Such research might also assist writing assessment scholars in understanding how professional articulations (e.g., secondary and postsecondary practitioner partnerships) affect the pathways students end up navigating in and between school sites. Relatedly, where writing assessment research is concerned, co-authorship with practicing teachers (Clark-Oates et al., 2015) is another way productive

secondary-postsecondary partnerships can be pursued—one that, perhaps, provides an additional means by which the micropolitical work of those teachers can be made more visible within our disciplinary conversations.

Moreover, as an existing space where secondary and postsecondary actors are partnered, teacher education work can also benefit from explicit adoption of a micropolitical perspective. As Cathy suggested, it might be helpful for teacher education to explicitly frame engagement with externally-mandated standards as a rhetorical, micropolitical process, training teachers to strategically curate and repurpose those standards, so that (in Cathy's words) they "can get around them if [they] need to." Foregrounding the micropolitical dimension of pathway-related reforms in this way could help secondary and postsecondary actors in developing what Kelchtermans and Ballet (2002) called "micro-political literacy," supporting them as they grapple with externally-mandated standards and assessments—pathway work that teachers (like those in our study) routinely participate in. While the future of the CCSS, like all reform initiatives, may be uncertain, it is worth remembering that the politics of pathways neither began with the CCSS, nor are these politics likely to end with it. For this reason, we have good cause to expect that—whatever reform initiatives the future holds—there will continue to be a need for teacher training that is sensitive to the micropolitics of pathways.

Discussing Teacher Engagements with Standards and Assessments

Our study recommends caution where this kind of equation is concerned. Many of our participants did *not* express a particular distaste for the CCSS or its associated state-mandated assessments—including those participants with special training. Indeed, Brenda—who was National Writing Project-trained and taught a high school class on college writing—stands out as perhaps the participant *most* enthusiastic about the SBAC. Moreover, our participants embraced externally-mandated standards while interpreting them in ways that matched their local instructional goals, assessment preferences, and the writing constructs they privileged. With these insights from teacher education actors in mind, our study suggests a view of teachers not as passive cogs within the political machinery of pathway-related reforms, but instead as micropolitical mediators who make strategic use of those reforms. Micropolitics are not only in play when teachers resist or subvert reform initiatives; teacher support for or reclaiming of standards and assessments can also be an informed, strategic matter (Dover et al., 2016; März & Kelchtermans, 2013). When we reframe teachers as micropolitical actors, we increase the likelihood that our ways of talking about teachers and their perceptions—even those we roundly disagree with—are ways that honor, rather than displace, the intellectual agency of teachers.

Limitations and Future Research

There are, of course, clear limitations to our work here. For one, the CCSS is by no means the only set of standards with which educators engage; more work can be done to discuss the ways multiple, overlapping sets of standards and programs (e.g., Advanced Placement, International Baccalaureate) complicate the politics of pathways in local school settings. Also, because our study focuses on teacher perceptions, we did not triangulate interview data with classroom artifacts and observation data. As a result of this limitation, our research does not examine whether or how participant accounts correspond to outsider/researcher perceptions of classroom realities. Beyond our small, demographically-unrepresentative sample and the resultant non-generalizability of our findings, our interviews were not longitudinal, and omitted perspectives from other important actors and stakeholders—including students. Where engagement with the CCSS and related assessments are concerned, more can and should be done to track the ways that perspectives of all relevant actors change or remain stable over time.

Additionally, while explicit consideration of social justice concerns has been beyond the scope of the present project, it is important to remember that any efforts to define student needs and pave educational pathways are freighted with ethical significance. Recent writing assessment scholarship has underscored the need to consider our assessment practices within a social justice framework, critically questioning how our practices define and structure opportunity (Elliot, 2016; Poe & Inoue, 2016). We believe there is a need for future work that brings micropolitics, social justice, and writing assessment literatures into closer conversation. Particularly promising in this respect would be research critically considering how teachers' local interpretations and repurposing of pathway-related standards participate in promoting (or impeding) educational opportunity (Dover et al., 2016).

We conclude with one suggestive avenue for future work we believe likely to serve as a compelling addition to writing assessment research agendas regarding pathway-related reforms. On the whole, our participants displayed a degree of nonchalance where the CCSS was concerned. As Brenda told us plainly: "I guess, in the scheme of all things to be concerned about, this [the CCSS] is just not high on my list." What *was* high on her list? Brenda reported her school's recent switch to a trimester system—purportedly to save money—had the kind of dramatic impact on writing instruction and assessment that we, as researchers, initially expected to hear about when participants discussed the CCSS:

> [The trimester system] caused us to decrease the amount we write. The class size has gone up, the time in which to teach writing has gone down, and unless you want to grade papers every single night and virtually give up your family life at home,

> during the school year, you're not teaching writing as much because with immediate feedback—how do you do that?

While time and course-load constraints might not be at the top of all K-12 teachers' concerns, we feel there is some promise in coupling our consideration of pathway-related reforms and their effects with questions calibrated to gauge those effects relative to (or as they intersect with) other local constraints and imperatives. Expanding our research in this way promises a means for more meaningfully appraising the impacts of standards like the CCSS. It also affords us a clearer sense of what additional micropolitically-relevant factors impact local writing instruction and assessment—factors that might otherwise be underemphasized in our conversations about the politics of pathways.

REFERENCES

Achinstein, B. (2002). Conflict amid community: The micropolitics of teacher collaboration. *Teachers College Record*, *104*(3), 421-455.

Adamson, B., & Walker, E. (2011). Messy collaboration: Learning from a learning study. *Teaching and Teacher Education*, *27*(1), 29-36.

Addison, J. (2015). Shifting the locus of control: Why the common core state standards and emerging standardized tests may reshape college writing classrooms. *Journal of Writing Assessment*, *8*(1). https://escholarship.org/uc/item/2w69x0w5

Addison, J. & McGee, S. J. (2015). To the core: College composition classrooms in the age of accountability, standardized testing, and common core state standards. *Rhetoric Review*, *34*(2), 200-218.

Ajayi, L. (2016). High school teachers' perspectives on the English language arts common core state standards: An exploratory study. *Educational Research for Policy and Practice*, *15*(1), 1-25.

Bailey, T. R., Jaggars, S. S., & Jenkins, D. (2015). *Redesigning America's community colleges: A clearer path to student success*. Harvard University Press.

Blase, J. (1991). The micropolitical perspective. In J. Blase (Ed.), *The politics of life in schools: Power, conflict, and cooperation* (pp. 1-18). Corwin Press.

Blase, J. (2005). The micropolitics of educational change. In A. Hargreaves (Ed.), *Extending educational change: International handbook of educational change* (pp. 264-277). Springer.

Bridges-Rhoads, S., & Van Cleave, J. (2016). #theStandards: Knowledge, freedom, and the common core. *Language Arts*, *93*(4), 260-272.

Clark-Oates, A., Rankins-Robertson, S., Ivy, E., Behm, N., & Roen, D. (2015). Moving beyond the common core to develop rhetorically based and contextually sensitive assessment practices. *Journal of Writing Assessment*, *8*(1). https://escholarship.org/uc/item/9325025k

Common Core State Standards Initiative. (n.d.). *Frequently asked questions*. http://www.corestandards.org/wp-content/uploads/FAQs.pdf

Common Core State Standards Initiative. (n.d.). *What parents should know*. http://www.corestandards.org/what-parents-should-know/

Common Core State Standards Initiative. (n.d.). *Read the standards*. http://www.corestandards.org/read-the-standards/

Dover, A. G., Henning, N., Agarwal-Rangnath, R. (2016). Reclaiming agency: Justice-oriented social studies teachers respond to changing curricular standards. *Teaching and Teacher Education*, *59*, 457-467.

Elliot, N. (2005). *On a scale: A social history of writing assessment in America*. Peter Lang.

Elliot, N. (2016). A theory of ethics for writing assessment. *Journal of Writing Assessment*, *9*(1). https://escholarship.org/uc/item/36t565mm

Gallagher, C. W. (2011). Being there: (Re)making the assessment scene. *College Composition and Communication*, *62*(3), 450-476.

Hillocks, G., Jr. (1999). *Ways of thinking, ways of teaching*. Teachers College Press.

Hillocks, G., Jr. (2002). *The testing trap: How state writing assessments control learning*. Teachers College Press.

Jacobson, B. (2015). Teaching and learning in an "audit culture": A critical genre analysis of common core implementation. *Journal of Writing Assessment*, *8*(1). https://escholarship.org/uc/item/9653s1fk

Kelchtermans, G., & Ballet, K. (2002). The micropolitics of teacher induction. A narrative-biographical study on teacher socialisation. *Teaching and Teacher Education*, *18(1)*, 105-120.

Kliebard, H. M. (2004). *The struggle for the American curriculum* (3rd ed.). Routledge.

Lazarín, M. (2014). *Testing overload in America's schools*. https://cdn.americanprogress.org/wp-content/uploads/2014/10/LazarinOvertestingReport.pdf

März, V., & Kelchtermans, G. (2013). Sense-making and structure in teachers' reception of educational reform. A case study on statistics in the mathematics curriculum. *Teaching and Teacher Education*, *29*, 13-24.

Matlock, K. L., Goering, C. Z., Endacott, J., Collet, V. J., Denny, G. S., Jennings-Davis, J., & Wright, G. P. (2016). Teachers' views of the common core state standards and its implementation. *Educational Review*, *68*(3), 291-305.

Murphy, A. F., & Haller, E. (2015). Teachers' perceptions of the implementation of the literacy common core state standards for English language learners and students with disabilities. *Journal of Research in Childhood Education*, *29*(4), 510-527.

Poe, M. (2008). Genre, testing, and the constructed realities of student achievement. *College Composition and Communication*, *60*(1), 141-152.

Poe, M., & Inoue, A. B. (2016). Toward writing assessment as social justice: An idea whose time has come. *College English*, *79*(2), 119-126.

Rose, M. (2016). Reassessing a redesign of community colleges. *Inside Higher Ed*. https://www.insidehighered.com/views/2016/06/23/essay-challenges-facing-guided-pathways-model-restructuring-two-year-colleges

Ruecker, T., Chamcharatsri, B., Saengngoen, J. (2015). Teacher perceptions of the impact of the common core assessments on linguistically diverse high school students. *Journal of Writing Assessment*, *8*(1). https://escholarship.org/uc/item/0pq793rq

Stein, Z. (2016). *Social justice and educational measurement: John Rawls, the history of testing, and the future of education*. Routledge.
Troia, G. A., & Graham, S. (2016). Common core writing and language standards and aligned state assessments: A national survey of teacher beliefs and attitudes. *Reading and Writing*, *29*(9), 1719-1743.
Wiggins, G., & McTighe, J. (2005). *Understanding by design* (2nd ed.). ASCD.

APPENDIX A. PARTICIPANT INFORMATION

Table 6.1. Participant Triads and School Sites

Triad	Field Instructor	Mentor Teacher	Student Teacher	School Site
A	Amanda	Anne	Alicia	Allendale High
B	Barbara	Brenda	Brandon	Bardstown High
C	Caleb	Cathy	Cal	Clayville High

Table 6.2. Field Instructor Demographic Data

Participant	Race	Gender	Age	School Site	Courses taught	Secondary experience
Amanda	White, Non-Hispanic	Female	45	Allendale High	7th and 8th ELA, speech	6 years
Barbara	White, Non-Hispanic	Female	64	Bardstown High	9th-12th ELA, world history, psychology, special education	35 years
Caleb	White, Non-Hispanic	Male	43	Clayville High	6th-12th* ELA	9 years

Table 6.3. Mentor Teacher Demographic Data

Participant	Race	Gender	Age	School Site	Courses taught	Secondary experience
Anne	White, Non-Hispanic	Female	35	Allendale High	10 and 12 grade ELA and public speaking	11 years
Brenda	White, Non-Hispanic	Female	44	Bardstown High	7th through 12th grade ELA, debate, college writing, public speaking	23 years
Cathy	White, Non-Hispanic	Female	32	Clayville High	6th through 12th grade ELA, drama, dance	10 years

Table 6.4. Student Teacher Demographic Data

Participant	Race	Gender	Age	Secondary Site
Alicia	White, Hispanic	Female	23	Allendale High
Brandon	Asian, Non-Hispanic	Male	23	Bardstown High
Cal	Black, Non-Hispanic	Male	22	Clayville High

APPENDIX B. SEMI-STRUCTURED INTERVIEW QUESTIONS

[NOTE: Questions marked with * were asked only of student teachers.]

Teacher training and experience

- How many years of experience do you have teaching?
- *Did you have any experience teaching/working in an educational context before entering your certification program? If so, tell me something about those experiences.
  - What were you responsible for in this educational context?
  - What kinds of guides did you use or were you given in this context?
  - In particular, were you responsible for planning lessons/activities?
  - What kinds of guides did you use or were you given to plan activities?
- Briefly describe the institution you are [*student] teaching at (large, small, urban, rural, demographically homogeneous or diverse).
- What grades/subjects have you taught (or do you plan to teach)?
- Briefly describe your teacher training experience.
  - *What is your program like? What do you feel your program is best preparing you to do?

Lesson Planning and Assessment

- Describe your lesson planning process for me.
- How do you go about planning what students will do each day?
- How do you decide what material(s) students will cover?
- How do you assess whether students have achieved the learning goals you set out for them?

Knowledge of CCSS

- *Were you familiar with the CCSS before you began your teacher certification program? If so, what did you know about them?
- How did you first hear about the Common Core State Standards?

- What do you know about the organization that created the CCSS?
- Have you read the document? In what form (online, printed, condensed, complete)?
- Have you received any training or professional development in using the CCSS? If so, describe what you took away from that experience.
- Beliefs and Attitudes Regarding the CCSS
- What value (if any) do you think the CCSS have for classroom teachers?
- What concerns (if any) do you have about how the CCSS might affect classroom teachers?
- What value (if any) do you think the CCSS have for students?
- What concerns (if any) do you have about how the CCSS might affect students?
- Briefly describe any additional ways in which you think the CCSS might be valuable.
- Briefly describe any additional concerns you have about the CCSS and its effects.
- Assessment and CCSS
- How do you evaluate (in class) whether students have met the CCSS?
- Do you think classroom evaluations of student performance are shaped or dictated by the CCSS? How?
- What do you know about the state-wide tests in development for measuring the CCSS?
- Have the state-wide tests for evaluating student progress changed in response to the CCSS? How?
- Have you ever implemented standards other than the Common Core? If so, do the CCSS seem the same or different from previous standards? Explain.
- Have the procedures evaluating you as a teacher been shaped or dictated by the CCSS? How?
- CCSS and Relevant Social Groups
- *How would you describe your field instructor's knowledge of the CCSS?
- *How would you describe your mentor teacher's knowledge and implementation of the CCSS?
- How useful do you think the CCSS are for new teachers versus experienced teachers?
- Do you think the CCSS play a different role in the education of students in different kinds of programs (like advanced placement, regular classes, or remedial classes)? If so, can you walk me through the differences?
- What else would you like me to know about your thoughts on the CCSS?

CHAPTER 7.

WRITING ASSESSMENT, PLACEMENT, AND THE TWO-YEAR COLLEGE

Christie Toth
University of Utah

Jessica Nastal
Prairie State College

Holly Hassel
North Dakota State University

Joanne Baird Giordano
Salt Lake Community College

Two-year colleges are experiencing rapid change, much of which is driven by reform-minded higher education researchers, philanthropists, and policymakers seeking to improve degree completion rates in the nation's open-admissions community colleges. As part of this broader push for reform, placement has come under increased scrutiny, and many two-year colleges are reevaluating and reimagining longstanding placement practices. To set the context for the 2018 special issue of Journal of Writing Assessment on Writing Placement at Two-Year Colleges, this introductory essay reviews five scholarly conversations essential for understanding the issues and stakes: 1) the distinctive histories, missions, demographics, and constraints and opportunities of open admissions two-year colleges; 2) the nature, problems, and possibilities of the reform pressures currently bearing on two-year colleges and placement; 3) the history of writing placement assessment and the theoretical debates surrounding its purposes and efficacy; 4) the recent ethical turn in writing assessment toward sociocultural models of validity and implications for writing placement at two-year colleges; and 5) emerging calls in two-year college writing studies for

DOI: https://doi.org/10.37514/PER-B.2024.2166.2.07

teacher-scholar-activism and critical reform that encourage faculty to take responsibility for challenging inequitable placement processes.

WHY TWO-YEAR COLLEGES?

Since the mid-20th century, two-year colleges—known historically as junior colleges, technical colleges, and community colleges, depending on the specific mission and programming of the institution—have served a critical function as an open-admissions pathway to postsecondary education for a wide range of students. These institutions provide several forms of local educational access, offering non-credit community education courses, "developmental" courses for those institutionally classified as "underprepared" for college coursework, vocational degrees and certificates (often with close ties to local industries), and transfer-oriented general education and associate programs for those pursuing bachelor degrees, as well as growing dual/concurrent enrollment and early college initiatives for high school students (Cohen, Brawer, & Kisker, 2014). In *Gateway to Opportunity?: A History of the Community College in the United States*, Beach (2012) reviewed scholarly perspectives on the function of two-year colleges and concluded that these institutions offer "a limited opportunity and a mixed blessing" (p. 128). Beach (2012) argued that the early mission of the community college was to "*limit* access to higher education in the name of social efficiency" (p. xx) but that students, faculty, and administrators galvanized by the democratic potential of open admissions "tried to refashion this institution into a tool for increased social mobility, community organization, and regional economic development" (p. xx). Tensions between these institutional missions, which reflect impulses of constraint and opportunity, have persisted through the demographic and economic upheavals of the twenty-first century, as two-year colleges became the focus of renewed scholarly debate, philanthropy-driven reform efforts, and state and federal policymaking aimed at increasing the percentage of Americans holding postsecondary credentials. These forces have been rapidly reshaping writing curricula and placement assessment at two-year colleges. At many institutions, however, neither English faculty nor the discipline of writing studies have been well-positioned to influence these reforms (Griffiths, 2017; Hassel et al., 2015; Toth, Griffiths, & Thirolf, 2013).

As a field of scholarly inquiry, writing assessment should have a significant interest in two-year colleges: In 2015, the 1,108 community colleges in the United States served 7.2 million credit-seeking students, which is 41% of all undergraduates nationwide (American Association of Community Colleges, 2017). However, two-year colleges and the faculty who teach in them have long been underrepresented in writing studies scholarship (Hassel & Giordano,

2013; Lovas, 2002; Nist & Raines, 1995; Toth & Sullivan, 2016), including emerging conversations about writing assessment, fairness, and social justice. This dynamic may be shifting. A 2016 special issue of *College English* on writing assessment as social justice, edited by Poe and Inoue, featured two essays focusing on community college students (Alexander, 2016; Naynaha, 2016). Chapters in Poe, Inoue, and Elliot's collection *Writing Assessment, Social Justice, and the Advancement of Opportunity* also begin to address these gaps (Moreland, 2018; Toth, 2018a; 2018b). However, many of these studies demonstrate little or no engagement with the scholarly literature in two-year college writing studies, and none were written by two-year college English faculty. While scholars at all institution types can advance this important scholarly conversation, the authors of this special issue of the *Journal of Writing Assessment* believe it is essential that two-year college faculty participate as knowledge-makers as well as beneficiaries of writing assessment research. Local context matters, and studies conducted at two-year college sites by two-year college faculty can directly inform institutional work and improve student experiences and outcomes. These studies can also make distinctive and important contributions to the broader scholarly conversation about writing assessment.

The underrepresentation of two-year colleges in the writing assessment literature is an urgent ethical issue given the racial, ethnic, and socioeconomic diversity of two-year college students. Nationwide, students of color attend community colleges at disproportionately high rates: These institutions enroll 56% of Native American undergraduates, 52% of Hispanic/Latinx students, and 43% of African American students (American Association of Community Colleges, 2017). Likewise, many "minority-serving"—or New Majority—institutions (e.g., historically or predominantly Black colleges, Hispanic-serving institutions, and tribally-controlled colleges) are primarily associate-granting. Two-year college students are more likely than students at selective-admissions institutions to come from low-income or working-class backgrounds and/or be among the first generation of their family to attend college. They are also more likely to be older/returning students, parents, veterans, immigrants or refugees, and/or students with disabilities (Cohen et al., 2014). These groups of students have long been systemically underrepresented, underserved, discouraged, and disadvantaged in postsecondary education, reflecting and reproducing broader structures of social inequality in the United States. Given these demographic realities, the scholarly conversation about writing assessment, social justice, and the advancement of opportunity must explicitly attend to two-year college contexts. Further, it must do so with an awareness of the distinctive conditions of teaching and administering writing in these settings, including the missions and student populations served, constraints on institutional resources, writing instructors' varying

disciplinary backgrounds and professional identities, limitations on faculty governance and academic freedom, and the current reform-minded policy context in which two-year college faculty are undertaking their work.

AN ERA OF REFORM

Community college researchers and reformers often invoke low and inequitable degree completion rates as a major motivation for enacting change (e.g., Bailey, Jeong, & Cho, 2010; Barnett & Reddy, 2017; Scott-Clayton, Crosta, & Belfield, 2014; Zaback, Carlson, Laderman, & Mann, 2016). In 2016, only 39% of students who enrolled at two-year colleges earned any kind of credential within six years, and nationally, just 16% of entering two-year college students go on to earn a bachelor's degree (Shapiro et al., 2016). There are also unjust racial disparities in these completion rates: Only 33% of Hispanic/Latinx students and 26% of African American students who enroll at two-year colleges earn a credential within six years, and just 11% of Hispanic/Latinx students and 9% of African Americans who begin at two-year colleges eventually complete bachelor degrees (Shapiro et al., 2017). Few argue that there is no need for reform; rather, debates hinge on the nature, goals, and underlying ideologies of those reforms.

As Sullivan (2008, 2017) has reminded us, measuring "student success" at open admissions institutions is a complex endeavor. Not all two-year college students aspire to transfer or even earn degrees: Many are pursuing two-year vocational, technical, or para-professional certifications, or they may be "testing the waters" to see whether college is for them; others are dual-enrollment/early college high school students or "reverse transfers" who have already attended four-year institutions and, for a variety of reasons, stopped out or changed their goals. Degree-seeking students may also shift their aspirations as they gain exposure to and experience with postsecondary education, and many students find themselves facing financial pressures, life crises, or family and community responsibilities that take priority over schooling, at least temporarily (Griffiths & Toth, 2017; Sullivan, 2008, 2017). Furthermore, longstanding federal measures of completion rates have penalized community colleges by not including part-time students or those who transfer to four-year-institutions in their success metrics. When the Department of Education revised these criteria in 2017, it found the 8-year combined graduation and transfer rate for community college students was 60% (Carey, 2017).

Although they face many limitations and constraints, local and comparatively affordable open admissions two-year colleges provide a crucial point of entry to students who would otherwise be unable to access (or re-access) public postsecondary education. Many of these students are not making "market"

choices between two- and four-year institutions, but rather between two-year colleges or no college at all, or between two-year colleges and for-profit institutions that may leave them deep in debt with unimproved employment prospects (Toth, Calhoon-Dillahunt, & Sullivan, 2016). To the extent that writing assessment—whether for placement, in the classroom, or as a requirement for exiting required course sequences—functions to support or undermine student success at two-year colleges, it plays a key role in either opening or foreclosing access to learning, credentials, and, ultimately, socioeconomic mobility for some of the least advantaged students in our postsecondary system.

Over the last few decades, calls among both state and federal policymakers to improve student retention and degree completion have increasingly been framed as a matter of institutional "accountability." As Toth et al. (2016) have observed, accountability measures often fail to acknowledge that "the academic playing field is not level. An institution's record of 'success' is largely shaped by its student demographics and resources. The performance metrics are stacked in favor of selective colleges and universities, particularly the most elite among them" (p. 401). This dynamic makes mounting pressures for performance-based funding problematic. Perversely, such policies risk punishing under-resourced institutions that serve under-resourced students by further denying them resources. They also incentivize heretofore *open* admissions institutions to begin refusing entry to students deemed unlikely to succeed (Toth et al., 2016), determinations typically made based on those students' performance on admissions or placement tests. In this situation, placement assessment and other forms of standardized testing can function to deny access—again, often the only available access to public postsecondary education—to already disadvantaged students. Thus, the stakes of writing assessment in the context of the accountability "movement" are high.

In recent years, the problem of degree completion at two-year colleges has attracted the attention of mega-philanthropies like the Lumina and Gates foundations, as well as higher education researchers who have made use of the influx of funding from such organizations. These parties have been a driving force behind many proposed policy reforms. Perhaps the most influential researchers have been those associated with the Community College Research Center (CCRC) at Columbia University's Teachers College. Over the last decade, the CCRC has produced a number of high-profile publications arguing that one major cause of departure prior to degree completion is the amount of time many two-year college students spend in developmental courses before they can enroll in credit-bearing college-level coursework (e.g., Bailey et al., 2010; Jaggars & Stacey, 2014): During the first decade of the twenty-first century, 68% of two-year college students enrolled in at least one developmental course (Chen, 2016). These researchers have found that, for many students, the costs of the time and resources spent in developmental courses

seem to outweigh the benefits to learning, with particularly negative impacts on students of color (Bailey & Cho, 2010; Bailey et al., 2010; Jaggars & Stacey, 2014; see also Nastal, 2019; Henson & Hern, 2019, in this special issue).

This line of research has fueled the now-robust movement for reducing enrollment in and/or accelerating developmental instruction at two-year colleges. It has also spawned heated debates between CCRC researchers and advocates of developmental education, who have questioned reformers' analyses and the political endgame of their research (for an illustrative exchange, see Bailey, Jaggars, & Scott-Clayton, 2013; Goudas & Boylan, 2012, 2013). The Council of Learning Assistance and Developmental Education Associations (CLADEA, n.d.), which includes most professional developmental education organizations, has responded to policy initiatives that reduce developmental education support with a statement on college access, arguing that "elimination or underfunding of learning assistance programs inevitably restricts college access in ways that lead to blatant educational disparities, very often with patterns related to race and socioeconomic status." The Council offered their own college completion plan in a white paper that the authors describe as "a call to action" for higher education institutions to provide access and support for all students through evidence-based practices (Casazza & Silverman, 2013).

While many two-year college English faculty have embraced—and, in some cases, been important leaders in—efforts to reduce the time students spend in developmental coursework (Adams, Gearhart, Miller, & Roberts, 2009; Cho, Kopko, Jenkins, & Jaggars, 2012; Hassel et al., 2015; Hern, 2012), many also share CLADEA's concern that broad-stroke critiques of developmental education are leading policymakers to cut resources and eliminate programs that provide necessary support for the most disadvantaged students, ultimately foreclosing their ability to access higher education (Hassel et al., 2015). Again, few of these faculty argue against the importance of enrolling students into college-level courses as quickly as possible. The debates center on what combination of reforms to curriculum, pedagogy, assessment, professional development, and resource allocation will best achieve that goal for the diverse student groups entering two-year colleges.

This broad rethinking of developmental education has drawn increased attention to the assessment practices used by two-year colleges to place incoming students into courses. CCRC researchers have released a series of studies suggesting that the common use of high-stakes, single-score purchased placement tests has led to widespread misplacement, and particularly "underplacement"—that is, placing students who are capable of succeeding in college-level coursework into developmental courses, which can negatively impact their persistence to degree completion (Bailey et al., 2010; Barnett & Reddy, 2017; Belfield & Crosta, 2012; Hodara, Jaggars, & Karp, 2012; Hughes & Scott-Clayton, 2011;

Scott-Clayton, 2012; Scott-Clayton & Belfield, 2015; Scott-Clayton et al., 2014; Nastal, 2019; Henson & Hern, 2019). These studies have led many two-year colleges to reconsider their reliance on commercial placement products like ACCUPLACER and the now-defunct COMPASS, and, following the recommendations of CCRC, many are adopting various forms of "multiple measures" placement that increase the range of ways that students can demonstrate readiness for college-level writing (Barnett & Reddy, 2017; Klausman et al., 2016).

The idea of multiple measures aligns with the Conference on College Composition and Communication's (CCCC) position statement on writing assessment (CCCC Executive Committee, 2009) and the Two-Year College English Association's (TYCA) "White Paper on Placement Reform" (Klausman et al., 2016). Hassel and Giordano (2011, 2015) have presented a successful two-year college model for multiple-measures placement grounded in disciplinary knowledge and values. Adopting multiple measures, however, does not automatically make a writing placement process valid, reliable, or fair. The field of writing assessment should continue to inform—and learn from—the development of new placement practices at two-year colleges. As the articles in this special issue of *JWA* demonstrate, two-year college English faculty across the country are seizing the national moment of reform as an opportunity to develop more equitable approaches to writing placement.

THEORIZING PLACEMENT

Placement is a writing assessment process unique to postsecondary education in the United States (Haswell, 2004). While other countries use proficiency testing for institutional admissions, many U.S. colleges use placement assessments once students have already been admitted. In the nation's open-admissions two-year colleges, where students enter from a wide range of academic trajectories and often have not taken any kind of admissions exam, placement assessment is nearly universal. The rationale for placement hinges on the following argument:

1. Placement testing identifies students with the weakest writing abilities.
2. In order to boost those abilities, placement tests funnel students into specific classes or sections where instruction can be more manageable and students can learn better.
3. Therefore, placement testing leads to improved student learning, retention, and completion.

This rationale is predicated on the algorithmic, decision-tree approach to placement advanced by Willingham (1974) more than four decades ago (Figure 7.1). This binaristic, decontextualized model has become the tacit theory undergirding most writing placement.

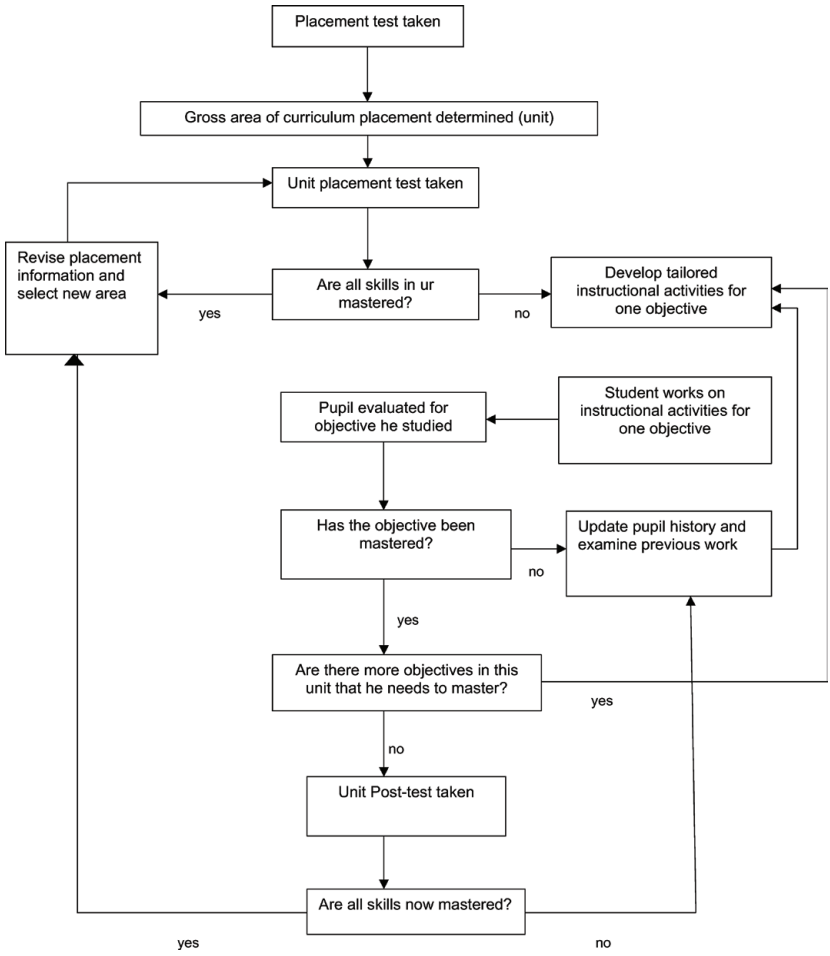

Figure 7.1. Willingham's placement model. Republished from College placement and exemption by W. W. Willingham, 1974, College Entrance Examination Board.

