When a Measure Becomes a Target: The Dangers of Using Grades in Writing Center Assessment

Bruce Bowles Jr
Texas A&M University–Central Texas


BRUCE BOWLES JR.

In the moment, it can sometimes be difficult to explain a visceral reaction to a comment from a colleague, especially when that colleague is an administrator. This happened to me a few years ago when the former Director of Institutional Research and Assessment and the then-Associate Provost at my institution expressed an interest in using student grades and/or GPAs in order to assess the effectiveness of the University Writing Center (UWC). My immediate instinct was to resist. Inventing an argument in the spur of the moment, I discussed the inherent statistical noise in such a process and how it would be nearly impossible to isolate the UWC's influence on students' grades. This was a fair enough argument in my estimation; however, I knew that there was something more behind my reluctance to embrace such an assessment. Unable to put my finger on it, I needed to better understand my intuitive, emotional reaction.

Over time, I came to realize that my greater concern was with the manner in which such an assessment could potentially incentivize problematic practices. For support of this notion, I turned to scholarship in the wider field of education assessment, drawing upon two useful concepts—assessment washback and consequential validity—both of which are connected to economist Charles Goodhart's famous maxim, referred to as Goodhart's Law: When a measure becomes a target, it ceases to be a good measure. According to this maxim, if grades become the measure for the effectiveness of a writing center, they will inevitably become a target, incentivizing directive tutoring practices that are more quickly able to improve students' grades, many of which are antithetical to best practices in writing center pedagogy. Yet, to understand this argument, it is critical to understand the complicated relationship writing center scholarship has with quantitative assessments, in particular those that employ students' grades.

QUANTITATIVE ASSESSMENT AND WRITING CENTERS: A MIX OF BOTH INTRIGUE AND SKEPTICISM

Although writing center assessment should, ideally, be tethered to research-based approaches that help to improve writing centers, unfortunately, as Miriam Gofine observes, justifying that writing centers are a worthwhile investment to higher-level administrators tends to be the primary driving force behind writing center assessment (40). Other scholars have called traditional measures–such as the number of tutorials, number of students supported, and student satisfaction surveys–into question. Along these lines, Neal Lerner believes that "justifying our existences based upon how many students we work with will never get us very far" ("Counting Beans" 60). Julie Bauer Morrison and Jean-Paul Nadeau have even shown that the scores from

DOI: https://doi.org/10.37514/WLN-J.2025.49.2.02

student satisfaction surveys tend to decrease after students receive their grades, with the scores on student surveys falling from a 4.81 to a 3.74 average out of 5 in their study. (The scores did go back up, interestingly, when students were surveyed a year later.)

Furthermore, Isabelle Thompson argues, "Having to settle for satisfaction as an outcome equivalent to success in tutorials demonstrates the importance of developing measures of student learning to push forward both assessment planning and research in writing centers" (37). Thompson also believes that grades and SAT scores (as a baseline for where students began their college careers)—with a large enough sample—can be used to provide evidence of writing center effectiveness. Lori Salem has also demonstrated that students with lower SAT scores tend to use the writing center more often. In order to win arguments with administrators, James Bell advocates more summative, quantitative approaches, noting that "While formative evaluation remains necessary for program improvement, summative evaluation answers accountability questions from people who hold the purse strings" (9). He believes that the more qualitative approaches writing center professionals tend to favor are not effective when working with senior administrators since these approaches are oftentimes viewed as highly subjective. Overwhelmingly, there are a lot of fair critiques of qualitative assessment practices, and the drive for more quantitative assessment methods is a valid one. However, quantitative assessment practices are not always as straightforward as they appear.

