Assessing Writing 42 (2019) 100418

Contents lists available at ScienceDirect

Assessing Writing

journal homepage: www.elsevier.com/locate/asw

Evidence of fairness: Twenty-five years of research in *Assessing Writing*

Mya Poe[a,*], Norbert Elliot[b]

[a] Northeastern University, United States
[b] University of South Florida, United States

ARTICLE INFO

Keywords:
Bias
Ethics
Fairness
Justice
Literacies
Validity

ABSTRACT

When *Assessing Writing* (*ASW*) was founded 25 years ago, conversations about fairness were very much in the air and illustrated sharp divides between teachers and educational measurement researchers. For teachers, fairness was typically associated with consistency and access. For educational measurement researchers, fairness was a technical issue: an assessment that did not identify the presence of β (the bias factor) was fair. Since its founding, *ASW* has continued to be a space where evolving discussions about fairness play out. In this article, we examine a selection of 73 *ASW* research studies published from 1994 to 2018 that use fairness as a category of evidence. In tracing the use of fairness and related terms across these research articles, our goal is to understand how the conversation about fairness has changed in the last quarter century. Following a literature review that situates fairness within generational, standards-based, and evidential scholarship, we analyze five trends in the journal: fairness as the elimination of bias; fairness as the pursuit of validity; fairness as acknowledgement of social context; fairness as legal responsibility; and fairness as ethical obligation. A tidy narrative that theoretical conceptualization of fairness has deepened over the *ASW* lifespan is not born out by our findings. Instead, evidence suggests that the disparate stances and methodological challenges that informed early research on fairness remain. As well, the textual record suggests that we have not developed or shared taxonomies for systematically investigating questions of fairness. In our desire to make the research we present actionable, we close by calling attention to the need for theorization of fairness, the advantages nuanced of research methods, and the benefits of non-Western perspectives.

1. Introduction

In the first editorial to *Assessing Writing* (*ASW*) in 1994, Brian Huot wrote about the genesis of the journal, which stemmed from the 1992 New Directions in Portfolios Conference (Black, Daiker, Sommers, & Stygal, 1994; Black, Helton, & Sommers, 1994). As Huot explained in his introduction, there were few venues for writing researchers to publish on assessment despite the wealth of scholarship presented at conferences. For Huot and co-editor Kathleen Blake Yancey, the journal was a space where "the crucial relationship between pedagogy and assessment" would be valued (Huot & Yancey, 1994, p. 143). There would not be a sole focus on educational measurement perspectives; instead, as the editors established in their first editorial of the second volume, *ASW* would "contribute to the increasingly interesting and divergent conversations about assessment currently taking place" (Huot & Yancey, 1995, p. 1).

Indeed, early issues of the journal created a space where teachers of writing, writing program administrators, and writing

* Corresponding author.
E-mail address: m.poe@northeastern.edu (M. Poe).

https://doi.org/10.1016/j.asw.2019.100418
Received 24 February 2019; Received in revised form 4 July 2019; Accepted 12 July 2019
Available online 30 August 2019

researchers could find a place to work through the issues that interested them in the ways they believed were viable for their programs and students. The traditions of measurement and composition would be in conversation. Notions of fairness would be important here. There was a sense of justice—that what researchers were publishing in the journal was not solely about assessing writing more efficiently or accurately; rather, the research was about making the teaching and assessment of writing fairer. In other words, the conversation was not just about technical advances in-and-of themselves but about examining the social conditions created by and through writing assessment. In fact, one might argue that because of its resonances with justice, democracy, and social good, the very notion of fairness has been integral to *ASW* since its inception.

Two articles from the first issue of *ASW* illustrate how fairness informed the journal from its inception. In "Validity in High Stakes Writing Assessment: Problems and Possibilities," Pamela Moss, a trained psychometrician then assistant professor at University of Michigan, wrote: "Extensive research has been conducted on how to develop and score standardized writing tasks to provide reliable, valid, and fair estimates of students' writing abilities (e.g., Breland, Camp, Jones, Morris, & Rock, 1987; Huot, 1990; Ruth & Murphy, 1988)" (1994, p. 109). In citing Hunter Breland, a former engineer and then researcher at Educational Testing Service (ETS), in relation to Brian Huot, then assistant professor of English at University of Louisville who advocated for linking writing assessment with learning, Moss was clearly making connections between the measurement and composition communities. Moss went on to connect questions of fairness to "consequential decisions about individuals and programs," opportunity to learn, and research on "the cognitive and social aspects of learning" (p. 110, p. 109).

If Moss was attempting to put competing notions of fairness in dialog, Michael Williamson (1994), an English professor who specialized in writing assessment and had studied with the psychometrician Michael J. Zieky at ETS, was interested in thinking about the history of education and the implications of fairness in testing beyond educational settings. In "The Worship of Efficiency: Untangling Theoretical and Practical Considerations in Writing Assessment," Williamson associated fairness with rise of rationalist methods at the beginning of the 20th century: "The positivist science of psychometrics that developed in the late nineteenth century connected to [a] shift in education began to provide assessment tools believed to be objective and fair because they were seen as independent of the bias of the human decisions of individual teachers" (Williamson, 1994, p. 151). Here, Williamson identified a concept of fairness associated with assessment methods designed to be distributed across settings—and, hence, often more reliable than valid or fair. He went on to locate another concept of fairness in the "bureaucratic model" of contemporary education, one in which there is a need for fairness in terms of equitable impact. In his description, fairness is related to treatment and the downstream implications of decisions related to access or exclusion (p. 151). Clearly, Williamson was thinking about the ramifications of assessment beyond a single practice and wanted teachers and researchers to think more expansively about the implications of experimental design and social consequence.

We provide as examples these articles in the first two issues of *ASW* to illustrate the complex discussions about fairness that researchers were already having in the field at the inception of the journal. Note that terminology about race, class, gender, or linguistic difference are not prominent in the quoted articles but that concerns about social inequality are certainly present. Emphasis on these two early articles is a useful lesson that current conversations about fairness, and the very meaning of fairness itself, are not new; rather, discussions of fairness in writing assessment are rooted in much deeper philosophical and methodological discussions in the field. These two articles, both published in 1994, therefore, identify an initial and enduring concern with evidence of fairness that continue to the present writing. That concern has taken many forms, as we will show: from the U.S. founding of the journal in 1994 under Huot and Yancey until 2000; through its internationalization under the editorship of Liz Hamp-Lyons from 2002 to 2017; and continuing under the current editorship of David Slomp and Martin East and their focus on the consequential dimensions of validity, reliability and fairness in international settings.

In the following article, written in response to the call from Slomp and East in their first issue as editors as they aimed to frame the future of writing assessment, we trace this evolving conversation in the pages of *ASW* from 73 selected research articles that use fairness and related terms as key words. After establishing our literature review (§2), research questions (§3) and methods (§4), we present five trends we identified in the 73 articles: fairness as the elimination of bias (§5.1); fairness as the pursuit of validity (§5.2); fairness as acknowledgement of social context (§5.3); fairness as legal responsibility (§5.4); and fairness as ethical obligation (§5.5). We then discuss our findings in terms of our research questions (§6) and conclude with recommendations (§7) and conclusions (§8) that may prove useful to writing assessment researchers.

2. Literature review

Three ways to review the scholarship related to fairness in writing assessment have informed the current study. The first way, as demonstrated by Dorans (2011), is to establish periodization through a generational approach to international educational measurement. A second way is to focus on U.S. educational measurement standards as they have changed from 1952 to 2014. A third way is to focus on changes in the way researchers have used evidence related to fairness. Other ways to review scholarship related to fairness could include mapping evolving methods, such as those used to describe demographic populations, or by mapping shifts in the teacher–researcher literature related to responding to student writing. In the case of demographic populations, for example, such a study would focus on the terminology used to describe different groups (e.g., by racial group or linguistic identity) to highlight shifting feedback concerns related to the question "fairness for whom?"

Table 1
Article Categorization (n = 73).

Definition of Fairness: *Standards for Educational and Psychological Testing*	
American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1985	No definition
American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999	"In testing, the principle that every test taker should be assessed in an equitable way" (p. 175).
American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014	"The validity of test score interpretations for intended use(s) for individual from all relevant subgroups. A test that is fair 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).

Evidence Related to Fairness: Appendix B				
Bias (n = 25)	Validity (n = 19)	Social (n = 18)	Legal (n = 8)	Ethics (n = 3)

2.1. Four generations of psychometric theory

Using periodization, Neil J. Dorans (2011) has identified four generations of educational measurement research. Because of the usefulness and brevity of the history, it is worth quoting Dorans in detail:

> The first generation, which was influenced by concepts such as error of measurement and correlation that were developed in other fields, focused on test scores and saw developments in the areas of reliability, classical test theory, generalizability theory, and validity. This generation began in the early twentieth century and continues today, but most of its major developments were achieved by 1970. The second generation, which focused on models for item level data, began in the 1940s and peaked in the 1970s but continues into the present as well. The third generation started in the 1970s and continues into today. It is characterized by the application of statistical ideas and sophisticated computational methods to item level models, as well as models of sets of items. The current fourth generation attempts to bridge the gap between the statistician/psychometrician role and the role of other components of the testing enterprise. It recognizes that testing occurs within a larger complex system and that measurement needs to occur within this larger context. (259)

Here we see the evolution of fairness as it has evolved: as the elimination of systematic error; as the examination of differential item functioning; as the use of item response theory; and as the exploration of sociocognitive models. Overall movement extends from a technical view of fairness (as bias) to an embedded view of fairness (as situated within contexts). In later work, Dorans (2017) has provided a review with research associated with each generational phase, save the earliest, which may be identified with the early nineteenth century work of Gauss on the elimination of error in physical measurement.

2.2. U.S. Educational measurement standards

Elliot (2015) has provided a complementary history to Dorans's narrative by focusing on United States' efforts to standardize measurement techniques associated with experimental research. While versions of *The Standards for Educational and Psychological Testing* were published from 1952 to 2014, Elliot notes, it is only with the most recent version that fairness as a form of evidence achieved equal standing (at least in the table of contents) with validity and reliability. During the publication period of *ASW*, three editions of the *Standards* are relevant. Definitions of fairness from the 1985, 1999, and 2014 editions are provided in Table 1.

As the definitions reveal, the 1985 edition may be firmly placed in what Dorans (2011) terms the second generation, with its emphasis on models for item level data. That is, while the glossary includes a definition for differential item functioning, there is no definition of bias. It is with the 1999 edition that views of fairness as equitable assessment first appear. In general, it may be said that it is the *fin de siècle* edition that ushers in third generation assessment. With the 2014 edition, we see a more fine-grained definition of fairness that attends to the impact of the interpretation and use argument discussed in §4.2.

2.3. Elimination of bias

Accompanying Dorans (2011) and Elliot (2015) is a third possible way to review the scholarship related to fairness in writing assessment—by examining evidence related to fairness. For example, an examination of the evidence related to Classical Test Theory (CTT) and Item Response Theory (IRT) would allow detailed analysis of bias research. Because there are important technical

considerations related to identification of bias, we provide a brief explication models related to CTT and IRT traditions in Appendix A.

The history of CTT can be dated to Charles E. Spearman (1904). His work serves periodization purposes to date the beginning of second generation research as characterized by Dorans (2011). In CTT, mathematical models are used in which an individual's observed score on a given assessment is the sum of the true score (the hypothetical average of scores that an individual would earn on an unlimited number of parallel test forms) and independent random error (a non-systematic error that has no relationship to the variables under examination in the test) (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014, p. 216, p. 225).

The first conceptualization of IRT appeared with Frederick M. Lord (1952). His work serves periodization purposes to date the beginning of third generation research as characterized by Dorans (2011). In IRT, mathematical models are used in which an individual's observed score is a functional relationship among test performance, the characteristics of test items, and the test-taker's rank on the variables under examination (AERA, APA, & NCME, 2014, p. 221; Yen & Fitzpatrick, 2006). With the advent of computer adaptive testing for students, IRT rose from a theory used mostly by specialists to one widely used internationally after 1980 (Carlson & von Davier, 2017, p. 133). While there are many forms of IRT analysis, Rasch techniques, i.e., those named after their Danish founder (Rasch, 1960), are of most interest to us. Rasch analysis can provide identification of item difficulty and test taker ability on the same scale, thereby offering precision in terms of interactions between individual items and individual student performance (Boone & Noltemeyer, 2017). For many researchers, including the authors of the present paper, the elision between bias and interaction that we see in our review of *ASW* article is troubling, as we explain in the Appendix A. While it is certainly true that the majority of research reported in *ASW* operates under CTT models, it is equally true that research in the journal has functioned under IRT models extended to Rasch analysis. Most commonly-used sources of evidence in this group include use of test scores, although mixed methods are also used to provide other forms of evidence related to score determination and difference

3. Research questions

Informed by the literature review, four questions guided our research as we investigated fairness as an evidential category in the journal:

1 How have writing assessment researchers constructed fairness?
2 How have those constructions either directly or indirectly revealed categories of evidence related to fairness?
3 Have there been major shifts in the use of fairness as a guiding research principle in the last 25 years?
4 Are there ways that fairness has not been used by writing assessment researchers that might prove fruitful?