The logic of this algorithmic model is often taken for granted. As Kane (1990) has discussed, the model relies on a linear, clearly-defined progression of attainable and demonstrable skills: Students demonstrate mastery of Skill A; they are then tested on Skill A; those who succeed on test of Skill A progress to Skill B (which relies on Skill A); those who fail on test of Skill A return to the beginning of the unit. One assumption of the algorithm is that the course is based on a series of discrete skills that can be mastered and that build on each other. Another is "that performance on the placement test is relevant to readiness for the . . . course" (Kane, 1990, p. 11). Over the last several decades, however, we have learned much about the recursive nature of writing. We know, for instance, that decontextualized

grammar-usage-mechanics instruction does not necessarily lead to improved writing; as a result, continuing to use placement assessments that rely on outdated notions of the writing construct are often neither valid, reliable, nor fair.

Thirty years ago, Morante (1987) argued that placement tests and their corresponding cut scores "play important roles in access, retention, and quality" (p. 63), asserting, "To dump everyone in the same level of course is significantly to increase the probability of lowering standards or of failing many students" (p. 63). A decade later, White (1995) claimed placement testing "[serves] to help underprepared students succeed instead of washing them out . . . [T]hese are the students for whom required placement and the required freshman course are necessary, for they are most in need of guidance and support" (pp. 76–77). At most institutions offering multiple levels of writing courses, including two-year colleges, which often offer two or more levels of "pre-college" developmental writing, these assumptions have gone largely unchallenged.

Indeed, placement has long been viewed as necessary to increase the productivity of both instructors and students in writing classes. Some institutions, for instance, segregate students who score highest on placement tests or entrance exams into honors-level courses where they receive more advanced instruction than a typical college writing course offers, benefit from smaller class sizes, and are surrounded by exceptional peers. In other contexts, they are exempted entirely from a college writing requirement. Students sorted in this way are, in the view of advocates, alleviated of the "burden" of assisting their peers who may have less preparation, and instructors are rewarded with teaching the best prepared and most motivated students. The "gateway" college writing courses (i.e., English 101 or Composition I) are then filled with students who are "average," and developmental courses are filled with students who need the most instruction, so teachers can target their lessons, assignments, and assistance appropriately for each group of students. While such sorting processes are not employed at every institution, when they are used, they are typically perceived as being necessary to "efficiently" shuttle students through their required writing courses. The perceived value of such efficiency relates directly to the material conditions of postsecondary writing instruction. Many composition programs nationwide face increasing class sizes while relying on often underprepared graduate student instructors and, particularly at two-year colleges, undercompensated and not-always-well-supported adjunct faculty. In these settings, sorting based on abilities is presumed to help ease the labor of teaching.

Historically—and, as Williamson (1994) has observed, problematically—writing assessment has often been driven by such questions of efficiency, or, as Yancey (1999) put it, "Which measure can do the best and fairest job of prediction with the least amount of work and the lowest cost?" (p. 489). This orientation treats composition courses as necessary but burdensome for both students and

the institution. In recent decades, writing program administrators and writing studies teacher-scholars have made headway in shifting the conversation about college composition from teaching "basic skills" to engaging students around disciplinarily-informed insights that help prime them for life-long development as critical readers and writers. However, at many institutions—and particularly at two-year colleges, where writing faculty often have less disciplinary authority over assessment—placement into composition courses is still viewed not as a pivotal educational moment for introducing students to local pedagogical orientations and the valued construct of writing, but rather a mechanism for putting students in their "proper" seats quickly, easily, and inexpensively. This perspective has led to the proliferation of methods that sort students cheaply and "accurately," often leaving unaddressed critical questions about what accuracy means, how it might shift depending on the stakeholder, and what messages placement conveys.

However, the placement processes employed by an institution do send powerful messages to incoming students, local high schools, and other concerned stakeholders. If high schools desire their graduating seniors to score well on the placement test used by their area community college, they almost inevitably will steer their curricula toward that test. Thus, for example, two-year college placement tests that require no writing will almost certainly encourage local high schools to emphasize multiple-choice testing and de-emphasize the difficult and often messy practice of teaching writing within purposeful rhetorical contexts. As Harrington (2005) argued, placement also plays a central role in representing our campuses and writing programs to students:

> Placement is more than a decision about coursework for students. It is most students' first contact with the theory and practice of first-year writing programs, and we would do well to make that first contact as inviting and theoretically sound as possible. To do so, we need to think less about placement as mechanism and more about placement as an opportunity to communicate. (p. 12)

Placement is an introduction to the institution and how it conceives of writing. It is not a neutral action. It communicates specific cultural values, language ideologies, expectations to test-takers and participants: In short, it communicates power. It can replicate or trouble inequitable social structures; it can support or challenge the current era of testing and assessment despair (Gallagher, 2007). Because students' encounters with placement are so central to their entry into postsecondary education, writing studies scholars argue that we should take that opportunity to communicate our most central values: rhetorical knowledge; critical thinking, reading, and composing; writing processes; and knowledge of—and capacity to challenge—conventions (Council of Writing Program Administrators, 2014).

Unfortunately, decontextualized algorithmic approaches to placement typically offer little helpful information about the ways most institutions and teacher-scholars conceive of writing. The widespread reliance on commercially produced tests that measure a very limited construct of writing has prioritized knowledge of Edited American English conventions at the expense of any other outcome, primarily because these are the skills that can be easily measured through multiple-choice tests (Huddleston, 1954; Stein, 2016; Williamson, 1994), quickly written paragraphs (Bereiter, 2003; Faigley, Cherry, Jolliffe, & Skinner, 1985), and automated writing evaluation (AWE) software (Burstein, 2012; Herrington & Moran, 2001, 2012; Perelman, 2012). Multiple-choice usage tests primarily reward familiarity with the conventions of a privileged written English variety most closely associated with White, middle-class, monolingual literacy practices. Even tests with an actual writing component assessed by AWE primarily measure length, an easily identifiable structure, and few linguistic or mechanical "errors," rather than meaning or rhetorical effectiveness; Perelman (2012) described these tests as "bullshit" (p. 427) because students may be rewarded when they include irrelevant or inaccurate information to answer short essay questions that have nothing to do with their knowledge or experience domains. In most actual college writing situations, students are expected to demonstrate knowledge based on course texts, assignments, and discussions or professional expertise. Placement assessments with such limited construct representation might work to shunt students into writing classes and allow them to check the box and finish their writing requirements. They do little, however, to expand the narrow conceptions of writing that many students bring with them based on prior assessment experiences or to prepare students for longer-term rhetorical awareness and writing knowledge transfer.

Despite writing placement's perceived necessity, Haswell (2004, 2005) has offered an astute critique of its efficacy: He claims reliability and predictability are poor enough to call into question the ubiquitous and long-standing use of placement testing. Most students have been found to change their score significantly the second time they took the test (Haswell, 2004). Furthermore, Haswell (2004) demonstrated that research conducted since placement testing began with the 1874 Harvard entrance exams shows that both indirect and direct methods of testing do little in the way of predicting student success. His analysis of studies from 1906, 1927, 1954, 1992, 1999, and 2004 suggested "that for decades college writing placements have been made on scores that leave unexplained, at best, two thirds of the variance in future course performance, and, on average, nine-tenths of it" (Haswell, 2004). Likewise, Smith (1993) analyzed the locally-designed test at University of Pittsburgh, which used a robust scoring method that relied on its expert teachers, and found 14% of students were under-placed. While this

may seem like a "good enough" number for some, Smith (1993) argued, "For the students and for the teachers, 'very few' is too many" (p. 192). This may be particularly true at open admissions two-year colleges, where underplacement into developmental courses can lengthen time to degree or discourage students from persisting or even enrolling (Adams, 1993; Bailey, 2009; Bailey et al., 2010; Henson & Hern, 2019; Nastal, 2019), while overplacement might increase the possibility of student failure, costing them time and tuition dollars and potentially resulting in academic probation or suspension.

According to Scott-Clayton (2012), high-stakes, single-score placement tests were being used by 92% of two-year colleges at the beginning of the decade. As the articles in this special issue of *JWA* demonstrate, we are only beginning to attend to what Messick (1989) called the *social consequences* of these longstanding placement practices. From the perspective of racial and socioeconomic equity, those consequences are often profoundly troubling. As Morris, Greve, Knowles, and Huot (2015) noted in their recent overview of book-length studies of writing assessment:

> While there is little or no scholarship focused specifically on two-year college writing assessment, it is important to recognize the important influence writing assessment can have for students' educational opportunities, especially at two-year colleges, which enroll the majority of postsecondary under-resourced students. (p. 120)

Furthermore, they argued, "Writing assessment can also be a critical issue for two-year college identity and legitimacy" (Morris et al., 2015, pp. 120–121). Over the course of our own careers, we have heard university-based colleagues speak dismissively of community colleges on the basis of their purportedly uncritical placement and "remediation" practices. Thus, the visible disconnect between writing assessment theory and on-the-ground placement practice has consequences for the reputations of two-year colleges, their instructors' professional status within the discipline, and the perceived value of the education their students receive.

In sum, we know the consequences of writing placement based on decontextualized algorithmic thinking and limited construct representation can be dire. It sends inaccurate and counter-productive messages about what we value in college writing; it appears to misplace students at unacceptable and often inequitable rates; it fails to assess key capacities necessary for college success; and it does not provide information about what kinds of supplementary supports might benefit students—something that contextualized, nonbinaristic measures with broader construct representation can offer (Hassel & Giordano, 2015). At two-year institutions, the consequences of poor placement practices are not simply a matter of how many credit-bearing writing courses a student will need to complete. In an

unreformed two-year college curriculum, misplacement can mean taking as many as three non-credit developmental courses before entering into credit-bearing composition (Patthey-Chavez, Dillon, & Thomas-Spiegel, 2005; Nastal, 2019). Many students will not have the time, money, or motivation to persist through a year of additional writing coursework—more if they do not pass a class. These barriers to educational access are a function of placement tests that sacrifice validity to reliability and underrepresent the writing construct; however, such barriers can be reduced or eliminated if we develop placement assessment processes that prioritize fairness.

THE ETHICS OF PLACEMENT

Over the last decade, writing studies scholars have been reexamining the ethics of assessment. Poe and Inoue (2016) have identified this theoretical movement as a turn toward "sociocultural model[s] of validity" (p. 118) that "provide a useful reworking of validity theory for the purposes of social justice" (p. 118). Scholars in this turn have drawn insights from a number of transdisciplinary critical fields, including:

- Critical race theory, whiteness studies, and anti-racism (e.g., Behm & Miller, 2012; Burns, Cream, & Dougherty, 2018; Hammond, 2018; Inoue, 2009a, 2009b, 2012, 2015, Inoue & Poe, 2012a, 2012b; Naynaha, 2016; Poe & Cogan, 2016; Poe, Elliot, Cogan, & Nurudeen, 2014)
- Decolonial theory (Cushman, 2016; Gomes, 2018)
- Translingual theory (Poe & Inoue, 2016)
- Queer theory (Alexander, 2016; Caswell, 2018)
- Philosophical work on ethics and social justice (e.g., Elliot, 2016; Poe & Inoue, 2016; Slomp, 2016b; Stein, 2016)

These scholars ask us to consider how writing assessments are shaped by dominant epistemological assumptions, values, and language ideologies that are raced, classed, gendered, and/or colonial/imperialistic, and often predicated on normativities regarding physical abilities, sensory processing, and neurotypicality. Such critical interrogation is essential even for assessments that appear on the surface to be neutrally "meritocratic." These assessment practices may still be enacting what Behm and Miller (2012), following Bonilla-Silva (2006), have called a "color-blind racist" assessment paradigm that continues to reproduce structures of social inequality. As we have noted, there is mounting evidence that longstanding writing placement practices at two-year colleges—institutions that are the major point of access to postsecondary education for many structurally disadvantaged groups—have been performing

precisely these inequitable functions. It is thus imperative that we bring the insights of this ethical turn in writing assessment to bear on the question of placement at two-year colleges.

New critical frameworks challenge algorithmic assessment models like Willingham's (1974). They offer valuable conceptual tools for analyzing the social consequences of two-year college assessment practices and ontological options for imagining fairer alternatives. These tools include *racial validity inquiry* (Inoue, 2012, 2015) and *disparate impact analysis* (Poe & Cogan, 2016; Poe et al., 2014), which encourage disaggregating assessment data by race and other legally protected categories. Extending these concepts, Slomp (2016b) has argued for "disaggregation of data so score interpretation can be clearly understood for all groups and each individual within those groups" (Slomp, 2016a), with particular attention to what Elliot (2016) has called the "least advantaged," to determine whether assessment practices are having an adverse impact on some groups. If so, these assessment practices can and should be redesigned to achieve more equitable outcomes. Such redesigns may require not only revising assessment processes and instruments, but a "fundamental rethinking" (Slomp, 2016b) of the values, goals, and practices driving writing assessment in the context of what Inoue (2015) calls our "local diversities" (p. 68). Both Cushman's (2016) argument for decolonizing the concept of validity and Alexander's (2016) suggestions for queering writing assessment ask us to question the epistemological universalism and normativities built into why and how we measure writing performance. They encourage us to develop assessments that value the plurality and diversity of our students' languages, literacies, and rhetorics. Such local re-valuation is particularly pressing at two-year colleges, given their diverse students, institutional missions, and community contexts.

The urgency of such rethinking is evident in *JWA*'s recent special issue on the ethics of writing assessment (Kelly-Riley & Whithaus, 2016). This special issue responds, in part, to the 2014 *Standards for Educational and Psychological Testing* articulated by the American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME, 2014)), which defined *fairness* in assessment as:

> The validity of test score interpretations for intended use(s) for individuals from all relevant subgroups. A test is fair that minimizes the construct-irrelevant variance associated with individual characteristics and testing contexts that otherwise would compromise the validity of scores for some individuals. (p. 219)

While the *Standards* document focused on subgroup difference, it framed fairness in service of validity (Elliot, 2015). As Slomp (2016b) has asserted, "cultural bias—such as subgroup differences being related to undemonstrated

assumptions about students rather than from reflective latent variable models validated under field-test conditions" was not explicitly attended to in the *Standards*, and as a result, students and practitioners may encounter technically sound assessment practices whose social consequences have been ignored.

The 2016 special issue contributors drew on foundational texts in social science, education, and assessment to arrive at their call to use writing assessment as a means of achieving social justice "as a principle of fairness so opportunities do not merely exist, but rather, so each individual has a fair chance to secure such opportunities" (Slomp, 2016b). Elliot (2016) identified "fairness" in writing assessment as "the identification of opportunity structures created through maximum construct representation under conditions of constraint—and the toleration of constraint only to the extent to which benefits are realized for the least advantaged." This rethinking of fairness in terms of opportunity structures has powerful implications for two-year colleges, which have a mission to provide access to educational opportunity for the "least advantaged."

As we have discussed, the commercial exams that have long dominated two-year college writing placement typically offer inadequate representation of local constructs of college writing. They also reproduce language and literacy ideologies that advantage students from White, middle-class communities. While we have long tolerated such constraints in the name of efficiency at often under-resourced open admissions institutions, it is now clear that those constraints have, in fact, harmed the least advantaged. Through systematic misplacement, particularly underplacement that delays enrollment in college-level courses, we have reduced those students' likelihood of degree completion. In the process, we have also sent them negative, destructive messages about their capacities as writers and learners and about the value of the rhetorical and literacy practices in their out-of-school communities. These disparate, adverse impacts are neither fair nor, in many cases, legal (Klausman et al., 2016; Poe & Cogan, 2016; Poe et al., 2014). The educational policy shifts of the last decade have created an opportunity to rethink "business as usual" (Klausman et al., 2016, p. 139) in two-year college writing placement. We will need all of the critical tools emerging from the field of writing assessment to reform these processes in ways that advance opportunity and social justice.

PLACEMENT REFORM AS TEACHER-SCHOLAR-ACTIVISM

In a recent special issue of *Teaching English in the Two-Year College* on academic freedom, Warnke and Higgins (2018) called on two-year college English faculty to become *critical reformers*. Critical reformers remain clear-eyed about the dangers of the neoliberal agenda that motivates some higher education reformers but take seriously the evidence higher education researchers have produced that

business-as-usual at community colleges is producing harmful and unjust inequities. "As critical reformers," Warnke and Higgins (2018) asserted:

> We are tasked with linking what we know empirically with our values and vision for the community college. When interests converge, we are responsible for reframing and reimagining ostensibly apolitical reform research . . . When our interests overlap partially with those in power, we may stand a chance of achieving progress through careful, structurally aware engagement. (p. 368)

We hope this special issue serves as a resource for two-year college faculty engaging in critical placement reform. We also hope it encourages the university-based writing assessment community to support two-year faculty in their critical reform efforts. Hassel and Giordano (2013) have called on writing studies to produce more scholarship that accounts for and responds to the needs of what they call the field's "teaching majority." Likewise, Toth (2018b) has urged writing assessment scholars to attend to the professional positioning of two-year college English faculty and produce scholarship these faculty can use to influence policy and practice at their institutions. Thus, this special issue both responds to and amplifies calls for cross-sector disciplinary "alliance" in our current era of reform (Toth, Sullivan, & Calhoon-Dillahunt, 2019).

Over the last five years, as reform pressures have spurred rapid change at community colleges across the country, two-year college writing studies has been undergoing a turn toward what Andelora (2013) and Sullivan (2015) have called *teacher-scholar-activism*. This turn is premised on "a conception of professional identity that foregrounds faculty members' responsibility to be public intellectuals and agents of change" (Toth et al., 2017, p. 31). Indeed, Warnke and Higgins (2018) explicitly situated their discussion of critical reform within the teacher-scholar-activist turn. TYCA, an organization within the National Council of Teachers of English (NCTE) with close ties to CCCC, is one important organization cultivating a professional community that fosters teacher-scholar-activism. Over the last decade, TYCA has become increasing engaged with policy-making and providing two-year college faculty with resources to be critical agents of change amid ongoing reform (Calhoon-Dillahunt, 2015; Toth et al., 2016).

These resources include two influential TYCA white papers (Hassel et al., 2015; Klausman et al., 2016) that have stated the importance of two-year college English instructors asserting disciplinarily-grounded professional authority in institutional processes that are increasingly being regulated by legislative mandates. Of particular relevance is the 2016 "White Paper on Writing Placement Reform" (Klausman et al., 2016), a document intended to provide an overview of the existing writing placement practices used across two-year colleges and to inform readers about the

disciplinary, professional, and political movements reshaping those practices. The white paper presents a synthesis of research on placement emerging from higher education reformers at the CCRC, as well as from writing studies.

The statement offers case studies of two promising approaches to two-year college writing placement—multiple measures and directed self-placement (DSP)—and articulates several key principles for designing, administering, and assessing placement practices. Those principles include (1) grounding in disciplinary knowledge, (2) involvement of English faculty in the development of placement processes, (3) sensitivity to the effects of placement processes on diverse groups of students, (4) ongoing local validation, and (5) integration of placement reform with other campus-wide efforts to support student success (Klausman et al., 2016, p. 126). The influence of this white paper is evident in several of the articles in this special issue of *JWA*.

These articles present critical discussions of a range of issues and options for two-year college writing placement in an era of reform. The first two pieces focus on theoretical issues. In "Beyond Tradition: Fairness, Placement, and Success at a Two-Year College," Nastal uses archival institutional data to interrogate long-standing approaches to writing placement at two-year colleges through emerging ethical conceptions of fairness. Based on evidence of racial disparities in her survival analysis of student persistence through her college's developmental writing course sequence, she critiques inherited placement traditions and calls for practices that better align with the commitments to democratic access that two-year colleges espouse. In "Are We Who We Claim to Be? A Case Study of Language Policy in Community College Writing Placement Practices," Gilman extends this line of critique through an examination of the tacit language policies embedded in her community college's placement assessment, policies that contradict the institution's stated commitment to diversity. Gilman calls for greater attention to the underlying language ideologies that drive two-year college writing placement.

The next two articles examine placement options in two-year colleges. In "Let Them In: Increasing Access, Completion, and Equity in English Placement Policies at a Two-Year College in California," Henson and Hern present a disparate impact analysis evaluating the effect of lowering placement cut scores on a purchased multiple-choice usage test at Henson's institution. They found strong evidence that the higher cut scores resulted in significant and inequitable underplacement that reduced the likelihood of persistence to degree completion for students of color. Based on these findings, they advocate for multiple measures placement that enables as many students as possible to enroll directly into credit-bearing college composition courses. Next, in "Directed-Self Placement in Two-Year Colleges: A Kairotic Moment," Toth presents findings from an interview-based study of DSP implementation in 12 two-year colleges, demonstrating that there is a

more extensive track record for DSP in open admissions settings than the scholarly literature has suggested. She finds that DSP offers a promising alternative to mandatory placement at two-year colleges, but that it also presents distinctive considerations for implementation that warrant deeper theorization and further research.

The special issue concludes with a collaboratively authored "Forum" that discusses how contributors see the special issue affirming, extending, and/or complicating the principles articulated in the 2016 TYCA white paper. This polyvocal conversation surfaces shared convictions as well as points of contention and unresolved questions that suggest areas for future activism, policy-making, and research. Informed by the articles in this special issue and critical questions raised in the body of the forum, Elliot, Poe, and Nastal offer a roadmap for two-year college placement reform that synthesizes the principles of the TYCA white paper with additional theoretical insights from the writing assessment and educational measurement literature. This document is designed to help facilitate local conversations about placement reform among faculty, administrators, and other stakeholders at two-year colleges. With this critical and theoretically-grounded yet practical resource for making institutional change, Elliot et al. offer a milestone example of cross-sector alliance in writing assessment that helps equip two-year college English faculty to assert professional authority in local policy decisions.

Taken together, the pieces in this special issue model the kind of critical reformer role that two-year college faculty can take on. We believe faculty at open admissions institutions need to be participants in conversations about writing assessment and social justice, and these articles demonstrate that two-year college faculty have much to offer those discussions. In addition to contributing to disciplinary knowledge, their efforts can provide colleagues at other two-year colleges with valuable insight and precedent for pursuing reform at their own institutions. Finally, these articles suggest that cross-sector scholarly alliances can strengthen our collective efforts to pursue more equitable approaches to writing assessment: approaches that honor open admissions students and the rhetorical resources of our communities. In sum, we hope this special issue persuades readers at *all* institution types that two-year colleges are important sites for making knowledge about writing assessment and for putting that knowledge to work as social justice-oriented praxis.

REFERENCES

Adams, P. D. (1993). Basic writing reconsidered. *Journal of Basic Writing*, *12*(1), 22–36.

Adams, P. D., Gearhart, S., Miller, R., & Roberts, A. (2009). The Accelerated learning program: Throwing open the gates. *Journal of Basic Writing*, *28*(2), 50–69.

Alexander, J. (2016). Queered writing assessment. *College English*, *79*(2), 202–205.

American Association of Community Colleges. (2017). *2017 FactsSheet*. American Association of Community Colleges. http://www.aacc.nche.edu/AboutCC/Pages/fastfactsfactsheet.aspx

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological measurement*. American Educational Research Association.

Andelora, J. (2013). Teacher/scholar/activist: A response to Keith Kroll's "The end of the community college English profession." *Teaching English in the Two-Year College*, *40*(3), 302–307.

Bailey, T. (2009). Challenge and opportunity: Rethinking the role and function of developmental education in community college. *New Directions for Community Colleges*, *145*, 11–30.

Bailey, T., & Cho, S. W. (2010). *Developmental education in community colleges* (Issue Brief Prepared for the White House Summit on Community Colleges). Teachers College, Columbia University.

Bailey, T., Jaggars, S. S., & Scott-Clayton, J. (2013). Commentary: Characterizing the effectiveness of developmental education: A response to recent criticism. *Journal of Developmental Education*, *36*(3), 18–34.

Bailey, T., Jeong, D. W., & Cho, S. W. (2010). Referral, enrollment, and completion in developmental education sequences in community colleges. *Economics of Education Review*, *29*(2), 255–270.

Barnett, E. A., & Reddy, V. (2017). *College placement strategies: Evolving considerations and practices* (CAPR Working Paper). Columbia University.

Beach, J. M. B. (2012). *Gateway to opportunity?: A history of the community college in the United States*. Stylus Publishing.

Behm, N., & Miller, K. D. (2012). Challenging the frameworks of color-blind racism: Why we need a fourth wave of writing assessment scholarship. In A. B. Inoue & M. Poe (Eds.), *Race and writing assessment* (pp. 127–140). Peter Lang.

Belfield, C., & Crosta, P. M. (2012). *Predicting success in college: The importance of placement tests and high school transcripts* (CCRC Working Paper No. 42). Columbia University.

Bereiter, C. (2003). Foreword. In M. D. Shermis & J. C. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. vii–x). Lawrence Erlbaum.

Bonilla-Silva, E. (2006). *Racism without racists: Color-blind racism and the persistence of racial inequality in the United States*. Rowman & Littlefield Publishers.

Burns, M., Cream, R., & Dougherty, T. R. (2018). Fired Up: Institutional critique, lesson study, and the future of antiracist writing assessment. In M. Poe, A. B. Inoue, & N. Elliot (Eds.), *Writing assessment, social justice, and the advancement of opportunity* (pp. 257-292). The WAC Clearinghouse; University Press of Colorado. https://doi.org/10.37514/PER-B.2018.0155.2.08

Burstein, J. (2012). Fostering best practices in writing instruction and assessment with E-rater®. In N. Elliot & L. Perelman (Eds.), *Writing assessment in the 21st century: Essays in honor of Edward M. White* (pp. 203–217). Hampton Press.

Calhoon-Dillahunt, C. (2015). "We've come a long way, baby": A former TYCA chair looks back. *Teaching English in the Two-Year College*, *42*(4), 352–358.

Carey, K. (2017). Revised data shows community colleges have been underappreciated. *New York Times*. https://www.nytimes.com/2017/10/31/upshot/revised-data-shows-community-colleges-have-been-underappreciated.html

Casazza, M. E., & Silverman, S. L. (2013). *Meaningful access and support: The path to college completion*. Council of Learning Assistance and Developmental Education Association. http://www.cladea.net/white_paper_meaningful_access.pdf

Caswell, N. I. (2018). Queering assessment: Fairness, affect, and the impact on LGBTQ writers. In M. Poe, A. B. Inoue, & N. Elliot (Eds.), *Writing assessment, social justice, and the advancement of opportunity* (pp. 353-378). The WAC Clearinghouse; University Press of Colorado. https://doi.org/10.37514/PER-B.2018.0155.2.11

Chen, X. (2016). *Remedial course taking at U.S. public 2- and 4-year institutions: Scope, experiences, and outcomes* (U.S. Department of Education No. NCES 2016-405). National Center of Education Statistics. http://nces.ed.gov/pubsearch

Cho, S.-W., Kopko, E., Jenkins, D., & Jaggars, S. S. (2012). *New evidence of success for community college remedial English students: Tracking the outcomes of students in the Accelerated Learning Program* (CCRC Working Paper No. 53). Community College Research Center, Columbia University.

Cohen, A. M., Brawer, F. B., & Kisker, C. B. (2014). *The American community college* (6th ed.). John Wiley & Sons.

Conference on College Composition and Communication Executive Committee. (2009). *Writing assessment: A position statement*. http://www.ncte.org/cccc/resources/positions/writingassessment

Council of Learning Assistance and Developmental Education Associations. (n.d.). *College access* (Policy Statement). Council of Learning Assistance and Developmental Education Associations.

Council of Writing Program Administrators. (2014). *WPA outcomes statement for first-year composition*. http://wpacouncil.org/node/4846

Cushman, E. (2016). Decolonizing validity. *Journal of Writing Assessment*, *9*(1). https://escholarship.org/uc/item/0xh7v6fb

Elliot, N. (2015). Validation: The pursuit. *College Composition and Communication*, *66*(4), 668–687.

Elliot, N. (2016). A theory of ethics for writing assessment. *Journal of Writing Assessment*, *9*(1). https://escholarship.org/uc/item/36t565mm

Faigley, L., Cherry, R., Jolliffe, D., & Skinner, A. (1985). *Assessing writers' knowledge and processes of composing*. Ablex Publishing Corporation.

Gallagher, C. W. (2007). *Reclaiming assessment: A better alternative to the accountability agenda*. Heinemann Educational Books.

Gomes, M. (2018). Writing assessment and responsibility for colonialism. In M. Poe, A. B. Inoue, & N. Elliot (Eds.), *Writing assessment, social justice, and the advancement of opportunity* (pp. 201-226). The WAC Clearinghouse; University Press of Colorado. https://doi.org/10.37514/PER-B.2018.0155.2.06

Goudas, A. M., & Boylan, H. R. (2012). Addressing flawed research in developmental education. *Journal of Developmental Education*, *36*(1), 2–13.

Goudas, A. M., & Boylan, H. R. (2013). A brief response to Bailey, Jaggars, and Scott-Clayton. *Journal of Developmental Education*, *36*(3), 28–32.

Griffiths, B. (2017). Professional autonomy and teacher-scholar-activists in two-year colleges: Preparing new faculty to think institutionally. *Teaching English in the Two-Year College*, *45*(1), 47-68.

Griffiths, B., & Toth, C. (2017). Rethinking "class": Poverty, pedagogy, and two-year college writing programs. In W. Thelin & G. Carter (Eds.), *Class in the composition classroom: Pedagogy and the working class* (pp. 231-257). Utah State University Press.

Hammond, J. W. (2018). Toward a social justice historiography for writing assessment. In M. Poe, A. B. Inoue, & N. Elliot (Eds.), *Writing assessment, social justice, and the advancement of opportunity* (pp. 41-70). The WAC Clearinghouse; University Press of Colorado. https://doi.org/10.37514/PER-B.2018.0155

Harrington, S. (2005). Learning to ride the waves: Making decisions about placement testing. *WPA: Writing Program Administration*, *28*(3), 9–29.

Hassel, H., & Giordano, J. B. (2011). First-year composition placement at open-admission, two-year campuses: Changing campus culture, institutional practice, and student success. *Open Words: Access and English Studies*, *5*(2), 29–39.