Lerner's odyssey with grade-based assessment is perhaps the most intriguing. In "Counting Beans and Making Beans Count" (published in 1997), Lerner investigated whether students coming to the writing center received higher grades than those students who did not. However, in a 2003 article, Lerner calls his own—along with Stephen Newmann's—grade-based writing center assessments into question. He notes that these studies were operating off of three primary, yet faulty, assumptions concerning the measures being used: that students with low SAT scores are at a disadvantage in first-year composition courses, that final grades in first-year composition courses accurately reflect writing ability, and that students will receive the same grade in first-year composition regardless of instructor. Lerner goes on to demonstrate how all three assumptions are quite faulty and, as he professes, "about as statistically and logically sound as the flat tax" ("Searching for the 'Proof'" 62). Beyond the tenuous statistical and logical soundness of such quantitative methods, which will vary predicated on assessment, another sinister force lurks. If grades become a major metric for assessing a writing center, problematic consequences are potentially on the horizon.

CONSEQUENTIAL VALIDITY, ASSESSMENT WASHBACK, AND INCENTIVIZING POOR PEDAGOGICAL PRACTICES

A specific component of validity theory needs to be considered in writing center assessment—consequential validity. Samuel Messick, a psychologist and assessment expert, contends that the validity of any assessment needs to consider several factors, one of the most important being the intended and unintended consequences of the use of the assessment and the results it produces. As Messick asserts, "To appraise how well a test does its job, we must inquire whether the potential and actual social consequences of test interpretation and use are not only supportive of the intended testing purposes, but at the same time are consistent with other social values" (8). When determining the validity of any assessment, it is crucial that we pay attention to these consequences, including—and especially—those that might not be intended.

Such consequences are what Michael Kane, an expert in educational measurement, refers to as unintended systemic effects. Kane observes how testing programs and assessments "can have

substantial, unintended effects on how institutions function (e.g., on what is included in school curricula)," further arguing that "such systemic effects have become major concerns, especially in education" (49). This is what is commonly referred to as assessment washback, a phenomenon in which tests can begin to dictate curriculum and influence what is taught as well as valued in an institution. Assessment washback can be as simple and seemingly innocent as teachers emphasizing certain content before an assessment to ensure their students perform well or as insidious as altering an entire curriculum to ensure a strong performance on an assessment. The former is potentially an example of positive washback; if the assessment is well-aligned with the curriculum and the construct it purports to measure, the teachers' focus can improve teaching and learning. However, negative washback occurs when the assessment is not well-aligned with the curriculum. In these instances, the assessment starts to actually dictate the curriculum itself.

In the case of writing center assessment, the potential unintended consequences are quite obvious—both writing center directors and tutors may become overly focused on improving students' grades on the texts they bring to writing centers. If grades and student GPAs become a prime point of emphasis in a writing center assessment, there is a chance they will influence writing center pedagogy, accompanied by more directive, less student-centered approaches to tutoring, especially if the administration of the institution is a more quantitatively-driven, outcome-focused group. Making sure the text will receive a better grade might become the priority over more effective—but time-consuming—pedagogical methods meant to improve students' writing abilities and habits over the long-term.

This is where Goodhart's Law comes into play. To reiterate, Goodhart's Law states that when a measure becomes a target, it ceases to be a good measure. What Goodhart's Law calls attention to is a strong tendency to optimize for what is going to be measured, particularly when that measurement carries high import. An apt example of Goodhart's Law in academia is the *U.S. News & World Report* college rankings. In 1983, *U.S. News & World Report* decided to begin evaluating colleges, ranking them on excellence. Without any definitive measures for educational excellence, the journalists at *U.S. News & World Report* chose proxies for excellence instead (O'Neil 52). These proxies included the SAT scores of incoming freshmen, student-teacher ratios, acceptance rates, retention rates, graduation rates, alumni donations, etc. (O'Neil 52-53). This algorithm has since had unintended—and rather devastating—consequences. Many colleges optimize solely for the proxies that affect their ratings while ignoring other practices that would better improve overall educational quality.