Through the answers to these questions, our larger goal was to suggest how the journal both reflects the state of knowledge in the field of writing assessment and how it might be an active agent in shaping the future of knowledge in the field.

4. Methods

Drawing on Petticrew and Roberts (2006), our systematic review was based on an analysis of *ASW* issues from 1994 (Vol. 1) to 2018 (Vol. 39). To answer our research questions, we used a three-phase approach that included keyword analysis, categorical analysis, and interpretative analysis.

4.1. Keyword analysis

We first conducted a keyword search using the Science Direct search tool to identify articles that included "fair" as a key word (Scott & Tribble, 2006). As shown in Table 2, the original "fair" search yielded 162 results, including 133 research articles. We also found under the term "fair" 14 book reviews, 10 editorials, as well as review articles and correspondence. Given the expansive number of findings, we limited our review to research articles.

In addition to the search term "fair," we expanded our search to include related keywords, such as "bias," "justice," and "ethics." These terms were selected because of their historical association with fairness. If one were searching more than 10 years ago for articles about fairness, the term "bias" would have likely been used. Today, the term fairness has been discursively tied to terms like ethics, law, human rights, and justice.

We then reviewed each of the articles to determine how the keyword was being used in the article. For example, the term "justice" might be used to refer to a study about criminal justice majors, rather than social justice. Likewise, some articles used a term, such as "fairness," generically. These articles were not included for further study because they were not grounded in a theoretical orientation to the use of the term.

While the keyword analysis told us about the frequency with which certain words were used in *ASW*, it did not tell us how those terms were being used, especially in relation to the kinds of arguments researchers were making in their research articles. To ascertain this information, we turned to recent validity research.

Table 2
Keyword Terms Used in Science Direct.

Term	Original Search	Research articles	Refined Search
Fair*	162	133	37
Bias*	139	119	25
Law or Legal	44	37	8
Justice	25	19	0
Ethics*	42	32	3
Democracy*	10	9	0
Human rights	4	4	0
Total articles	426	353	73

Note: Asterisks indicate that varied terms were used in the keyword search (e.g., fair, fairness, fairly, and so forth).

4.2. Categorical analysis

In 1988, Cronbach invited researchers to think of a "validity argument" instead of "validation research" (4). Recognizing the contingencies involved in validation processes, researchers began referencing what has come to be known as an interpretation and use argument (IUA). In the IUA approach, claims are based on the "network of inferences and assumptions inherent in the proposed interpretation and use" (Kane, 2013, p. 2). Evidence is collected according to categories. Such categorical analysis informed how we parsed our findings in the second step of our methods through categorical analysis. Drawing on three editions of the *Standards for Educational and Psychological Testing* that informed research in *ASW* from its origin to the present (AERA, APA, & NCME 1985, 1999, 2014), we were able to group the 73 studies into five categories of evidence, as shown in Table 1: bias (n = 25), validity (n = 19), social (n = 18), legal (n = 8), and ethics (n = 3). Authors, titles, dates, and categorization are provided in Appendix B.

Within each of the categories we identified in the second step in our methods, we further analyzed the articles using an interpretative content analysis approach, emphasizing how the term fairness or the related term was used. As necessitated by our literature review, within each of these groups we looked for shifts over time in how writers were using ideas related to fairness.

4.3. Interpretative analysis

As Levine (in press) demonstrated in an analysis of New York State English Language Arts examinations from 1900 to 2018, interpretative analysis of content has been used in many forms to map the landscape of teaching and assessing writing in the United States. In point of fact, this special issue is part of interpretative analysis historiographic traditions in writing assessment (Behizadeh & Engelhard, 2011; Elliot, 2005; Haswell & Elliot, 2019). Such studies have also been used to examine U.S. composition journals from 1979 to 2018 (Wood & Elliot, in press) and from 1984 to 2019 (Hesse, 2019). Content analysis has also been used to study demographic categories related to disaggregation of data (Poe, 2009) as well as analysis of genre in writing assessment (Beck & Jeffery, 2007). We, therefore, proceed essayistically, an approach drawn from tradition and necessity. As is clear from the second phase of our methods, methodological techniques such as coding at the nodes (Strauss & Corbin, 1998) would not have served us well due to the shifting nature of measurement demonstrated in the literature review and, within it, the categories of evidence identified in Table 1 and made granular in Appendix B.

5. Five trends

Our findings related to the presence and shifts in fairness research in *ASW* are best presented visually as well as textually. First, as Fig. 1 reveals, the 1999 edition of the *Standards* captured the educational measurement zeitgeist that had an impact of the journal—that discussions about fairness were nascent but would have a lasting impact on the research published in *ASW*. As Fig. 1 shows, though, there has been a dearth in philosophical articles related to fairness. Fig. 1 also points to the internationalization of the journal under the editorship of Liz Hamp-Lyons from 2002 to 2017 in which there was a demonstrated sharp rise in fairness research related to bias, validity, social, and legal evidence. As we will demonstrate, and as the Fig. 1 shows, however, different methods and models would come to inform fairness research. Our review of research articles published in *ASW* over the last 25 years demonstrates that references to fairness in *ASW* articles were often associated with validity, especially in more recent work that drew on validity as manifested in interpretation and use arguments, IUA approaches discussed in §4.2 and §5.2. This finding was not unexpected. Likewise, the large corpus of articles that relied on the terms "bias" to describe issues related to fairness was not surprising. What was surprising, however, and what we describe below, was the use of the term "bias" was often not used in sense of classical test theory.

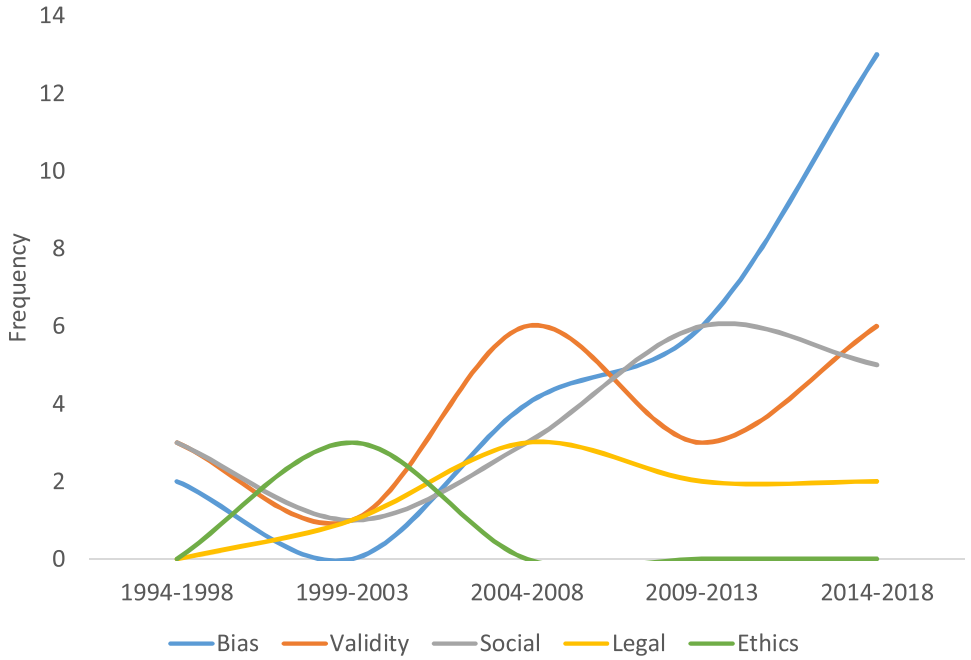

Fig. 1. Illustration of Five Trends of Fairness Research in *Assessing Writing*, 1994–2018.

Beyond this surprise, we also found that sociocultural theories of writing have played far less of a role in the journal from a fairness perspective—a surprising finding, given the call within the journal to include such theories in writing assessment scholarship (Camp, 2012).

Beyond our visual analysis, we discuss the five trends of research in *ASW* from 1994 to the present that relate to fairness: fairness as the elimination of bias; fairness as the pursuit of validity; fairness as acknowledgement of social context; fairness as legal responsibility; and fairness as ethical obligation. For each of these trends, we offer a brief theoretical orientation and then discuss representative articles.

5.1. Fairness as the elimination of bias

In analyzing the 25 articles we categorized in Appendix B as related to bias, we observed the influence of Classical Test Theory and Item Response Theory.

5.1.1. Classical test theory to examine bias

In a U.S. study published in the early years of the journal, Haswell and Haswell (1996) used assumptions of Classical Test Theory to examine the way that knowledge of gender may impact reader response. Their sampling plan included 32 post-secondary writing instructors and 32 first-year college students. Using relational modeling, they studied interactions between four independent variables (*X*): gender of reader, gender of interviewer who prompted response during a recorded review session with students and teachers, status of the reader as student or teacher, and prior knowledge of gender. These were manipulated to examine their relationship with five dependent variables (*Y*): essay rating, positive critique, authorial agency classified according to gender knowledge or gender assumptions, discourse as related to critique categories, and gender cues identified by the reader. Methodologically, the study used general linear modeling: multiple analysis of variance, with Wilks's likelihood ratio criterion providing the inference testing to determine whether the multiple effect would be further explored through univariate analysis using *F*-tests. Identified were statistically significant interactions between gender and each of the independent variables. Of notable relevance, half of the readers formed an image of the writer's sex even when the writer's name was masked and the writer's sex was not provided by the writer in the essay. As Haswell and Haswell concluded, "[T]raditional gender research has probably underestimated the presence of gender bias in compositional settings" (p. 70). Referencing feminist theorist Sandra Harding (1986) and her claim that "there are no social activities that escape gendering," the researchers find parallels between their empirical research and sociological traditions.

The CTT research tradition remains central to the journal, and the targets of interest remain central. In their recent study of post-

secondary student perceptions of writing error, text quality, and writer characteristics Johnson, Wilson, and Roscoe (2017)) use general linear modeling to identify what they categorize as "a constant concern" for any assessment: the assigning of scores and ratings that systematically differ based on some factor unrelated to the construct of interest (p. 83). As they found in their post-hoc correlational analysis, mean trait judgments of text quality and author characteristics were strongly correlated (from 0.65 to 0.82) at high levels of statistical significance ($p < .001$) with absence or presence of error. As they conclude, "the presence of writing errors, and particularly lower-level errors, resulted in perceptions of authors as less intelligent, creative, and hard-working, as well as less kind, generous, or loyal. These personal judgments could reciprocally influence perceptions of the text" (p. 83). Over emphasis on knowledge of conventions as the main determinant of score assignment can, in fact, result in evidence of bias. As used in the journal, CTT has been especially useful in substantiating the fact that rank order inferences made from scores may, in fact, be due to bias.