Hassel, H., & Giordano, J. B. (2013). Occupy writing studies: Rethinking college composition for the needs of the teaching majority. *College Composition and Communication*, *65*(1), 117–139.

Hassel, H., & Giordano, J. B. (2015). The blurry borders of college writing: Remediation and the assessment of student readiness. *College English*, *78*(1), 56–80.

Hassel, H., Klausman, J., Giordano, J. B., O'Rourke, M., Roberts, L., Sullivan, P., & Toth, C. (2015). TYCA white paper on developmental education reforms. *Teaching English in the Two-Year College*, *42*(3), 227–243.

Haswell, R. (2004). Post-secondary entrance writing placement: A brief synopsis of research. *CompPile.Org*. http://comppile.org/profresources/writingplacementresearch.htm

Haswell, R. (2005). Post-secondary entrance writing placement. *CompPile.org*. http://comppile.org/profresources/placement.htm

Henson, H. & Hern, K. (2019). Let them in: Increasing access, completion, and equity in English placement policies at a two-year college in California. *Journal of Writing Assessment, 12*(1). https://escholarship.org/uc/item/3nh6v5d0

Hern, K. (2012). Acceleration across California: Shorter pathways in developmental English and math. *Change: The Magazine of Higher Learning*, *44*(3), 60–68.

Herrington, A., & Moran, C. (2001). What happens when machines read our students' writing? *College English*, *63*(4), 480–499.

Herrington, A., & Moran, C. (2012). Writing to a machine is not writing at all. In N. Elliot & L. Perelman (Eds.), *Writing assessment in the 21st century: Essays in honor of Edward M. White* (pp. 219–232). Hampton Press.

Hodara, M., Jaggars, S. S., & Karp, M. M. (2012). *Improving developmental education assessment and placement: Lessons from community colleges across the country* (CCRC Working Paper No. 51). Community College Research Center, Columbia University.

Huddleston, E. M. (1954). Measurement of writing ability at the college entrance level: Objective vs. subjective testing techniques. *Journal of Experimental Education*, *22(3)*, 165–213.

Hughes, K. L., & Scott-Clayton, J. E. (2011). Assessing developmental assessment in community colleges. *Community College Review*, *39*(4), 327–351.

Inoue, A. B. (2009a, Summer). Self-assessment as programmatic center: The first-year writing program and its assessment at California State University, Fresno. *Composition Forum*, *20*.

Inoue, A. B. (2009b). The technology of writing assessment and racial validity. In C. Schreiner (Ed.), *Handbook of research on assessment technologies, methods, and applications in higher education* (pp. 97–120). Information Science Reference.

Inoue, A. B. (2012). Racial methodologies for composition studies: Reflecting on theories of race in writing assessment research. In L. Nickoson & M. P. Sheridan (Eds.), *Writing studies research in practice: Methods and methodologies* (pp. 125–139). Southern Illinois University Press.

Inoue, A. B. (2015). *Antiracist writing assessment ecologies: Teaching and assessing writing for a socially just future*. Parlor Press.

Inoue, A. B., & Poe, M. (2012a). *Race and writing Assessment. Studies in composition and rhetoric.* Peter Lang.

Inoue, A. B., & Poe, M. (2012b). Racial formations in two writing assessments: Revisiting White and Thomas's findings on the English Placement Test after 30 years. In N. Elliot & L. Perelman (Eds.), *Writing assessment in the 21st century: Essays in honor of Edward M. White* (pp. 341–359). Hampton Press.

Jaggars, S. S., & Stacey, G. W. (2014). *What we know about developmental education outcomes. Research Overview*. Community College Research Center, Teachers College, Columbia University.

Kane, M. T. (1990). *An argument-based approach to validation* (ACT Research Report Series No. 90–13). American College Testing Program.

Kelly-Riley, D., & Whithaus, C. (2016). Introduction to a special issue on a theory of ethics for writing assessment. *Journal of Writing Assessment*, *9*(1). https://escholarship.org/uc/item/8nq5w3t0

Klausman, J., Toth, C., Swyt, W., Griffiths, B., Sullivan, P., Warnke, A., . . . Roberts, L. (2016). TYCA white paper on placement reform. *Teaching English in the Two-Year College*, *44*(2), 135–157.

Lovas, J. C. (2002). All good writing develops at the edge of risk. *College Composition and Communication*, *54*(2), 264–288.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher*, *18*(2), 5–11.

Morante, E. A. (1987). A primer on placement testing. *New Directions for Community Colleges*, *1987*(59), 55–63.

Moreland, C. (2018). Chasing transparency: Using disparate impact analysis to assess the (in)accessibility of dual enrollment composition. In M. Poe, A. B. Inoue, & N. Elliott (Eds.), *Writing assessment, social justice, and the advancement of opportunity* (pp. 171-200). The WAC Clearinghouse; University Press of Colorado. https://doi.org/10.37514/PER-B.2018.0155.2.05

Morris, W., Greve, C., Knowles, E., & Huot, B. (2015). An analysis of writing assessment books published before and after the year 2000. *Teaching English in the Two-Year College*, *43*(2), 118–140.

Nastal, J. (2019). Beyond tradition: Writing Placement, Fairness, and Success at a Two-Year College. *Journal of Writing Assessment, 12*(1). https://escholarship.org/uc/item/4wg8w0ng

Naynaha, S. (2016). Assessment, social justice, and Latinxs in the US community college. *College English*, *79*(2), 196–201.

Nist, E. A., & Raines, H. H. (1995). Two-year colleges: Explaining and claiming our majority. In J. Janangelo & K. Hansen (Eds.), *Resituating writing: Constructing and administering writing programs* (pp. 59–70). Boynton/Cook.

Patthey-Chavez, G. G., Dillon, P. H., & Thomas-Spiegel, J. (2005). How far do they get? Tracking students with different academic literacies through community college remediation. *Teaching English in the Two-Year College*, *32*(3), 261–277.

Perelman, L. (2012). Mass-market writing assessments as bullshit. In N. Elliot & L. Perelman (Eds.), *Writing assessment in the 21st century: Essays in honor of Edward M. White* (pp. 425–438). Hampton Press.

Poe, M., & Cogan, J. A. (2016). Civil rights and writing assessment: Using the disparate impact approach as a fairness methodology to evaluate social impact. *Journal of Writing Assessment*, *9*(1). https://escholarship.org/uc/item/08f1c307

Poe, M., Elliot, N., Cogan, J. A., & Nurudeen, T. G. (2014). The legal and the local: Using disparate impact analysis to understand the consequences of writing assessment. *College Composition and Communication*, *65*(4), 588–611.

Poe, M., & Inoue, A. B. (2016). Toward writing as social justice: An idea whose time has come. *College English*, *79*(2), 119–126.

Scott-Clayton, J. (2012). *Do high-stakes placement exams predict college success?* (Working Paper No. 41). Columbia University.

Scott-Clayton, J., & Belfield, C. (2015). *Improving the accuracy of remedial placement* (Research Overview). Columbia University.

Scott-Clayton, J., Crosta, P. M., & Belfield, C. R. (2014). Improving the targeting of treatment: Evidence from college remediation. *Educational Evaluation and Policy Analysis*, *36*(3), 371–393.

Shapiro, D., Dundar, A., Huie, F., Wakhungu, P., Yuan, X., Nathan, A., & Hwang, Y. (2017). *Completing college: A national view of student attainment rates by race and ethnicity- Fall 2010 cohort* (Signature Report No. 12b). National Student Clearinghouse Research Center.

Shapiro, D., Dundar, A., Wakhungu, P., Yuan, X., Nathan, A., & Hwang, Y. (2016). *Completing college: A national view of student attainment rates- Fall 2010 cohort* (Signature Report No. 12). National Student Clearinghouse Research Center.

Slomp, D. (2016a). An integrated design and appraisal framework for ethical writing assessment. *Journal of Writing Assessment*, *9*(1). https://escholarship.org/uc/item/4bg9003k

Slomp, D. (2016b). Ethical considerations and writing assessment. *Journal of Writing Assessment*, *9*(1). https://escholarship.org/uc/item/2k14r1zg

Smith, W. L. (1993). Assessing the reliability and adequacy of using holistic scoring of essays as a college composition placement technique. In M. M. Williamson & B. A. Huot (Eds.), *Validating holistic scoring for writing assessment: Theoretical and empirical foundations* (pp. 142–205). Hampton Press.

Stein, Z. (2016). *Social justice and educational measurement: John Rawls, the history of testing, and the future of education*. Routledge.

Sullivan, P. (2008). Measuring "success" at open admissions institutions: Thinking carefully about this complex question. *College English*, *70*(6), 618–632.

Sullivan, P. (2015). The two-year college teacher-scholar-activist. *Teaching English in the Two-Year College*, *42*(4), 327–50.

Sullivan, P. (2017). *Economic inequality, neoliberalism, and the American community college.* Palgrave MacMillan.

Toth, C. (2018a). Directed self-placement at "democracy's open door": Writing placement and social justice in community colleges. In A. B. Inoue, M. Poe, & N. Elliot (Eds.), *Writing assessment, social justice, the advancement of opportunity* (pp. 137-170). The WAC Clearinghouse; University Press of Colorado. https://doi.org/10.37514/PER-B.2018.0155.2.04

Toth, C. (2018b). Toward writing assessment that accounts for (and to) community colleges. In M. Poe, A. B. Inoue, & N. Elliot (Eds.), *Writing assessment, social justice, and the advancement of opportunity* (pp. 395-398). The WAC Clearinghouse; University Press of Colorado. https://doi.org/10.37514/PER-B.2018.0155.2.12

Toth, C., Calhoon-Dillahunt, C., & Sullivan, P. (2016). A dubious method of improving educational outcomes: Accountability and the two-year college. *Teaching English in the Two-Year College*, *43*(4), 391–410.

Toth, C., Sullivan, P., & Calhoon-Dillahunt, C. (2019). Two-year college teacher-scholar-activism: Reconstructing the disciplinary matrix of writing studies. *College Composition and Communication, 71*(1), 86-116.

Toth, C., Griffiths, B., & Thirolf, K. (2013). "Distinct and significant": Professional identities of two-year college English faculty. *College Composition and Communication*, *65*(1), 90–116.

Toth, C., Jensen, D., Reynolds, M., Suh, E., Spiegel, C. L., Blaauw-Hara, M., & Andelora, J. (2017). Symposium: Guidelines for preparing teachers of English in the two-year college. *Teaching English in the Two-Year College*, *45*(2), 29-46.

Toth, C., & Sullivan, P. (2016). Toward local teacher-scholar communities of practice: Findings from a national TYCA survey. *Teaching English in the Two-Year College*, *43*(3), 247–273.

Warnke, A., & Higgins, K. (2018). A critical time for reform: Empowering interventions in a precarious landscape. *Teaching English in the Two-Year College*, *45*(4), 361-384.

White, E. M. (1995). The importance of placement and basic studies: Helping students succeed under the new elitism. *Journal of Basic Writing*, *14*(2), 75–84.

Williamson, M. (1994). The worship of efficiency: Untangling theoretical and practical considerations in writing assessment. *Assessing Writing*, *1*(2), 147–173.

Willingham, W. W. (1974). *College placement and exemption*. College Entrance Examination Board.

Yancey, K. B. (1999). Looking back as we look forward: Historicizing writing assessment. *College Composition and Communication*, *50*(3), 483–503.

Zaback, K., Carlson, A., Laderman, S., & Mann, S. (2016). *Serving the equity imperative: Intentional action toward greater student success*. State Higher Education Executive Offices Association. http://www.sheeo.org/sites/default/files/2016_SHEEO_CCA_ServingEquityImperative.pdf

PART 3.

IMPLICATIONS OF AUTOMATED SCORING OF WRITING

RETROSPECTIVE.

IMPLICATIONS OF AUTOMATED SCORING OF WRITING

Laura Aull
University of Michigan

Popular notions of automated scoring are often oversimplified, and grim. They bring to mind product-oriented treatments of writing. They summon images of machines replacing teachers. They conjure inhumane outsourcing—an un/necessary evil, the equivalent of getting a robot on the phone who cannot understand you and leaves you desperately annunciating, *operator!*

The implications of being misunderstood are far more serious than an unsuccessful phone call, of course. Scoring algorithms are unable to parse some students' creative ideas because they are in language deemed nonstandardized by schools that reinscribe prescriptive and oppressive histories (Hammond, 2019; Perryman-Clark, 2013). Automated scoring that is most able to focus on machine-readable text does not focus on 'languaging': the mental processes of meaning-making that surround the text produced (Ivanič, 2004). In both of these examples, automated scoring belies the practices and principles most of us support in our writing courses.

Yet we cannot ignore automated scoring any more than we can ignore any widespread writing assessment approach. It is part of student writing today, and it touches on all the major themes in this collection: technical issues; evolving ideas about writing; teachers' and students' lived experiences; policy; and embedded concerns about reliability, validity, bias, and fairness. Automated scoring requires our engagement, even as it can be hard to know what to think, between media representations of automated scores, outsourced automated tools, high stakes for students and instructors, and the varied demands on writing students, educators, and administrators to use automated scoring tools.

This essay strives to offer an overall look at automated scoring in writing assessment over the past two decades, particularly how it constructs student writing and writers and how writing educators might engage with it. To do so, I draw on three articles from the *Journal of Writing Assessment* that illustrate writing educators' critical engagement with automated scoring:

1. Validity of Automated Scoring: Prologue for a Continuing Discussion of Machine Scoring Student Writing by Michael Williamson (2003)

DOI: https://doi.org/10.37514/PER-B.2024.2166.2.23

2. Critique of Mark D. Shermis & Ben Hamner, "Contrasting State-of-the-Art Automated Scoring of Essays: Analysis" by Les C. Perelman (2013)
3. Globalizing Plagiarism & Writing Assessment: A Case Study of Turnitin by Jordan Canzonetta and Vani Kannan (2016)

As the titles suggest, the articles have different perches and points of entry vis-à-vis automated scoring. Williamson describes two stakeholder groups implicated in automated scoring, writing educators and measurement professionals, in order to explore each group's perspective and the dangers of keeping them separate. Perelman offers a data-driven appraisal of automated scoring by critiquing a foundation-funded study's unfounded claims—in turn, offering an illustration of the important possibilities of the cross-pollination called for by Williamson. Canzonetta and Kannan write about *Turnitin.com*, a plagiarism detection software, as it contends to move into global formative assessment, bringing with it particular ideologies.

Together, the articles help us consider five questions as they have been answered over time:

1. What is automated scoring?
2. What does automated scoring do?
3. What is the role of automated scoring (or, are humans good at scoring)?
4. What responsibility do writing educators have vis-à-vis automated scoring?
5. What might the future of automated scoring of writing look like?

Considering the three articles in light of these questions allows us to interrogate automated scoring, its implications, and its future possibilities for student-centered assessment.

WHAT IS AUTOMATED SCORING?

While each article focuses on various components of automated scoring, together they illustrate its expanding scope over the past two decades. In 2004, Williamson defined automated scoring as "the use of computer algorithms to simulate holistic ratings of student writing" (p. 86).[i] In this definition, Williamson's pairing of algorithms and holistic scoring, if obvious, is interesting: computer algorithms measure discrete, direct features—say, the number of words per sentence, or the frequency of certain content words—and holistic ratings are overall arguments about a piece of writing (e.g., a rating of whether the writing is successful in light of the task). Here, I use *arguments* in Kane's (2013) sense: a score is an interpretive argument about the writing. In the

case of an automated score, a holistic rating for a full piece of writing is an algorithm-based interpretive argument about how the discrete features come together in a single score. The evidence for the argument includes the features the algorithm is designed to measure.

Depending on the target writing feature(s), algorithms can use a formula consistently, more easily than humans (for instance, a computer algorithm can measure lexical sophistication, per use of rare and varied content-related words, consistently), a point to which we'll return below. As these three articles suggest, the key question for automated scoring centers on what algorithms can actually identify and evaluate, and to what end. In other words, "the question is whether the task itself is computable" (Williamson, 2003, p. 96)—the extent to which writing can be analyzed and evaluated by machines.

On this, the articles show consensus overall: automated scoring cannot evaluate the full complexities of writing as situated rhetorical action, though it poses important opportunities for writing educators and assessment. Williamson argues that an automated score cannot replace a human rater score, but automated scoring can augment the teaching of writing. Perelman shows the risks involved when automated scoring tools are "judged like the answer to a math problem or GPS directions" (para. 6); he also argues that automated scoring demands rigorous statistical analysis and offers important information. Canzonetta and Kannan likewise note some limitations of automated scoring tools and call for critical engagement, particularly with the "global cultural work" of plagiarism detection tools.

As we can see, the three articles imply a broader definition of automated scoring, as the use of algorithms to evaluate various aspects of writing, whether or not the expectation is to simulate human scorers. For those cases in which an automated scoring tool alone is used to determine a writing score (e.g., ACCUPLACER, WritePlacer, and WritePlacer ESL), Williamson's early definition still applies. In all cases, we can think of automated scoring as an approach that uses algorithm-based evaluations of writing as evidence for an interpretive argument about said writing.

WHAT DOES AUTOMATED SCORING DO?

Embedded in our definition above is that automated scoring algorithms determine what counts, and doesn't count, in a piece of writing. An overall score based on a given algorithm draws inferences from those aspects determined to count. In so doing, any given automated tool constructs writing, assessment, and writers in particular ways.

Constructing Writing

Computer algorithms are written by humans and carried out by computers; they are limited to (and enhanced by, depending on your perspective) what computers can do. Any given automated scoring algorithm implies what matters most, constructing writing according to what it measures, such as mechanical choices, length (Perelman), text-matching, and/ or standardized written academic English (Canzonetta and Kannan). In turn, it implies that certain aspects of writing *don't count*—the things not measured, if the automated score is the only score used.

All the articles emphasize that like any kind of writing, academic writing is an integration of many processes, and this is part of concerns about automated scoring. Using the same automated scoring tool on each one—or comparing automated scores across them—belies the situated rhetorical action entailed in each one. The articles underscore that automated scoring does not construct writing as situated language use: it cannot account for writing as rhetorical action (Perelman), writing as situated literacy (Williamson), and writing and source use as culturally-specific action (Canzonetta & Kannan). The use of different automated tools, in turn, emphasizes different conceptions of assessment validity, which is also always situated (Williamson, 2003).

Constructing writing through design and use of automated scoring can also point to the lack of a clear writing construct. This is a problem Perelman delineates in his critique of the Shermis and Hamner study: "Without [any explicit construct of *writing*], it is, of course, impossible to judge the validity of any measurement" (para. 6). An illustration Perelman offers is the use of multiple constructed response tasks compared in the same way, e.g., some that require understanding and incorporation of included reading texts, and some that do not.

Constructing Assessment

Williamson traces ideas about validity with attention to the situated nature of literacy. Williamson's attention to the "complexity of validity inquiry" (p. 259) illuminates differences in conceptions of assessment "between English Studies and educational measurement, the difference between social science and humanistic disciplines" (p. 260). To date in 2004, Williamson argued, many researchers in English Studies subscribed to "an older notion of validity," thereby "unwittingly missing" more contextualized, less rigid conceptions of validity for writing assessment (p. 262). In other words, early questions about validity focused on a given test, and whether it measured what it purported to measure. In those cases, assessment is constructed as a process of consistent measurement,

regardless, for instance, of the validity of the writing construct, or the particular abilities emphasized in an assessment task.

Assessment tools construct writing assessment through what they do *not* evaluate as well. Deane et al., all ETS researchers, (2013) show that "[automated scoring] systems do not explicitly evaluate the validity of reasoning, the strength of evidence, or the accuracy of information" (para. 2). They illustrate how this poses a risk because it can mean a disconnect between what assessors value and what an automated tool can measure. In such cases, the scoring makes an interpretive argument about a piece of writing based only on partial evidence about the piece of writing.

Perelman shows that the *scoring* part of automated scoring is important not only in what it interprets but also in how it is represented. His article reviews a study that represents its findings as though they suggest that automated scorers are as accurate as human raters. Yet as Perelman shows, the automated scores were rounded to integer values in ways that favored the automated scores. Perelman writers, "Essays scores, be they holistic, trait, or analytical, always are continuous variables, not discrete variables (integers), even though graders almost always have to give integer values as scores" (para. 14). Interrogating automated scoring and representations thereof have high stakes for how assessment is constructed vis-à-vis policy decisions. The paper Perelman critiques, for instance, was sent to the Partnership for Assessment of Readiness of College and Careers and the Smarter Balanced Assessment Consortium. Thus Perelman argues that automated scoring demands rigorous statistical analysis and offers important information in light of potential policy decisions.

Constructing Writers and Assessors

While the first three articles highlight how automated scoring constructs writing and its measurement, Canzonetta and Kannan's article more specifically illuminates how writers are constructed by automated scoring. They question, "How is the student plagiarist being discursively constructed? What are the implications of these constructions as Turnitin rolls out its assessment platform?" (p. 298). Their answers to these questions show how Turnitin.com's strategies construct those who assess writing, too.

Canzonetta and Kannan discuss "three primary rhetorical strategies for advancing Turnitin" that construct writers and assessors in culturally-specific, hierarchical ways. They include: (1) plagiarism detection as social improvement that forms modernized, idealized, western students; (2) plagiarism as a national concern with ramifications for citizenship, economy, and character; and (3) plagiarism detection as requiring standardized, western policies through private/

public partnerships (para. 11). These strategies construct writing as rule mastery, writers as needing to be regulated, and instructor-assessors as regulators.

In sum, the articles remind us that vis-à-vis the question *what does automated scoring do?*, we can answer: it constructs writing, writers, and assessors in particular ways. As a writing assessment tool, any automated scoring tool is constitutive; all assessment activities are complex and value-laden activities, whether baked into an algorithm or required of a human (Aull, 2017). Thus the authors remind us that, as is any writing assessment approach, automated scoring is an opportunity to see what we prioritize in writers and writing—a chance to understand and question what that is and what's left out.

WHAT IS THE ROLE OF AUTOMATED SCORING (OR, ARE HUMANS GOOD AT SCORING?)

Underpinning questions about automated and human scoring is the fundamental question *what is "good" scoring?* In early writing assessment research especially, good scoring was often described in terms of reliability, in other words, consistent scoring. It was also often described as scoring with agreement between scores (human, automated, or both). We will consider agreement and reliability below, but a broader answer is that even pondering "good" automated and human scoring entails more questions posed by the three articles: questions about the widespread use of automated scoring and about the strengths and limitations of both machine and human scoring.

The Inevitability of Automated Scoring

"Two things are certain," writes Williamson. "One, automated scoring programs can replicate scores for a particular reading of student writing, and this technology is reliable, efficient, fast, and cheap. Two, automated scoring has been and will continue to be used in various large-scale assessments of student writing" (p. 256). Canzonetta and Kannan describe how Turnitin.com frames automation in similar ways, emphasizing "efficiency, adaptability, and individual choice" (p. 305).

We cannot escape from this answer: automated scoring is widespread. It is used in large part because it is an inexpensive way to assess student writing, which is otherwise labor- and time-intensive. That, and a longer history of testing dating back to the early 20th century made writing a skillset for ranking individuals against one another in a single test (Elliot, 2005; Hammond, 2019; Aull, 2024). At the same time, the three *Journal of Writing Assessment* articles help frame automated scoring not only as a large-scale inevitability, but also as a reminder that we should not accept any assessment approach without ongoing questions about its limitations.

Automated Scoring Limitations, Human Strengths

Computers cannot interpret or infer; they can only analyze the observable components of a text, which are based on the preferences and designs of the developers or the graders whose models were used. Thus there are always limitations—in the basic example above, essay length is easy for automated scoring tools to detect, but it is not always a sign of valued written detail. Perelman also reminds us that automated scoring tools can be "gamed"; for instance, writers can use conventionally prioritized academic writing choices, including cohesive ties ("however,") and large words (lexical complexity). They can likewise be gamed by *avoiding* choices conventionally devalued in standardized usage preferences, including "sentence fragments" and sentences beginning with conjunctions like *and*. None of these choices necessarily have to do with genre-specific idea development. But they could lead to a positive automated score because automated tools can directly measure these features. Deane et al. (2003) describe how automated scoring systems rely "on measures of such things as the structure and elaboration of student essays, the sophistication of vocabulary, or the number of errors in grammar, usage, mechanics or style" (para. 2), defined according to the scoring system designers.

By contrast, automated tools cannot look at writing as a situated act, contingent on audience, genre, and purpose of a specific rhetorical situation. In the case of automated plagiarism detection, Canzonetta and Kannan show, *Turnitin.com* developers talk about writing in a way that "displaces composition and rhetoric's arguments for situated pedagogical approaches" (p. 305). Perelman shows how automated tool design can elide common expectations for writing such as coherence, e.g., by not accounting for paragraph breaks.

Alternatively, human raters are capable of assessing writing tasks as situated actions. Deane et al. (2013) show that human raters can attend to genre-specific critical thinking on complex writing tasks involving source text integration in ways that automated scoring cannot. In the case of exams purportedly measuring students' preparation for college writing, construct validity, or the relevance of the constructed response task, depends on having human raters. Human raters can consider how writers have engaged with source text use, a common expectation of college writing.

Indeed, human scorers can read for audience-specific choices, idea development with or without explicit transitional phrases, and integration of others' ideas. These practices support today's sociocultural conceptualizations of writing as rhetorical action among language users in a specific text and community. As Perelman puts it, "Writing is foremost a rhetorical act, the transfer of information, feelings, and opinions from one mind to another mind . . . The essence of

writing, like all human communication, is not that it is true or false, correct or incorrect, but that it is an action, that it does something in the world" (para. 4). Humans more easily read and write in this sociocultural way than machines. Yet Perelman's point reminds us that human and automated scorers alike can belie this sociocultural conceptualization when they focus on error-hunting rather than situated meaning-making.

It follows that like writing and literacy, writing assessment and its validity are situated. We can see in Williamson's 2003 article that ideas about validity needed expanding at the time: he cautions against seeing validity only as reliability. In other words, he decries the notion that validity means "a test has to measure what it purports to measure," because validity is situated "in a particular use of a test, in a particular context, at a particular time" (p. 267). This emphasis on use anticipates Kane's work on interpretation and use arguments that are part of any test: scores represent inferences drawn from assessments. Those scores (and use of those scores, such as admissions or course placement) make interpretive arguments (Kane, 2013). Likewise, recent notions of reliability have expanded, calling into question prevailing standards built on narrow testing constructs, moving instead toward reliability based on the measurement, conditions, and objectives of complex writing performances (Ross & LeGrand, 2017).

Human Limitations, Automated Scoring Strengths

The last section makes it easy to see why it is common to assume that human scoring is ideal—that, to use an example from the opening, automated scoring is at best a necessary evil. Indeed, trust in humans underpins the prevailing practice of training an automated scoring model against human scores. Canzonetta and Kannan write that Turnitin's "intelligent assessment" alleges to grade papers like humans. More generally, Williamson calls for assurance that "the goals of people . . . drive the development of automation, not the automation itself" (p. 272).

But as writing and research remind us, humans also have limitations; like automated scoring, human scoring merits critical investigation. Human scorers working in the U.S., Canzonetta and Kannan emphasize, have overwhelmingly inherited "a culture of standardized testing" and "hegemonic cultural expectations about writing and authorship." In particular, humans have subjective associations with usage choices and with particular constructed response tasks (Johnson & VanBrackle, 2012; Perryman-Clark, 2016). Most U.S. language arts and writing instructors have taught rules based on 18th century usage preferences rather than "what language is and allows human beings to do" (Gere et al., 2021; Smitherman, 2017, p. 6). They have learned to "know good writing when they see it" rather than to analyze language patterns (Aull, 2021; Lea & Street,

1998). Research shows that when presented with an expression of non-standard English, a typical rater will undervalue the essay even though an answer may be functionally equivalent to a response given in standard English (Shermis, Burstein, Higgins, & Zechner, 2010, p. 5). It is hard to train human scorers to consider language diversity without bias, even though research shows that alternatives to conventional structural choices, for instance, vis-à-vis articles, verb morphology, nouns, and verbs, do not inhibit meaning or cause lower scores for many human raters (Allen, Crossley, Kyle, & McNamara, 2014).

Research also shows lack of agreement between humans in terms of what matters most and when. Studies show that style errors, for instance, are context dependent, and agreement on when a style error occurs may differ from person to person (Crossley, Bradfield, & Bustamante, 2019). The writing task can influence not only human scores, but also the extent to which humans agree on scores.

Automated scoring technology, then, can be one way to strive to mitigate socially-constructed bias against nonstandardized usage preferences, making it easier for a blueprint to read for ideas expressed in diverse ways. Shermis et al. describe that automated scoring tools would have the capacity to overcome this human limitation if the relevant affected variables associated with nonstandard English can be isolated and adjusted (Shermis, Burstein, Higgins, & Zechner, 2010).

Human and automated scoring together. Agreement between human and automated scoring is much discussed in research on automated scoring methods. Yet the three articles suggest that the more important question is, instead, what each can read and interpret. According to Williamson, different approaches to agreement highlight different disciplinary approaches of writing educators versus education measurement professionals. While for an assessment professional trained as a social scientist, the "immediate question is whether the procedures used by automated scoring engines simulate the scoring process of human raters," a writing educator trained in English studies expects that literacy entails various readings on one text (Williamson, p. 265). One way to read this tension is as a productive one. With a view of writing as complex written action contingent on rhetorical situations, what aspects of academic writing are rigid, and what expectations need not be?

In a critique of the Shermis and Hamner study, Perelman shows that exploring agreement includes important questions related to constructed response tasks, resolving disagreement, and evaluation criteria. On the first point, Perelman shows that constructed response task influenced agreement: human scorer agreement in response to one essay task (essay set 2) was stronger relative to other essay set scores, and relative to machine scores. On the second point, Perelman shows that use of integer scores necessarily belies the fact that writing

scores are continuous. In human scoring, resolving scores involves adjudicators rather than rounding; in the Shermis and Hamner study, Perelman shows how scoring resolution procedures were used in ways that privileged machine scorers and penalized human readers. Finally, Perelman shows that in certain evaluation categories, humans agreed more than automated scorers: those categories which measured ideas, content, organization, style, and voice, had an exact agreement value of 0.76, compared to the range of machine values of 0.55-0.70. All of the above provide support for Perelman's argument that in contrast to the widely-reported Shermis and Hamner claims about the reliability of machine scoring, actually, "the data provide some, although not conclusive, support for the assertion that human scorers performed more reliably than the machines, especially on longer papers that were scored for writing ability rather than solely on content" (para. 3). Perelman's work suggests another implication for calibrating for agreement between automated scoring and human raters is that, while automated scoring may be developed to imitate human assessors, human raters, too, become calibrated to evaluate like automated scoring tools.