If writing centers use grades and/or GPAs in their assessment practices (a proxy for the effectiveness of tutorials—I would argue) they are not necessarily creating an explicit incentive program, but they may be incentivizing pedagogical approaches that will attempt to improve students' grades during tutorials. This is much more likely in educational environments that become focused on hitting targets and tether funding and continued support to particular metrics. Ideally, writing center professionals would not succumb to such temptations, yet if the performance of a writing center on such an assessment were tethered to funding, tenure for the director, etc., the incentivization is apt to be strong. Despite the fact that improvement in grades can be beneficial for students to a certain extent, the tutoring practices that would be used to achieve them could undermine students' long-term growth as writers in an effort to obtain short-term improvement in grades by employing more directive tutoring practices. And, as philosopher Ruth Grant claims, "An incentive that serves a legitimate purpose must be judged ethically illegitimate when it undermines a more important competing purpose" (63). Incentivizing such

behavior may undermine the instructional nature of writing centers and shift writing centers away from a more process-oriented approach.

SATISFYING THE DESIRE FOR MEANINGFUL DATA WITH A MORE QUALITATIVE APPROACH

I am also leery of completely focusing on proving effectiveness as the primary goal of writing center assessments. Rather, it might be better to focus on improving effectiveness. Over the last few years, I drew upon the work of R. Mark Hall in order to enact an assessment of our tutoring practices in the UWC. The staff and I developed a list of 10 Valued Practices for our UWC. As Hall notes, the work that goes into generating such a list is rewarding in and of itself. The UWC staff had lively discussions as we took an initial list generated by a graduate student tutor and me (as part of a project for his independent study) and revised it, cutting certain values, adding new ones, and arguing over seemingly miniscule particularities that actually proved quite important when we got to the core of the issue. For instance, the emphasis on positive reinforcement in tutorials (i.e., Value #4: Identified, or had the student identify, at least three positive elements of the text and/or writing process that were useful for praise and encouragement) and student agency (i.e., Value #9: Ensured student was granted primary ownership for revisions made during the session) that came out of these conversations drove at core principles we discussed throughout tutor training and staff meetings; however, once they were codified as values, everyone was more aware of them and—in particular—whether they were actually being enacted. An equally lively conversation emerged three years later when we repeated the assessment as we revised the first list for the second cycle, connecting to Hall's observation that "shared principles and propositions for observing might lead us to unearth—and, perhaps, critically examine—underlying values and assumptions guiding tutoring routines" (16). Our discussions definitely proved rather fruitful in this regard; the generation and revision of the 10 values actually served to define—and at times reinforce—what we truly valued in the UWC.

Both times the UWC conducted the assessment, I worked with the veteran tutors (those with more than one year experience) to norm how we would evaluate tutorials based on the scoring sheet we generated. Throughout the year, we collaboratively observed 100 tutorials, often when the veteran tutors had downtime or as part of my own formal observations of the tutors. The data were completely anonymous; no tutor was held accountable for a poor performance. However, when the data were collected and analyzed, it did allow the UWC staff and me to see where we were performing admirably and where we might not be doing as well as expected. Three years later, when we repeated the assessment, we were able to track our growth across the 10 values. For instance, the UWC saw a remarkable improvement on Value #10, which focused on creating revision goals for the student for after the consultation (or before the next consultation). This was encouraging since I made this a major area of focus in tutor training and staff meetings after the performance during the initial assessment cycle was not as impressive as the UWC staff and I would have hoped.

I use this example not as a form of self-congratulation nor as a model I believe everyone should replicate. Far from it. (The model is not even mine.) This assessment was successful, though, since it tethered to the rhetoric of the institution itself. Continuous improvement is a major point of emphasis when discussing assessment at my institution. Rather than using assessment to demonstrate our effectiveness, we were able to demonstrate how effective we were at striving for continuous improvement. Additionally, we demonstrated the value of our qualitative approach to assessment. (To be fair, it also helped that the UWC excelled on traditional metrics—students visited us quite frequently and valued our services, which is evidenced through our surveys and stories the administration had heard themselves.)

HOW RAD DO WE WANT TO BE?

Calls for, and the implementation of, RAD (replicable, aggregable, and data-driven) research and assessment strategies abound in academia and in writing center studies. Such approaches can be immensely beneficial and provide insight into institutional trends, educational practices, etc. Nevertheless, they can also come with a host of unintended consequences. In the end, assessment tells us as much about what we value in our programs as it does about the performance of our programs. When considering using grades and/or GPAs in writing center assessment, the concept of assessment washback and Goodhart's Law demonstrate that there is a significant risk of creating a target out of such a measurement, of making grades the valued priority over learning.