5.1.2. Item response theory to examine bias

While CTT remains the main tradition for the examination of bias in the journal, the journal has also hosted innovative IRT designs. Notable among the early research is a comparison of generalizability theory and Rasch analysis in a study of U.S. college sophomores. Sudweeks, Reeve, and Bradshaw (2005) published the results of a pilot study to evaluate and improve scoring procedures using two models associated with IRT: generalizability theory and Rasch modeling. While the first method provides a way of partitioning the total variance in a set of ratings into separate, uncorrelated parts that are each associated with a different source of variability, the second method (in this case, a technique called Many-faceted Rasch Measurement) provides a way to examine the effects of other sources of systematic error such as rater inconsistency, differences in ratings across time, and level of difficulty differences in the among writing tasks. Examining 15 sources of variation in a sample of 48 essays scored on a 9-point holistic scale, the study demonstrated that G theory provided a broad analysis that identified single variance, while Rasch measurement provides data on a more granular level. Sudweeks, Reeve, and Bradshaw demonstrated the potential of IRT models as complementary and established a basis for such research that would, over time, be used to identify varied sources of variability that are identified as sources of bias if designated logit scales are exceeded. (See the Appendix A for more on Rach scales and their logic)

Examination of different sources of variation continues to be the main use of IRT studies in the journal. In "Going online: The effect of mode of delivery on performances and perceptions on an English L2 writing test suite," Brunfaut, Harding, and Batty (2018) used assumptions of IRT to investigate differences between delivering examinations in paper and online formats. Using the Trinity College London Integrated Skills in English (ISE) suite across three proficiency levels of the Common European Framework of Reference for Languages, a total of 459 test-takers completed both paper and online forms. The sampling plan of test-takers also completed a Likert scale survey to establish their personal backgrounds and gauge perceptions of the impact, usability and fairness of the delivery modes. Using Many-facet Rasch measurement (MFRM), the researchers used a four-facet model that was constructed for each ISE level: test-takers, order (first or second sitting), raters, and rating criteria on the tasks. With this design, the effect of delivery mode on ISE writing scores was examined in two ways: through a bias/interaction analysis of mode of delivery and task (thus providing a general description of the extent to which mode affected task performance); and through a bias/interaction analysis of mode and rating scale category (thus providing an exact view of where any effect was located). To examine background and perception, classical models were used to provide descriptive and inferential statistics. MFRM score analysis revealed that delivery mode had no discernible effect across the three proficiency levels. Analysis of the Likert-scales revealed that test-takers held more positive perceptions of the online delivery mode than of paper-based delivery. As used in studies such as this, IRT has been especially useful in alerting *ASW* readers to the many sources of variability in a given study. As well, emphasis on these intersections, rather than on a single source of bias, has proven useful documenting the fluid nature of bias.

5.2. Fairness as the pursuit of validity

In identifying the 19 articles we categorized in Appendix B as evidence of validity related to fairness, we analyzed each within the context of shifting views on validity over the last 25 years. It might accurately be said that mention of the 1992 New Directions in Portfolios Conference in the first editorial of the journal represented a need to refresh research on written communication with attention to robust construct validity, especially after the U.S. accountability movement had adopted restricted forms of assessment (such as timed, impromptu writing combined with multiple-choice testing) during the 1980s that had demonstrated differential validity for diverse student groups (White & Thomas, 1981).

In 2013, Michael T. Kane substantively advanced the relationship between fairness and validity. In writing about IUA in which claims are based on test scores, Kane paid considerable attention to score use (pp. 45–63). He concluded:

> I think that the evaluation of consequences of score-based decision rules should be included under the heading of validity, but in taking this position I have indicated the kinds of consequences that I think should be included (those with a potential for substantial impact in the population of interest, particularly adverse impact and systemic consequences). Negative consequences play a direct and immediate role in the evaluation of score uses, and they have a more limited and less direct role in the validation of underlying interpretations associated with score uses. (p. 61)

The 2014 *Standards,* designed with Kane and colleagues on the joint committee, provided guidance on fairness, and the connection to the unified model validity was clear. As the committee explained,

> Fairness is a fundamental validity issue and requires attention throughout all stages of test development and use…fairness to all individuals in the intended population of test takers is an overriding foundational concern, and that common principles apply in responding to test-taker characteristics that could interfere with the validity of test score interpretation. (2014, p. 49)

The *Standards* detail how fairness should be folded into validity: (1) through attention to the construct being measured, (2) through considerations of test-takers' response processes, and (3) through the disaggregated reporting of data. In short, the 2014 revision made apparent fairness as an evidentiary category.

Studies in the journal focusing on fairness as the pursuit of validity are usefully classified into two groups: those that deal with single targets of variance and those that deal with interactions among study variables. As in the case with studies framing fairness as the elimination of bias, the most commonly-used sources of evidence in this group include scores, although corpus-based methods have recently been used to shed light on interaction effects.

5.2.1. Validity, fairness, and rater accuracy

As is the case with the study reported in Sudweeks et al. (2005), Wang and colleagues used IRT theory as it is applied to Rasch modeling. In this case, two of the leading U.S. Rasch researchers—George Engelhardt and Edward W. Wolfe—work with colleagues in linking inter-rater reliability to fairness in large-scale writing assessments. Using a mixed methods approach, the researchers attempted to identify evidence that some writing samples were more difficult to score than others—and to suggest why that may be so based on rater justification—in assessments integrating reading comprehension and writing ability. Twenty raters scored 100 randomly selected student essays with a trait rating scale. Using a many-facet Rasch model—an extension of Rasch measurement models—rater accuracy model revealed essays receiving higher scores were more difficult to score accurately. Further, among the top difficult-to-score essays, there was a tendency for raters to use the middle categories, thus resulting in inaccurate ratings.

While we continue to note challenges involved with small sample sizes—and the accompanying generalization inferences—the study is impressive in its use of many-facet Rasch modeling to identify the different facts of rater response that contribute to difficult to score essays. In many ways, the study confirms the value of IRT modeling in considering an observed score as a functional relationship with writing performance, the characteristics of written response, and the test-taker's rank on examined variables. Because targeting rater accuracy provides information on both rater performance and rater cognition, Rasch modeling is uniquely suited to examine intricate relationships. Using IRT models, similar studies focusing on rather accuracy have been conducted by Huang (2012) and by Wind and Engelhard (2013).

5.2.2. Validity, fairness, and genre interaction

As is the case with the study reported by Haswell and Haswell (1996), variables associated with the construction of valid assessments yield important fairness inferences when performance is disaggregated by participant group. Disaggregation by genre also yields valuable information. Genre disaggregation is examined in an early study in the journal of expressive forms (story, poem, play, or other form) on the Maryland School Performance Assessment Program. Goldberg, Roswell, and Michaels (1998) examined issues of choice as they were related to "fairness of the assessment instrument" (p. 41). Using a random sample of approximately 120 student responses at each grade levels 3, 5, and 8, the researchers employed a methodology that used close textual analysis to determine the elements of writing that contributed to scores—and, hence, to relative difficulty—of the genres students uses. Surveys of the scoring team were used to determine perceived difficulty in making score decisions based on genre use, rubric design, use of anchor papers, and links to classroom assessment. Using Cohen (1988), effect size calculations were used to examine difference in performance due to student genre choice. The effect sizes were generally small once raw and scales scores were equated. In analyzing student demographic difference (gender and race/ethnicity) and performance levels (proficiency) by genre, however, effect size differences were larger between poems and stories than between plays and stories. This finding led to an important observation about choice: "limiting an expressive writing task to poetry alone might disadvantage students who are less able in other areas of English language arts" (p. 50). Analysis of reader surveys revealed that the expressive writing rubric was suited for scoring stories—but not for scoring plays. Readers also revealed that they had the most difficulty in scoring poetry. Differences in genre were examined to provide the reasons behind the effect size differences and the reader responses. For example, while readers are trained to honor students' topic choice, sentimental poems that used cliched phrases and images appeared to have interfered with the ability of readers score according to rubric-specific issues of development, order, and language.

In recent work, Barkaoui and Knouzi (2018) have used corpus-based methods to investigate the effects of writing mode (computer vs. paper) and computer ability on the scores as well as related linguistic characteristics of essays written in response to a second language writing test. In terms of interaction, they found that writing mode had significant effects on measures of fluency, lexical complexity, cohesion, and content—a reminder that the study of validity is best understood as the study of interactions as they appear in distinct contexts. It is this more recent tradition to which we now turn.

5.3. Fairness as acknowledgement of social context

Bias and validity were not the only traditions that informed research published in *ASW*. As we discover in the 18 articles identified in Appendix B, social theories of learning would also have an important influence on the journal. We note that social theories of learning used in teaching and assessing writing preceded fourth generation psychometric research described in §2.1.

A social view on fairness is informed by three overlapping changes in the study of writing. First, a social action view of genre had significant research impact. Beginning in the 1950s, social perspectives on communication advanced by British researcher J. L. Austin (1962) and later by American John Searle (1969) would challenge older linguistic theories about language use and cognition. In addition, the rediscovery of Mikhail Bakhtin's works on addressivity and interdiscursivity (e.g., "Discourse in the Novel" (1934-1935/1981) and "The Problem of Speech Genres" (1952–1953/1986) would also influence how writing researchers theorized genre. For researchers working from a Rhetorical Genre Studies orientation toward the study of writing, genres, as tools for social action, rarely (if ever) function in isolation; instead they interact with other genres to form genre sets and systems (Bazerman, 1994; Devitt, 1993). Moreover, because texts are always negotiated in social contexts, textual realizations, though typified, are not static entities; instead, they are better described as "stabilized-for-now" (Schryer, 1994). These typified rhetorical actions are not neutral; they are laden with ideology (Gee, 1990, 2008).

Second, in addition to social theories of texts, a number of other forces would shape how writing researchers would come to understand changes in writers themselves—i.e., writing development. One moment of impact was the 1966 Dartmouth Seminar where James Britton and colleagues proposed a growth model of writing, based on a process-based view of writing. Parsing the results of student writing gathered from 65 schools into sense of audience and sense of function (p. 112), Britton, Martin, McLeod, and Rosen in *The Development of Writing Abilities (11–18)* (1975) traced how student writing changed over seven years in terms of audience relationships and function. Based on their findings, they advocated for a "developmental role for writing in school" (p. 201). Their work, along with Janet Emig's *Composing Processes of Twelfth Graders* (1971), which relied on think-aloud protocols, would change forever how writing researchers would theorize notions of writing development. We note that the process movement in writing assessment preceded the validation movement in psychometric discussed in § 4.2. The shift from static targets of evidence to research acknowledging construct complexity is evident throughout the quarter-century of the journal.

Third, if social theories of communication would influence how writing researchers understood the texts that people produce, then it would be innovative research designs that would influence how writing researchers were to understand writing in cultural contexts. The influence of qualitative approaches to the study of writing would be felt in research on writing in communities, such as Shirley Brice Heath's *Ways with Words: Language, Life and Work in Communities and Classrooms* (1983) and Victoria Purcell-Gates's (1995) *Other People's Words: The Cycle of Low Literacy*. Combined with social theories of learning (Vygotsky, 1978) and communities of practice perspectives (Lave & Wenger, 1991), researchers would study, among other topics, children's introduction and enculturation into schooling literacies (Taylor, 1983), students' writing development in the disciplines (Beaufort, 2008), the ways that societal structures shape the use and meaning of literacy (Pritchard, 2016; Veira, 2016), and the interaction of unequal social and institutional structures on students' lives (Sternglass, 1997). Together, social theories of writing development and the methods developed to study writing through such an orientation would not just illuminate the development of individual writers; they would also show how writers' identities, including race, socioeconomic status, and gender, are deeply part of their writing development. Moreover, social approaches to the study of writing would bring systemic inequality to the foreground, showing how racism, for example, shaped writing development.

Socio-cultural theories of writing are evident in this group of articles primarily through two sub-themes: that scoring practices are negotiated through social context; and that student prior learning and cultural context shape how students respond to test prompts. Most commonly-used sources of evidence in this group include interview data, but transcripts of scoring sessions, coding of student writing, and quasi-experimental forms of evidence were also found.

5.3.1. Social context as a place for negotiated teacher knowledge

The relationship between rater expertise and fairness was first examined by Broad (1997) in his study of a two-level portfolio assessment program that relies on socially-agreed upon values for scores. Broad called this approach that relies on balancing the perspectives of raters from different standpoints "communal writing assessment" in which "two or more judges working to reach a joint decision on the basis of a writing performance" (p. 134). Such an approach, Broad explained, is in contrast to a hierarchical measurement scoring process that relies on a calibrated score. In the communal assessment, the score is negotiated, allowing a place for "teachers' special knowledge" (p. 150).

The influence of socio-cultural theory is evident in Li and Barnard's (2011) New Zealand study of academic tutors' beliefs and practices of providing feedback on students' written assignments. Li and Barnard drew on Vygotsky as well as fourth generation evaluation in studying the beliefs and practices of a group of untrained and inexperienced part-time tutors. Drawing on data from 28 surveys, 16 individual interviews, 9 "think aloud" and stimulated recall sessions, and focus group meetings, they concluded that while "tutors initially stated their belief that the purpose of providing feedback was to assist the students to improve their academic writing skills;. . . it emerged that their primary concern was to justify the grades that they awarded" (p. 137). As the Li and Barnard study showed, teacher knowledge as a basis for fairness works best when teachers are highly trained and can work in a community context where dialog encourages the negotiation of evaluation standards.