Canzonetta and Kannan stress that a role of automated scoring is imparting culturally-specific ideas: there is no global, human agreement on a definition of plagiarism and fair use, thus any automated plagiarism tool imposes its moral and pedagogical conceptualizations on its users. If we connect this point to the other articles, we can consider the extent to which agreement itself is a culturally-specific value—a value situated in field, place, and time. In the case of Turnitin.com, Canzonetta and Kannan describe, it is a Western view of character, citizenship, and economics that is imposed. Under the guise of plagiarism detection, they write, these views are imposed but under-acknowledged. Canzonetta and Kannan call for critical attention to the views underpinning scoring and cases of automated and human scoring dis/agreement. Their call poses important considerations for culturally-specific ideas in solely human scoring as well.

Together, all of these ideas point to *a* role, but not *the only* role, for automated scoring in writing assessment. Automated scoring may provide a fuller picture of performance in complex assessment tasks that involve both reading and writing. Perelman's article underscores that the reliability of humans and of automated scoring is a site for investigation. While evidence shows that humans can read with genre-specific attention, evidence also shows they can read with socially-constructed bias against certain kinds of language use. Thus one way to respond to the questions *what is the role of automated scoring?* and *are humans good at scoring?* is to say that any scoring, and how it is represented, merit critical, ongoing attention from writing educators and educational measurement professionals alike.

WHAT RESPONSIBILITY DO WRITING EDUCATORS HAVE VIS-À-VIS AUTOMATED SCORING?

All three articles illuminate the risks of keeping writing and measurement specialists apart. First, Williamson notes a lack of cross-talk between "writing teachers" and the "assessment community," the two stakeholder groups most implicated in the scoring of writing. He explores "beliefs and assumptions held by each side" (p. 254), including disciplinary differences entailing different goals: more humanistic approaches emphasizing social context in college writing studies; more scientific approaches emphasizing aggregated patterns for assessment professionals.

Separation between the stakeholder groups, Williamson notes, means a lack of "productively learn[ing] to talk to each other about automated scoring." It means insufficient questioning of key assessment concepts such as validity, which becomes a "taken-for-granted ubiquity" in the lived experiences of students and teachers by becoming a "totalizing" and "naturalized concept and invisible instrument of rigor." Describing the early 21st century, Williamson writes, "English teacher response to automated scoring has been limited and . . . does not refer to any of the evidence presented by the developers of automated scoring programs" (p. 254).

Alternatively, Williamson describes, debate and understanding across the groups can challenge those in English Studies to address basic procedural issues of social science with questions of validity. It can challenge those in social science to consider validity as a situated construct, one that must observe the same situatedness that literacy theorists have been articulating for some time. Ultimately, the groups can come together around "the shared goal of moving toward more reliable and efficient ways to measure educational achievement and writing ability" (p. 254). At minimum, Williamson writes, "automated scoring is an incredible research opportunity through which we can explore the many different ways student writing can be read, valued, and sanctioned" (p. 254).

All three articles are examples of how productive examination of automated scoring is facilitated by engaging educational measurement and writing education. Perelman's article exemplifies the possibility of multiple methods and perspectives coming together. He uses statistical tests to expose unsupported claims in Shermis and Hamner's paper in ways valued by social scientists, and he attends to constructed response task as a situated written action in ways valued by those in English studies. Perelman's critique also shows the importance of assessment and writing researchers' engagement with publicly-rendered claims about writing and scoring. Shermis and Hamner's study findings, critiqued by Perelman, circulated widely: following Shermis's presentation of their findings at the National Council on Measurement in Education's annual meeting, the study

was cited in *Inside Higher Ed* and *The New Scientist,* and by a press release from the University of Akron.

Canzonetta and Kannan add additional emphasis, calling for cross-talk about the role of large corporations in global automated assessment services. In so doing, they take up Grabill's call for more attention to automated writing technologies by composition and rhetoric scholars, stressing that "[g]lobally, millions of students are subjected to writing technologies that writing experts did not design" (p. 296). Canzonetta and Kannan specifically outline plagiarism detection tools (PDSs), using Turnitin.com's success as an example of corporate influence in U.S. universities. They analyze how PDSs constitute instructors (presumed to be members of the "Turnitin.com educational community") as preservers of ethical and moral standards, positioned antagonistically against students, and assumed to be consistent across institutions and geographic locations. They call for direct engagement, so that there is greater understanding of the global cultural work of automated plagiarism and assessment tools.

In sum, the articles underscore that writing specialists have a responsibility to engage critically with automated scoring. The alternative, they imply, constitutes risks and missed opportunities. They bring to mind White's urging: "Assessment of writing can be a blessing or a curse, a friend or a foe, an important support for our work as teachers or a major impediment to what we need to do for our students. Like nuclear power, say, or capitalism, it offers enormous possibilities for good or ill, and, furthermore, it often shows both its benign and destructive faces at the same time" (White, 1994, p. 137). In none of these metaphors is there an option for writing educators to leave alone automated scoring as part of writing assessment.

WHAT MIGHT THE FUTURE LOOK LIKE?

Like all assessment technologies, automated scoring of writing highlights and precludes particular constructs of writing and beliefs about writers. These can be implicit and taken for granted, particularly when they are established, normalized, and widespread. Alternatively, these articles make automated scoring a site of critical engagement in ways that expose implicit ideas. They help us question: What kinds of writing and writers are valued in automated scoring? What kinds of writing and writers are valued in human scoring? What can we learn from each one?

This kind of critical engagement helps us, in Williamson's words, "study automated assessment in order to explicate the potential value for teaching and learning, as well as the potential harm" (p. 270). With one more article in the *Journal of Writing Assessment*, Ellen Cushman's (2016) "Decolonizing Validity," they pave the way for more inquiry-oriented, student-centered assessment. To

close, I consider related possibilities for automated scoring, for making it a site of investigating rather than damning language difference, a site of collective exploration rather than a site of top-down design and individual mastery.

Decolonizing Validity

The articles by Williamson, Perelman, and Canzonetta and Kannan conceptualize automated scoring as a site for ongoing inquiry into available tools and decisions about their use. They call for understanding writing, writing assessment, and validity as situated in rhetorical situations. They show how the use of automated scoring tools can entail a cyclical approach to validity: if we define validity narrowly and measure it in narrow tests, we learn only about those narrow conceptualizations. If validity is a matter of narrowly-defined, consistent scoring, then an automated scorer can measure length, mechanics, and lexical cohesion in a timed writing task, for instance, and be valid. Alternatively, if validity is instead a matter of fairness defined as equitable distribution of scores across different student groups, then a valid test must do very different things.

Cushman argues that to date, the "concept of validity creates the colonial difference as it maintains social, epistemic, and linguistic hierarchies." It does so by "identifying what is objective and what is evidence," and by hiding its social construction: validity "is a naturalized concept and invisible instrument of rigor that totalizes the realities of students and researchers" (para. 7). Drawing on Tiostanova and Mignolo's (2012) phrasing ("dwelling in the borders" in order to '"change the construct" itself), Cushman offers an alternative conception of validity, one in which "[d]welling in the borders begins with the knowledge, languages, histories, and practices understood and valued by the people who live these realities" (para. 25). In this conceptualization, validity is collectively constructed and navigated.

In turn, validity evidence tools work "not as a way to maintain, protect, conform to, confirm, and authorize the current systems of assessment and knowledge making, but rather as a way to better understand difference in and on its own terms." In other words, validity could be seen not as a way to hold individuals to one set of metrics determined by an external group—not as "a way to maintain, protect, conform to, confirm, and authorize the current systems of assessment and knowledge making." Rather, validity could be seen in terms of descriptive power: what it helps us learn about difference. In this approach, validity measures do not "mak[e] [one] experience into a universal one, the baseline against which all Others are tested and their knowledges and languages are deemed deficit to" (para. 26). A valid assessment approach would thus be one that "seek[s] to identify understandings in and on the terms of the peoples who experience them" (para. 26).

These ideas, in turn, carry implications for students' learning alluded to by Perelman, Williamson. They carry implications for linguistic diversity alluded to by Perelman and Canzonetta and Kannan. To unpack these ideas, let's briefly consider what they entail in terms of exploring difference, formative assessment, and critical language analysis.

Exploring Difference

Williamson underscores that "fluent adult reading" expects different views from different readers (Williamson p. 265). Generally, automated scoring and inter-rater agreement expect the opposite: they expect "convergent reading." Likewise, Perelman demonstrates the dangers of disparate approaches to resolving difference. And like other studies focused on agreement between human scorers, and/or between human and machine scoring, Deane et al. prioritize what Williamson calls convergence: the smaller the difference, the better (and "ideal for operational use" are very small differences in scores inferred from reading).

Here we see a good example of Cushman's point: in this case, disparate reading is, in a sense, invalid; validity rests on agreement in reading and inferences. What if we could construct writing, and reading one's own and other's writing, not as a site of deciding whether it was right or wrong, but of exploring the inferences drawn from machines and from humans? What would it take to create that world? Canzonetta and Kannan underscore that this is not easy, because rhetorics of standardization and consistency are beneficial for Turnitin's business model.

Formative Assessment

Canzonetta and Kannan describe that "Formative assessment necessitates that teachers respond to students' needs, personalities, struggles, and strengths; and get to know them apart from their writing." Ultimately, Canzonetta and Kannan caution against automated plagiarism tools' role in formative assessment, but they do point to that role as a site for critical investigation: "it is important to critically interrogate Turnitin's rhetorics of formative assessment, which obscure the company's cooptation of student data and potential to undermine writing program goals" (para. 27). This is all the more important because the message from plagiarism tools can promote formative uses that ultimately "aim to quell critique and breed a compliant, submissive population of students" instead of a more student-centered invitation of the students' active questioning of ideas about plagiarism and writing.

One way that automated scoring could be used would be in formative reports for students' use. Following Cushman, these reports could be a site of exploring

difference. What range of responses were there? What did these differences achieve? How did they respond to or change the rhetorical situation of the task?

Critical Language Analysis

Cushman writes that "Validity is on the one hand instrumental tool, which was established to manage peoples, knowledges, lands, governments, and institutions, and on the other hand, a meta-discourse which reified the social, linguistic, and epistemological hierarchies that made it possible, hence further securing its own position of authority to identify what counts as valid" (para. 11). Any use of automated scoring framed as evaluating whether writing is "good" is an example of such metadiscourse. Alternatively, framing language patterns as situated opens space for existing and valuable language diversity.

In other words, the use of automated tools and human reading can be used to identify and explore patterns descriptively, if only we can frame difference this way. Descriptive labels are possible when we chart linguistic patterns. For instance, language can be formal and informational, with many noun and prepositional phrases, such as academic writing. Language can be informal and interpersonal, with many verbs and pronouns, such as informal internet writing. This kind of charting of micro-linguistic features beyond mechanics could support Critical Language Awareness pedagogy illustrated in years of studies in student writing (Fairclough, 2014; Sanchez & Paulson, 2008; Shapiro, 2022)—pedagogy that supports students' exploration of "the social and linguistic rules" of their own language use (Smitherman, 2017, p. 6). Williamson alludes to the opportunities for automated tools to chart language patterns, noting the emerging developments in natural language processing.

Research in writing studies and assessment offers examples of how automated tools might support these efforts. Based on corpus linguistic analysis, research tests popular writing advice offered to students, showing that it doesn't always bear out (Lancaster, 2016). Similar studies map linguistic features in order that students can use linguistic patterns in analysis of their own writing (Aull, 2015, 2020). Other research questions, more broadly, the usefulness of automated tools to map out student writing features (Crossley, Kyle, & McNamara, 2015).

CONCLUSION

In the sections above, we've seen automated scoring framed as the use of algorithms to evaluate aspects of writing and as a site for exploring ideas about writing and writers embedded in any given approach to writing assessment. We've seen that automated scoring constructs writing, writers, and the practices of

assessment in particular ways. We've seen the limitations of automated scoring and of human raters. All of these ideas point to the value of integrated discussions that make both human and automated scoring a site for ongoing, critical attention. Keeping automated scoring as a site of inquiry opens possibilities for mapping writing difference not for the sake of ranking but for greater understanding and exploring.

[i] To define "holistic ratings," we can turn to Haswell and Elliot: "the use of a scale to assign a single value mark to a whole essay and not separately to separate aspects of the essay, with scorers trying to apply the scale consistently, and with the final score for each essay derived from two or more independent ratings" (p. 1). Williamson also notes studies that established that holistic scoring is a limited form of reading.

REFERENCES

Allen, L., Crossley, S., Kyle, K., & McNamara, D. S. (2014). The importance of grammar and mechanics in writing assessment and instruction: Evidence from data mining. *Grantee Submission*. [Paper Presentation] International Conference on Educational Data Mining.

Aull, L. L. (2015). *First-year university writing: A corpus-based study with implications for pedagogy*: Springer.

Aull, L. L. (2017). Tools and tech: A new forum. *Assessing Writing*. *33*, A2-A7.

Aull, L. L. (2020). *How students write: A linguistic analysis*. MLA.

Aull, L. L. (2021). What is "Good Writing?": Metadiscourse as civil discourse. *Journal of Teaching Writing 36*(1), 37-60.

Aull, L. L. (2024). *You can't write that . . . 8 myths about correct writing.* Cambridge University Press.

Canzonetta, J., & Kannan, V. (2016). Globalizing plagiarism & writing assessment: a case study of Turnitin. *The Journal of Writing Assessment, 9*(2). https://escholarship.org/uc/item/5vq519dr

Crossley, S. A., Bradfield, F., & Bustamante, A. (2019). Using human judgments to examine the validity of automated grammar, syntax, and mechanical errors in writing. *Journal of Writing Research, 11*(2), 251-270.

Crossley, S. A., Kyle, K., & McNamara, D. S. (2015). To Aggregate or not? Linguistic features in automatic essay scoring and feedback systems. *Journal of Writing Assessment, 8*(1). https://escholarship.org/uc/item/1f21q8ck

Cushman, E. (2016). Decolonizing validity. *Journal of Writing Assessment, 9*(1). https://escholarship.org/uc/item/0xh7v6fb

Deane, P., Williams, F., Weng, V., & Trapani, C. S. (2013). Automated essay scoring in innovative assessments of writing from sources. *Journal of Writing Assessment, 6*(1), 40-56. https://escholarship.org/uc/item/3nf6r4kv

Elliot, N. (2005). *On a scale: A social history of writing assessment in America*. Peter Lang.

Fairclough, N. (2014). *Critical language awareness*: Routledge.

Gere, A. R., Curzan, A., Hammond, J., Hughes, S., Li, R., Moos, A., Smith, K., Van Zanen, K., Wheeler, K., & Zanders, C. J. (2021). Communal justicing: Writing assessment, disciplinary infrastructure, and the case for critical language awareness. *College Composition and Communication, 72*(3), 384-412.

Grabill J. (2016) Do we learn best together or alone? Your life with robots. *Computers & writing conference, [Keynote address]* Rochester, New York. http://elireview.com/2016/05/24/grabill-cw-keynote

Hammond, J. (2019). *Composing progress in the United States: Race science, social justice, and the rhetorics of writing assessment, 1845-1859.* [Doctoral Dissertation, University of Michigan].

Ivanič, R. (2004). Discourses of writing and learning to write. *Language and Education, 18*(3), 220-245.

Johnson, D., & VanBrackle, L. (2012). Linguistic discrimination in writing assessment: How raters react to African American "errors," ESL errors, and standard English errors on a state-mandated writing exam. *Assessing Writing, 17*(1), 35-54.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1-73.

Lancaster, Z. (2016). Do academics really write this way? A Corpus investigation of moves and templates in "they say/I say". *College Composition and Communication, 67*(3), 437-464.

Lea, M. R., & Street, B. V. (1998). Student writing in higher education: An academic literacies approach. *Studies in Higher Education, 23*(2), 157-172. https://doi.org/10.1080/03075079812331380364

Perelman, L. C. (2013). Critique of Mark D. Shermis & Ben Hammer, "Contrasting state-of-the-art automated scoring of essays: Analysis." *Journal of Writing Assessment, 6*(1). https://escholarship.org/uc/item/7qh108bw

Perryman-Clark, S. M. (2013). African American language, rhetoric, and students' writing: New directions for SRTOL. *College Composition and Communication*, *64*(3), 469-495, https://www.jstor.org/stable/43490767

Perryman-Clark, S. M. (2016). Who we are (n't) assessing: Racializing language and writing assessment in writing program administration. *College English, 79*(2), 206-211.

Ross, V., & LeGrand, R. (2017). Assessing writing constructs: Toward an expanded view of inter-reader reliability. *Journal of Writing Analytics, 1*, 227-257. https://doi.org/10.37514/JWA-J.2017.1.1.09

Sanchez, D. M., & Paulson, E. J. (2008). Critical language awareness and learners in college. *Teaching English in the Two-Year College, 36*(2), 164-176.

Shapiro, S. (2022). *Cultivating critical language awareness in the writing classroom*. Routledge.

Shermis, M., Burstein, J., Higgins, D., & Zechner, K. (2010). Automated essay scoring: Writing assessment and instruction. *International encyclopedia of education, 4*(1), 20-26.

Shermis, M. D., Burstein, J., Higgins, D., & Zechner, K. (2010). Automated essay scoring: Writing assessment and instruction. In P. Peterson, E. Baker, & B. McGaw (Eds.), *International Encyclopedia of Education* (3rd ed.) (pp. 20-26). Elsevier.

Shermis, M. D., & Burstein, J. C. (2003). *Automated essay scoring: A cross-disciplinary perspective*: Routledge.
Shermis, M. D., & Hamner, B. (2012). Contrasting state-of-the-art automated scoring of essays: Analysis. http://www.scoreright.org/NCME_2012_Paper3_29_12.pdf
Shermis, M. D., & Hamner, B. (2013). Contrasting State-of-the-Art Automated Scoring of Essays. In M. D. Shermis, & J. Burstein (Eds.), *Handbook of automated essay evaluation* (pp. 213-246). Routledge.
Smitherman, G. (2017). Raciolinguistics,"mis-education," and language arts teaching in the 21st century. *Language Arts Journal of Michigan, 32*(2), 4-12.
Tiostanova, M. V., & Mignolo, W. (2012). *Learning to unlearn: Decolonial reflections from Eurasia and the Americas*. Ohio State University Press.
White, E. M. (1994). *Teaching and assessing writing: Recent advances in understanding, evaluating, and improving student performance. Revised and expanded.* Jossey-Bass.
Williamson, M. M. (2003). Validity of automated scoring: Prologue for a continuing discussion of machine scoring student writing. *Journal of Writing Assessment, 1*(2), 85-104. https://escholarship.org/uc/item/8nv3w3w8

CHAPTER 8.

VALIDITY OF AUTOMATED SCORING: PROLOGUE FOR A CONTINUING DISCUSSION OF MACHINE SCORING STUDENT WRITING

Michael Williamson
Indiana University of Pennsylvania

Writing assessment has developed along two separate lines, one centered in professional organizations for writing teachers and the other centered in professional organizations for the broader assessment community. As the controversy about automated scoring continues to develop, it is important for writing teachers and researchers to become fluent in the discourse of the broader assessment community. Continuing to label the work of the broader assessment community as positivist and continuing to ignore it will only result in a continuing sense of defeat as automated assessment is adopted more widely. On the other hand, an examination of the literature on educational assessment will reveal that the theoretical base for assessment is quite consistent with the principles adopted by the writing assessment community.

> Grading essays by computer seems to have entered an explosive new phase, and I hope that, by the end of this talk, you folks will be excited, too, about all the changes this may mean for testing. After all, essay grading has been done for perhaps 4 thousand years. But now we seem to face a brand-new opportunity: Not simply to help in human essay grading, but to firm it up with actual objective data, of the kind never really used.
>
> – Ellis Batten Page (1995)

Anson (2003), reflecting on developments in artificial intelligence (AI), suggests it has provided little to serve any useful purpose in the English classroom because software has not been sufficiently sophisticated. Earlier, Herrington and

DOI: https://doi.org/10.37514/PER-B.2024.2166.2.08

Moran (2001) examined an emerging application of AI in English Studies, automated scoring, and the use of computer algorithms to simulate holistic ratings of student writing. Although they are concerned about the adequacy of the feedback provided by such programs, the greater concern is the implications for students' learning when computers are the basis for grades. However, automated scoring technologies are finding wider acceptance among educators. The Commonwealth of Pennsylvania recently made a commitment to the use of Intellimetric, the scoring engine reviewed by Herrington and Moran (2001). Some reports suggest that this engine was to be used in 2003 to score the writing of students on the mandatory Pennsylvania state achievement examinations (Indiana Gazette, 2003). Other states and individual school districts are either implementing or exploring the implementation of one of the available engines.

This obvious conflict suggests that some may see valid applications for automated scoring, whereas others see none, suggesting that a deeper examination of the available inquiry about the validity of automated scoring is necessary. English teacher response to automated scoring has been limited and such response (Anson, 2003; Herrington & Moran, 2001) does not refer to any of the evidence presented by the developers of automated scoring programs. There remains a need to examine the claims made by test developers about the validity of automated scoring and to determine whether any possible objections have been addressed.

Initially, I hoped to write an article that picked between the various arguments and claims and contended for certain use of automated scoring in writing assessment. Unfortunately, my reading of the literature around this issue left me feeling that other precursor work needed to be done before the two camps, what Moss (1998) first labeled college writing assessment and educational measurement, could productively learn to talk to each other about automated scoring. In this article, I explore various beliefs and assumptions held by each side. Looking at the history of test development in general and writing assessment in particular, I examine the drive toward more reliable and efficient ways to measure educational achievement and writing ability. Additionally, I consider the various epistemological orientations of those who work in social science and the humanities, noting how each disciplinary area has changed over the last several years with the influence of postmodern theories of knowing and making meaning. I hope that this article can establish a common ground for future scholarship and discussion. At the very least, automated scoring is an incredible research opportunity through which we can explore the many different ways student writing can be read, valued, and sanctioned.

Automated scoring is not new. It first appeared in 1966, in the work of Ellis Page (1995). The response to this early work from the English-teaching community was similar to current responses. Reviewed in *Research in the Teaching of*

English, Page's original work drew a response similar to Anson and Herrington and Moran from Macrorie (1969). On the other hand, Coombs (1969) was skeptical, but not entirely dismissive of the potential demonstrated in Project Essay Grade. However, automated scoring does not seem to have been wholeheartedly embraced by anyone in English Studies publishing in typical outlets, such as *College English* or *Research in the Teaching of English*.

On the other hand, a recently burgeoning literature on automated scoring has appeared in the literature typically examined by the broader assessment community, much of it suggesting that automated scoring does have valid applications for the assessing of writing.

AUTOMATED SCORING AS WRITING ASSESSMENT

Although it has a new face, the controversy over automated scoring reflects the constant struggle over writing assessment and the apparent stasis in achieving a resolution (Williamson, 1993). Until recently, the controversy focused on movement from indirect to direct measurement of writing (Williamson, 1993), as reflected in Yancey's (1999) history of the last 50 years of writing assessment. Currently, writing assessment seems to be caught in a three-way tug of war involving the introduction of portfolio assessment in the teaching and assessing of writing. Yancey suggests a shift in focus from reliability in the dispute over direct and indirect assessment, to validity, a dispute over how much writing is necessary to make a valid judgment about students' writing. From the beginning, there has been an explicit concern about the effects of particular approaches to assessment on the teaching and learning of writing, in effect, a question about the validity of assessment. Yancey's view reflects a trend in the literature by and for writing teachers and researchers to respond primarily to the challenges posed by systems developed to ensure the reliable scoring of student writing. The proposal to replace essay examinations with objective examinations, based in multiple-choice technologies began the controversy.

> The most recurrent criticism of essay tests, and the one about which the most has been written, concerns the unreliability of evaluating essay answers. If a test is to be worth while [sic] as a measuring instrument, it must measure what it purports to measure consistently and dependably (Stalnaker, 1951, p. 498).

As Yancey points out, the response to objective testing was the development of direct assessment approaches using writing, justified in terms of their reliability, just as the justification for indirect assessment, using multiple-choice items, was grounded in its reliability compared to the earlier use of writing as a tool for assessment. Although the battleground itself was seen as reliability, the larger struggle was about validity, thought it was focused at the time in terms of reliability.

All educational measurements are generally intended to elicit information regarding the structure, dynamics, and functioning of the student's mental life as it has been modified by a particular set of learning experiences. The special problem in the case of the achievement test is to obtain information which is reliable and pertinent, and to do so efficiently (Stalnaker, 1951, p. 496). These concerns evolved into the traditional claim that a test had to measure what it purports to measure and that reliability is a necessary but insufficient claim for validity.

Validity has two aspects, which may be termed relevance and reliability. "Relevance" concerns the closeness of agreement between what the test measures and the function that it is used to measure. "Reliability" concerns the accuracy and consistency with which it measures whatever it does measure in the group with which it is used. To be valid—that is, to serve its purpose adequately—a test must measure something with reasonably high reliability, and that something must be fairly closely related to the function it is used to perform (Cureton, 1951, p. 622).

Although some developers have made claims about potential pedagogical uses of automated scoring programs, I only focus on their validity as it pertains to writing assessment. The larger, and perhaps more important issue of their pedagogical value is another question, one that does not seem of immediate relevance for writing assessment. I begin with Herrington and Moran's (2001) exploration because it reflects my own examination of particular programs. There is, however, a paucity of research beyond such informal examinations. Second, for the most part, feedback to students is based on boilerplate rubrics, some quite complex and sophisticated. Rubric-based feedback in any kind of scoring may not address the particular reason an essay was placed in a score category (Broad, 2003; Huot, 1993; Pula & Huot, 1993; Smith, 1993). The qualities and bases for human judgment of complex performances cannot be explained by a rubric. Two things are certain. One, automated scoring programs can replicate scores for a particular reading of student writing, and this technology is reliable, efficient, fast, and cheap. Two, automated scoring has been and will continue to be used in various large-scale assessments of student writing.

VALIDITY

As early as 1951, validity was defined by Edward Cureton in the first edition of what would become a periodic definition of the state of the art in educational measurement, *Educational Measurement*.

The essential question of test validity is how well a test does the job it is employed to do. The same test may be used for several different purposes, and its validity may be high for one, moderate for another, and low for a third (p. 621).

An important and forward-looking aspect of this definition is that it is grounded in the use of a test, not in the test itself. The definition of validity evolved with both formal and informal meanings, as can be noted in Cronbach's (1971) leading text on the theory and practice of educational measurement.

We defined validity as the extent to which any measuring instrument measures what it is intended to measure. However, as we pointed out in Chapter 1, strictly speaking, "One validates, not a test, but an interpretation of data arising from a specified procedure" (p. 447).

While conforming with this general definition, Anastasi (1976) presents three primary forms of validity, each defined by the procedures used to determine them.

Fundamentally, all procedures for determining test validity are concerned with the relationships between performance on the test and other independently observable facts about the behavior characteristics of under consideration (p. 134).

She also provides a separate treatment of validity as an issue for interpreting test results through the use of decision theory, a further coupling of validity with particular uses of a test.

In one of the seminal works on writing assessment produced by writing researchers, Cooper and Odell (1977) define validity with a slightly different focus, one that may ultimately be responsible for spreading the informal definition of validity as the dominant meaning in writing assessment.

If a measure or measurement scheme is valid, it is doing what we say it is doing. We want to insist on a careful distinction between predictive validity and other kinds of validity, content and construct validity (p. xi).

This definition reflects what I am labeling the informal definition of validity. Later definitions of validity tend to adopt this informal definition, for instance:

> Although validity is a complex concept—colleges offer advanced courses in it—one simple concept lies behind the complexity: honesty. Validity in measurement means that you are measuring what you say you are measuring, not something else, and that you have really thought through the importance of your measurement in considerable detail. (White, 1994, p. 10)

White's definition of validity is metaphorical, and although metaphor is not unknown in social sciences research, the redefining of validity in this case moves two fundamentally different definitions of the same concept further apart.

The essential, crucial difference between these two definitions lies in the distinction between defining validity as procedure and validity as a property of a test. This distinction emerges from the difference between understanding the mathematical basis for assessment and the application of assessment in what

Stalnaker (1951) labels achievement testing. Tests like statistical operations are conducted to make informed educational judgments. The simplicity of distinction between a procedural and conceptual understanding of validity is not always as clear and separate as it might seem. The fundamental nature of validity can be rendered confusing by educational researchers themselves.

While the definition of validity seems simple and straightforward, there are several different types of validity that are relevant in the social sciences. Each of these types of validity takes a somewhat different approach in assessing the extent to which a measure measures what it purports to (Carmines & Zeller, 1980, p. 17).

Broad (2003) labels one stance in writing assessment "positivist," a stance that can be traced to Berlin's (1984) history of writing instruction. Positivism as a theoretical approach to the philosophy of science certainly characterizes early psychometric theory and its attempt to define psychology and educational and psychological assessment as a science. Guilford (1954) traces the emergence of statistical investigation in psychology and grounds his approach to the field in mathematics, as well as statistical inquiry, "The progress and maturity of a science are often judged by the extent to which it has succeeded in the use of mathematics" (p. 1). Gulliksen (1950) specifically limits his description of mental testing to those defined by quantitative methods, while specifically noting the difference between statistics and mathematics. In Guilford's terms, mathematics is a "universal language that any discipline may use with power and convenience" (p. 1). That this movement toward the use of mathematics and quantification may be positivist is one that deserves larger exploration in the literature of the field. However, there is an interesting contrast to what may be perceived as the problem of quantification in writing assessment.

As early as the 1950s, at least, such issues as validity were seen less as defined by the results of a statistical test than as a matter of disciplinary disputation, the assembling of evidence, not the simple results of a statistical test (Cureton, 1951). In a related example, in discussing educational evaluation, one of the primary applications of educational measurement, Cooley and Lohnes (1976), both eventually to become president of the American Educational Research Association, suggest that the scrutiny of the field and not objectivity is the issue. Moss (1998) calls her response essay to a study of writing assessment validation, "The Test of the Test." For Moss, validation is a practice in turning the gaze toward the construct of the assessment itself. It is a form of reflective practice, or as Ellen Schendel (1999) claims, "social action."

What tends to keep researchers honest is the publicly available record of what they did and what they found, and not a godlike objectivity which some people seem to feel those doing evaluations should exhibit. Scientists doing basic

research know that if their work is to have any value whatsoever, it will be closely read and critically examined by their colleagues in the field (Cooley & Lohnes, 1976, p. 2).

These and other perspectives of validity are rooted in the ideas of Cronbach (1988, 1989) and Messick (1989). Cronbach (1989) characterizes validity as a form of disciplinary argumentation, one that is never finished and that evolves with each new use of an assessment in a new locale: "Validation is a lengthy, even endless process" (p. 151). Such a definition is supported by Cureton (1951) and Anastasi (1976) as well. It is this definition that leads Huot (2002) to characterize assessment as a continuing form of research. Thus, writing assessment should be viewed as a continuing examination of the available tools for assessment, as they are used for making new decisions. New developments will inevitably bring new tools, all of them requiring validity inquiry of their own.

Smith (1993) is probably the first researcher in writing assessment who fully reflects the complexity of validity inquiry. Although his work is some of the first substantive research that looks at the validity and not reliability of a writing assessment (Huot, 1996), ironically, he eschewed the word validity because he wanted to avoid any baggage associated with such a term. He used accuracy of placement as the goal for his placement testing program at the University of Pittsburgh. With collaborators, he designed a series of studies on the procedures that structure the way teachers make decisions based upon their reading of placement essays. Each of the studies led to a modification of the procedures that allowed a stronger claim to the validity of the assessment, the accuracy of placement of students in the writing program. This not only demonstrates more accurate placement of students over time, but it also led to a modification of the scoring procedures themselves. The end result was a less costly system because the reading and decision making were rooted in the context about which the teachers were expert.

