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