The importance of community context was most evident in Lindhardsen's 2018 study of raters' decision-making behaviors in an established communal writing assessment (CWA) context. Bringing us full circle to Broad's 1997 study of communal writing assessment, Lindhardsen drew on the perceived fairness of CWA as an assessment method. Situated in the context of high school written EFL exit exam (HHX1) in Denmark, Lindhardsen used transcripts of recorded scoring sessions, retrospective questionnaires, and think aloud protocols of independent scoring sessions to study the decision-making behaviors of 20 raters, "tracing their behaviors all the way from independent rating sessions, where initial images and judgments are formed, to communal rating sessions, where final scores are assigned on the basis of collaboration between two raters" (p. 12). She discovered the following:

> [I]n working with their co-raters, the raters would deliberate and validate their scores and assessment strategies against each other by making explicit their specific assessment strategies, by exemplifying directly form the student scripts and by revisiting their notes and sometimes the student scripts during the process. They would here also articulate their general impressions, compare scripts and envision the personal situation of the students. (p. 22–23)

As Lindhardsen makes clear, CWA follows from a hermeneutic transaction in which dissensus is recognized and incorporated into the scoring process. Rather than viewing discrepancy as a source of error, negotiated meaning is a viewed as a method to increase score validity.

5.3.2. Social context as location for student knowledge

The socio-cultural context for students surrounding assessment contexts was also an area of interest for researchers connecting socio-cultural concerns to fairness. For example, He and Shi's 2008 study of ESL students' perceptions and experiences of standardized English writing tests relied on Chinese and Taiwanese student reports of fairness on the Test of Written English (TWE) and the English Language Proficiency Index (LPI). Many of the students in the study had passed the TWE but failed to pass the LPI because of the tests relied on different constructs of writing: "many participants passed the TWE by relying on memorization of writing samples whereas they failed LPI because they lacked skills in constructing their own texts" (p. 143). But the challenges of the LPI lay beyond the construct of writing. For international test-takers, the LPI essay prompts were perceived as "unfair and culturally biased. . . . The participants' complaints about culturally biased essay prompts in LPI and a lack of understanding of what is expected of LPI also raise questions about the validity of the test from the students' perspectives." (p. 141). In concluding that "the contrasting experiences of the participants imply that essay prompts should be culturally fair and relate to general experiences of all rather than disadvantage any particular group" (p. 143), He and Shi invoked contemporary theories of fairness related to subgroup scores and opportunity to learn.

In a study of timed writing, Petersen (2009) also investigated resistant students' responses—in this case, to Washington State University's then new timed writing test based on a reflective prompt of the program's learning goals. In offering a rationale for the study, Petersen stated that in comparison to compliant ("yes") writers who were likely to score at "exceptional" or "acceptable" levels and "no" writers who less likely to score at "exceptional" or "acceptable" levels (and more likely at the "needs work" level), resistant writers who actively or covertly resisted the test prompts scored lower *or* higher overall on average. In other words, although the number of resistant writers was small, they were more likely to score at "exceptional" and "needs work" than the other groups. After coding the essays, Petersen found that some students resisted the prompt itself while others challenged the "educational value of the writing task" or "testing situation of timed-write" (p. 185). In interpreting the findings, Petersen pointed to the ways that students' prior knowledge of timed testing situations likely shaped their generic responses and that "the rhetorical awareness of skilled writers seems to make the effort of challenging the prompt less risky, especially when compared to attempts made by less skilled writers who (whether out of anger, frustration, confusion, or other reasons) engage in critique with less deliberation and awareness of reader expectations" (p. 187). Like He and Shi, Petersen connects questions of fairness to student knowledge, arguing "the timed-write is far from perfect, but efforts to understand the various factors that make some tests 'better' or more 'fair' to students is a continuous process that writing programs cannot and should not resist" (p. 192).

5.4. Fairness as legal responsibility

The fourth thread of research we traced over 25 years of *ASW* was related to law in 8 articles identified in Appendix B. The overwhelming number of articles related to fairness published in *ASW* have focused on U.S. legal requirements for writing assessment—namely through various reauthorizations of the 1965 Elementary and Secondary Education Act (ESEA). As part of President Lyndon Baines Johnson's War on Poverty initiative, ESEA was an important addition to the support of public education in the U.S, which is primarily financed through local taxes. It was also significant in addressing racial inequalities in schools in the American south that served African American children as such schools were often under-supported. Since 1965, reauthorizations of ESEA have ushered in more government oversight of elementary and secondary education. In 1994, the reauthorization of ESEA known as the Improving America's Schools Act mandated that students be assessed in grades 3 through 5, grades 6 through 9, and grades 10 through 12; and that schools demonstrate "adequate yearly progress" (H.R. 6–7). The No Child Left Behind Act of 2001 (2002), which was another reauthorization of ESSA, continued the push for large-scale testing of every student in U.S. public schools, mandating testing of students in specific subject areas–math and reading or language arts—and sanctioning schools that did not show "adequate yearly progress" (No Child Left Behind Act, 2001, 115 Stat 1445).

Considerations of fairness in educational contexts extend beyond policies directly related to education. For example, the 14th amendment (U.S. Const. amend. XIX) provides equal protection under the law:

> All persons born or naturalized in the United States, and subject to the jurisdiction thereof, are citizens of the United States and of the state wherein they reside. No state shall make or enforce any law which shall abridge the privileges or immunities of citizens of the United States; nor shall any state deprive any person of life, liberty, or property, without due process of law; nor deny to any person within its jurisdiction the equal protection of the laws. (§3)

In addition to federal protection of individual rights, state and local laws also offer protection of individual rights. Specifically, the Civil Rights Act of 1964 and Americans with Disabilities Act of 1990 have expanded protections for individuals in society. In educational contexts, legal protections are provided under the Individuals with Disabilities Education Act (Family and Medical Leave Act of 1993, originally known as the Education for All Handicapped Children Act of 1975). These laws related to civil rights have had an important impact on educational testing in the U.S. because of the protections they provide against intended and unintended discrimination (see Poe & Cogan, 2016, for a legal discussion; see Appendix A for the origin of linear modeling in CTT used to detect bias following 1964 U.S. federal legislation.)

Articles that used a legal framework for discussion tended to either be concerned with documenting the current state of testing policy or the effects of such policies on students and teachers. In almost every case, legal imposition was seen as a barrier to learning. Most commonly-used sources of evidence in this group include textual analysis and interview protocols.

5.4.1. Documenting testing policy

Two studies employing textual analysis have attempted to document the effect of educational policy in the design of state mandated writing assessment. In a study of construct variability in US. state and national writing assessments, Jeffery (2009) took up the thorny issue of alignment between prompt-genre demands and assessment scoring criteria in 41 U.S. state and national high school direct writing assessments. She found that the direct writing assessments were not informed by what was had become a body of knowledge in writing studies theory and research. Unexpectedly, she found that the direct writing assessments did not vary as much as she expected; while they did vary in terms of time allotted and test design, she found the assessment "largely converge in emphasizing persuasive, argumentative and explanatory writing" (p. 13). In comparing state and national direct writing assessments, she found that "national assessments are more coherent than state DWAs in that genre expectations are consistently associated with rubric criteria (p. 13). In noting the value of commonality, especially given student mobility, Jeffery argues, "a collective approach may be both more efficient and fairer" (p. 16)

Behizadeh and Eun Pang (2016) took a more global look at state responses to nationally-mandated testing of writing. Using a document analysis of websites in the 50 U.S. states, they sought to determine writing assessment formats and scoring practices. They found "46 out of 50 states (92%) were primarily using on-demand essay assessment, often in conjunction with multiple choice and short answer items, and no state was utilizing portfolios for writing assessment" (p. 32). Eighteen after years the Spalding and Cummins 1988 study of portfolio assessment in Kentucky, Behizadeh and Eun Pang found no states using portfolio assessment as part of its mandated testing policy. Although "many states are engaged in multi-genre, multi-sample assessment," they caution there is a conflation of reading and writing tasks on the assessments and the readings may not be culturally relevant (p. 37). In regard to scoring, they found "98% of state writing assessment was scored externally with no involvement of the classroom teacher" (p. 25). For Behizadeh and Eun Pang, the legal requirements of large-scale testing have led critics to assume that sociocultural theory is incompatible with accountability and led to the belief that "conflates quantification with fairness and reliability" (p. 27)

5.4.2. Documenting the effects of testing policy on students and teachers

As the previous articles reveal, tracing the impact of educational policies are not neat affairs, and several studies looked at the effects of educational policy on students and teachers. In their study of students' views of writing under the Kentucky Education Reform Act (KERA), Spalding and Cummins (1998) document student views on the Kentucky state-mandated portfolio system that was implemented in the 1990s. They describe the important rights issue that promoted the legislation:

> In 1985, sixty-six poor school districts tiled a lawsuit "claiming the way Kentucky funded public education was inadequate and unequal" (Bishop, 1998). The Kentucky Supreme Court agreed with them, declared the whole public school system unconstitutional, and in 1989 ordered the Kentucky General Assembly to establish a more equitable system and to "'monitor it on a continuing basis so that it will always be maintained in a constitutional manner"' (Rose v. the Council for Better Education, Inc., 1989 quoted in Guskey, 1994, p. 1). The 1990 Kentucky Education Reform Act was the legislature's response. (p. 170)

Yet, beyond mentioning the legal impetus for the new legislation, the authors do not engage with the significance of the aim. Instead, they focus their study on "students' views of their high school writing experiences since the passage of KERA and the implementation of its controversial testing and accountability component, the Kentucky Instructional Results Information System (KIRIS)" (p. 169). In interviewing 450 students, Spalding and Cummins found that while students reported a number of positive writing activities in relation to the new portfolio requirement, "some two-thirds of the students stated that compiling the portfolio was not a useful activity" (p. 191).

Dappen, Isernhagen, and Anderson (2008) document a more optimistic examination of statewide assessment in their analysis of the Nebraska Statewide Writing Assessment. Known as STARS, the Nebraska School-based Teacher-led Assessment and Reporting System was then "based on the philosophy described by the National Research Council (2001) that the effectiveness of a state assessment system must be judged by the extent to which it promotes student learning (p. 47). According to the authors, the teacher-led efforts of the STARS system, which relied on "the involvement of Nebraska classroom teachers, who each year were selected by the NDE upon recommendation of their district superintendent or assessment contact person to participate in a writing development

task force" (p. 49), meant that teachers were not merely involved in scoring but that they were also deeply involved in prompt design and field testing those writing prompts. The researchers conclude that STARS worked in the years they studied (2002–2005), showing statistically significant gains at almost each administration at all grade levels tested (4, 8, and 11). And impressively,

> the positive responses of Nebraska teacher raters on the scoring experience provide evidence that teachers are empowered by their involvement and its positive impact upon classroom practices. Teachers are more confident and they have a better understanding of how writing relates to thinking and reasoning. Teacher involvement in the writing process contributes to their ability to help students become better writers. Nebraska's writing results support the value of teacher empowerment and may serve as a model to other states using statewide assessments to improve student writing." (p. 57)

The STARS system was dismantled in 2008 in lieu of the Nebraska State Accountability Assessments (NeSA), "a statewide initiative to support greater reporting for AYP as required by NCLB" (Ruff, 2019, p. 20).

5.5. Fairness as ethical obligation

The final thread of research we traced over 25 years of *ASW* is related to ethics and human rights in 3 articles. We identified these articles as ethics in Appendix B. The concept of fairness in writing assessment is an appealing concept because of its resonances in western philosophy with systems of ethics, morality, and virtue. In Western philosophy, it is common to draw on the classical origins of Socrates, Plato, and Aristotle. In pointing to this western philosophical basis for fairness, we are not suggesting that fairness does not resonate in other philosophical or rhetorical traditions outside the west. Our scope is limited to a western orientation because the field of writing assessment has very much been rooted in a western disposition toward knowledge-making, a limitation which scholars have only recently begun to discuss (Cushman, 2016; Poe et al., 2018).