The notion of validity as argument and the nature of professional judgment is related to Bleich's (1975) view of interpretive communities and Kuhn's (1996) view of the way that science changes through changes in the worldview of the members of the discipline. The meaning of a text, be it a poem or a validity inquiry, lies with the community of readers in the field and their intertextual experiences with the field. Such a position reflects a more postmodern view than the positivism cited as the basis for psychometric theory.

An additional consideration for validity is the impact of the assessment (Messick, 1989). The consequences of decisions made on behalf of a test is a core concern for validity inquiry because the use of a test may impact what is learned and how that learning takes place. This concern for the impact of a test is one of the ethical bases for validity theory. Thus, validity inquiry must examine how

learning changes as a result of the implementation of an assessment. Although this sounds like an ethical way to proceed, English professionals might question the existence of studies of the consequences of high-stakes testing on individual students taking high-stakes tests. Interaction among various fields is important if we are to understand complex phenomena. In particular, measurement theory in education and psychology has to respond to developments in psychology and education if the field is to remain viable. The impact of theoretical changes is not universally distributed in a field (Kuhn, 1996). If there are specialists in educational measurement still working with a variety of validities, there are still writing teachers and researchers who pursue grammar study as a prescriptive methodology in the teaching of writing. If validity theory has not coalesced into a univocal stance in measurement, the meaning of error is equally problematic for many teachers who are not able to grasp or who are unfamiliar with the complexity of disciplinary discourse on error.

After all, members of any academic field are part of both the paradigm that is disappearing and the new paradigm that provides a new synthesis for the field (Kuhn, 1996). That some may quote the contemporary definition and unwittingly include older definitions is not surprising. An interpretive community does not need to be, indeed is unlikely to be, univocal about any reading of any text. Importantly, if early theories of assessment were deterministic in the positivist sense that they were seen as objective explanations of reality, the postmodern influence in assessment publicly acknowledges the debate that always existed, and provides a new understanding of the meaning of such debate.

The core of my concern in the different representations of validity has to do with the difference between English Studies and educational measurement, the difference between social science and humanistic disciplines. A science depends on a clearly defined methodology as the basis for disciplinary disputation. Although English Studies depends largely on a hermeneutic form of inquiry, one based in close reading, assessment depends on evidence defined by the procedures that are used to collect it. For instance, the heart of the definition provided by Carmines and Zeller (1980) highlights the defining of each of the various types of validity as a procedure, despite the fact that it misses the more important concern that validity is contextualized.

Two conflicting views of research methodology are the primary problem for humanists as they attempt to represent their views outside of English Studies because any argument about validity will have to face the need to address the basic procedural issues of social science. Furthermore, if validity is seen as a unitary construct that involves the consequences of the test's use in context, validity can be seen as a situated construct, one that must observe the same situatedness that literacy theorists have been articulating for some time.

As a student of English Studies, I am concerned by the claims of Herrington and Moran (2001). As any good scholars in the field, they read the text of the automated scoring engine and see the rhetorical implications of its use in English classrooms. However, as a student of educational assessment, I know that their review of automated assessment does not provide the kind of structured inquiry necessary to convince a member of the community of readers in assessment. It is easy to adopt the stance that all psychometricians are positivists if one does not understand the fundamental role of scientific procedure in defining inquiry. However, the label itself has no meaning outside of English Studies because any form of quantification is labeled positivist. The label itself is, therefore, one that does not make the case against claims by psychometricians about the validity of particular approaches to assessment. In fact, most first-year composition texts would probably characterize such an argument as ad hominem.

PRINCIPLES GUIDING THE EXAMINATION OF THE VALIDITY OF AUTOMATED SCORING

All of the following statements are derived from the literature on education assessment and follow from my characterization of validity as it is defined by the following:

1. The validity of an assessment lies in the decision that is made on the basis of the test, not the test itself.
2. Validity is a form of scholarly argumentation, based on research, which subjects the assessment to open discussion about both its substance and its meaning.
3. Validity is not a substantial or concrete set of claims, the argument is open to question with each use of the assessment and as developments in various theories, both within and outside of assessment provide new perspectives on assessment, what is being assessed, and how the assessment is being used.
4. Validation research is never a closed circle. Each use of an assessment, whether in the same or different contexts must be examined to ascertain and revalidate the validity argument for the assessment, its uses, and the meaning of its uses.
5. In addition to examining the adequacy of the assessment for the decisions that are to be made from its use, assessment developers and users also have an ethical responsibility to examine the consequences of an assessment, to examine the effects of the assessment on both immediate contexts and broader cultural contexts.

Notice that each of these statements contains a procedural definition of validity. I argue that the definitions of validity that are common in English Studies are static, indeed, are positivist in the sense that they suggest we can know that a test is valid in objective terms, because we can know it is doing what we say it is doing. In other words, because many in English Studies ascribe to an older notion of validity (White, 1994; Yancey, 1999), they are unwittingly missing an opportunity to apply postmodern theories to validity inquiry and are, instead, promoting a rigid, decontextualized "positivist" concept of validity for writing assessment.

ARTIFICIAL INTELLIGENCE

Automated scoring is based in the technology of AI, and claims to bring the relative efficiency of automation to scoring essays. These two concepts need to be defined as part of the process of validity inquiry. AI is a research paradigm built around several sciences. The primary goal of the emergent paradigm has been the simulation of human intelligence and behavior in the electronic system of a computer. Developments in each of these sciences, from linguistics to psychology and mathematics to computer science, have allowed a nearly continuous development of demonstrations of intelligent machines. The emergent technologies have resulted in a variety of applications that both enhance and simulate human performance in a variety of fields. Thus, it seems that the use of such technologies would inevitably lead to their application in English studies. The first such application—Project Essay Grade—was seen by its developers as a method of relieving writing teachers of the burden of grading, leading also to more objective grades (Ajay, Tillett, & Page, 1973; Page, 1966, 1967a, 1967b, 1995; Page & Fisher, 1968). After an initial ambiguous response (Coombs, 1969; Macrorie; 1969), the concept of computer grading seems to have had little attention from researchers in composition and rhetoric for some time (Huot, 1996).

The development of the personal computer in the 1980s led to an outburst of enthusiasm for the use of computers in the writing classroom. The cutting edge of the field of computers and composition was initially defined by the seminal work of Hugh Burns (1979) with rhetorical invention and the rapid growth of word processing, among other business and personal applications. Burns' work reflected the early applications of artificial intelligence to English Studies. His work demonstrated the programming theories of artificial intelligence pioneered by Joseph Weizenbaum in the development of Eliza, a computer program designed to simulate the psychotherapeutic interviews of Carl Rogers. Eliza was considered to be a failure because the program did not meet Turing's criterion for a computing machine simulating human behavior,

a primary consideration in judging the validity of computer programs that "artificially" simulate human intelligence.

Alan Turing was one of the pioneers of digital computing at Bletchley Park in England during World War II. As a very early theorist in computing, he suggested that a successful demonstration of human intelligence by a computer would be indistinguishable from the performance of an actual human. In other words, Liza would be successful if the program were able to provide counseling to a human client without the client being able to determine whether the advice came from a machine or another human. Neither *Eliza* nor Burn's invention programs meet the criterion because they were unable to respond coherently to aberrant statements. The result of aberrant statements or questions about questions from the human user resulted in meaningless responses from the programs. Although the programming had a rudimentary syntactic parser, enabling it to extract relevant words from the input, it had no means of examining the meaning of any of the input. Therefore, it was easily "fooled" into giving unintelligible or meaningless responses. Subsequently, demonstrations of AI have been based on successively sophisticated approximations of human intelligence. Most of these early demonstrations were intended only to model what was possible, not necessarily to meet Turing's criterion.

Since the early demonstrations of machine intelligence, researchers working in the multidisciplinary field of natural language processing were busy with both basic research into computer simulation of language and immediate applications of this technology. With each new demonstration of the emerging technology, more sophisticated responses to human language were possible, as were more sophisticated applications. The accessibility of computers to those outside of computer science owes as much to the developments in AI as to the developments in the electronics side of computing.

AUTOMATION

Automated scoring—the use of computers to simulate holistic ratings of English essays—is quite accurately described as automation in the original sense of the word—the use of technology to relieve humans of repetitive work, work that taxes the limits of our abilities. It is, simply, the performance of tasks by machines, tasks that were originally performed by skilled humans, made skilled humans more productive, or created less skilled work from more complex work. Early automation is represented by the agricultural machines that first improved tilling the soil and subsequently harvesting. The original Luddites of 1811-1812 were weavers in England, members of a craft guild who attempted to destroy the newly invented machinery that left fewer jobs for unskilled workers. Mechanical

developments in automation began to skyrocket with the introduction of computer technology. Today, labor unions representing the interests of workers have been watching the emergence of automation with considerable concern because industrial, production line workers have been replaced by electronically operated machines that perform repetitive tasks with greater precision and accuracy than humans, at least in the view of industries that have adopted this technology. The motivation underlying electronic automation, even as it was in the planning, viewed constant repetition as a weakness in humans. Industrial automation was motivated by efficiency. To the extent that computers can make any task more efficient, they will be of interest in industry. Although workers in AI may not perceive the impact of their work, much of the research and development for applications of the emergent theories in AI have been funded by governments and industry looking for ways to operate more efficiently, even if only to get beyond errors and other problems that reflect the limits of human performance.

In the case of automation, the concern for a computer's performance is not on whether it meets Turing's criterion. Instead, the question is whether the task itself is computable. According to Johnson-Laird (1977), computability depends on being able to specify a task with sufficient precision to develop a programming algorithm, based in the computational structure of computer software. For instance, welding an exact spot on a car body involves only a question of space and time—the movement of the machine to the location of the weld and the length of the welding time. Although the relative quality of human labor and automation is certainly one issue, the real question lies with the sufficiency of the performance of the machine. If sufficient quality can be achieved by a computer program or robot, operating at greater speed and less cost, clearly, the programming is successful. The cost reduction and increased efficiency of machine operation, when seen only in terms of the costs of production and profit margins, are clearly a business issue.

CAN HOLISTIC SCORING BE AUTOMATED?

In an earlier essay, I discussed in some detail the underpinning of much assessment practice in the "Worship of Efficiency" (Williamson, 1994). Further demonstration of the role of efficiency in assessment is provided by some of the sources cited earlier in this text (Cureton, 1951; Stalnaker, 1951). The question of validity for automated scoring turns, in this circumstance, on whether automated scoring can provide results at least as trustworthy as human raters with greater efficiency and less cost. From this perspective, the question of validity for automated scoring can be answered in the same way that questions of quality are determined for other forms of automation. Although cost accounting may

be more relevant to business, the mathematical apparatus of assessment theory is employed in demonstrating the quality and validity of automated assessment as it compares to holistic raters. For others, the question of any automation of the work of writing teachers and assessors is a question of the computability of human language in the first place. In other words, can a computer using AI *read*?

Validity, as it is related to the comparability of holistic rating, is considerably more limited than some of the larger questions that have been raised about holistic scoring itself, such as the adequacy of the criterion definition of writing represented by a single essay and the adequacy of the criterion definition of reading represented in standardized rating sessions. The distinction between these two views of validity inquiry about automated scoring has important consequences for how specific investigations into its validity will be understood.

For example, the key issue for those creating automated scoring is whether the program can predict holistic ratings of more than six raters (Burstein, 2003), many more than the number typically employed in a holistic scoring session. To support their claim, automated scoring needs to demonstrate that it is more efficient and costs considerably less than rating sessions.

However, the discussion, within English Studies, seems to be dominated by a very different definition of the activity of holistic rating. The criterion that Herrington and Moran (2001), as well as Anson (2003) appear to be using is whether a computer can read. At least three studies (Huot, 1993; Huot & Pula, 1993; Wolfe, 1997) established that holistic scoring is a limited form of reading. In the Huot and Pula and Huot studies, holistic raters made rapid decisions about the placement of students reflected in the writing, and then spent time responding to other aspects of the writing. Wolfe found that raters who agree at a high rate with each other have a more focused reading process.

For a social scientist, the immediate question is whether the procedures used by automated scoring engines simulate the scoring process of human raters. This question is more difficult to answer because holistic rating is not reading as is usually defined in literacy research where the goal is to produce various readings; the push for writing assessment has been toward a single reading (Elbow & Yancey, 1994). Holistic scoring, by definition, limits the features of a text that the rater attends to. The scoring process also limits the purposes for which a text is read. Such convergent reading is not what is typically represented as fluent adult reading, an act of making meaning that typically leads to divergent views of a text.

IS HOLISTIC SCORING VALID?

There have been two large studies of the validity of holistic scoring, as applied to individual essays (Gottshalk, Swineford, & Coffman, 1966) and to multiple

essays from the same writers, intended as a form of portfolio assessment (Breland et al., 1987). The earlier study suggested that multiple-choice tests of grammar predict a student writer's performance more accurately than an essay when it is scored using a holistic procedure by two or three raters. Consequently, the relatively cheaper and more efficient indirect approach was justified because it could predict an individual's score on a criterion with the greater precision and accuracy than a writing sample. The claim for the validity of indirect assessment is based on a form of criterion validity known as concurrent validity that compares an examinee's performance on two different valued measures. Veal and Hudson (1983) dispute that result in another study of the use of holistic scoring using state assessment data from Georgia in which students' performances on multiple-choice tests of usage and grammar do match well with a holistic score.

Breland et al. (1987), stipulating that direct writing assessment is more efficient and less costly, demonstrated that one essay read two or three times could attain the reliability of indirect measures. Their criterion definition of writing was six essays from each writer. They conclude that the best approach to writing assessment is a combination of both direct and indirect assessment because the two work together to provide both a broader and more reliable picture.

Although psychometric theory clearly supports the need for studying validity in particular applications of a test, in practice multiple-choice tests were considered adequate when used "off the shelf" by educational institutions. Thus, although the theory was suggesting the need for more study of assessment procedures in particular applications, conventional wisdom allowed for their use as ready made instruments for student, teacher, and program evaluation. This was equally true in the use of holistic scoring. For the most part, writing assessments used holistic scoring without much examination of the validity of its actual use, because the understanding of assessment theory prevalent in the field was that a test using writing is more valid on its face and in its content than any form of indirect test (Yancey, 1999).

White (1994) recounted the political struggles involved with the adoption of direct assessment. However, the extant theory in measurement could have been used to support the argument against indirect assessments had more writing assessment developers, like Veal and Hudson (1983), used the theory to argue their position. Hence, with greater fluency in the theory that was used, writing teachers and researchers would likely have been able to develop assessments that could be demonstrated to have the same kinds of properties that were valued in the validation of indirect assessment. One good example is the study by Breland et al. (1987), which suggests that holistic scoring of a writing portfolio leads to more accurate predictions than the score of any single essay in the portfolio. As early as Terman's 1916 book on the measurement of intelligence, statistical

procedures were well defined for an examination of the contributions of test length to overall test reliability. Item validity was really the only focal concern, because as Gulliksen (1950) points out:

> We see that the validity coefficient is the square root of that for the reliability coefficient. . . . Since the validity coefficient is usually considerably smaller than the test reliability, this usually means that changing the length of the test can be expected to have a very slight effect on the validity of the test. (p. 90)

Thus, the ultimate focus in measurement theory is on reliability to the extent that it is defined statistically. A test reliability of 0.9 will provide a test validity of 0.3, for instance. Little wonder that the traditional debate over holistic scoring confuses reliability and validity.

If, as I have argued, validity is seen as existing in a particular use of a test, in a particular context, at a particular time, validity reflects the situatedness of literacy as most researchers and teachers of writing have been claiming. Thus, validity does not lie in statistical procedures alone. However, test developers themselves rarely study the validation of decisions. Furthermore, the kind of study undertaken by Smith (1993) is costly and lengthy, and requires both experience and training in empirical research. Because efficiency is valued in applications of assessment theory (Williamson, 1994) and not very many involved with writing assessment have training in empirical research (White 1994), it is not surprising that there is very little validation research available for particular uses of writing assessment. Exceptions are seen in the work of Blakesley (2003) with Directed Self-Assessment and Herrington on the use of technology using Smith's (1993) and Haswell's (2001) approach to scoring.

VALIDATION STUDIES OF AUTOMATED SCORING

There is really only a single automated scoring engine that has a consistent record of validation research, eRater as it is used to score essays for the Graduate Management Aptitude test. Until recently, the essay portion of the test was read by a group of holistic raters, trained by Educational Testing Service (ETS), the test developer and vendor. The scores are used by graduate programs in business to determine admission to their programs. Like the SAT and the Graduate Record Examination (GRE), the scores are used as one indication of performance in a program of study, along with other indicators, such as class rank, grade-point average (GPA), and the school graduating the applicant. However, the responsibility of the actual validation of each of those examinations lies with the institutions that use them to make decisions about admissions. ETS cannot provide

validation data for any of those examinations because they do not have relevant local data to determine the suitability of each examination for the decision to admit or deny admission to an applicant to a particular program. Validation data, such as national norms and performance of students with self-reported characteristics such as GPA are frequently part of these examinations. But, the only place to determine the validity of admissions decisions is within the institution using the scores. In the case of the SAT, most admissions departments use the scores in formulas to predict such things as first-year GPA. Similarly, ETS reports the success of similar predictions for a number of schools as part of their validation research.

The GRE is now scored by one human rater and eRater. For the most part, ETS has been examining the accuracy of *eRater* in predicting holistic scores from human raters. Their research suggests that eRater is able to predict the scores of six raters with greater accuracy than two human raters. The question, then, is, are the eRater scores any more or less accurate than the scores provided by the two human raters typically used? If the criterion is the more raters the better, then the answer is obviously, yes. The science of psychometrics depends on the sheer magnitude of neumbers in order to statistically prove anything. A traditional direct writing assessment like holistic scoring generates a single score, technically a one item test. Because reliability is greatly improved by the number of scores, it is easy to see how subtly and quickly the question can turn to reliability. In the case of Smith's (1993) accuracy of placement, accuracy focuses on the decision and the underlying principle that all decisions are not equal. *eRater*, however, focuses on the predictive power of one set of procedures compared to another. For validity, the real question for *eRater* is whether the scores help make better decisions about students than the current procedures used by a particular college or university.

For those of us who use traditional holistic scoring procedures, the answer is likely to be that it does, because *eRater* is going to provide more stable scores than two holistic raters. However, the real test of the validity of *eRater* may lie in a comparison with procedures like Smith's that focus on the expert knowledge of teachers who determine whether the student who wrote the essay belongs in their course or the one above or below it. In this case, it is not clear that one procedure has an advantage over the other because there has never been an attempt to examine the relative value of eRater compared to the expert placement model defined by Smith.

Because the immediate question of the validity of automated scoring turns on reliability, as Huot (2002) asserts, reliability has always been the focus of the debate about writing assessment. Thus, the question of which assessment provides the best judgment of a student's placement into a writing program has still

not been answered. As various new assessments have been created (e.g., Broad, 2003; Haswell, 2001; Murphy & Underwood, 1998; Royer & Gilles, 2003), there has been a pressing need to document that these assessments promote valid and reliable educational decisions about students, teachers, and programs. Unfortunately, systematic and rigorous attention is not always given to things like consequences for various participants in the assessment.

For placement, the study of the validity of writing assessment should be focused, like Smith's, on the decision about the best course for a student to enter the writing program at a particular college. Writing exit examination validation research should be focused on a decision about a student's mastery of the curriculum, for both college and school students. Furthermore, there is little reporting of validation research in the assessment literature, in part, I suspect, because writing assessment is a field marginalized by most writing teachers and researchers. Most teachers, with good reason, fear any use of assessment, because assessment has become highly politicized by federal and state government, as well as by local school boards and administrators.

CAN COMPUTERS REPLACE ENGLISH TEACHERS?

Ultimately, one question that may cause an implicit fear is the unspoken potential for the role of automation in education as a whole, not just assessment. Does the future suggest that teachers can be replaced by computers or some evolutionary mutation of them or that one teacher via distance education technology can instruct innumerable students at various locations? One primary question I am attempting to examine is whether automated assessment should be seen as a potential threat or benefit. This fear has been the root of response to automation because automation has typically reduced the workforce in any industry. The curriculum research of the 1970s and 1980s is best summarized as an attempt to find the holy grail of education, a curriculum that is teacher proof, in the sense that the training and experience of a teacher are irrelevant to its success. The tepid results of that search are probably the reason experimental comparison of curriculums disappeared. The most valuable lesson that emerged is the importance of the teacher. Trained, experienced, and motivated teachers are the heart of successful education, despite the public furor over teachers' qualifications. Darling-Hammond and Youngs (2002) examine hundreds of studies about educational progress of various kinds of students and found that the overriding variable, more than ethnicity or income, that predicted student success was the teacher.

Many futurists, both utopian and dystopian, have seen the future filled with intelligent machines. At this stage, Anson's (2003) suggestion may be the best view, there is little that AI can offer a writing teacher. However, our real concern

should be how AI might augment the teaching of writing in the future. Explicit views of the future are not of much value, particularly because the likelihood of automation replacing some aspects of teaching writing is already evident, as we have been seeing, the continuing use of electronic technology to compliment or replace some of the work of teachers. As we have also experienced, there will be those who claim that computers allow for greater efficiency, justifying increasing the numbers of students working with individual teachers. It seems clear that computers are here to stay in English Studies, even if only as word processors to make the production of paper text easier and as communication devices to connect writers to one another for responding. We have to expect that the future will also hold some developments that can help us and some that can be hurtful. Some developments will be faddish, oversold by developers and producers of the technology, whereas others will enter our toolbox with the potential to help students learn if used properly. My answer to the problem of automated assessment is precisely the last point. Its potential suggests that it might have some value in writing classrooms, but it is not clear what that may be. Second, if it does have value, it will take continuing study understand the consequences and to establish the value through validity inquiry.

I am suggesting a stance on automated assessment that can best be characterized as carefully directed critique toward the developers of automated assessment. Because Pennsylvania has adopted automated assessment and the results of that automation will be used to determine funding for school districts, there is no question it is being used in regulatory ways. Why should we expect anything different? Assessment has been used as a gate keeper for as long as assessment has resulted in excluding some and including others in schooling.

Out-of-hand or outright rejection of automated assessment, a blanket condemnation, can only be self-serving. More importantly, we need to examine the use of automated scoring as we would any other assessment, according to the criteria of the most current theories on validating educational assessment. Arguing that theories of literacy do not justify the use of automated assessment, is similar to earlier arguments that indirect assessment does not have content validity. This argument is not going to be compelling with an educational measurement audience, not to mention policymakers and regular citizens. Furthermore, without an understanding of the common language of assessment as it is grounded in the social sciences research methodology, we will find that our righteous indignation, our hermeneutic arguments about the meaning of new types of assessment, are met by a wondering stare, at best, and a dismissive glare, at worst.

What I am arguing we do is to study automated assessment in order to explicate the potential value for teaching and learning, as well as the potential harm. The theory of the developers can itself be used as a ground for validity

arguments. However, we have to be willing to look outside our field to understand the theory of another, a theory that has clearly been at the heart of assessment practices in our culture for more than a century, and an industry that has become embedded in education in America over the last 50 years. The practices are accepted by most Americans as valid for use in education. If educators have not been successful in opposing those practices, it may be that we have not been able to understand what drives them and to be able to offer critiques that have been seen as questioning that validity.

CONCLUSION

Writing assessment in American education has two professional groups with developed bodies of theory and practice. The first group, whose primary interest is assessment, is the membership of the two professional organizations, the American Education Research Association (AERA) and the American Psychological Association (APA). They far outnumber the members of the second group, the membership of the National Council of Teachers of English and College Composition and Communication. For a number of years, APA and AERA were loosely allied through members with dual memberships. More recently, recognizing their common concerns and shared field, they began to work together. The result is a clearly defined statement of definitions and standards for test development and validation (AERA, 1999). Although the measurement community is not inherently hostile to the concerns of writing teachers, its members will be looking for the kinds of evidence articulated in the standards, applying the technology of validation research to the discussion of implementing automated scoring. Furthermore, their direct involvement with public education, as the primary source for assessment tools, lends them a strong voice in the federal, state, and local politics of assessment.

The contrasts between English Studies and educational assessment are many, running beyond concepts or methodology. The common ground is also quite large. One important point of comparison lies in the question of what constitutes important research in the two fields. In English, researchers are typically expected to demonstrate their mastery of the field in publications that are authored by a single individual. In assessment, as in most scientific fields, important research can only be conducted by a team of people, each contributing to the conceptualization and execution of the study. If it is time to examine the research methodology or social sciences as it impinges on assessment, it may also be time to explore the potential for collaborative research, not just within either a social science or humanistic tradition (see Huot, 2002, for a discussion of a unified field of writing assessment). If we continue to espouse outmoded

views of assessment, to fail to understand the complexity of validity theory, for instance, we are going to be frustrated at every turn. If for no other reason, as Sun Tsu (1994) observed, one has to know the enemy to defeat him. In this case, I hope knowing one's enemy might lead to a productive alliance.

A student of mine was attempting to articulate a complex problem for her dissertation project, one involving the value of historical study of the field. She finally told me that she recognized she was approaching the project with the wrong attitude. She said that she had forgotten a couple of the basic things she tries to teach her students: Who is the audience and what kinds of rhetorical practices are expected?

Who is our audience for our critique of automated scoring? If it is ourselves, we can continue to confront assessment developers with the challenge that their work does not conform to contemporary theories of literacy. However, when they suggest that contemporary theories of literacy are at the basis of their work, our best critique lies in a close examination of the theory, as opposed to an examination of the practice itself. Surely, well-directed critique is more successful than blanket condemnation. But, such critique emerges from the study of assessment theory, validity theory in particular. Such a critique is supported by those theories, if we take the time to use our own research skills, interpretive reading of culture icons, such as the texts of the field.

I will leave you with a story that has guided my work in the use of technology in my classroom and the suggestions that I give to others: In graduate school, I shared an apartment with a fellow student. At the time, he was working as a welder for a local company building automobile transport trailers. One day, he come in from work telling me that he had been let go. His schedule was flexible, built around his class schedule at the university. His boss had told him that the computer was not able to work with his schedule, so he had to either work full time or leave. He left and went on to accomplish some fine work in our field. However, I have adopted as a basic principle of working with computer analysts and programmers, "If your program does not do what we need it to do, you have done a poor job, go back and fix it!" The goals of people must drive the development of automation, not the automation itself. We have to find the right way to say, "Fix it!" The real trick is to get the right people to listen. As inheritors of the tradition of rhetoric, writing teachers should know more about how to speak to their audiences.

REFERENCES

American Educational Research Association, American Psychological Association, & National Council of Measurement in Education. (1999). *Standards for educational and psychological testing*. American Educational Research Association.

Ajay, H. B., Tillett, P. I., & Page, E. B. (1973). Analysis of essays by computer (AEC-II). Final report to the National Center for Educational Research and Development (Project No. 80101), p. 231.

Anastasi, A. (1976). *Psychological testing* (4th ed.). Macmillan.

Anson, C. R. (2003). Responding to and assessing student writing: The uses and limits of technology. In P. Takayoshi & B. Huot (Eds.), *Teaching writing with computers* (pp. 234- 246). Houghton Mifflin.

Berlin, J. A. (1984). *Writing instruction in nineteenth-century American colleges*. Southern Illinois University.

Blakesley, D. (2003). Directed self-placement in the university. In D. Royer & R. Gilles (Eds.), *Directed self-placement: Principles and practices* (pp. 31-48). Hampton Press.

Bleich, D. (1975). *Readings and feelings: An introduction to subjective criticism*. National Council of Teachers of English.

Breland, H. M., Camp, R, Jones, R. J., Morris, M. M., & Rock, D. A. (1987). *Assessing writing skill.* College Entrance Examination Board Research Report No. 11. Educational Testing Service.

Broad, B. (2003). *What we really value: Beyond rubrics in teaching and assessing writing.* Utah State University Press.

Burns, H. L. (1979). *Stimulating rhetorical invention in English composition through computer-assisted instruction*. Dissertation Abstracts International, DAI-A 40/70, p. 3734, January 1980, DAI Order number AAT 7928268.

Burstein, J. C. (2003). The E-rater® Scoring Engine: Automated essay scoring with natural language processing. In M. D. Shermis & J. C. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 113-121). Erlbaum.

Carmines, E. G., & Zeller, R. A. (1980). *Reliability and validity assessment*. Sage Publications.

Cooley, W. W., & Lohnes, P.R. (1976). *Evaluation research in education: Theory, principles, and practice*. Irvington.

Coombs, D. H. (1969). Review of *The Analysis of Essays by Computer*, by Ellis B. Page and Dieter H. Paulus. *Research in the Teaching of English*, *3*(2), 222-228.

Cooper, C. R., & Odell, L. (1977). *Evaluating writing: Describing, measuring, judging*. National Council of Teachers of English.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed.) (pp. 443-507). American Council on Education.

Cronbach, L. J. (1988). Five perspectives on test validity argument. In H. Wainer (Ed.), *Test validity* (pp. 3-17). Erlbaum.

Cronbach, L. J. (1989). Validity after thirty years. In R. Linn (Ed.), *Intelligence: Measurement theory and public policy* (pp. 147-171). University of Illinois Press.

Cureton, E. (1951). Validity. In E. F. Lindquist (Ed.), *Educational measurement* (pp. 621-694). American Council on Education.

Darling-Hammond, L., & Youngs, P. (2002). Defining "highly qualified teachers": What does "scientifically-based research" actually tell us? *Educational Researcher*, *31*(9), 13-25.

Elbow, P., & Yancey, K. B. (1994). On the nature of holistic scoring: An inquiry composed on email. *Assessing Writing, 1(1)*, 91-108.

Gottshalk, F. I., Swineford, F., & Coffman, W. (1966). *The measurement of writing ability.* College Entrance Examination Board Research Monograph N. 6. Educational Testing Service.

Guilford, J. P. (1954). *Psychometric methods.* New York: McGraw-Hill.

Gulliksen, H. (1950). *Theory of mental tests*. Wiley.

Haswell, R. (Ed.). (2001). *Beyond outcomes: Assessment and instruction in a university writing program.* Ablex.

Herrington, A., & Moran, C. (2001). What happens when machines read our students' writing? *College English, 63(4)*, 480-499.

Huot, B. (1993). The influence of holistic scoring procedures on reading and rating student essays. In M. M. Williamson & B. A. Huot (Eds.), *Validating holistic scoring for writing assessment* (pp. 206-236). Hampton Press.

Huot, B. (1996). Computers and assessment: Understanding two technologies. *Computers and Composition, 13(2),* 231-244.

Huot, B. (2002). *(Re)Articulating writing assessment for teaching and learning.* Utah State University Press.

Johnson-Laird, P. N. (1977). *Mental models: Towards a cognitive science of language, inference, and consciousness.* Harvard University Press.