Although it is tempting to think writing center professionals can avoid such perils, incentivization is one of the most powerful forces on human behavior. In particular, as Grant argues, "we need to remember that incentives are a form of power as well as a form of trade" (41). They can exert a strong influence over people and control behavior, even if they are offering something in return. By giving in to demands, whether explicit or implicit, to tether writing center assessments to students' grades, writing center professionals leave themselves vulnerable in a variety of ways. The assessment can backfire, and grades might not correlate, or—even worse—negatively correlate, with writing center attendance. Grade improvement could become the sole or primary currency by which the writing center is evaluated. And, even if the results are positive, if a writing center is demonstrated to improve students' grades, the tendency will only further the demand for such results.

When designing writing center assessments, then, we need to carefully contemplate one question in particular: What are our assessment practices incentivizing? Consequential validity matters substantially when assessing a writing center; the wrong measurement can skew goals and priorities in unintended ways. The dangers these unintended systemic effects can create are often difficult to deal with once they manifest. For this reason, consequential validity needs to be of paramount concern when designing assessments for writing centers. And, ideally, considerations of consequential validity should occur in the planning stages as well as after the assessment has been enacted. Similar to medicine, prevention is often better—and less costly—than treatment.

WORKS CITED

Bauer Morrison, Julie, and Jean-Paul Nadeau. "How was Your Session at the Writing Center? Pre- and Post-Grade Student Evaluations." *Writing Center Journal*, vol. 23, no. 2, 2003, pp. 25-42, https://doi.org/10.7771/2832-9414.1516.

Bell, James. "When Hard Questions are Asked: Evaluating Writing Centers." *Writing Center Journal*, vol. 21, no. 1, 2000, pp. 7-28, https://doi.org/10.7771/2832-9414.1458.

Gofine, Miriam. "How Are We Doing? A Review of Assessments within Writing Centers." *Writing Center Journal*, vol. 32, no. 1, 2012, pp. 39-49, https://www.jstor.org/stable/i40135922.

Grant, Ruth. *Strings Attached: Untangling the Ethics of Incentives*. Princeton UP, 2011.

Hall, R. Mark. *Around the Texts of Writing Center Work: An Inquiry-Based Approach to Tutor Education*. Utah State UP, 2017.

Kane, Michael. “Validating the Interpretations and Uses of Test Scores.” *Validity, special issue of Journal of Educational Measurement*, vol. 50, no. 1, 2013, pp. 1-73, https://doi.org/10.1111/jedm.12000.

Lerner, Neal. “Counting Beans and Making Beans Count.” *Writing Lab Newsletter*, vol. 22, no. 1, 1997, pp. 1-4, https://wac.colostate.edu/docs/wln/v22/22.1.pdf.

---. “Writing Center Assessment: Searching for the ‘Proof’ of Our Effectiveness.” *The Center Will Hold*, edited by Michael Pemberton and Joyce Kinkead, Utah State UP, 2003, pp. 58-73.

Messick, Samuel. “Meaning and Values in Test Validation: The Science and Ethics of Assessment.” *Educational Researcher*, vol. 18, no. 2, 1989, pp. 5-11, https://doi.org/10.2307/1175249.

Newmann, Stephen. "Demonstrating Effectiveness." *Writing Lab Newsletter*, vol. 23, no. 8, 1999, pp. 8-9.

O’Neil, Cathy. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Broadway Books, 2016.

Salem, Lori. “Decisions…Decisions: Who Chooses to Use the Writing Center?” *Writing Center Journal*, vol. 35, no. 2, 2016, pp. 147-171, https://doi.org/10.7771/2832-9414.1806.

Thompson, Isabelle. “Writing Center Assessment: Why and a Little Bit of How.” *Writing Center Journal,* vol. 26, no. 1, 2006, pp. 33-61, https://doi.org/10.7771/2832-9414.1592.