Only three studies engaged more than superficially with questions of ethics and fairness. In these cases, deliberation was in relation to construct representation or consequence. Informed by a postmodern ethical framework that inquires "into the practices, the assumptions, the theories, and the consequences of requiring students to assess themselves, their writing, and their performances," Schendel and O'Neill (1999) connected ethics to recent advances in validity that underscored the "impact on the community in which it takes place" to determine validity (p. 202). Drawing on examples of university-wide portfolio assessment and directed self-placement, they seek to open a conversation in the field about the uses of self-assessment, specifically questioning if such methods are fairer. Concerned about the use of self-assessment with limited empirical evidence to support its use, Schendel and O'Neill wrote about consequences, notably that "self-assessment can serve a gatekeeping function because by participating in the assessment, students may expose their own weaknesses (p. 200). They argued that contrary to popular belief, the internalized gaze of the evaluator self-assessment allows little room for resistance; thus the "confessional and self-regulatory nature of self-assessments does not automatically contribute to the 'empowering' of students" (p. 207).

Cumming (2002) offered a reflective discussion of ethical considerations in large-scale writing assessment design in his article, "Assessing L2 writing: Alternative constructs and ethical dilemmas." Like Schendel and O'Neill, Cumming was concerned with ethical issues that arise in large-scale testing contexts. In drawing on his experiences designing the new task types for TOEFL (Test of English as a Foreign Language®), Cumming reflected on how "certain ethical considerations, now conventional in practices for high stakes language tests, necessarily influence our expectations for fairness and feasibility in such tests" (p. 74). He interrogated two assumptions that have traditionally informed large-scale testing—that there should be a uniform context to assess examinee's performance and that performances should be comparable—and the ethical concerns that have traditionally followed—for example, ethical obligations toward human subjects, confidentiality and personal privacy of test-takers, accessibility concerns, and timely reporting (p. 74). Cumming concludes with a caution—that as our aspirations for large-scale writing assessment grow, so must the validity evidence used to support those advances. Without evidence of score impact, the ethics of progress in large-scale testing programs is questionable; and the traditional conceptions of fairness that depend on uniform context to assess examinee's performance may no longer be applicable. As Cumming keenly observes, large-scale tests establish their own contexts that may be susceptible to bias against subgroups of students. (p. 80–81).

Hamp-Lyons (2002) extended emphasis on obligation by attending to postmodernism (viewing writing assessment as an "implicitly political act") and construct representation (recalling that writing is "less well understood than many other constructs") (p. 5). Situating her role historically—this article appeared in her first issue as *ASW* Editor—she categorized writing assessment into four generations: single sample direct, single sample limited response, multiple sample portfolio, and consequential. The fourth generation shift is dramatic, calling for ethical responses to the restricted fixation with technique in the first three generations. She writes,

> The ethical dilemmas and challenges we face in balancing society's need for assessments with our determination to do our best for learners are very great. Accepting a shared responsibility for the impact of writing assessment practices will put consideration of our own ethical behaviour at the top of our agenda. (p. 14)

In world grown larger and more complex, Hamp-Lyons was determined to make the ethics of test practices part of her editorship. Yet, because a journal is not an edited collection, the accomplishment of that aim depended on others.

6. Discussion

Based on the 73 articles we examined for this review, we offer a discussion about each of our research questions concerning the ways fairness has been defined, and has evolved, in the 25 years that *ASW* has been published. Based on our findings, we see that while foundational research has begun, much work remains. Before discussing our findings, we acknowledge that there are other possible ways of concluding a study such as ours. For example, using a reference corpus of the 291 articles selected for analysis by Slomp and East in the introduction of this special issue, researchers could use the 73 articles provided in Appendix B (25 percent of the total) using lexical, stance, and topic modeling techniques. This technique could confirm, expand, or complicate the present study.

Research question 1: How have writing assessment researchers constructed fairness?

Huot and Yancey's establishment of *ASW* was cast as "a space that does not yet exist" (1994, p. 4). The journal has provided that space for researchers from diverse theoretical and methodological backgrounds. In regard to discussions of fairness, the journal has welcomed a variety of perspectives on fairness from educational measurement and writing studies. As Fig. 1 illustrates, evidence of fairness was most often framed as bias research. In studies involving validity as contributing to fairness, interrater reliability remained a common form of evidence.

In terms of social, legal, and ethical studies, we found that the concept of fairness was constructed in ways distinct from the 1985 and 1999 editions of the *Standards* identified in Table 1. In is only with the 2014 edition that we find acknowledgement of context as a factor in score interpretation and use—a recent recognition of situated language use from the measurement community that had been part of *ASW* since its inception.

Research question 2: How have those constructions either directly or indirectly revealed categories of evidence related to fairness?

Given the variety of definitions that we found, it is not surprising that researchers in *ASW* have attended to diverse methods in gathering evidence of fairness, including statistical analysis of scores, student and rater reports, and qualitative transcript analysis. This emphasis on methodology, rather than on principled evidence gathering based on fairness, reveals that evidential categories related to fairness are either implied or absent. This, of course, is not to say that the methods themselves are not related to the drive for fairness. Indeed, the overwhelming evidence in the articles we reviewed came not from multiple-choice tests of writing but from student writing performance. In this regard, the conclusion to the Goldberg et al. (1998) study stands as an anthem for writing studies researchers:

> The field of composition studies has prided itself on a concern with actual student texts and has as its hallmark the careful analyses of those texts. It is important to maintain these strengths and traditions not only when developing and refining large scale instruments to assess writing, but also when conducting the research that ought to infuse decisions about the interpretation and use of assessments. (p. 66)

Although we were heartened to see that student writing formed the basis of evidentiary claims about performance, we also noted two consistent challenges in collecting evidence: reliance on small sample size and elision of interaction effects with evidence of bias. The study by Haswell and Haswell (1996) is indicative of challenges related to small sample size for sub-groups: the sampling plan included only 16 female and 16 male students, with equal numbers of paired instructors. Common in the articles we reviewed, small sample sizes potentially violate the Gaussian distribution, makes random assignment impossible in experimental research, and limits generalization inferences across populations and study sites. The IRT study by Brunfaut et al. (2018) is indicative of problems related to the formulation of "bias/interaction." In a footnote, the authors write that "the concept of fairness in this study was restricted to a psychometric dimension" understood as test fairness and not as consequence (p. 7). Nevertheless, there are complexities with the psychometric dimension due to the elision identified above: absence of interaction is not evidence of absence of bias.

In social cultural research, researchers generally do not theorize fairness—for example, by pointing to the *Standards* requirements for accessibility, human rights definitions, or cultural theory—although there is a very strong social justice tradition external to writing assessment. Moreover, in terms of methods, we found that researchers do not discuss how various kinds of qualitative research (think-aloud protocols, discourse-based interviews, semi-structured interviews, and focus groups) contribute to documenting fairness.

In the legal articles, fairness as a human right is not to be found. This important omission in the literature is a crucial gap in the literature that omits discussions of students' rights to education, the right of opportunity to learn, and the right to accessible testing conditions. Accessibility is not only an ethical concern; it is also a legal one in many countries.

In the small sample of we have classified as ethical, there is a notable absence of theory-building. That is, there is little or no attention to typology, construct representation, learning sequence, or types of consequences. Discussions of ethics are related exclusively to obligation and consequence.

Research question 3: Have there been major shifts in the use of fairness as a guiding research principle in the last 25 years?

In our analysis presented in this article, we saw shifts in the conceptualization of fairness, but a tidy narrative that theoretical conceptualization of fairness has deepened over 25 years is not born out by our findings. This finding is similar to that of Haswell and Elliot (2019) in their study of over 1000 research studies, handbooks, and policy documents related to holistic scoring in the U.S and

U.K. from the mid-1930s to the mid-1980s. As they observe, researchers had access to these categories of validity: concurrent, construct, content, criterion, and predictive. In terms of reliability, researchers offered evidence that included inter-rater agreement, inter-rater reliability, intra-rater reliability, test reliability, and writer reliability. In terms of fairness, however, evidence was inferred from validity and reliability, and differential prediction techniques appeared quite late in the 1970s. Concern for fairness were related almost exclusively to evidence regarding consistency and consequence.

The absence of a deepened narrative at the present is not due to a simple split between educational measurement and writing studies researchers. Given our findings, we wish to nuance complicate the claim that "measurement theory has had a strong influence on writing assessments, while writing theory has had minimal influence on writing assessments" (Behizadeh & Engelhard, 2011, p. 189). What we see in our analysis is that as studies involving forms of evidence that are social, legal, and ethical come into view within the pages of *ASW*, writing theory has impacted measurement theory. Indeed, it is not too far a reach to claim sociocultural theories may yet determine, fourth generation research identified by Dorans (2011) as it occurs within larger, complex contexts such as writing.

However, this is not to say that research programs devoted to fairness have emerged in the journal. We remain cautious as the influence of U.S. psychometric traditions remains strong. In this context, the influence of the educational measurement community is particularly notable given its global reach. Undue influence was a concern that Liz Hamp-Lyons raised in a 2014 special issue *Research in the Teaching of English*:

> Beyond a doubt, US psychometric work…has had a strong influence on how writing is tested around the world, as is shown by the popularity of large-scale standardized tests such as the SAT, GRE, GMAT, etc., as well as the TOEFL and TOEIC in the field of English as a Second Language.. . . These tests have shaped those countries' perceptions and expectations of what makes a good "writing test" and what makes for good practice in assessing writing but have narrowed the construct of what "good writing" is. (Hamp-Lyons, 2014, p. 357)

If *ASW* is going to be an international journal that holds to its origins "to create a space which does not yet exist" (Huot, 1994, p. 4), then we must consider how considerations of fairness must account for more than the views of a few.

Research question 4: Are there ways that fairness has not been used by writing assessment researchers that might prove fruitful?

While we are hopeful about the categories used to gather evidence in the journal, our optimism is tempered. When we dig down granularly into *ASW* publications, the textual record suggests that we do not develop or share taxonomies for researching questions of fairness. In the existing record, there are substantial gaps and evident disjunctures. In reviewing the work published in *ASW* over the last 25 years, we find that a single fact remains: Most large scale assessments remain grounded in first-generation concepts and methods, even while drawing on second and third generation machinery such as IRT theory to improve quality.

We also see how legal and ethical orientations toward fairness have played only a minor role in the journal, despite their importance in education and measurement. While legal standards related to fairness in educational contexts have had a major impact on policy makers, teacher, and students, there is a noticeable absence of these laws in scholarship on writing assessment as evidenced in *ASW*. None of the *ASW* articles that work within the U.S. context of mandated testing address the legal requirements for fairness.

7. Recommendations

In the afterword to *The Practical Past* (2014), Hayden White wrote that we need a new understanding of history lest we wind up submitting "to the authority of those claiming the right to tell us who we are, what we are supposed to do, and what we should strive for in order to be at all" (p. 103). His recommendation is that we consider a practical past alongside the historical past. In that imaginative consideration, we gain scope, depth, and awareness. If the historical past is a scientific enterprise such as we have presented in the present work—the literal truth that is clear, unambiguous, and validated—then the practical past sits along beside it as an imaginative act. Based on our analysis, what can an imaginative view of the past afford? What identities can be formed? To address such questions, we conclude with three hopeful recommendations.

Recommendation 1: While diverse categories of evidence have been used to support fairness in writing assessment, fairness needs further theorization.

While there is foundation for a body of knowledge regarding fairness in writing assessment, at the present writing there is only one theory of fairness in writing assessment (Elliot, 2016), and only one taxonomy of fairness (Slomp, 2016). Dorans (2017) has identified one reason for this absence in educational measurement: "Not all fairness considerations can be reduced to quantitative evaluations" (p. 222). Indeed, as he recognized, even differential item functioning—perhaps the most broadly used way of establishing absence of bias in multiple-choice testing—is an unreliable measure because a test item "is an unreliable measure of the construct of interest" (p. 223). A second reason arises from the disjuncture that presently exists among CTT, IRT, and social theories of texts presented in §5.3. One promising point of resolution between quantitative evaluation and model use is Robert M, Mislevy's

socio-cognitive foundation for educational measurement. Mislevy (2018), Mislevy and Elliot (in press), and Oliveri, Mislevy, and Elliot (in press) have advanced resonances among socio-cognitive views of the construct of written communication, category of evidence models, and principled views of fairness. As we advance in theorizing fairness, we must also recognize that a single theory of fairness is neither required nor need take precedence. In fact, the establishment of a single theory of fairness will always be exclusive of other ways of knowing outside the Western tradition. For this reason, we underscore the importance of culturally responsive evaluation and indigenous evaluation for their attention to contextual factors, social relevance, historical injustice (Cram, 2016; Hood, Hopson, & Kirkhart, 2015; LaFrance & Nichols, 2010). In terms of fairness, such complementary models could launch new programs of research in writing assessment.