Kuhn, T. S. (1996). *Structure of scientific revolutions* (3rd ed.). University of Chicago Press.

Macrorie, K. (1969). Review of The Analysis of Essays by Computer by Ellis B. Page and Dieter H. Paulus. *Research in the Teaching of English, 3(2)*, 228-236.

Messick, S. (1989). Test validity. In R. Linn (Ed.), *Educational measurement* (3rd ed.) (pp. 13- 103). American Educational Research Association; National Council of Measurement in Education.

Moss, P. (1998). The role of consequences in validity theory. *Educational measurement: Issues and practices, 17*(2), 6-12.

Murphy, S., & Underwood, T. (1998). Interrater reliability in a California middle school English/language arts portfolio assessment program. *Assessing writing, 5*(2), 201-230.

Page, E. B. (1966. The imminence of grading essays by computer. *Phi Delta Kappan,* 238-243.

Page, E. B. (1967a). Grading essays by computer: Progress report. Proceedings of the 1966 Invitational Conference on Testing (pp. 87-100). Educational Testing Service.

Page, E. B. (1967b). Statistical and linguistic strategies in the computer grading of essays. Proceedings of the Second International Conference on Computational Linguistics. Grenoble, France.

Page, E. B. (1985). Computer grading of student essays. In T. Husén & Postlethwaite (Eds.), International Encyclopedia of Educational Research (pp. 944-946). Pergamon.

Page, E. B. (1993. New computer grading of student prose, using a powerful grammar checker. [Paper presentation]. Annual meeting of the North Carolina Association for Research in Education. Greensboro, NC.

Page, E. B. (1995). Computer grading of essays: A different kind of testing? [Invited address] American Psychological Association, Divisions 5, 7, 15, 16.

Page, E. B., Fisher, G. A., & Fisher, M. A. (1968). Project Essay Grade: A FORTRAN program for statistical analysis of prose. *British journal of mathematical and statistical psychology, 21*(1), 139.

Page, E. B., Tillett, P. I., & Ajay, H. B. (1989). Computer measurement of subject-matter essay tests: Past research and future promise. Proceedings of the First Annual Meeting of the American Psychological Society, Alexandria, VA.

Pula, J., & Huot, B. (1993). A model of background influences on holistic raters. In M. M. Williamson & B. A. Huot (Eds.), *Validating holistic scoring for writing assessment* (pp. 237- 265). Hampton Press.

Royer, D., & Gilles, R. (2003). *Directed self-placement: Principles and practices.* Hampton Press.

Schendel, E. (1999). Exploring the theories and consequences of self-assessment through ethical inquiry. *Assessing writing, 6*(2), 199-227.

Smith, W. L. (1993). Assessing the reliability and adequacy of using holistic scoring of essays as a college composition placement technique. In M. M. Williamson & B. A. Huot (Eds.) *Validating holistic scoring for writing assessment* (pp. 142-205). Hampton Press.

Stalnaker J. M. E. (1951). The essay type of examination. In E. F. Lindquist (Ed.), *Educational Measurement* (pp. 495-532). American Council on Education.

Sun Tsu. (1994). *The art of war.* Westview.

Terman, L. M. (1916). *The measurement of intelligence.* Houghton Mifflin.

Veal, R. A., & Hudson, S. A. (1983). Direct and indirect measures for the large-scale evaluation of writing. *Research in the Teaching of English, 17*(3), 285-296.

White, E. M. (1994). *Teaching and assessing writing: Recent advances in understanding, evaluating, and improving student performance.* Jossey-Bass.

Williamson, M. M. (1993). An introduction to holistic scoring: The social, theoretical, and historical context for writing assessment. In M. M. Williamson & B. A. Huot (Eds.), *Validating holistic scoring for writing assessment* (pp. 1-43). Hampton Press.

Williamson, M. M. (1994). The worship of efficiency: Untangling theoretical and practical consideration in writing assessment. *Assessing writing, 1*(2), 147-174.

Wolfe, E. W. (1997). The relationship between essay reading style and scoring proficiency in a psychometric scoring system. *Assessing writing, 4*(1), 83-106.

Yancey, K. B. (1999). Looking back as we look forward: Historicizing writing assessment. *College Composition and Communication, 50*(3), 483-503.

CHAPTER 9.

CRITIQUE OF MARK D. SHERMIS AND BEN HAMNER, "CONTRASTING STATE-OF-THE-ART AUTOMATED SCORING OF ESSAYS: ANALYSIS"

Les C. Perelman
MIT

Although the unpublished study by Shermis & Hamner (2012) received substantial publicity about its claim that automated essay scoring (AES) of student essays was as accurate as scoring by human readers, a close examination of the paper's methodology demonstrates that the data and analytic procedures employed in the study do not support such a claim. The most notable shortcoming in the study is the absence of any articulated construct for writing, the variable that is being measured. Indeed, half of the writing samples used were not essays but short one-paragraph responses involving literary analysis or reading comprehension that were not evaluated on any construct involving writing. In addition, the study's methodology employed one method for calculating the reliability of human readers and a different method for calculating the reliability of machines, this difference artificially privileging the machines in half the writing samples. Moreover, many of the study's conclusions were based on impressionistic and sometimes inaccurate comparisons drawn without the performance of any statistical tests. Finally, there was no standard testing of the model as a whole for significance, which, given the large number of comparisons, allowed machine variables to occasionally surpass human readers merely through random chance. These defects in methodology and reporting should prompt the authors to consider formally retracting the study. Furthermore, because of the widespread publicity surrounding this study and because its findings may be used by states and state consortia in implementing the Common Core State Standards, the authors should make the test data publicly available for analysis.

DOI: https://doi.org/10.37514/PER-B.2024.2166.2.09

On April 16, 2012, Mark D. Shermis, Dean of the School of Education at the University of Akron, presented a paper at the annual meeting of the National Council on Measurement in Education on "Contrasting State-of-the-Art in Automated Scoring of Essays: Analysis." Despite its fairly nondescript title, the paper claimed that machines graded essays as well as expert human raters, a claim that was publicized in various press releases and newspaper articles. A press release from the University of Akron, for example, stated, "A direct comparison between human graders and software designed to score student essays achieved virtually identical levels of accuracy, with the software in some cases proving to be more reliable, a groundbreaking study has found" (Man and Machine, 2012). A headline in Inside Higher Ed read, "A Win for the Robo-Readers," and the story included statements such as the following:

> The study, funded by the William and Flora Hewlett Foundation, compared the software-generated ratings given to more than 22,000 short essays, written by students in junior high schools and high school sophomores, to the ratings given to the same essays by trained human readers. The differences, across a number of different brands of automated essay scoring software (AES) and essay types, were minute. (Kolowich, 2012)

Even the venerable British publication, *The New Scientist*, reported "The essay marks handed out by the machines were statistically identical to those from the human graders, says [Jaison] Morgan. 'The result blew away everyone's expectations,' he says." (Giles, 2012) Yet these reports and other statements can best be characterized as unsubstantiated overstatement. The study, however, employs an inconsistent and questionable methodology that favors the machines over the human graders. Even with these biased procedures and results, the data still give some, but lacking the full test essay sets, inconclusive indication that in actual assessments of writing, human scorers were more reliable than machines.

The study derived from the Automated Student Assessment Prize (ASAP), a competition sponsored by the William and Flora Hewlett Foundation, to assess the efficacy of automated scoring engines. The competitions involved evaluating essays from statewide assessments that had already been scored by human readers. Phase One, which dealt with "long-form constructed responses" (although over half of the responses were essentially paragraphs) had two parts. The first involved scoring engines developed by nine testing companies such as the Educational Testing Service, Pearson Knowledge Technologies, and CTB/McGraw-Hill. The second competition was an open contest among software developers. The Shermis and Hamner paper reports on only the first part, the performance of the nine vendors.

A version of the study has subsequently been published as a chapter (Shermis & Hamner, 2013) in the *Handbook of Automated Essay Scoring* co-edited by Mark D. Shermis, the lead author of the original paper, and Jill Burstein (2013). Other contributors in the same volume state explicitly that the study showed Automated Essay Scoring is capable of producing scores similar to those of human readers. All of these studies referenced the original Shermis and Hamner paper. It is the paper's central claim, articulated in the abstract, that has elicited so much publicity: The results demonstrated that overall, automated essay scoring was capable of producing scores similar to human scores for extended-response writing items with equal performance for both source-based and traditional writing genre [sic].

That central claim, however, is clearly not supported by the data. Conversely, the data provide some, although not conclusive, support for the assertion that human scorers performed more reliably than the machines, especially on longer papers that were scored for writing ability rather than solely on content.

NO DEFINITION OF WRITING CONSTRUCT

One major problem with the study is the lack of any explicit construct of writing. Without such a construct, it is, of course, impossible to judge the validity of any measurement. Writing is foremost a rhetorical act, the transfer of information, feelings, and opinions from one mind to another mind. The exact nature of the writing construct is much too complex to outline here; suffice it to say that it differs fundamentally from the Shermis and Hamner study in that the construct of writing cannot be judged like the answer to a math problem or GPS directions. The essence of writing, like all human communication, is not that it is true or false, correct or incorrect, but that it is an action, that it does something in the world. That is what sophists like Protagoras and philosophers, most notably, Aristotle, noted in classical times, and more recently "ordinary language" philosophers like J. L. Austen (1962) and H. Paul Grice (1989), and linguists such as Dan Sperber and Deirdre Wilson (1990), have made apparent in current discussions of language use.

The seriousness of this lack of definition of the writing construct in the study manifests itself in various defects and confusions, beginning with the heterogeneous mix of papers that constitute the study. The study is based on a corpus of eight different essay sets that come from six different states. Each essay set contains a prompt, training information that includes rubrics and in some cases annotated or unannotated training samples along with other materials. Sixty percent of the total papers from each data set were publicly available with scores as training sets. The remaining 40% of papers were divided into two test sets of

20% each. Consequently, of the total sample size of 22,029 papers, only 4,343 papers in eight different essay sets comprised the actual test sets, ranging from 304 to 601 papers each.

However, half the essay sets and over half the aggregate number of papers in the test set were not evaluated on any construct connected to writing. The study defines the four essay sets #3-#6 as source-based writing assignments. Source-based writing assessments measure student writing in response to specific texts or data. A prominent example of this kind of assessment is the Document-Based Question in the Advanced Placement Language and Composition Examination (Perelman, 2008). Students are given a passage or a short essay and then asked to write an argument or analysis about it. Unlike the rubrics that govern the scoring of essay-sets in the Shermis study, the rubrics for these AP Examinations emphasize writing skills such as organization, argument, and expression as well as a student's mastery of content. Essay sets #3-#6, on the other hand, contain prompts and rubrics that are not based on document or source-based writing, however, but are content-based or content dependent exercises that are scored solely on the understanding of content rather than any assessment of writing ability.

Two of these essay sets, #3 and #4, are focused solely on literary analysis. Essay set #3 consists of responses to a prompt based on "Rough Road Ahead" by Joe Kurmaskie: "Write a response that explains how the features of the setting affect the cyclist. In your response, include examples from the essay that support your conclusion." Essay set #4 consists of responses to a prompt based on "Winter Hibiscus" by Minfong Ho. The prompt repeats the last paragraph of the story and then asks students to "Write a response that explains why the author concludes the story with this paragraph. In your response, include details and examples from the story that support your ideas" (see Appendix A).

The rubrics, based on a scale from 0-3, are identical for each of these two essay sets (#3 and #4). The score of 3, the highest score, is defined in each of the rubrics by the following language:

Score 3: The response demonstrates an understanding of the complexities of the text.

- • Addresses the demands of the question
- • Uses expressed and implied information from the text
- • Clarifies and extends understanding beyond the literal

The rubrics and other materials for essay sets #5 and #6 explicitly define them as reading tests, while defining a different scale for writing tests. (See Appendix A; Kaggle-Data 2012.) The prompt for essay set #6 is based on an excerpt

discussing the obstacles to putting a mooring mast for dirigibles on top of the Empire State Building:

> Based on the excerpt, describe the obstacles the builders of the Empire State Building faced in attempting to allow dirigibles to dock there. Support your answer with relevant and specific information from the excerpt.

Immediately following the rubric, which focuses on ability to understand the content, not on ability to write, were these scoring notes.

The obstacles to dirigible docking include:

- Building a mast on top of the building
- Meeting with engineers and dirigible engineers
- Transmitting the stress of the dirigible all the way down the building; the frame had to be shored up to the tune of $60,000
- Housing the winches and other docking equipment
- Dealing with flammable gases
- Handling the violent air currents at the top of the building
- Confronting laws banning airships from the area
- Getting close enough to the building without puncturing

I took my first and last programming class in 1967, learning Fortran IV. Even with my rusty recollection of an antique programming language, I am fairly confident that I could construct a program that could also do very well scoring these essays simply by counting strings of key words and phrases, including synonyms. Such a program, however, would not in any way be assessing writing.

LACK OF A CONSTRUCT FOR THE ESSAY

The majority of essay sets in this study are single paragraphs, not essays. Although the study explicitly stated that it was exploring how well machines could grade extended-response writing, (i.e., essays), only three of the eight data sets consisted of what is commonly defined as extended-response writing. The mean number of words for essay sets #3, 4, 5, 6, and 7, ranged from 98.70 to 173.43 words. A typed double-spaced page of prose with 12-point type is approximately 250 words. (Note: the number of words in this paragraph up to and including this note is 96.)

Only three of the eight studies have average word lengths of over 360 words, approximately 1.5 typed double spaced pages. One of the senior managers of the project reported in an email that the team spent three months asking every state for sets of long-form constructed response essays, and even requested

papers from international sources (J. Morgan, personal communication, June 19, 2012). He admitted that some of the essay sets defined as "essays" in the study were shorter than the length the team desired, but defended the sample by arguing that even these short pieces of writing are categorized as essays by the states, meaning that at least one of the fifty states defines it as an essay. These essays, however, and the way that they are scored, may be unrepresentative of state writing assessments as a whole, especially since the training materials for essay sets #5 and #6 specifically differentiate these assessments from assessments of writing ability (See Appendix A). To further muddle the study, all of the paragraph markings were removed (Shermis & Hamner, p. 8), converting even the longer, multi-paragraph essays into single paragraphs, making any evaluation of key essay features such as paragraph coherence impossible. In sum, only the scores of three collections of essays out of eight in this study represent actual measures of multi-paragraph writing ability. Moreover, even in those three collections of essays, there were no paragraph markings for computers to use to judge development and coherence - only block text.

FLAWED EXPERIMENTAL DESIGN I: APPLES AND ORANGES

The study employs different methods for humans and machines in computing reliability for all of its measures. This highly unorthodox and statistically indefensible approach substantially and unfairly biases the measures in favor of the machines, artificially increasing their reliability compared to that of the human readers. For all of the measures beginning with those displayed in Table 8 in Appendix B, the study uses the measure H1H2, the comparison between the two human readers, as the measure for reader reliability, while the measure for machine reliability is the comparison between the machines and a construct, the resolved score (RS). (To avoid confusion, tables in this Critique are labeled with letters while tables in Appendix B taken from the Shermis and Hamner Report are labeled with numbers.)

In most essay testing situations, the standard practice is that the resolved score is the sum of the two reader scores if the scores are identical or adjacent; or, if the scores vary by more than one point, the resolved score is established by one or two supervisors rereading the essay. Yet, only one of the essay sets in the study, #1, follows the standard best practice of combing two equal or adjacent scores to compute the RS. Essay set #7 combines composite scores regardless of the size of the difference between them. Essay set #8 also appears to combine composite scores regardless of the size of the difference but has a third reader adjudicating 17.7% of the RS's randomly. Essay set #2 uses the score of only the first reader as the RS, regardless of the second reader's score. The remaining four essay sets,

#3, #4, #5, and #6, all compute the RS as the higher of the two scores.2 This procedure, followed by four essay sets - half of the total number - and containing 55% of the essays in the aggregate data sample, skews many of the measures used in favor of the AES scores.

Before going into a more technical explanation, the bias produced by using these two different measures can be best illustrated by a hypothetical example. A company is hiring an additional reader to score essays. It has two applicants for the position and will select the applicant that has the greatest reliability in scoring. A reader who already works for the company has scored all of the essays. The first applicant scores the essays and her reliability is determined by comparing her scores to those of the first reader. The second applicant, however, is told before scoring the essays that his reliability will be determined by comparing his score to the higher of the two previous readers' scores if the scores differ. He realizes that his chances improve dramatically simply by always selecting the higher score in any case in which he is wavering between two scores. He does so, scores more reliably, and gets the job. Clearly, the procedure was biased in favor of the second applicant. This example is completely analogous to the procedure used in the study for essay sets #3, #4, #5, and #6, which are biased towards the machines. If such a procedure as described in the scenario were actually implemented in the real world, it would clearly be an unfair hiring practice. Similarly, this practice used in the Shermis and Hamner study unfairly biases and therefore invalidates half the results.

Essays scores, be they holistic, trait, or analytical, always are continuous variables, not discrete variables (integers), even though graders almost always have to give integer values as scores. The report recognizes this fact in the observation on page 24 that values for the Pearson r "might have been higher except that the vendors were asked to predict integer values only." Each reader has to select a single integer value even though some essays might be on the border between two adjacent integers. Some 3's on a 4-point-scale might be very high 3's bordering on a 4, while other 3's may be very low 3's bordering on a 2. Significantly, some of the training materials for the essay sets included essays scores with plus and minus signs. In the terminology of Classical Test Theory, the True Score might be 3.3 or 2.8. Consequently, adjacent agreement in the correct direction between two readers (e.g. one rater gives an essay a score of 3 and the second rater gives the essay a score of 4) will more closely approximate a True Score of 3.4 than two scores of 3.

Resolving scores merely by selecting the higher one ignores the continuous nature of the scores being measured and penalizes human raters while giving AES algorithms a substantial advantage by allowing them to optimize agreement with the RS by rounding up just like the example of the second job applicant. In the case of the essay that has a True Score of 3.4, for example, there are four likely pairs of scores that would be produced by two human raters: 3-3, 3-4,

4-3, and 4-4. Note that in three of these four cases, selecting the higher score makes 4 the resolved score, and that in two of these three instances one of the two reader scores will be lower than the resolved score. This unjustified score bias can be observed in the data in the tables at the end of the report. Table 9.1 and Table 4 of the study (Appendix B) display the means for the five score sets that ether combine rater scores to compute RS (#1, #7, and #8) or use a single score as the RS (#2A and #2B). In these essay sets, the resolved score does not bias the results against the human readers. Significantly, the means of the human reader scores match the means of the RS much more closely than those of the means of the machine scores. Table 9.2 displays the means for the essay sets that used the higher human rater score as the resolved score. The contrast in the differences between the human reader means and the resolved score means in the two tables is striking and provides a powerful illustration of how computing the RS as the higher scores skews the results against human readers.

Table 9.1. Test Set Mean for Resolved Scores=Sum of Scores or Single Score

	H1	H2	RS	Diff. Avg. Human Rater Means from RS Means	Range of AES Mean Scores	Range of Diff. of AES Mean Scores from RS Means
1	8.61	8.62	8.62	-0.01	8.49-8.80	-0.13-0.18
2A	—	3.39	3.41	-0.02	3.33-3.41	-0.08-0.00
2B	—	3.34	3.32	0.02	3.18-3.37	-0.14-0.05
7	20.02	20.24	20.13	0.00	19.46-20.05	-0.67- -0.08
8	36.45	36.70	36.67	-0.09	37.04-37.79	.037-1.12

Separating the essay sets into two groups, those that use a single human score or a sum of two human scores to compute the resolved score and those that use the higher score as the resolved score, present two very different sets of values of the metrics used in the study. Tables 9.1 & 9.2 also demonstrate the substantial difference for means.

Table 9.2. Test Set Mean for Resolved Scores=Higher Human Rater Scores

	H1	H2	RS	Diff. Avg. Human Rater Means from RS Means	Range of AES Mean Scores	Range of Diff. of AES Mean Scores from RS Means
3	1.79	1.73	1.90	0.14	1.84-1.95	-0.06-0.05
4	1.38	1.40	1.51	0.12	1.34-1.57	-0.17-0.06
5	2.31	2.35	2.51	0.18	2.44-2.54	-0.07-0.03
6	2.57	2.58	2.75	0.18	2.54-2.83	-0.04-0.08

Similar distinctions can be shown in the other tables. Indeed, in the five measures of agreement, exact agreement (Table 8), exact and adjacent agreement (Table 10), Kappas (Table 12), Quadratic Weighted Kappas (Table 14) and the Pearson r (Table 16), the human raters in the group of essay sets clearly outperform the AES engines in the first three and have mixed results for the Quadratic Weighted Kappa and Pearson r. Curiously, for the Quadratic Weighted Kappa (Table 14) the relationship of the two groups is inverted - human raters in two of the four essay sets that use the higher score as the resolved score (#3 and #4) as well as score sets #2A and #2B outperform the AES engines while AES engines outperform human raters in the other essay sets. This anomaly may partially be an artifact of the Quadratic Weighted Kappa measuring correspondence not between two raters, as is its intended use, but between a rater (i.e., the machine score) and the artificial construct of the resolved score as higher of the two scores. Another possible explanation is offered by Brenner & Kliebsch (1996) who noted that that quadratically weighted kappa coefficients tend to increase with larger scales while unweighted kappa coefficients decrease. They noted that "variation of the quadratically weighted kappa coefficient with the number of categories appears to be strongest in the range from two to five categories" (p. 201). As displayed in Table 3 of the report, the scales for essay sets #3 and #4 consisted of a scale of four (0-3), while essay sets #5 and #6 consisted of a scale of five (0-4). With the exception of the four point scale for score set #2B, all the other essay sets had scales greater than five. For essay set #1 the range of the rubric was 1-6 and the range of the resolved score was 2-12. For scoring set #2A, the range was 1-6; for scoring set #7, the range of the rubric was 0-12, and the range of the resolved score was 0-24. For essay set #8, the range of the rubric was 0-30, and the range of the resolved score was 0-30.

The confusion between human scores and resolved score is found throughout the text. The report states, for example, on page 22, "all vendor engines generated predicted means within 0.10 of the human mean for Essay set #3 which had a rubric range of 0-3." The report, however, is referring to the mean of the resolved score not the mean of the human raters, which were, in actuality, lower than the resolved score by 0.11 and 0.17 respectively. (See Table 9.2)

The standard method for comparing the reliability of machine scores to human scores is to compare the reliability of the machine scores to each of the two human scores and then compare those scores to reliability of the human scorers to each other (McCurry, 2010). In McCurry's study, as in many others, humans clearly outperformed machines. Yet the Shermis and Hamner study instead chose to use different variables for humans and machines.

The two readers' individual scores compared to the resolved score (H1 and H2) are consistently higher than those of the machine scores for all of the metrics displayed in all of the tables (Appendix B). This phenomenon could well be an artifact of the individual reader score being a contributing element to the resolved score. However, of the nine score sets, the two scores of H2, the second human reader, for #2A and #2B are completely independent of the resolved score because reader H1 defined the resolved score and H2's scores were used only for computing grading reliability. Consequently, in essay sets #2A and #2B the human reader score and the machine scores are compared to the same measure. That the human rater in essay sets #2A and #2B outperformed all of the machines in every metric except for one machine in Pearson r correlation offers some evidence that the high individual reader scores compared to the resolved score are not solely an artifact of their being a part of the whole. As shown in Table 9.3, #2A, which measured ideas, content, organization, style, and voice, had an exact agreement value of 0.76, compared to the range of machine values of 0.55-0.70. Its Kappa was 0.62, compared to the range of machine values of 0.30-0.51. Its Quadratic Weighted Kappa was 0.80, compared to the range of machine values of 0.62-0.74. And its Pearson r was 0.73, compared to the range of machine values of 0.62-0.74. Similarly, #2B, which measured conventions of grammar, usage, punctuation, and spelling, had an exact agreement value of 0.73 compared to the range of machine values of 0.55-0.69. Its Kappa was 0.56, compared to the range of machine values of 0.27-0.49. Its Quadratic Weighted Kappa was 0.76, compared to the range of machine values of 0.62-0.74. And its Pearson r was 0.76, compared to the range of machine values of 0.55-0.71. Significantly, the prompt in essay set #2 was a traditional argumentative prompt.

Table 9.3. Essay Set #2-H2 Score Compared to Resolved Score vs. Machine Scores

Metric	**2A**		**2B**	
	H2	**2A Range of Machine Scores**	**H2**	**2B Range of Machine Scores**
Exact Agreement	0.76	0.55-0.70	0.73	0.55-0.69
Kappa	0.62	0.30-0.51	0.56	0.27-0.47
Quadratic Weighted Kappa	0.80	0.62-0.74	0.76	0.62-0.74
Pearson *r*	0.73	0.62-0.74	0.76	0.55-0.71

In sum, the use of the artificially inflated resolved scores skews any meaningful analysis. This is particularly serious in the case of the quadratic weighted kappa, which is meant to compare the scores of two autonomous readers, not a reader score and an artificially resolved score. (Sim & Wright, 2005). In a subsequent report (Morgan, Shermis, Van Deventer, & Vander Ark, 2013), this same team uses the quadratic weighted kappa as the single measure of "the concordance between hand scores and machine scores" (p. 11), apparently unaware that they were measuring the concordance between resolved scores and machine scores and that in addition to all the other problems associated with employing resolved scores, the quadratic weighted kappa is an inappropriate measure for such a comparison.

The design of the study allows random chance to produce some seemingly impressive machine scores. Having a pair of readers compete against nine scoring engines is, in essence, like running multiple T-tests or any other kind of multiple comparisons. An occurrence can appear significant but might just be a lucky random occurrence. Any single high machine score among the nine scores by nine different vendors compared by five different metrics - that is 405 individual measures - could possibly be a random anomaly, or to put it in more colloquial terms, a lucky guess, especially since the size of the individual test essay sets were relatively small, ranging from 304 to 601. Statisticians have long known the dangers of producing what is called a Type I Error or False Positive when there are multiple comparisons without any overall testing of the entire model. When deciding if the difference between two variables is significant or possibly due to random chance, the standard statistical practice is to require that the probability of the difference being a product of random chance less than 5% or 1 in 20. But with repeated instances or comparisons, the probability of producing one or more statistically significant events increases. The chance of rolling two dice and getting two sixes is one in thirty six or 2.7%, but if I roll the dice thirty times, there is over a 50% chance I will roll two sixes. Although the case of comparisons in the study is slightly different, the basic analogy holds. The comparison of 405 measures to the resolved scores will produce some high correlations merely by chance.

The standard methodology to prevent these kinds of errors is to perform a test of the model as a whole. Unfortunately, no such tests were performed in the Shermis and Hamner study. Indeed, although various claims were made in the paper, no tests of statistical significance were reported by the authors. Instead, the authors present impressionistic assertions such as "In general, performance on kappa was slightly less with the exception of essay prompts #5 & #6. On these data sets, the AES engines, as a group, matched or exceeded

human performance" (p. 23). There are no parameters given on what constituted matching or exceeding human performance.

Eleven months after the paper was presented and widely publicized, the lead author was quoted in the press as stating that he did not perform a regression analysis or any other statistical tests on the data in his study because that was one of conditions imposed upon him by major vendors of essay grading software, including McGraw-Hill and Pearson (Rivard, 2013). Such conditions were not disclosed in the Methods Section of the original paper, even though disclosures of such externally imposed restraints is standard practice in academic publications and especially in empirical studies such as this one.

FAULTY ANALYSIS: SMOKE AND MIRRORS

Overall, the analysis minimizes the accuracy of the human scorers and over-represents the accuracy of machine scoring. The clearest omission is the failure of the authors to report the fairly large percentage of machine values for the Pearson r and the Quadratic Weighted Kappa that fell below the minimum standard of 0.7. This value is used as the absolute minimum because the shared variance or what the machine clearly predicts is the square of that number, or approximately half of the population (Schultz, 2013; Ramineni & Williamson, 2013; Williamson, Xi, & Breyer, 2012). Any value below 0.7 will be predicting significantly less than half the population and, because this is an exponential function, small decreases in value produce large decreases in the percentage accurately predicted. A Pearson r of 0.6, for examples, yields a shared variance or predictive value of 0.36 or slightly more than one-third. A value of 0.5 yields 0.25 or one-quarter of the population. Yet for the Quadratic Weighted Kappa, 28 of the 81 machine scores, 35.6%, are below the minimally acceptable level of 0.7, even though the machines had the advantage in half of the essay sets of matching an inflated Resolved Score. In contrast, the human readers, who had to match each other with no artificial advantage, had only one Quadratic Weighted Kappa below 0.7, for the composite score on essay set # 8 or only 1 out of 9 or 11.1%. Similarly, for the Pearson r, the human readers again had a value below 0.7 for essay set #8 or 11.1% while 23 out of 91 of the machine scores or 28.4% were below the minimum threshold of 0.7.

The claim that the analysis in the paper unfairly underrepresents the performance of the human readers is further substantiated by the comparison between human readers and machine scores condensed in Tables 9.4-9.7. I did not include a table for adjacent and exact agreement because with many of the scales being 1-4, adjacent and exact agreement was often at 0.99 for both humans and machines.

Table 9.4. Exact Agreement Summary

Essay Set	Human Readers			Machines		
	H1	H2	H1H2	Median	Mean	Range
1	0.64	0.64	0.64	0.44	0.43	.31-.47
2a	—-	0.76	0.76	0.68	0.66	.55-.70
2b	—-	0.73	0.73	0.66	0.65	.55-.69
3	0.89	0.83	0.72	0.69	0.67	.61-.72
4	0.87	0.89	0.76	0.65	0.64	.47-.72
5	0.77	0.79	0.59	0.68	0.66	.47-.71
6	0.8	0.81	0.63	0.64	0.64	.51-.69
7	0.28	0.28	0.28	0.12	0.12	.07-.15
8	0.35	0.35	0.29	0.16	0.16	.08-.23

Exact agreement is summarized in Table 9.4. The report aggregates the ranges of agreement for the two human readers H1H2 among all eight essay sets and all nine rows of data, stating on page 22 that "The human exact agreements ranged from 0.28 on essay set #8 to 0.76 for essay set #2." The report then states, "the predicted machine score and had a range from 0.07 on essay set #2 [sic] to 0.72 on essay sets #3 and #4. An inspection of the deltas on Table 9 shows that machines performed particularly well on essay sets #5 and 6, two of the source-based essays."

The report ignores how human scorers performed better than the machines for most of the essay sets. Of the nine scores, the human rater agreement coefficients exceeded the top score of the machines in six of them, tying in a seventh. In essay set # 1 both readers performed .17 better than the best performing machine. In essay set #2A, the single "read-behind" reader performed .06 better than the best performing machine. In essay set #2B, the single "read-behind" reader performed .04 better than the best performing machine. The next four essay sets are content-based reading tests. For essay sets #3 and #4, the agreement of the two readers outperforms all but one of the machines and ties that one. The report also makes the careless error of incorrectly attributing the 0.07 exact agreement to essay set #2 instead of to essay set #7.