Recommendation 2: If we are to identify complementary models of fairness that integrate various research traditions, we must be open to methods of greater nuance.

As complementary models emerge, researchers need not just attend to evolving views of fairness; they must also be open to new methods that potentially advance fairness. Recent research, for example, has demonstrated that IRT person-fit models are promising in providing examination of differential functions of groups of items that, taken together, can provide a portrait of student abilities based on multi-dimensional response models (Carlson & von Davier, 2017; Meijer & Sijtsma, 2001; Mislevy, 2018; Rupp, 2013). Such computerized models could, at least in theory, provide formative assessments for students on constructed response writing tasks built on socio-cognitive modeling; indeed, such models exemplify fourth generation assessment. Sociocultural forms of evidence might also be developed to account for intersectional or layered hierarchies of influence within a student's writing development that then inform performance outcomes. The work of Guillermo Solano-Flores on cultural validity in language testing holds promise for developing novel writing assessment methods (e.g., Solano-Flores, 2008; Solano-Flores & Li, 2013; Solano-Flores, Backhoff, Contreras-Niño, & Vázquez-Muñoz, 2015). New visual modeling methods and robust software that allows for analysis of complex data sets could be integral to this work. Network analysis could allow researchers to see connections of variables across time and context that, in turn, could result in new opportunities to advance student learning through formative assessment.

Recommendation 3: While an international presence exists in current writing assessment programs of research, U.S. trends exert a disproportionate methodological influence—one that must be countered by expansive non-Western perspectives.

It is indeed true that U.S. psychometric models have had an undue influence on how writing is tested around the world. Modernist in design and capitalist in efficiency, U.S. distributed assessments almost universally rely on a web of standardization that runs from item development to aggregated result reporting. While the Western orientation has been acknowledged previously in other evaluation research, only recently in writing assessment literature has attention been paid to the Western cultural construction of validity (Cushman, 2016; Poe et al., 2018). Because culturally diverse philosophical views of assessment fairness express concerns beyond deficit views—that is, beyond the assumption that once construct-irrelevant variance is removed that all will be well—it will be increasingly important to understand how non-Western cultures construct, interpret, resist, and transform evidence of fairness.

8. Conclusion

The history of a journal like *Assessing Writing*—one that came about as a field was emerging—is ultimately an intellectual history. In this capacity, *ASW* serves as both a testament to the changing landscape of the field of writing assessment and the imperative of academic journals to lead conversations in a field. The space that the original editors, Huot and Yancey, created for making connections between pedagogy and assessment has proved to be generative over time. The generative nature of fairness research in the journal over the last 25 years, however, has not necessarily resulted in shared taxonomies across disciplinary orientations, led to a deepening of theoretical conceptualization of fairness, or brought about innovative classroom assessment approaches. Much therefore remains to be done in the next 25 years. As we have shown, doors remain open for next-generation writing assessment. Based on its rich past, there is every reason to believe that *Assessing Writing* will lead future, new programs of research associated with fairness.

Acknowledgements

The authors would like to thank Cherice Escobar Jones for preparation of Fig. 1 and Appendix B. We would also like to thank Robert J. Mislevy for his review of Appendix A and his consultation regarding new uses of IRT models used to provide student ability portraits. We also thank our two anonymous reviewers.

Appendix A. Fairness as the Elimination of Bias

The study of bias is based on mathematical models. As such, a brief articulation of models used under CCT and IRT is important if we are to understand articles published in *Assessing Writing* from 1994 to 2018 that focus on evidence related to the study of bias.

We begin by conceptualizing bias for a single trait, such as writing ability. Under such an assumption, we would begin with a model proposed by Scheuneman (1984) as an extension of CTT:

$$X = \theta + \beta + \delta \quad (1)$$

in which

X = the observed score
θ = the true score[1]
β = the bias factor, which is not present in the CTT model
δ = the zero-centered measurement error independent of membership in a group

While the expected value for X would be θ for all persons under CTT, the mean score for a sub-group under examination could be less than zero due to β. The subgroup mean score would then tend to result in observed scores below the true mean, constituting bias in the statistical sense of the term.

As Scheuneman and other mindful researchers note, it is not possible to distinguish values of β, or even averages of β, within a given subgroup g for a given task or for a test as a whole, from observations of X alone. External information is required, in the form of expert opinion, other (also fallible) measures (such as holistic or trait scores), or information about task demands and response processes. It is possible, though, to estimate *differences* in group-average βs for individual items on a test. Ideally, for a given item j, one would compare average scores on item j in each group g at the same levels of θ. Since θ is not observable, a common technique is to compare the performance of groups on an item by comparing item j averages among persons *at the same total score levels*. By doing so we learn whether item j is relatively harder for men students than for women students, for example, who have the same overall score. These item-by-group interactions indicate that the average βs are different for the groups and properly call attention to the item in question. Note that these analyses focus on differences among groups at similar levels of performance, and not on possible score differences for the groups as a whole (i.e., "impact").

Analyses of item-by-group interactions do not, however, provide information about absolute levels of β. That is, a uniform shift downward of β for all items for a particular group would not produce group-by-item interactions, but would, under Scheuneman's framing, constitute bias. More technically sophisticated versions of group-by-item interaction have been developed in classical theory, and analogues have been extended to IRT (Carlson & von Davier, 2017). While more flexible, powerful, and accurate, they share the fundamental property that absence of item-by-group interactions is not equivalent to absence of group-related bias.

Historically, in 1968 Cleary had extended bias studies to include prediction based on general linear modeling written as follows:

$$Y = \beta_{0g} + \beta_g X + \varepsilon \quad (2)$$

in which

Y = the outcome variable
X = the predictor variables
β_g = the coefficients for the predictor variables as pertain to group g
β_{0g} = the intercept as pertains to group g
ε = measurement error

Use of this prediction model allowed Cleary to use two hypothesis tests to examine bias of a given test X, with respect to a criterion variable Y, with respect to given subgoups:

- *Equality of Slopes*: This first hypothesis states that the relationship between the predictor and outcome variables are the same for all groups. For example, the slopes (i.e., the change in *y* for a change in *x* of one unit) of a writing sample (predictor variable) would be equal for all groups. If the hypothesis is true, then the only remaining factor unique to the individual groups is the intercept term.
- *Equality of Intercepts*: Given that the slopes are equal, if the intercepts are not equal, then consistent errors of prediction are being made for one or more groups. The test must then be considered biased under Cleary's approach for an under-predicted sub-group, at least for this criterion variable and the distributions of scores of the population at issue.

Extending this concept of differential prediction first made during the era of US Civil Rights, Berry (2015) made an important recent distinction between differential prediction and differential validity in the use of general linear modeling. In the Cleary model, differential prediction suggests a difference in regression lines between subgroups. In differential validity, as Berry observed, if correlations among predictor and outcome variables are different for subgroups, the assessment may or may not have equal predictive validity—and thus not equal meaning in terms of drawn inferences—across all subgroups. The Berry distinction is especially important to writing studies researchers who may, or may not, have access to criterion variables used to establish concurrent or predictive validity and, therefore, rely on differential validity evidence.

We discuss the Cleary model in some detail because it was the first, and is still probably the most widely used, method for studying prediction bias in test uses (Poe, Elliot, Cogan, & Nurudeen, 2014). The Cleary description of bias for prediction and selection problems is both intuitive and plausible because it incorporates the relationship between dependent and independent variables in terms of outcomes.[2]

[1] Note that the interpretation of true score θ in (1) is no longer the definition given in Lord & Novick (1968, p. 30), as the expectation of observed score. Rather, it is a Platonic interpretation, such that a true score is a value on an existing attribute that is the target of measurement. The two concepts and definitions do not coincide. The Platonic definition suits Scheuneman's purposes, however. It allows us to conceive of θ as an ability that is the construct intended to be measured and β as an effect that is related to a person's sub-group membership but unrelated to θ.

[2] Subsequent research has shown that other equally plausible definitions for fair selection and prediction have been proposed, which need not coincide with Cleary's, and in some cases are incompatible with Cleary's or one another. A coherent framework requires the more technical framework afforded by Bayesian inference and utility theory (Petersen & Novick, 1976).

In general, these conceptualizations of bias are formulated under CCT. It is important to remember that the use of the term "bias/interaction" in IRT is conditional on overall levels of performance as in the CTT methods noted above. Under assumptions of IRT related to Rash techniques—those most common in the journal—bias is identified under the formulation of "bias/interaction." While a useful definition, the elision between bias and interaction is logically troubling. If interaction occurs beyond a specified interval in a log-odds unit (called a logit), then evidence of bias is identified. This model is problematic for three reasons: the target is a specific interaction, and so other sources of bias may remain unexamined; absence of the specified logit is assumed to be evidence of absence of bias; and actionable directions for elimination of bias are not readily apparent from the identification of interactions unless subgroups of students are identified. For more on IRT and Rasch models used in the identification of bias, see Dorans (2017), especially pp. 213-214.

Appendix B. Fairness Articles Analyzed in Assessing Writing, 1994-2018

Articles Analyzed (n = 73)

	Reference	Group
1	Aull, L. (2015). Connecting writing and language in assessment: Examining style, tone, and argument in the U.S. Common Core Standards and in exemplary student writing. *Assessing Writing*, *24*, 59-73.	Legal
2	Baker, B.A. (2010). Playing with the stakes: A consideration of an aspect of the social context of a gatekeeping writing assessment. *Assessing Writing*, *15*, 133-153.	Social
3	Ball, A.F. (1997). Expanding the dialogue on culture as a critical component when assessing writing. *Assessing Writing*, *4*, 169-202.	Bias
4	Barkaoui, K., & Knouzi, I. (2018) The effects of writing mode and computer ability on L2 test-takers' essay characteristics and scores. *Assessing Writing*, *36*, 19-31.	Validity
5	Behizadeh, N., & Engelhard, 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*, 189-211.	Social
6	Behizadeh, N., & Eun Pang, M. (2016). Awaiting a new wave: The status of state writing assessment in the United States. *Assessing Writing*, *29*, 25-41.	Legal
7	Bridgeman, B., & Ramineni, C. (2017). Design and evaluation of automated writing evaluation models: Relationships with writing in naturalistic settings. *Assessing Writing*, *34*, 62-71.	Validity
8	Broad, B. (1997). Reciprocal authorities in communal writing assessment: Constructing textual value within a "new politics of inquiry." *Assessing Writing*, *4*, 133-167.	Social
9	Brunfaut, T., Harding, L., & Batty, A.O. (2018). Going online: The effect of mode delivery on performances and perceptions on an English L2 writing test suite. *Assessing Writing*, *36*, 3-18.	Bias
10	Burke, J., & Cizek, G. (2006). Effects of composition mode and self-perceived computer skills on essay scores of sixth graders. *Assessing Writing*, *11*, 148-166.	Validity
11	Chen, J., White, S., McCloskey, M., Soroui, J., & Chun, Y. (2011). Effects of computer versus paper administration of an adult functional writing assessment. *Assessing Writing*, *16*, 49-71.	Bias
12	Cherry, R.D., & Witte, S.P. (1998). Direct assessments of writing: Substance and romance. *Assessing Writing*, *5*, 71-87.	Social
13	Condon, W. (2013). Large-scale assessment, locally-developed measures, and automated scoring of essays: Fishing for red herrings? *Assessing Writing*, *18*, 100-108.	Social
14	Cumming, A. (2002). Assessing L2 writing: Alternative constructs and ethical dilemmas. *Assessing Writing*, *8*, 73-83.	Ethics
15	Dappen, L., Isernhagen, J., & Anderson, S. (2008). A statewide writing assessment model: Student proficiency and future implications. *Assessing Writing*, *13*, 45-60.	Legal
16	Denny, C. (2008). Dangerous liaisons: Reflections on a pilot project for state-mandated outcomes assessment of written communication. *Assessing Writing*, *13*, 26-44.	Legal
17	Deygers, B., Branden, K.V., & Peters, E. (2017). Checking assumed proficiency: Comparing L1 and L2 performance on a university entrance test. *Assessing Writing*, *32*, 43-56.	Bias
18	di Gennaro, K. (2013). How different are they? A comparison of Generation 1.5 and international L2 learners' writing ability. *Assessing Writing*, *18*, 154-172.	Bias
19	Elbow, P. (2006). Do we need a single standard of value for institutional assessment? An essay response to Asao Inoue's "community-based assessment pedagogy." *Assessing Writing*, *11*, 81-99.	Validity
20	Ferris, D.R., Evans, K., & Kurzer, K. (2017). Placement of multilingual writers: Is there a role for student voices? *Assessing Writing*, *32*, 1-11.	Social
21	Goldberg, G.L., Roswell, B.S., & Michaels, H. (1998). A question of choice: The implications of assessing expressive writing in multiple genres. *Assessing Writing*, *5*, 39-70.	Validity
22	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.	Bias
23	Hamp-Lyons, L. (2002). The scope of writing assessment. *Assessing Writing*, *8*, 5-16.	Ethics
24	Haswell, R.H. (1998). Rubrics, prototypes, and exemplars: Categorization theory and systems of writing placement. *Assessing Writing*, *5*, 231-268.	Bias
25	Haswell, R.H., & Haswell, J.T. (1996) Gender bias and critique of student writing. *Assessing Writing*, *3*, 31-83.	Validity
26	He, L., Shi, L. (2008). ESL students' perceptions and experiences of standardized English writing tests. *Assessing Writing*, *13*, 130-149.	Social
27	Huang, J. (2008). How accurate are ESL students' holistic writing scores on large-scale assessments? –A generalizability theory approach. *Assessing Writing*, *13*, 201-218.	Validity
28	Huang, J. (2012). Using generalizability theory to examine the accuracy and validity of large-scale ESL writing assessment. *Assessing Writing*, 17, 123-139.	Bias
29	Isbell, D.R. (2017). Assessing C2 writing ability on the Certificate of English Language Proficiency: Rater and examine age effects. *Assessing Writing*, 34, 37-49.	Validity
30	James, C. L. (2008). Electronic scoring of essays: Does topic matter? *Assessing Writing*, *13*, 80-92.	Validity

31	Jeffery, J.V. (2009). Constructs of writing proficiency in U.S. state and national writing assessments: Exploring variability. *Assessing Writing, 14*, 3-24.	Legal
32	Jeong, H. (2017). Narrative and expository genre effects on students, rater and performance criteria. *Assessing Writing, 31*, 113-125.	Bias
33	Jiuliang, L. (2014). Examining genre effects on test takers' summary writing performance. *Assessing Writing*, 22, 75-90.	Bias
34	Johnson, A.C., Wilson, J., & Roscoe, R.D. (2017). College students' perceptions of writing errors, text quality, and author characteristics. *Assessing Writing, 34*, 72-87.	Bias
35	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*, 35-54.	Social
36	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*, 35-54.	Legal
37	Klein, J., & Taub, D. (2005). The effect of variation in handwriting and print on evaluation of student essays. *Assessing Writing, 10*, 134-148.	Bias
38	Klobucar, A., Elliot, N., Deess, P., Rudniy, O., & Joshi, K. (2013). Automated scoring in context: Rapid assessment for placed students. *Assessing Writing, 18*, 62-84.	Validity
39	Knoch, U., Read, J., & von Randow, J. (2007). Re-training writing raters online: How does it compare with face-to-face training? *Assessing Writing, 12*, 26-43.	Bias
40	Lallmamode, S.P., Daud, N.M., & Abu Kassim, N.L. (2016). Development and initial argument-based validation of a scoring rubric used in the assessment of L2 writing electronic portfolios. *Assessing Writing, 30*, 44-62.	Bias
41	Lam, R. (2017). Taking stock of portfolio assessment scholarship: From research to practice. *Assessing Writing, 31*, 84-97.	Social
42	Lee, H. (2008) The relationship between writers' perceptions and their performance on a field-specific writing test. *Assessing Writing, 13*, 93-110.	Social
43	Lee, H.K. (2004). A comparative study of ESL writers' performance in a paper-based and a computer-delivered writing test. *Assessing Writing, 9*, 4-26.	Validity
44	Lee, Y. (2002). A comparison of composing processes and written products in timed-essay tests across paper-and-pencil and computer modes. *Assessing Writing, 8*, 135-177.	Social
45	Li, J., & Barnard, R. (2011). Academic tutors' beliefs about and practices of giving feedback on students' written assignments: A New Zealand case study. *Assessing Writing, 16*, 137-148.	Social
46	Li, J., & Lindsey, P. (2015). Understanding variations between student and teacher application of rubrics. *Assessing Writing, 26*, 67-79.	Bias
47	Lindhardsen, V. (2018). From independent ratings to communal ratings: A study of CWA raters' decision-making behaviors. *Assessing Writing, 35*, 12-25.	Social
48	Ling, G. (2017). Are TOEFL iBT writing test scores related to keyboard type? A survey of keyboard-related practices at testing centers. *Assessing Writing, 31*, 1-12.	Validity
49	Mahfoodh, O.H.A. (2017). "I feel disappointed": EFL university students' emotional responses towards teacher written feedback. *Assessing Writing, 31*, 53-72.	Social
50	Marefat, F., & Heydari, M. (2016). Native and Iranian teachers' perceptions and evaluation of Iranian students' English essays. *Assessing Writing, 27*, 24-36.	Bias
51	Mendoza, A., & Knoch, U. (2018). Examining the validity of an analytic rating scale for a Spanish test for academic purposes using the argument-based approach to validation. *Assessing Writing, 35*, 41-55.	Bias
52	Moss, P. (1994). Validity in high stakes writing assessment: Problems and possibilities. *Assessing Writing, 1*, 109-128.	Validity
53	Murphy, S. (1994). Portfolios and curriculum reform: Patterns in practice. *Assessing Writing, 1*, 175-206.	Social
54	Penny, J.A. (2003). Reading high stakes writing samples: My life as a reader. *Assessing Writing, 8*, 192-215.	Legal
55	Petersen, J. (2009). "This test makes no freaking sense": Criticism, confusion, and frustration in times writing. *Assessing Writing, 14*, 178-193.	Social
56	Peterson, S., Childs, R., & Kennedy, K. (2004). Written feedback and scoring of sixth-grade girls' and boys' narrative and persuasive writing. *Assessing Writing, 9*, 160-180.	Bias
57	Ramineni, C., & Williamson, D.M. (2013). Automated essay scoring: Psychometric guidelines and practices. *Assessing Writing, 18*, 25-39.	Validity
58	Rezaei, A.R., & Lovorn, M. (2010). Reliability and validity of rubrics for assessment through writing. *Assessing Writing, 15*, 18-39.	Bias
59	Schendel, E., & O'Neill, P. (2000). Exploring the theories and consequences of self-assessment through ethical inquiry. *Assessing Writing, 6*, 199-227.	Ethics
60	Slomp, D. (2008). Harming not helping: The impact of a Canadian standardized writing assessment on curriculum and pedagogy. *Assessing Writing, 13*, 180-200.	Social
61	Spalding, E., & Cummins, G. (1998) It was the best of times. It was a waste of time: University of Kentucky students' views of writing under KERA. *Assessing Writing, 5*, 167-199.	Legal
62	Sudweeks, R.R., Reeve, S., & Bradshaw, W.S. (2005). A comparison of generalizability theory and many-facet Rasch measurement in an analysis of college sophomore writing. *Assessing Writing, 9*, 239-261.	Bias
63	Trace, J., Meier, V., & Janssen, G. (2016). "I can see that": Developing shared rubric category interpretations through score negotiation. *Assessing Writing, 30*, 32-43.	Bias
64	Wang, J., Engelhard, G., Raczynski, K., Song, T., & Wolfe, E. (2017). Evaluating rater accuracy and perception for integrated writing assessments using a mixed-methods approach. *Assessing Writing, 33*, 36-47.	Validity
65	Weigle, S.C. (2004). Integrating reading and writing in a competency test for non-native speakers of English. *Assessing Writing, 9*, 27-55.	Validity
66	Weigle, S.C. (2013). English language learners and automated scoring of essays: Critical considerations. *Assessing Writing, 18*, 85-99.	Validity
67	Willard-Traub, M., Decker, E., Reed, R., & Johnston, J. (2000). The development of large-scale portfolio placement assessment at the University of Michigan: 1992-1998. *Assessing Writing, 6*, 41-84.	Validity
68	Williamson, M. (1994). The worship of efficiency: Untangling theoretical and practical considerations in writing assessment. *Assessing Writing, 1*, 147–173.	Social
69	Wind, S.A., & Engelhard G. (2013). How invariant and accurate are domain ratings in writing assessment. *Assessing Writing, 18*, 278-299.	Bias
70	Wind, S.A., Stager, C., & Patil, Y.J. (2017). Exploring the relationship between textual characteristics and rating quality on rater-mediated writing assessment: An illustration with L1 and L2 writing assessments. *Assessing Writing, 34*, 1-15.	Validity
71	Wiseman, C. S. (2012). Rater effects: Ego engagement in rater decision-making. *Assessing Writing, 17*, 150-173.	Bias
72	Wolfe, E.W., Song, T., & Jiao, H. (2016). Features of difficult-to-score essays. *Assessing Writing, 27*, 1-10.	Bias
73	Zhang, J. (2016). Same text different processing? Exploring how raters' cognitive and meta-cognitive strategies influence rating accuracy in essay scoring. *Assessing Writing, 27*, 37-53.	Bias

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1985). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

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

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

Americans with Disabilities Act of 1990, Pub. L. No. 101-336, § 2, 104 Stat. 327 (1999). Retrieved http://library.clerk.house.gov/reference-files/PPL_101_336_AmericansWithDisabilities.pdf.

Austin, J. L. (1962). *How to do things with words.* Cambridge, MA: Harvard University Press.

Barkaoui, K., & Knouzi, I. (2018). The effects of writing mode and computer ability on L2 test-takers' essay characteristics and scores. *Assessing Writing, 36*, 19–31.

Bakhtin, M. (1934-1935/1981). Discourse in the novel. In M. Holquist (Ed.), *The dialogic imagination: Four essays* (pp. 262-349). Austin, TX: University of Texas Press.

Bakhtin, M. (1952-1953/1986). The problem of speech genres. In C. Emerson & M. Holquist (Eds.), *Speech genres and other late essays* (pp. 60-102), Austin, TX: University of Texas Press.

Bazerman, C. (1994). Systems of genres and the enactment of social intentions. In A. Freedman, & P. Medway (Eds.). *Genre and the new rhetoric* (pp. 67–85). London, UK: Taylor and Francis.

Beaufort, A. (2008). *College writing and beyond: A new framework for university writing instruction.* Logan, UT: Utah State University Press.

Beck, S. W., & Jeffery, J. (2007). Genres of high-stakes writing assessments and the construct of writing competence. *Assessing Writing, 12*, 60–79. https://doi.org/10.1016/j.asw.2007.05.001.

Behizadeh, N., & Engelhard, G., Jr (2011). Historical view of the influences of measurement and writing theories on the practice of writing assessment in the United States. *Assessing Writing, 11*, 189–211. https://doi.org/10.1016/j.asw.2011.03.001.

Behizadeh, N., & Eun Pang, M. (2016). Awaiting a new wave: The status of state writing. *Assessing Writing, 29*, 25–41. https://doi.org/10.1016/j.asw.2016.05.003.

Berry, C. M. (2015). Differential validity and differential prediction of cognitive ability tests: Understanding test bias in the employment context. *Annual Review of Organizational Psychology and Organizational Behavior, 2*, 435–465. https://doi.org/10.1146/annurev-orgpsych-032414-111256.

Bishop, B. (1998). *Equality was heart of reform.* Lexington Herald-Leader pp. Fl, F3.

Black, L., Daiker, D. A., Sommers, J., & Stygal, G. (1994). *New directions in portfolio assessment: Reflective practice, critical theory, and large-scale scoring.* Portsmouth, NH: Heinemann.

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

Boone, W. J., & Noltemeyer, A. (2017). Rasch analysis: A primer for school psychology researchers and practitioners. *Cogent Education, 4*, 1–13. https://doi.org/10.1080/2331186X.2017.1416898.

Breland, H. M., Camp, R., Jones, R. J., Morris, M. M., & Rock, D. A. (1987). *Assessing writing skill. College Board research monograph No. 11.* New York: College Entrance Examination Board.

Britton, J., Burgess, T., Martin, N., McLeod, A., & Rosen, H. (1975). *The development of writing abilities.* London, UK: Macmillan11–18.

Broad, B. (1997). Reciprocal authorities in communal writing assessment: Constructing textual value within a "New politics of inquiry.". *Assessing Writing, 4*, 133–167. https://doi.org/10.1016/S1075-2935(97)80010-4.