Table 9.5. Kappa Summary

Essay Set	Human Readers			Machines		
	H1	H2	H1H2	Median	Mean	Range
1	0.53	0.53	0.45	0.29	0.28	.16-33
2a	—-	0.62	0.62	0.48	0.46	.30-.51

Essay Set	Human Readers			Machines		
	H1	H2	H1H2	Median	Mean	Range
2b	—-	0.56	0.56	0.45	0.42	.27-.49
3	0.83	0.77	0.57	0.53	0.52	.45-.59
4	0.82	0.84	0.65	0.50	0.50	.30-.60
5	0.69	0.71	0.44	0.55	0.52	.28-.59
6	0.70	0.71	0.45	0.46	0.46	.31-.55
7	0.23	0.23	0.18	0.07	0.07	.03-.09
8	0.26	0.26	0.16	0.09	0.08	.04-.13

Table 9.5 summarizes the Kappa scores. On page 23, the report states that "in general, performance on kappa was slightly less with the exception of essay prompts #5 & #6. On these essay sets, the AES engines, as a group, matched or exceeded human performance." While this last claim is true for essay set #5, it was not true for essay set #6, where the value for H1H2 fell right in the middle of the machine scores. Moreover, the machine performance was not "slightly" lower than human performance measured by H1H2, it was substantially lower for all essay sets except 5 & 6 as can be observed simply by comparing H1H2 with the median and range values of the machine scores in Table 9.5.

Table 9.6. Quadratic Weighted Kappa Summary

Essay Set	Human Readers			Machines		
	H1	H2	H1H2	Median	Mean	Range
1	0.77	0.78	0.73	0.78	0.77	.66-.82
2a	—-	0.80	0.80	0.70	0.70	.62-.74
2b	—-	0.76	0.76	0.66	0.65	.55-.69
3	0.92	0.89	0.77	0.72	0.71	.65-.75
4	0.93	0.94	0.85	0.76	0.77	.67-.81
5	0.89	0.90	0.74	0.81	0.79	.64-.82
6	0.89	0.89	0.74	0.76	0.74	.65-.81
7	0.78	0.77	0.72	0.77	0.75	.58-.84
8	0.75	0.74	0.61	0.68	0.67	.60-.73

Tables 9.6 and 9.7 summarize the scores on the quadratic weighted kappa and the Pearson r. As mentioned previously, the machines do better on the quadratic weighted kappa except for score sets #2A and #2B and the literary analysis questions, essay sets #3 and #4. The performance of H1H2, the comparison of the two readers' scores, is mixed against the machine scores.

Table 9.7. Pearson r Summary

Essay Set	Human Readers			Machines		
	H1	H2	H1H2	Median	Mean	Range
1	0.93	0.93	0.73	0.80	0.77	.76-.82
2a	—-	0.80	0.80	0.71	0.70	.62-.74
2b	—-	0.76	0.76	0.67	0.66	.55-.71
3	0.92	0.89	0.77	0.72	0.71	.65-.75
4	0.94	0.94	0.85	0.76	0.77	.68-82
5	0.89	0.90	0.75	0.81	0.79	.65-.84
6	0.89	0.89	0.74	0.77	0.75	.65-.81
7	0.93	0.93	0.72	0.78	0.76	.58-.84
8	0.87	0.88	0.61	0.70	0.68	.62-.73

These results, as stated previously, may simply be the artifact of using different measures for machines and human readers as well as the improper use of the quadratic weighted kappa.

CONCLUSION

The study's numerous and substantial defects clearly undermine its conclusions. Only three of the eight essay sets used in the study contained scores that assessed students' ability to write more than a paragraph, and only one of the five other essay sets contained scores that were concerned with writing ability at all. Even more disturbing was that, with the exception of essay set #2, the study did not measure the correspondence between human readers and machine scores but used different measures for human and machine reliability that artificially inflated machine performance in half the essay sets. For the one essay set, #2, in which the study directly compared human and machine reliability, human readers were clearly more reliable than all of the machines for both of the writing scores contained in this essay set. Moreover, the study failed to follow standard statistical practice to guard against false positives and also made its assertions in the absence of any statistical tests, only based on the impressions of the authors. Consequently, Professor Shermis and Mr. Hamner should consider formally retracting all versions of this study in print or, at a minimum, respond in print to the criticisms enumerated in this article. Even with the flawed overall design of the study, further and rigorous statistical analysis of data may yield some interesting and extremely important information. Moreover, there are pressing policy decisions that argue for further analysis of these data. This paper has been reported to both the Partnership for Assessment of Readiness of College and Careers

and the Smarter Balanced Assessment Consortium. The data and conclusions in this report may inform decisions by these two consortia about the use of automated essay scoring in the high stakes testing connected to the Common Core Standards, therefore, it is imperative that the authors publicly post the raw test set data from this study for rigorous statistical analysis.

NOTE

Although the adjudication rules given for the essay set descriptions for Essay Sets #3 and #4 do not mention it, examination of the training set revealed that, like Essay Sets #5 and #6, the resolved score was computed by taking the higher of two adjacent scores. There were no sets of scores in the training sample for Essay Sets #3 and #4 that contained pairs of scores that differed by more than one point, and no third rater scores. Consequently, four of the data sets from, at most, two states computed the resolved score by taking the higher score if the two rater scores were not identical. The authors mention, on page 9, instances in which the higher of the two scores in one essay set (#5) was not the resolved scores. In the two instances I identified, the two readers' scores were not adjacent and the resolved score was probably an adjudicated score.

REFERENCES

Austin, J. L. (1962). *How to do things with words.* Harvard University Press.

Brenner, H., & Kliebsch, U. (1996). Dependence of Weighted Kappa Coefficients on the Number of Categories. *Epidemiology , 7*(2), 199-202.

Giles, J. (2012). AI graders get top marks for scoring essay questions. *The new scientist, 2861.* http://www.newscientist.com/article/mg21428615.000-ai-graders-get-top-marks-for-scoring-essay-questions.html

Grice, H. P. (1989). *Studies in the way of words.* Harvard University Press.

Kaggle. (2012). Data—The Hewlett Foundation Automated Essay Scoring. http://www.kaggle.com/c/asap-aes/data

Kolowich, S. (2012). A Win for the Robo-Readers. I*nside Higher Ed.* http://www.insidehighered.com/news/2012/04/13/large-study-shows-little-difference-between-human-and-robot-essay-graders

Man and machine: Better writers, better grades. (2012). The University of Akron News. http://www.uakron.edu/im/online-newsroom/news_details.dot?newsId=40920394-9e62-415d-b038-15fe2e72a677

McCurry, D. (2010). Can machine scoring deal with broad and open writing. *Assessing Writing, 15(2),* 118-129.

Morgan, J., Shermis, M. D., Van Deventer, L., & Vander Ark, T. (2013). Automated Student Assessment Prize: Phase 1 & Phase 2. http://gettingsmart.com/wp-content/uploads/2013/02/ASAP-Case-Study-FINAL.pdf

Perelman, L. (2008). Information illiteracy and mass market writing assessments. *College Composition and Communication , 60*(1)*,* 128-141.

Ramineni, C. A., & Williamson, D. A. (2013). Automated essay scoring: Psychometric guidelines and practices. *Assessing Writing, 18(1),* 25-39.

Rivard, R. (2013). Professors at odds on machine-graded essays. *Inside Higher Ed.* http://www.insidehighed.com/news/2013/03/15/professors-odds-machine-graded-essays

Schultz, M. T. (2013). The IntelliMetric Automated Essay Scoring Engine—A Review and an Application to Chinese Essay Scoring. In M. D. Shermis & J. Burstein (Eds.), Handbook of automated essay evaluation (pp. 89-98). Routledge.

Shermis, M. D., & Burstein, J. (Eds.). (2013). *Handbook of automated essay evaluation.* Routledge.

Shermis, M. D., & Hamner, B. (2012). Contrasting State-of-the-Art Automated Scoring of Essays: Analysis. http://www.scoreright.org/NCME_2012_Paper3_29_12.pdf

Shermis, M. D., & Hamner, B. (2013). Contrasting State-of-the-Art Automated Scoring of Essays. In M. D. Shermis, & J. Burstein (Eds.), *Handbook of automated essay evaluation* (pp. 213-246). Routledge.

Sim, J., & Wright, C. C. (2005). The Kappa Statistic in Reliability Studies: Use, Interpretation, and Sample Size Requirements. *Physical therapy, 85*(3)*,* 257-268.

Sperber, D. (1990). *Relevance: communication and cognition.* Basil Blackwell.

Williamson, D. A., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practices , 31*(1), 2-13.

APPENDIX A

Prompts, Rubrics, and Other Materials

From The Hewlett Foundation: Automated Essay Scoring [Data and training material]. http://www.kaggle.com/c/asap-aes/data. Copyright 2012 by Kaggle. Reprinted with permission.

APPENDIX B

Selected Tables

Source: Mark D. Shermis & Ben Hamner, "Contrasting State of the Art Automated Scoring of Essays: Analysis." https://www.semanticscholar.org/paper/Contrasting-state-of-the-art-automated-scoring-of-Hamner-Shermis/cad818cdb3b8bd2e283743161857826854820 9e1

CHAPTER 10.

GLOBALIZING PLAGIARISM AND WRITING ASSESSMENT: A CASE STUDY OF TURNITIN

Jordan Canzonetta
Syracuse University

Vani Kannan
Syracuse University

This article examines the plagiarism detection service Turnitin.com's recent expansion into international writing assessment technologies. Examining Turnitin's rhetorics of plagiarism alongside scholarship on plagiarism detection illuminates Turnitin's efforts to globalize definitions of and approaches to plagiarism. If successful in advancing their positions on plagiarism, Turnitin's products could be proffered as a global model for writing assessment. The proceedings of a Czech Republic conference partially sponsored by Turnitin demonstrate troubling constructions of the "student plagiarist." They demonstrate, too, a binary model of west and nonwest that stigmatizes nonwestern institutions and students. These findings support an ongoing attention to the global cultural work of corporate plagiarism detection and assessment.

There is nothing immutable about the cheating culture that now exists in many educational settings worldwide. On the contrary, *we know the values of students can be changed when institutions invest in the right strategies.* This has happened in areas related to diversity, gender relations, and substance abuse—both in the U.S. and overseas. So far, though, promoting integrity has not commanded adequate attention or resources. This session will explore key drivers of the cheating culture and outline what it will take to *dismantle that culture*. It will examine cases where *education institutions have changed how young people think and behave*—and how these lessons can be applied to promoting integrity.

In the keynote address at the 2016 Computers and Writing conference, Jeff Grabill argued automated writing technologies need to be at the forefront of disciplinary conversations and actions within the field of composition and rhetoric.

DOI: https://doi.org/10.37514/PER-B.2024.2166.2.10

His speech marks a clear exigence: Globally, millions of students are subjected to writing technologies that writing experts did not design. Grabill argued disciplinary action is urgent because "students whose community and home languages are not mainstream are being given bad robots"; because Turnitin is the most popular writing technology deployed globally; and because so many of these programs advance "writing as a fundamentally individualized activity involving a student, a computer, and an algorithm" (2016). Popular automated assessment programs have been decried by writing experts because they "align with the narrow view of writing that was dominant in the more recent era of testing and accountability, a view that is increasingly thrown into question. New technologies . . . are for the most part being used to reinforce old practices" (Vojak et al., 2011, p. 99). Further, these programs fail to use technology that promotes an understanding of core concepts writing experts believe about writing: "that it is a socially-situated practice; that it is a functionally and formally diverse activity; and that it is increasingly multimodal" (Vojak et al., 2011, p. 108).

Grabill's keynote emerges in a kairotic moment in higher education, as for-profit assessment companies like Turnitin expand their global reach and begin to deploy "formative" and "summative" writing assessment programs. We adopt NCTE's definition of formative assessment: "the lived, daily embodiment of a teacher's desire to refine practice based on a keener understanding of current levels of student performance, undergirded by the teacher's knowledge of possible paths of student development within the discipline and of pedagogies that support such development" (NCTE, 2013b, p. 2). Summative assessment, then, for the purposes of our framework, refers to "final evaluative judgment" of student writing (NCTE, 2013b, p. 2). However, we should mention that Turnitin's use of these terms does not appear to align with NCTE's definitions.

Turnitin's artificial intelligence for writing assessment, a program called "adaptive technology," is now marketed as a cutting-edge product for assessing student writing. The "Turnitin Scoring Engine" website claims the platform can "Us[e] your previously-graded sample essays . . . [to identify] patterns to grade new writing like your own instructors would. Give the Engine a set of samples, and it will accurately score an unlimited number of new essays quickly and reliably" ("Turnitin Scoring Engine," n.d).[1] This scoring engine offers to mimic the behavior of teachers by using algorithmic technology to analyze a teacher's prompts and grading comments to produce an evaluative response to student writing ("Features: Overview," n.d). Thus, Turnitin's "intelligent assessment" alleges to grade papers like humans can on categories of "lexical, syntactic, and

1 Because Turnitin is in the process of testing its new assessment platforms, the company's technology, language, and website are constantly changing. Thus, the information we refer to may appear on the website under different headings or may have been otherwise altered.

stylistic features of writing, such as word choice and genre conventions. It uses these features to assess content mastery and genre awareness ("Turnitin Scoring Engine," n.d). According to Grabill, such corporate assessment programs are influencing vast student populations—as Turnitin boasts, "30 million" students—across the globe ("Homepage," n.d).

Turnitin's success in the U.S. is deeply connected to corporate influence in U.S. universities, heavy reliance on contingent labor, a culture of standardized testing, hegemonic cultural expectations about writing and authorship, and the complex web of material factors that shape writing assessment (Chatterjee & Maira, 2014; Giroux, 2007; Herrington & Moran, 2001; Vie, 2013a; Vojak et al., 2011). We have three central concerns in this article: Turnitin's institutionalized plagiarism detection, its move to writing assessment, and its global expansion. Prominent and respected organizations in the field of composition and rhetoric, including the CCCC Intellectual Property Committee [CCCC-IP], the Council of Writing Program Administrators [CWPA], and the National Council of Teachers of English [NCTE], have aligned themselves against the detrimental pedagogical practices advanced by Turnitin (CCCC-IP, 2006; CWPA, 2003; NCTE, 2013a). Of particular concern is that PDSs demonize nonnative English speakers and "unwittingly construct international students as plagiarists" (Hayes & Introna, 2005, p. 55). This important scholarship asks the discipline to pay particular attention to the rhetorical construction of the student-plagiarist by PDSs, and the values ascribed to plagiarism, authorship, and intellectual property. Additionally, now that Turnitin offers an assessment platform, plagiarism detection technology must be understood in conjunction with such platforms, as they are now (or will be) packaged and sold together.

This move toward "scalable" assessment, as Grabill suggested, has global implications; from Turnitin's inception, it has linked integrity, values, and honesty to its global community of users:

> Turnitin.com is currently helping high school teachers and university professors everywhere bring academic integrity back into their classrooms . . . We encourage any educator who values academic honestly to help us take a stand against online cheating and become a member of the Turnitin.com educational community. ("About Us," March 31, 2001)

Although the company now adopts more nuanced rhetorical approaches to sell their product, this original language is likely still familiar to those who teach, work, and study in educational institutions. This familiarity is part of its insidiousness—it situates instructors (presumed to be members of the "Turnitin.com educational community") as preservers of ethical and moral standards,

positioned antagonistically against students, and assumed to be consistent across institutions and geographic locations. This language also foreshadows the global initiatives that the company would pursue years later. In 2015, Turnitin's website claimed that the program was "[u]sed by over 1.6 million instructors at more than 10,000 institutions in 135 countries, [and] is the world's leading cloud-based software for evaluating student work" ("Features: Overview," n.d). In the year since we began writing this article, the number of institutions has jumped from 10,000 to 15,000 ("Homepage," n.d).

The company now globally markets its plagiarism detection program as an aid to overworked teachers by offering services that 1) "streamline" grading, 2) offer a solution to "deteriorating" student ethics, and 3) serve as a placement/evaluation program for newly matriculated students (Janssens & Tummers, 2015, p. 12; "What We Offer," n.d; "Why Turnitin," n.d). The "Global Effectiveness" page on Turnitin's website boasts the company "impact[s] levels of unoriginal writing and promote[s] the use of online feedback globally," and the "Third-Party Academic Research" page draws from peer-reviewed articles from all over the world (2015). The company grants "Global Innovation Awards" to educators and technology administrators "who demonstrate a commitment to academic integrity, excellence in enhancing student learning, or champion the innovative and effective use of Turnitin to support learning at their school or institution," offering recipients "professional opportunities to become content contributors and be leaders in the Turnitin community"; the 2015 awardees were chosen from 400 nominations in 50 countries ("Global Effectiveness," 2015).

In the context of Turnitin's globalization, we ask which countries, regions, and peoples are being defined as having correct or incorrect values of authorship. In invoking the rhetoric of globalization, we find Scholte's conception of globalization as internationalization, liberalization, universalization, westernization, and respatialization to be useful (2000, p. 2). We focus specifically on universalization and westernization, as these seem to be the primary features of Turnitin's global rhetorics, where a "culture" of plagiarism requires intervention so that "integrity" can be restored worldwide through the implementation of values that are presumed to be universal but in reality reassert western hegemony. How is the student plagiarist being discursively constructed? What are the implications of these constructions as Turnitin rolls out its assessment platform?

METHODS

To attend to these questions, we first offer an overview of Turnitin's plagiarism detection software, mapping the company's movement towards writing assessment. Then, we situate Turnitin within disciplinary critiques of plagiarism detection

services (Howard, 1999; Purdy, 2005; Vie, 2013a & 2013b). Throughout, we draw from the proceedings of the biennial academic "Plagiarism Across Europe and Beyond" conference (2013, 2015), building on Poe and Inoue's work on racial formations related to standardized test scores to ask "what writing constructs reward which group of students" (Poe & Inoue, 2012, p. 358). We conclude by extending Grabill's call to focus collective disciplinary efforts on interrogating corporate writing assessment platforms, stressing the exigency for critical awareness of how PDSs such as Turnitin are constructing the student-plagiarist globally, with the acknowledgment that binary divisions of west/nonwest obscure the heterogeneity of both.

While we do not suggest that Turnitin's sponsorship means direct endorsement all of the policies and ideas that were presented at these conferences, we do argue that the presentations in these proceedings align with and reflect the rhetoric the company has adopted. Thus, this article draws out linkages between PDSs and the knowledge production around plagiarism and assessment happening worldwide in sites where such programs invest money. Turnitin's direct support of this conference is notable particularly because many presentations promote PDSs in diverse geographic regions. These arguments then lay the groundwork for plagiarism and assessment standardization via automated protocols like Turnitin's.

Our coding and interpretation approach, because it is contextualized within disciplinary critiques of Turnitin and PDSs more broadly, can be characterized as Values Coding, wherein our orientation towards rhetorical constructions of the student-plagiarist serve as a lens of analysis (Saldaña, 2009, p. 7). Following grounded theory methods, we broke the texts into small units of information and developed codes to describe "word[s] or short phrase[s] that symbolically assig[n] a summative, salient, essence-capturing, and/or evocative attribute for a portion of language-based or visual data" (Saldaña, 2009, p. 3). We categorized the list of codes into themes, then "abstract[ed] out beyond the codes and themes to the larger meaning of the data," linking to and contextualizing the findings within existing literature (Creswell, 2012, p. 187). Finally, we classified the codes into larger themes, or "broad units of information aggregated to form a common idea" (Creswell, 2012, p. 186; methods adapted from Kannan, 2014).

Our analysis revealed three primary rhetorical strategies for advancing Turnitin—and PDSs more broadly—within the conference proceedings as services that should be implemented not only at institutional and state levels, but across the whole European Union and globally: (1) Plagiarism detection represents social improvement and formation of model, modernized, idealized, western students; (2) Plagiarism is a national concern with ramifications for citizenship, economy, and character; and (3) Approaches to plagiarism detection need to be standardized and aligned with western institutions and states; public/private

partnerships and linked state policies are the best way to do so. In the following sections of this article, "Critiques of Plagiarism Detection Services" and "Turnitin, Assessment, & Globalization," we draw from our findings.

CRITIQUES OF PLAGIARISM DETECTION SERVICES

Scholars in composition and rhetoric have long worked to overturn the individualistic constructions of authorship and stigmatization of student-plagiarists advanced by PDSs like Turnitin. What are the implications as Turnitin expands across the globe? How do these definitions of authors and plagiarists construct different student populations and geographic regions?

Conflict surrounding plagiarism often relates to definitional tension. In this study, we adopt the CWPA's understanding of plagiarism: "[i]n an instructional setting, plagiarism occurs when a writer deliberately uses someone else's language, ideas, or other original (not common-knowledge) material without acknowledging its source" (2003). However, the word *plagiarism* also has quite violent connotations; its Latin roots are tied to words like "stealing" and "rape," which links the word and its history to ideologies of property, theft, and bodily violation (Howard, 2000, pp. 479-483). Rhetorics of plagiarism are often linked to "metaphors of gender, weakness, collaboration, disease, adultery, rape, and property that communicate a fear of violating sexual as well as textual boundaries" (Howard, 2000, p. 474; for an extension of the metaphor, see also Robillard, 2009; Vidali, 2011). Scholars contend PDSs advance singular conceptions of authorship (Howard, 1999; Vie, 2013a); create an adversarial relationship between students and teachers (NCTE, 2013a); sign over intellectual property rights to the company's database and/or force instructors to use these programs (Canzonetta, 2014, p. 39; Purdy, 2005, p. 278); and mask deeper pedagogical and political economic concerns by offering a "corporate solution" to teaching problems (Marsh, 2004, p. 428). PDSs arose as a technological response to catching violators who, according to the creators of Turnitin, were increasing with alarming rapidity as students began to do more and more research online (Vie, 2013b). Indeed, in the "Plagiarism Across Europe and Beyond" proceedings, the availability and globalization of digital information is cited as a reason for the rise of plagiarism, along with the "deteriorating ethical values of students" (Janssens & Tummers, 2015, p. 12).

This emphasis on integrity and the specter of waning values masks Turnitin's cooptation of students' intellectual property. Indeed, scholars and writing teachers are not the only groups to take issue with Turnitin; students and parents in the U.S. have led efforts to both petition against and sue the company, citing concerns about intellectual property. In a 2007 case, students at McLean High

School in McLean, VA, and Desert Vista High School in Phoenix, AZ, filed a lawsuit against Turnitin (Zimmerman, 2007). The events that led up to the eventual filing of the lawsuit in March 2007 began in September of 2006, when a group of students at McLean High School circulated a petition to oppose the mandatory submission of their work to a newly adopted Turnitin.com: "[t]he petition, which garnered 1,190 student signatures of the approximately 1800 students that attend the school requested that the mandate to submit work to Turnitin be removed and that an 'opt-out' option be allowed" (Zimmerman, 2007).While students did not win the case, their work to contest Turnitin's use of student intellectual property, and the call for the student choice to "opt-out" of Turnitin (mirroring movements to "opt-out" of standardized testing) drew attention to the negative impact of PDSs, and the corporatization of education more broadly, on students. Unfortunately, neither these lawsuits nor repeated criticisms of PDSs have impacted Turnitin's widespread adoption by educational institutions, but the company has shifted its marketing rhetoric from "catching plagiarists" to "meet[ing] exigencies" in our field to both deflect criticism and respond to the labor crisis in higher education (Vie, 2013b).

In the current iteration of the website, the word *plagiarism* only appears on the main page twice (in smaller text than other language on the page) under subheadings; this is a departure from its early website iterations, which foreground anti-plagiarism zeal ("About Us," Wayback Machine, March 31, 2001; "Homepage," n.d). Despite Turnitin's move towards broader writing assessment technologies, it still uses problematic plagiarism detection software. Its plagiarism detection "tool" can only provide students and teachers with a report containing percentages of text that corresponds to various sources on the Internet, sources in its database, and periodicals, journals and publications, and cannot infallibly identify plagiarism ("FAQ," n.d; Purdy, 2009, pp. 65-67).With Turnitin's increased presence in global writing assessment technology, PDSs become more problematic when we consider the effects they have on nonnative English speakers. Hayes and Introna (2005) suggest PDSs may inhibit some ELL students who are trying to participate in the writing process, but are stymied in their attempts because the detective component of the programs "limit[s] the opportunities and time that students have to learn how to write in the new western, not to mention subject specific, educational context" (p. 67). The use of PDSs at the onset of the composing process implies students have higher stakes for writing in new cultural contexts. Without having the chance to learn about new practices in those environments, students are discouraged from taking risks, "experiment[ing]," or "observ[ing]" (p. 67).

Current PDS platforms, then, are shaping educational space so that students are castigated for departing from Edited American English (EAE) and western

ideals about singular authorship, as Introna and Hayes (2011) explained:

> Plagiarist practices are often the outcome of many complex and culturally situated influences . . . [E]ducators need to appreciate these differing cultural assumptions if they are to act in an ethical manner when responding to issues of plagiarism among international students. (p. 215)

Originality/singularity is not globally accepted as the primary theory of authorship; not all students are asked to produce original work, and imitation can often be a staple in some writing processes (Hayes & Introna, 2005, p. 59). Thus, Turnitin's emphasis on originality/singularity elides a complex cultural understanding of plagiarism and authorship.

These underlying ideologies of original/singular authorship were laid bare and explicitly connected to culture in Turnitin's "Plagiarism Education Week" event "Copy/Paste/Culture." Held April 20-24, 2015, the conference was marketed as investigating "how current global trends are affecting our values, especially those related to education, and proposing strategies on how we can address these challenges. #integrity2015." The conference focused on how to dismantle the "culture of plagiarism," variously described as a "mindset" of narcissism and entitlement (Hoyt, 2015). As the conference description shows, "our values" are presumed to align with western constructions of authorship. Indeed, something as banal and familiar as the hashtag "integrity"—a word that students and teachers are likely used to seeing mobilized in discussions of plagiarism—immediately connects intellectual property to character, and by extension, plagiarism to poor character.

The "Plagiarism Across Europe and Beyond" conference proceedings echo these stark character judgments, and explicitly situate them in terms of a geographic binary of west and nonwest, including designations of "high trust" versus "low trust" societies and populations (Burkatzki, Platje, & Gerstlberger, 2013, p. 171). In this framework, it becomes the duty of the west (and PDSs) to counter tolerance towards plagiarism, export knowledge, and modernize culture. Through this mapping of nonwest, the proceedings constitute and consolidate geographic sites for corporate/state-level plagiarism detection intervention, with the assumption that Turnitin possesses the correct values of authorship. Howard (1999) explained such rhetorics are largely related to archaic constructions of plagiarism, and don't allow much space for cultural variance in writing processes:

> For the past century and more, [western] academic textual values have been relatively unified, ascribing four properties to the "true" author: autonomy, originality, proprietorship, and morality . . . The writer who is not autonomous and original

> demonstrates an absence of morality, earns the label "plagiarist" and deserves punishment. (p. 58)

In the remainder of this section, we identify and deconstruct moments in Turnitin's conferences when these notions of authorship were upheld by presenters, and discuss the ideological implications of such ideas.

Following Hesford and Schell (2008), we aim to engage critically with the idea of nationhood by examining the way particular nations—and student bodies within nations—are described within the conference. We do so with recognition that the concept of "the west" is a monolithic consolidation of multifarious languages, cultures, communities, and histories—and "the west" is being defined in very specific ways in these proceedings, erasing indigenous, diasporic, and non-standard American English-speaking students in the process. In a presentation on cultural understandings of plagiarism, "the west" was defined as "mainly English speaking countries: UK, USA, Canada, Australia, New Zealand," while "the east" was defined as "particularly Confucian Heritage Cultures: China, Japan, Singapore, and Vietnam" (Gow, 2013, p. 27). By linking "the west" to the English language, and distinguishing it from "Confucian Heritage Cultures," this presentation not only delineates nations along racial and cultural markers, but also suggests that the English language and non-Confucian cultural values are inherently more ethical. Further, in an article on South Asian MBA students studying in the UK at Cardiff School of Management, India was folded into the "eastern" region, and the MBA students' "academic malpractice" was cited as a problem which parallels the university's "similar issues with other cultures" (Wellman & Fallon, 2013, p. 71). Thus, "the east" can be understood to include international students attending "western" universities. In the process, populations including South Asians who grew up the UK are erased, as are complexities and distinctions within and across "nonwestern" cultures.

Further, the "west" is described as practicing appropriate and punitive measures in dealing with plagiarism, and distinguished from Eastern European countries, where "plagiarism is not considered to be a big problem"—an accusation that is duly framed as problematic and as a potential market (Foltýnek, Rybička, & Demoliou, 2013, p. 127). For example, Lithuania was described as a corrupt, post-Soviet country with a "high level of tolerance toward cheating" (Novelskaitė & Pučėtaitė, 2013, p. 238). Similar arguments were made about "developing" countries including Brazil, where cultural knowledge about plagiarism was framed as "rudimentary" (Krokoscz & Putvinskis, 2013, p. 281), and Nigeria, where problems were cited in a "student plagiarism culture"—the subtext being plagiarism is not taken seriously or punished appropriately, impacting students' "experience when they study elsewhere" (Orim, Borg, & Awala-Ale

2013, p. 66). This totalizing attitude toward academic integrity stands at odds with localized, context-specific understandings of plagiarism and pedagogies of authorship.

Drawing on these geographic delineations, the conference presenters advocated the global applicability of PDSs, and outlined the social impacts of plagiarism. For example, in a presentation about Lithuania, the speakers connected plagiarism and the social ills of late capitalism:

> Plagiarism is not only an academic issue. It concerns *public interest at large*. . . . it discredits the acknowledgments given by higher education institutions to their graduates, *diminishes public trust* in professional qualifications and social institutions in general . . . [plagiarism can] *incite society's feeling of social injustice* and, in radical cases, *cynicism and alienation among its members*. (Novelskaitė & Pučėtaitė, 2013, p. 237, emphasis added)

This direct correlation between plagiarism, trust, cynicism, and alienation is extended into economic success in another presentation: "The Academic Integrity Maturity Model (AIMM) was developed to measure the level of academic integrity maturity for particular country . . . *the more mature the academic integrity in particular country, the richer the country*" (Foltýnek & Surovec, 2015, p. 121, emphasis added). Conversely, infractions in academic integrity are directly linked to long-term unemployment and rising crime, thus linking poverty to moral failure, and moral failure to plagiarism. As one speaker noted, "if there is high long-term unemployment rate in [a] particular country, people tend to be less satisfied with their lives, crime increases and people give up an honest way of life and tend to dishonesty including academic integrity breaches" (p. 129).

Complementing this emphasis on character, citizenship, and economics, proceedings celebrated courageousness as the goal of academic work, as is demonstrated in a keynote address:

> Courage is an element of character that allows learners to commit to the quality of their education by holding themselves and their fellow learners to the highest standards of academic integrity even when doing so involves risk of negative consequences or reprisal. Being courageous means acting in accordance with one's convictions. (Bretag, 2015, p. 6)

In adopting this moral agenda in a keynote presentation, a clear tone was established for the 2015 conference: Plagiarism isn't a pedagogical issue, it's about virtue. In framing plagiarism as a problem that is bound to economics,

citizenship, morality, and integrity, these presenters created a need for the solution Turnitin purports to offer.

These alarming links drawn between plagiarism, integrity, character, and geographical location have deep implications in light of Turnitin's global reach and venture into writing assessment. Turnitin is positioning itself to become the global plagiarism police. In promoting western writing values internationally, programs like Turnitin are poised to standardize writing globally in alignment with EAE and western conceptions of authorship, which reinforces harmful and ideologies that affect writing teachers' authority to determine our pedagogies and assess our students' work.

TURNITIN, ASSESSMENT, & GLOBALIZATION

Turnitin's venture into writing assessment is troubling. As we have seen, although Turnitin boasts that their new algorithmic technology is adaptive (i.e, artificially intelligent) and can accommodate each teacher's behavior and grading practices, the conference proceedings suggest a move toward promoting consistency and standardization in students' writing practices, and ascribing negative character value to those who plagiarize based on a hierarchical, colonizing, and fallacious west/nonwest binary.

Turnitin's latest projects involving adaptive technology offer "formative" and "summative" assessment platforms tailored to the "needs of 21st century classrooms" ("Features: Overview," n.d). However, the company's long-term use of an algorithm to carry out its text-matching services stands at odds with its efforts to persuade the public that its pedagogy and formative assessment are in students' and teachers' interests (Turner, 2014). "Intelligent assessment," as Turnitin's marketing calls it, claims to incorporate formative and summative writing assignments "with a range of feedback tools, including automated feedback, originality check, online grading and peer review," and offers "a solution that improves student writing, saves instructors' time and enhances the quality of feedback to student and provides institutions with insights into how students learn over time" ("Lightside Labs," n.d). In order to sell this "adaptable technology," Turnitin claims its "Scoring Engine" will use an algorithm that is trained to "[use] your previously-graded sample essays, [to] identif[y] patterns to grade new writing like your own instructors would. Give the Engine a set of samples, and it will accurately score an unlimited number of new essays quickly and reliably" ("What We Offer," n.d). "Adaptation," here, displaces composition and rhetoric's arguments for situated pedagogical approaches with a neoliberal rhetoric of efficiency, adaptability, and individual choice. Turnitin's rhetoric, a clear response to scholarship in writing assessment that urges

local and contextualized assessment (Barlow et al., 2007), alleges it can analyze teachers' prompts and comments on students' papers the way a teacher would 90% of the time ("Lightside Labs," n.d).

Formative assessment necessitates that teachers respond to students' needs, personalities, struggles, and strengths; and get to know them apart from their writing. Revision Assistant, a new feature of Turnitin's software, claims to offer formative assessment and is meant to provide holistic responses to student work. In practice, it produces a signal score that shows students how their teachers would score their work and provides feedback about how to achieve their desired scores in the areas of "Analysis, Focus, Language, Evidence" ("Revision Assistant," n.d). Turnitin described an earlier model of this program as an aid to "marginalized students" who "take great advantage of this student-driven process, bringing Revision Assistant's feedback to the teacher and proactively asking for help" ("Lightside Labs," n.d). Through Revision Assistant, Turnitin offers what Condon (2013) explicitly cautioned against: "systems of writing . . . subject to the fallacy of surrogation—the substitution of a statistical artifact—a number—in place of the need for complex information" (p. 101). While Revision Assistant's more substantive feedback on analysis, focus, language, and evidence might seem less alarming than Turnitin's plagiarism detection scoring algorithm, a machine is still assigning students a signal score based on an arbitrary scale to convey information about students' literate and rhetorical abilities.

Automated assessment platforms Turnitin offers also allow institutions unprecedented levels of surveillance over their students' work. The website boasts that the program is an opportunity for teachers to garner a composite image of how all their students are writing, which is an appealing offer to those who engage in program-wide assessment. Zwagerman (2008) claimed that, through comparing and viewing thousands of pages of student work, reports of student work lend themselves to "the panoptic logic that a structure of examination and documentation does not preclude individuality but rather accounts for it and renders it intelligible" (p. 691). Students are watched to ensure their originality and individuality, which is then legitimized by the machine that polices them. Another problem with PDSs—which becomes even more serious as PDSs venture into assessment—is the unfettered access teachers, institutions, and governments gain to student data. Spellmeyer (1996) has long argued that, rather than offering unlimited data to agencies that may not prioritize pedagogy and best practices for students, we need to

> guard against . . . any effort to exclude programs, departments, and universities from the collecting and interpreting of data on their own classes, since the parties that control the

> spin put on this information will have the last word in every forum. (p. 180)

Thus, it is important to critically interrogate Turnitin's rhetorics of formative assessment, which obscure the company's cooptation of student data and potential to undermine writing program goals.

Furthermore, Deborah Harris Moore (2013) contends that the fear caused by surveillance can be disempowering to students: "Using fear as a deterrent . . . is unethical because it forces students into behaviors based on their perceived powerlessness . . . [S]tudents may see [this technology] as an all-seeing, determining, and surveying mechanism" (pp. 110-111). After the McLean High School lawsuit, this culture of surveillance now appears to be taken for granted by many students, who, according to instructors, view Turnitin as either an "arbitrary hoop" to jump through to submit their papers, or as a "psychological deterrent" and "authority" on plagiarism (Canzonetta, 2014, pp. 21-33). Turnitin's database was initially designed for this purpose—to deter students from plagiarizing by invoking its vast, national collection of student writing (Zimmerman, 2007).

Beyond serving as a deterrent to plagiarism, Turnitin has seized the opportunity to exploit the current labor crisis in higher education.[2] As Herrington and Moran (2001) noted, "when human labor is in crisis, we often turn toward technology to mitigate human stress and loss of funding to alleviate insufficient staffing" (p. 220). Indeed, the company has positioned Revision Assistant as an ally and resource for overworked teachers, arguing that it "takes many of the challenges of continuous feedback out of the teaching equation, such as the pressure on instructors to provide consistent, timely feedback for all of their students . . . teachers are provided with a better picture of each student's progress when making a final assessment" ("Features: Overview," n.d). By offering a tool to lighten workloads and the pressures of promptly returning students' work with feedback ("Customers," n.d), the company appeals to administrators whose instructional staffs are either overburdened or understaffed; for those who may not share composition and rhetoric's critiques of PDSs, Turnitin is proffered as a solution to the complex problem that grading writing presents. The artificial intelligence Turnitin is testing claims to be for students, and for teachers who need more time; it instead appears to be a band-aid for upper-level university administrators who would rather put money into a technological "panacea," as Marsh (2004) wrote, than contend with hiring more faculty. Instead of learning about students, Turnitin's formative assessment

2 In the U.S. in 2012-2013 academic year, approximately 76% of higher education's instructional staff consisted of contingent laborers (Curtis & Thornton, 2013, p. 8).

model learns teachers and their behaviors, assesses generic writing processes, and supplies an automated response to a perceived problem. Considering the contingent positions that many writing instructors occupy, and the money-saving imperative of corporatizing universities, Turnitin's formative assessment model poses a major threat for agency and autonomy within writing programs. The data produced through this program could have serious implications for instructors' job security if students aren't achieving scores administrations approve of—scores that could be set and established by Turnitin.

What, then, are the implications of these moves in light of Turnitin's expansion abroad? Rhetorical links between adaptability, assessment, plagiarism, and pedagogy are visible in the "Plagiarism Across Europe and Beyond" conference proceedings, and Turnitin is cited by many presenters as a positive pedagogical tool that offers opportunities for teachers to craft formative assessment pedagogies that directly result in lowered instances of plagiarism. Indeed, formative assessment is implicitly used to justify the use of Turnitin (Meacheam & Faifua, 2015, p. 45). Our analysis of the conference proceedings reveals a particular emphasis on rhetorics of integrity and consistency, linking western values of authorship with standardization across institutions and geographies. Of particular note is a reference in a keynote address to the monetary investment (€ 300,000) the European Union designated for the project *Impact of Policies for Plagiarism in Higher Education Across Europe* (IPPHEAE), conducted between 2010 and 2013. In this discussion, presenters asked:

> What impact did the project have on national and institutional policies for academic dishonesty and plagiarism? What evidence is there that policies for academic integrity in higher education in different parts of Europe are fit for purpose? How can institutions be sure their policies are effective and being applied consistently? What more needs to be done? (Glendinning, 2015, p. 7)

Through this neoliberal rhetoric of fitness (Dingo, 2012), we see a clear call for uniformity in coping with plagiarism—a pedagogical problem that, as composition and rhetoric scholarship shows, is highly contextual and occurs on a "continuum," not in a vacuum (Sutherland-Smith, 2008, p. 8). Similarly, presentations in both 2013 and 2015 advocated worldwide implementation of an "ANTIPLAG system" that has been adopted in Slovakia and is now enforced there by law:

> the SK ANTIPLAG system (a central repository of theses and dissertations, a plagiarism detection system, a comparative

> corpus, local repositories of theses and dissertations) started routine operation after a preparatory phase. Pursuant to the amendment to the Higher Education Act from October 2009, the use of SK ANTIPLAG . . . is *mandatory* for all Slovak higher education institutions operating under the Slovak legal order. It is an unparalleled and unprecedented implementation of such a system on a *national level***.** A relevant milestone has been built not only on the Slovak scale, *but also worldwide.* (Kravjar, 2015, p. 147, emphasis added)

A policy in which PDS use is mandated by the state is ideal for companies like Turnitin; the presenters urged such a model to be implemented worldwide. In an article on the Czech Republic PDSs are defined as "a unique solution in Europe and very likely in the world" (Kravjar & Noge, 2013, p. 212). Similarly, in a presentation on plagiarism in Cyprus, concerns are raised about "the extent of plagiarism practiced by students worldwide" (Kokkinaki, Iacovidou,& Demoliou, 2013, p. 192). Not only does this state-mandated plagiarism check globally advance models of authorship that are compatible with the use of PDSs, but it also allies such companies with powerful governmental agencies to which institutions of higher education are often beholden.

Turnitin applies its rhetorics of consistency to plagiarism policies as well as formative assessment components, claiming that "a consistency of approach" for using PDSs as formative tools should be implemented on a wider scale. Conference presentations suggest the need to change students themselves, and the need for a strict institutional culture:

> Perhaps a *consistency of allowing formative use of originality checking systems* . . . might produce the needed *behavioural changes* needed in our student populations. This is presuming that any institution has a *backbone of policy and practice that supports action* in relation to plagiarism. (Meacheam & Faifua, 2015, p. 47, emphasis added)

Such rhetorics of consistency are a growth strategy for corporate assessment in the context of neoliberal globalization. The "behavioral changes" these scholars and teachers promote (and seek to enforce with legal measures) aim to quell critique and breed a compliant, submissive population of students. Once students sign over their intellectual property to PDSs, an agency that legally enforces such a system would create ideal conditions for assessment companies. Standardizing writing processes, practices, and assessment sets the stage to minimize or eliminate any opposition to their products from scholars and teachers.

Interestingly, in the proceedings, calls for consistency are paired with presentations calling for contextual understandings of plagiarism, incorrectly suggesting that the conference represents a fair debate and echoing Turnitin's uptake of disciplinary critiques of PDSs: "Every single instance of plagiarism is unique and requires careful examination of all the circumstances and facts, but universal standards on the systematic level also should exist and serve as a prevention of plagiarism and other types of research misconduct" (Vasiljevienė & Jurčiukonytė, 2015, p. 164). However, these gestures mean little when set alongside the framing of the conference and its broad geographical consolidation, "across Europe and beyond." Considering Turnitin's partial sponsorship of this conference, paired with their new initiative to implement formative assessment technology, we have to consider that rhetorics of standardization and consistency are beneficial for Turnitin's business model, and promises of contextual specificity will be necessary in order to persuade fields like composition and rhetoric to adopt its assessment program.

CONCLUSION

Scholarship in composition and rhetoric defines plagiarism as a highly contextual, case-by-case pedagogical issue. Turnitin's assessment platform could be used to execute standardized assessment of writing "across Europe and beyond," as the conference title indicates. Corporate PDSs, thus, have the potential to standardize student writing itself, potentially on a global scale. Turnitin has set the stage for and monopolized the plagiarism detection market—the end results of which are promoting singular, original conceptions of authorship globally. Now more than ever, it is time for rhetoricians and compositionists to "use our own pedagogies and technologies . . . [and] fix our gaze on the millions of learners who are being taught with technologies made by people who know very little about writing and learning to write" (Grabill, 2016). Scholars within the field have begun to develop new writing technologies, such as Eli Review, a program created by writing experts—Grabill, Hart-Davidson, and McLeod—at Michigan State University ("About Eli Review"). Other scholars have endorsed emergent technologies; for example, Les Perelman, famed debunker of the robo-graders supports the technology WriteLab (Berdik, 2015). However, the discipline still faces the problem of globalized models for standardized plagiarism detection and writing assessment.

Who will benefit from globalized programs like Turnitin's, and who will be left out? Herrington and Moran (2001) warned:

> The marketing muscle of these testing companies, and the

> concurrent expansion of the computer-as-reader of students' classroom writing, writing teachers need to understand what is happening here and take a careful look at its substance and likely consequences, lest we be seen as irrelevant and be 'sent out of the room' by the other stakeholders. (p. 220)

And so, as Turnitin is poised to debut its new assessment platform internationally, we need to ask and to challenge what the company values in order to remain in the room.

Our research suggests that debates on assessment must attend to the definitions of plagiarism and authorship that are being implemented globally by Turnitin. The field of composition and rhetoric should build on important existing work (Howard, 1999; Marsh, 2004; Poe & Inoue, 2012; Purdy, 2005; Vie, 2013a & 2013b) to examine how globalizing technologies are changing what we know about plagiarism policies, pedagogy, and writing in a global context, and how corporations like Turnitin are profiting from racist, deficient discursive constructions of the nonwestern student plagiarist and nonwestern countries. Given Turnitin's emerging formative assessment model and globally deployed PDS model, what are the national and international implications for assessment? As educators and researchers of writing, we must think deeply and critically about the links between automation, plagiarism, and assessment, and foreground the global implications of automated plagiarism and assessment protocols.

REFERENCES

Please note that all citations that link to Turnitin in this article are no longer live or retrievable.

Barlow, L., Liparulo, S. P., & Reynolds, D. W. (2007). Keeping assessment local: The case for accountability through formative assessment. *Assessing writing, 12*(1), 44-59.

Berdik, C. (Sept., 2015). A critic's second thoughts on robo-grading. Bostonglobe.com.

Bretag, T. (2015). Enacting academic integrity—it takes courage. In *Plagiarism Across Europe and Beyond 2015 Conference Proceedings* (p. 6). Brno: Mendel University.

Burkatzki, E., Platje, J. & Gerstlberger, W. (2013). Cultural differences regarding expected utilities and costs of plagiarism between high-trust and low-trust societies—preliminary results of an international survey study. In *Plagiarism Across Europe and Beyond 2013 Conference Proceedings* (pp. 171-191). Brno: Mendel University.

Canzonetta, J. (2014). Plagiarism detection services: Instructors' perceptions and uses in the first-year writing classroom (Master's thesis). http://search.proquest.com/docview/1553839828

CCCC-IP Caucus recommendations regarding academic integrity and the use of plagiarism detection services. (2006). http://culturecat.net/files/CCCC-IPpositionstatementDraft.pdf

Chatterjee, P. & S. Maira. (2014). *The imperial university: Academic repression and scholarly dissent.* University of Minnesota Press.

Condon, W. (2013) Large-scale assessment, locally-developed measures, and automated scoring of essays: Fishing for red herrings? *Assessing Writing, 18*(1), 100-108.

Copy/Paste/Culture Week. 3rd Annual Plagiarism Education Week. http://www.turnitinuk.com/en_gb/about-us/press/turnitin-hosts-3rd-annual-plagiarism-education-week-april-20-24

Council of Writing Program Administration. (2003). Defining and avoiding plagiarism: The WPA statement on best practices. http://wpacouncil.org/node/9

Creswell, J. W. (2012). *Handbook of qualitative research: Choosing among five approaches.* Thousand Oaks: Sage Publications.

Curtis, J. W., & Thornton, S. (2013). Here's the news: The annual report on the economic status of the profession, 2012-2013. *Academe, 99*(2), 4-19.

Customers. http://turnitin.com/en_us/customers/overview

Dingo, R. (2012). *Networking arguments: Rhetoric, transnational feminism, and public policy writing.* University of Pittsburgh Press.

Eli Review. About Eli Review. http://elireview.com/about/

FAQs. http://turnitin.com/en_us/what-we-offer/faqs

Features: Overview. (2015). http://turnitin.com/en_us/features/overview

Foltýnek, T., Rybička, J., & Demoliou, C. (2013). Do students think what teachers think about plagiarism? In *Plagiarism Across Europe and Beyond 2013 Conference Proceedings* (pp. 127-135). Brno: Mendel University.

Foltýnek, T. & Surovec, M. (2015). Promoting academic integrity helps national economy. In *Plagiarism Across Europe and Beyond 2015 Conference Proceedings* (pp. 121-133). Brno: Mendel University.

Giroux, H. (2007). *The university in chains.* Paradigm Publishers.

Glendinning, I. (2015). Ippheae: Past, present and future. In *Plagiarism Across Europe and Beyond 2015 Conference Proceedings* (p. 7). Brno: Mendel University.

Global Effectiveness. http://turnitin.com/assets/en_us/media/global-effectiveness/

Grabill J. (2016). Do we learn best together or alone? Your life with robots. *Computers & writing conference,* May 20, 2016. Web. http://elireview.com/2016/05/24/grabill-cw-keynote/

Grow, S. (2013). A cultural bridge for the academic concept of plagiarism: A comparison of Chinese and British cultural concepts of plagiarism by Chinese master's graduates of UK institutions employed by Sino-foreign joint ventures in Shanghai, China. In *Plagiarism across Europe and beyond 2013 conference proceedings* (pp. 27-41). Brno: Mendel University.

Hayes, N., & Introna, L. (2005). Systems for the production of plagiarists? The implications arising from the use of plagiarism detection systems in UK universities for Asian learners. *Journal of academic ethics, 3*(1), 55-73. https://doi.org/10.1007/s10805-006-9006-4

Herrington, A., & Moran, C. (2001). What happens when machines read our students' writing? *College English 63*(4), 480-499.

Hesford, W. & Schell, E. E. (2008). Introduction: Configurations of transnationality: Locating feminist rhetorics. *College English, 70*(5), 461-470.

Homepage. (2016). turnitin.com.

Howard, R. (1999). *Standing in the shadow of giants: Plagiarists, authors, collaborators*. Ablex.

Howard, R. (2000). Sexuality, textuality: The cultural work of plagiarism. *College English*, (*62*)4, 473-491

Hoyt, S. M. (2015). Copy/Paste/Culture: Plagiarism education week at K-state libraries. *K-State Today,* http://www.k-state.edu/today/announcement.php?id=19646

Introna, L. D., & Hayes, N. (2011). On sociomaterial imbrications: What plagiarism detection systems reveal and why it matters. *Information and organization, 21*(2), 107-122. https://doi.org:10.1016/j.infoandorg.2011.03.001

Janssens, K. & Tummers, J. (2015). A pilot study on students' and lecturer's perspective on plagiarism higher professional education in Flanders. In *Plagiarism across Europe and beyond 2015 conference proceedings* (pp. 12-23). Brno: Mendel University.

Kannan, V. (2014). Rhetorics of song: Critique, persuasion, and education in Woody Guthrie and Martin Hoffman's "deportees" (Master's thesis). UMI Dissertation Services from ProQuest.

Kokkinaki, A., Iacovidou, M. & Demoliou, C. (2015). Students' perceptions on plagiarism and relevant policies in Cyprus. In *Plagiarism across Europe and beyond 2015 conference proceedings* (pp. 192-200). Brno: Mendel University.

Kravjar, J. (2015). SK antiplag is bearing fruit. In *Plagiarism across Europe and beyond 2015 conference proceedings* (pp. 147-163). Brno: Mendel University.

Kravjar, J. & Noge, J. (2013). Strategies and responses to plagiarism in Slovakia. In *Plagiarism across Europe and beyond 2013 conference proceedings* (pp. 201-215). Brno: Mendel University.

Krokoscz, M. & Putvinskis, R. (2013). Analysis of the perceptions of undergraduate students in business administration on the occurrence of academic plagiarism in Brazil. In *Plagiarism across Europe and beyond 2013 conference proceedings* (pp. 281-282). Brno: Mendel University.

LightSide Labs. (2015). http://turnitin.com/en_us/lightside-labs

Marsh, Bill. (2004). Turnitin.com and the scriptural enterprise of plagiarism detection. *Computers and Composition, 21(4)*, 427-438.

Meacheam, D. & Faifua, D. (2015). Perspectives on turnitin use in an Australian setting. In *Plagiarism across Europe and beyond 2015 conference proceedings* (pp. 37-53). Brno: Mendel University.

Moore, D. H. (2013). Instructors as surveyors, students as criminals: *Turnitin and the culture of suspicion*. In M. Donnely & R. Ingalls (Eds.), *Critical conversations about plagiarism* (pp. 101-118). Parlor Press.

National Council of Teachers of English. (2013a). (2013). Resolutions & sense of the house motions. Resolution 3. National Council of Teachers of English. http://www.ncte.org/cccc/resolutions/2013

National Council of Teachers of English. (2013b). Formative assessment that truly informs instruction. National Council of Teachers of English. http://tinyurl.com/4m6nb2ac

Novelskaitė, A. & Pučėtaitė, R. (2013). Plagiarism in Lithuanian academia: Formal definition and informal attitude. In *Plagiarism across Europe and beyond 2013 conference proceedings* (pp. 236-247). Brno: Mendel University.

Online grading. (2015). turnitin.com/en_us/what-we-offer/online-grading

Orim, S.M., Borg, E. & Awala-Ale, I. (2013). Students' experience of institutional interventions on plagiarism: Nigerian case. In *Plagiarism across Europe and beyond 2013 conference proceedings* (pp. 54-69). Brno: Mendel University.

Plagiarism across Europe and beyond 2013 conference proceedings. (2013). Brno: Mendel University. http://plagiarism.pefka.mendelu.cz/files/proceedings.pdf

Plagiarism Education Week: Copy/Paste/Culture. (2016). http://go.turnitin.com/webmail/45292/333498835/384a973cff3cbbd380925f4fd272e9ec

Poe, M., & Inoue, A. (2012). Racial formations in two writing assessments: Revisiting White and Thomas' findings on the English placement test after 30 years. In N. Elliot & L. C. Perelman (Eds.), *Writing assessment in the 21st century: Essays in honor of Edward M. White* (pp. 343-361). Hampton Press

Purdy, J. (2005). Calling off the hounds: Technology and the visibility of plagiarism. *Pedagogy, (5)*2, 275-296.

Purdy, J. (2009). Anxiety and the archive: Understanding plagiarism detection services as digital archives. *Computers and Composition*, *26(2)*, 65-77.

Revision Assistant. (2016). http://turnitin.com/en_us/what-we-offer/revision-assistant

Robillard, A. (2009). Pass it on: Revising the "plagiarism is theft" metaphor. *JAC, 29*(1/2), 405-435.

Saldaña, J. (2009). *The coding manual for qualitative researchers.* Sage.

Scholte, J. A. (2000). *Globalization: A critical introduction.* St. Martin's Press.

Spellmeyer, K. (1996). Testing as surveillance. In E. M. White, W. Lutz, & S. Kamusikiri, *Assessment of writing: Politics, policies, practices* (pp. 174-184). Modern Language Association.

Sutherland-Smith, W. (2008). *Plagiarism, the internet, and student learning: Improving academic integrity.* Routledge.

Third-Party Academic Research. (2016). http://turnitin.com/en_us/resources/research

Turner, C. (2014). Turnitin and the debate over anti-plagiarism software. *NPR.* https://tinyurl.com/3525t7z7

Turnitin.com. (2001). About us. *Wayback Machine*, https://web.archive.org/web/20010331090743/http://www.turnitin.com/new.html

Turnitin Scoring Engine. (2016). http://turnitin.com/en_us/what-we-offer

Vidali, A. (2011). Embodying/disabling plagiarism. *JAC, 31*(3/4), 752-769.

Vie, S. (2013a). A pedagogy of resistance toward plagiarism detection technologies. *Computers and Composition*, *30(1)*, 3-15.

Vie, S. (2013b). Turn it down, don't Turnitin: Resisting plagiarism detection services by talking about plagiarism rhetorically. http://cconlinejournal.org/spring2013_special_issue/Vie/

Vasiljevienė, N. & Jurčiukonytė, A. (2015). The problems of legal and ethical regulation: A case study of the plagiarism lawsuit. In *Plagiarism Across Europe and Beyond 2015 Conference Proceedings* (pp. 164-179). Brno: Mendel University.

Vojak, C., Kline, S., Cope, B., McCarthey, S., & Kalantzis, M. (2011). New spaces and old places: An analysis of writing assessment software. *Computers and Composition*, *28*(2), 97-111. https://doi.org/10.1016/j.compcom.2011.04.004

Wellman, N. & Fallon, J. (2013). International MBA students' academic malpractice: A quantitative survey. *Plagiarism across Europe and beyond 2013 conference proceedings* (pp. 70-91). Brno: Mendel University.

What We Offer. (2016). http://turnitin.com/en_us/what-we-offer/

Why Turnitin. (2016). http://turnitin.com/en_us/why-turnitin

Zimmerman, T. (2007). McLean students file suit against Turnitin.com: Useful tool or instrument of tyranny? *National Council of Teachers of English.* https://cccc.ncte.org/cccc/committees/ip/2007developments/mclean

Zwagerman, S. (2008). The scarlet P: Plagiarism, panopticism, and the rhetoric of academic integrity. *College Composition and Communication*, *59*(4), 676-710.

EDITORS AND RETROSPECTIVE CONTRIBUTORS

Laura Aull is Associate Professor and Writing Program Director at the University of Michigan, where she teaches English linguistics and writing pedagogy. She is editor of the *Assessing Writing* Tools & Tech Forum, and she is the author, most recently, of *How Students Write: A Linguistic Analysis* and the forthcoming book *You Can't Write That: 8 Myths about Correct English*.

Carolyn Calhoon-Dillahunt, a former CCCC Chair and TYCA Chair, teaches writing at Yakima Valley College, an open admissions Hispanic-serving Institution. She also helps coordinate program and institutional assessment within the Arts & Sciences division of the college and is engaged in departmental and college-wide equity work. Her scholarly interests center on pedagogy, assessment, and education policy. She has published articles in *The WPA Journal, TETYC,* and *CCC* and has co-authored chapters in *New Directions for Community Colleges* and the recently published collection, *Writing Placement in Two-Year Colleges: The Pursuit of Equity in Postsecondary Education.*

Brian Huot has been a full time writing teacher and writing program administrator since 1980. Currently he is Professor of English at Kent State University. He is past chair of the College Section Committee and Member of the NCTE Executive Committee (2006-2008) and a current member of the Council of Writing Program Administrator Executive Board. He is a contributing scholar to the literature on the teaching and assessing of writing and has served as consultant for various institutions. He is currently a member of the NCTE Consulting Network.

Diane Kelly-Riley is Professor of English and Vice Provost for Faculty at the University of Idaho. She studies writing assessment theory and practice, validity theory, race and writing assessment, public humanities and multimodal composition. She was editor of the *Journal of Writing Assessment* from 2011-2022. She published *Improving Outcomes: Disciplinary Writing, Local Assessment and the Aim of Fairness* with Norbert Elliot (MLA, 2021).

Ti Macklin is the Director of First-Year Writing at Boise State University where she teaches courses in composition and rhetoric. Her research interests lie largely in First-Year Writing and writing assessment with a particular focus on assessment at the individual, classroom, and programmatic levels. Her most recent work examines the experiences of graduate and undergraduate students in first-year writing. She served on the editorial staff of the *Journal of Writing Assessment* for nine years.

David H. Slomp is Professor of Literacy and Assessment at the University of Lethbridge where he is also serving as Associate Dean of Graduate Studies and Research in Education. Since 2017 he has been serving as Co-editor-in-chief of *Assessing Writing*. His research focuses on the ethics of writing assessment, the consequences of assessment design and use, and the development of writing ability.

Carl Whithaus is Professor of Writing and Rhetoric at the University of California, Davis. He studies the impact of information technology on literacy practices, writing assessment, and writing in the sciences and engineering. His books include *Multimodal Literacies and Emerging* Genres (University of Pittsburgh Press, 2013), *Writing Across Distances and Disciplines: Research and Pedagogy in Distributed Learning* (Routledge, 2008) and *Teaching and Evaluating Writing in the Age of Computers and High-Stakes Testing* (Erlbaum, 2005).

Kathleen Blake Yancey is Kellogg Hunt Professor and Distinguished Research Professor Emerita at Florida State University. She has served as president/chair of several US literacy organizations, including the Council of Writing Program Administrators, the Conference on College Composition and Communication, and the National Council of Teachers of English. She participates in US and global assessment efforts, including as faculty for the WASC Leadership Academy and the AAC&U ePortfolio Institutes, and as a board member for the Association of Authentic, Experiential, and Evidence-based Learning. Author/co-author of 100+ refereed articles and book chapters and author/editor/co-editor of 16 scholarly books, she has received multiple awards, including the FSU Graduate Teaching Award (twice); the Purdue University Distinguished Woman Scholar Award; and the CCCC Exemplar Award.

CONSIDERING STUDENTS, TEACHERS, AND WRITING ASSESSMENT

The editors and authors in this edited collection, available in two volumes, consider the increasing importance of students' and teachers' lived experiences within the development and use of writing assessments. Presenting key work published in the *Journal of Writing Assessment* since its founding in 2003, the collection explores five major themes: technical psychometric issues; politics and public policies shaping large scale writing assessments; automated scoring of writing; fairness; and the lived experiences of humans involved in assessment ecologies. The book also provides reflections from leading writing assessment scholars who examine how these themes continue to shape current and future directions in writing assessment.

Diane Kelly-Riley is Professor of English and Vice Provost for Faculty at the University of Idaho. She studies writing assessment theory and practice, validity theory, race and writing assessment, public humanities and multimodal composition. She was editor of the *Journal of Writing Assessment* from 2011-2022. **Ti Macklin** is the Director of First-Year Writing at Boise State University, where she teaches courses in composition and rhetoric. Her research interests lie largely in first-year writing and writing assessment. Her most recent work examines the experiences of graduate and undergraduate students in first-year writing. She served on the editorial staff of the *Journal of Writing Assessment* for nine years. **Carl Whithaus** is a Professor of Writing and Rhetoric at the University of California, Davis. He studies the impact of information technology on literacy practices, writing assessment, and writing in the sciences and engineering. His books include *Multimodal Literacies and Emerging Genres* (University of Pittsburgh Press, 2013) and *Teaching and Evaluating Writing in the Age of Computers and High-Stakes Testing* (Erlbaum, 2005).

Perspectives on Writing
Series Editors: Rich Rice and J. Michael Rifenburg

The WAC Clearinghouse
Fort Collins, Colorado 80523
wac.colostate.edu

University Press of Colorado
Denver, Colorado 80202
upcolorado.com

ISBN 978-1-64215-216-6