Brunfaut, T., Harding, L., & Batty, A. O. (2018). Going online: The effect of mode of delivery on performances and perceptions on an English L2 writing test suite. *Assessing Writing, 36*, 3–18. https://doi.org/10.1016/j.asw.2018.02.003.

Camp, H. (2012). The psychology of writing development—And its implications for assessment. *Assessing Writing, 17*, 92–105.

Carlson, J. E., & von Davier, M. (2017). Item response theory. In R. E. Bennett, & M. von Davier (Eds.). *Advancing human assessment: The methodological, psychological and policy contributions of ETS* (pp. 133–178). Cham, Switzerland: Springer.

Civil Rights Act of 1964, Pub.L. 88-352, 78 Stat. 241 (1964). Retrieved https://legcounsel.house.gov/Comps/Civil%20Rights%20Act%20Of%201964.pdf.

Cleary, T. A. (1968). Test bias: Prediction of grades of Negro and White Students in integrated colleges. *Journal of Educational Measurement, 5*, 115–124. https://doi.org/10.1111/j.1745-3984.1968.tb00613.x.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Cram, F. (2016). Lessons on decolonizing evaluation from Kaupapa Māori evaluation. *Canadian Journal of Program Evaluation, 30* (Special Issue), 296–312.

Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer (Ed.). *Test validity* (pp. 3–17). Hillsdale, NJ: Erlbaum.

Cumming, A. (2002). Assessing L2 writing: Alternative constructs and ethical dilemmas. *Assessing Writing, 8*, 73–83. https://doi.org/10.1016/S1075-2935(02)00047-8.

Cushman, E. (2016). Decolonizing validity. *The Journal of Writing Assessment, 9*. Retrieved http://journalofwritingassessment.org/article.php?article=92.

Dappen, L., Isernhagen, J., & Anderson, S. (2008). A statewide writing assessment model: Student proficiency and future implications. *Assessing Writing, 13*, 45–60. https://doi.org/10.1016/j.asw.2008.04.00s.

Devitt, A. J. (1993). Generalizing about genre: New conceptions on an old concept. *College English, 44*, 573–586.

Dorans, N. J. (2011). Holland's advice for the fourth generation of test theory: Blood tests can be contests. In N. J. Dorans, & S. Sinharay (Eds.). *Looking back: Proceedings of a conference in honor of Paul W. Holland* (pp. 259–272). New York, NY: Springer.

Dorans, N. J. (2017). Contributions to the quantitative assessment of item test, and score fairness. In R. E. Bennett, & M. von Davier (Eds.). *Advancing human assessment: The methodological, psychological and policy contributions of ETS* (pp. 201–230). Cham, Switzerland: Springer.

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

Elliot, N. (2015). Validation: The pursuit. [Review of *Standards for educational and psychological testing,* by American Educational Research Association, American Psychological Association, and National Council on Measurement in Education]. *College Composition and Communication, 66*, 668–685.

Elliot, N. (2016). A theory of ethics for writing assessment. *Journal of Writing Assessment, 9*. Retrieved http://journalofwritingassessment.org/article.php?article=98.

Emig, J. (1971). *The composing processes of twelfth graders.* Urbana, IL: National Council of Teachers of English.

Family and Medical Leave Act of 1993, 29 U.S.C. § 2601 (1993). Retrieved https://www.govinfo.gov/content/pkg/STATUTE-107/pdf/STATUTE-107-Pg6.pdf.

Gee, J. P. (1990). *Social linguistics and literacies: Ideology in discourses.* New York: Routledge.

Gee, J. P. (2008). A sociocultural perspective on 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. 76–108). Cambridge, UK: Cambridge University Press.

Goldberg, G. L., Roswell, B. S., & Michaels, H. (1998). A question of choice: The implications of assessing expressive writing in multiple genres. *Assessing Writing, 5*, 39–70.

Guskey, T. R. (1994). Introduction. In T. R. Guskey (Ed.). *High stakes performance assessment: Perspectives on Kentucky's educational reform* (pp. l–6). Thousand Oaks, CA: Corwin Press.

Hamp-Lyons, L. (2002). The scope of writing assessment. *Assessing Writing, 8*, 5–16.

Hamp-Lyons (2014). Writing assessment in a global context. *Research in the Teaching of English, 48*, 353–362.

Harding, S. (1986). *The science question in feminism.* Ithaca, NY: Cornell University Press.

Haswell, R., & Haswell, J. (1996). Gender bias and critique of student writing. *Assessing Writing, 3*, 31–83. https://doi.org/10.1016/S1075-2935(96)90004-5.

Haswell, R., & Elliot, N. (2019). *Early holistic scoring of writing: A Theory, a history, a reflection.* Logan, UT: Utah State University Press.

He, L., & Shi, L. (2008). ESL students' perceptions and experiences of standardized English writing tests. *Assessing Writing, 13*, 130–149. https://doi.org/10.1016/j.asw.2008.08.001.

Heath, S. B. (1983). *Ways with words: Language, life and work in communities and classrooms.* New York, NY: Cambridge University Press.

Hesse, D. (2019). Journals in composition studies, thirty-five years after. *College English, 81*, 367–396.

Hood, S., Hopson, R., & Kirkhart, K. (2015). Culturally responsive evaluation: Theory, practice, and future. In K. Newcomer, H. Hatry, & J. Wholey (Eds.). *Handbook of practical program evaluation* (pp. 281–311). (4th edition). Hoboken, NJ: Wiley.
Huang, J. (2012). Using generalizability theory to examine the accuracy and validity of large-scale ESL writing assessment. *Assessing Writing, 17*, 123–139.
Huot, B. (1990). The literature of direct writing assessment: Major concerns and prevailing trends. *Review of Educational Research, 60*, 237–263. https://doi.org/10.3102/00346543060002237.
Huot, B. (1994). Editorial: An introduction to assessing writing. *Assessing Writing, 1*, 1–9.
Huot, B., & Yancey, K. B. (1994). From the editors. *Assessing Writing, 1*, 143–145.
Huot, B., & Yancey, K. B. (1995). From the editors. *Assessing Writing, 2*, 1–4.
Jeffery, J. V. (2009). Constructs of writing proficiency in US state and national writing assessments: Exploring variability. *Assessing Writing, 14*, 3–24. https://doi.org/10.1016/j.asw.2008.12.002.
Johnson, A. C., Wilson, J., & Roscoe, R. D. (2017). College student perceptions of writing errors, text quality, and author characteristics. *Assessing Writing, 34*, 72–87. https://doi.org/10.1016/j.asw.2017.10.002.
Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*, 1–73. https://doi.org/10.1111/jedm.12000.
LaFrance, J., & Nichols, R. (2010). Reframing evaluation: Defining an indigenous evaluation framework. *Canadian Journal of Evaluation, 23*, 13–31.
Lave, J., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation.* New York, NY: Cambridge University Press.
Levine, S. (in press) A century of change in high school English assessments: An analysis of 110 New York Regents Exams, 1900 – 2018. *Research in the Teaching of English*, 54.
Li, J., & Barnard, R. (2011). Academic tutors' beliefs about and practices of giving feedback on students' written assignments: A New Zealand case study. *Assessing Writing, 16*, 137–148. https://doi.org/10.1016/j.asw.2011.02.004.
Lindhardsen, V. (2018). From independent ratings to communal ratings: A study of CWA raters' decision-making behaviors. *Assessing Writing, 35*, 12–25. https://doi.org/10.1016/j.asw.2017.12.004.
Lord, F. (1952). *A theory of test scores. Psychometric monograph No. 7.* Retrieved from Richmond, VA: Psychometric Corporation. https://www.psychometricsociety.org/.
Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Menlo Park, NJ: Addison-Wesley.
Meijer, R. R., & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement, 25*, 107–135.
Mislevy, R. J. (2018). *Sociocognitive foundations of educational measurement.* New York, NY: Routledge.
Mislevy, R. J., & Elliot, N. (in press). Ethics, psychometrics, and writing assessment: A conceptual model. In J. Duffy & L. Agnew (Eds.), *Rewriting Plato's legacy*. Logan, UT: Utah State University Press.
Moss, P. (1994). Validity in high stakes writing assessment: Problems and possibilities. *Assessing Writing, 1*, 109–128. https://doi.org/10.1016/1075-2935(94)90007-8.
National Research Council (2001). *Knowing what students know. The science and design of educational assessment.* Washington, DC: National Academy Press.
No Child Left Behind Act of 2001, P.L. 107-110, 20 U.S.C. § 6319 (2002). Retrieved from https://www2.ed.gov/policy/elsec/leg/esea02/107-110.pdf.
Oliveri, M. E., Mislevy. R. J., & Elliot, N. (in press). After admissions: What comes next in higher education. In M. E. Oliveri & C. Wendler (Eds.), *Higher education admission practices: An international perspective*. Cambridge, UK: Cambridge University Press.
Petersen, J. (2009). "This test makes no freaking sense": Criticism, confusion, and frustration in times writing. *Assessing Writing, 14*, 178–193. https://doi.org/10.1016/j.asw.2009.09.006.
Petersen, N. S., & Novick, M. N. (1976). An evaluation of some models for culture-fair selection. *Journal of Educational Measurement, 13*, 3–29. https://doi.org/10.1111/j.1745-3984.1976.tb00178.x.
Petticrew, M., & Roberts, H. (2006). *Systematic reviews in the social sciences: A practical guide.* Malden, MA: Blackwell.
Poe, M. (2009). Reporting race and ethnicity in international assessments. In C. Schreiner (Ed.). *Handbook of research on assessment technologies, methods, and applications in higher education* (pp. 368–385). Hershey, PA: IGI Global Books.
Poe, M., & Cogan, J. A. (2016). Civil rights and writing assessment: Using the disparate impact approach as a fairness methodology to evaluate social impact. *The Journal of Writing Assessment, 9*. Retrieved http://journalofwritingassessment.org/article.php?article=97.
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*, 588–611.
Poe, M., Inoue, A. B., & Elliot, N. (Eds.). (2018). *Writing assessment, social justice, and the advancement of opportunity*Fort Collins, CO: The WAC Clearinghouse and University Press of Colorado. Retrieved https://wac.colostate.edu/books/perspectives/assessment/.
Pritchard, E. D. (2016). *Fashioning lives: Black queers and the politics of literacy.* Carbondale, IL: Southern Illinois University Press.
Purcell-Gates, V. (1995). *Other people's words: The cycle of low literacy.* Cambridge, MA: Harvard University Press.
Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* CPH: Danish Institute for Educational Research.
Ruff, R. (2019). State-level autonomy in the era of accountability: A comparative analysis of Virginia and Nebraska education policy through No Child Left Behind. *Education Policy Analysis Archives, 27*, 1–27.
Rupp, A. A. (2013). A systematic review of the methodology for person fit research in item response theory: Lessons about generalizability of inferences from the design of simulation studies. *Psychological Test and Assessment Modeling, 55*, 3–38.
Ruth, L., & Murphy, S. (1988). *Designing writing tasks for the assessment of writing.* Norwood, NJ: Ablex.
Schendel, E., & O'Neill, P. (1999). Exploring the theories and consequences of self-assessment through ethical inquiry. *Assessing Writing, 6*, 199–227. https://doi.org/10.1016/S1075-2935(00)00008-8.
Scheuneman, J. D. (1984). A theoretical framework for the exploration of causes and effects of bias in testing. *Educational Psychologist, 19*, 219–225. https://doi.org/10.1080/00461528409529298.
Schryer, C. (1994). the lab vs. the clinic: Sites of competing genre. In A. Freedman, & P. Medway (Eds.). *Genre and the new rhetoric* (pp. 105–124). London: Taylor and Francis.
Scott, M., & Tribble, C. (2006). *Textual patterns: Key words and corpus analysis in language education.* Amsterdam, Netherlands and Philadelphia, PA: John Benjamins.
Searle, J. (1969). *Speech acts: An essay in the philosophy of language.* Cambridge, UK: Cambridge University Press.
Slomp, D. (2016). An integrated design and appraisal framework for ethical writing assessment. *The Journal of Writing Assessment, 9*. Retrieved from http://journalofwritingassessment.org/article.php?article=91.
Solano-Flores, G. (2008). Who is given tests in what language by whom, when, and where? The need for probabilistic views of language in the testing of English language learners. *Educational Researcher, 37*, 189–199. https://doi.org/10.3102/0013189X08319569.
Solano-Flores, G., & Li, M. (2013). Generalizability theory and the fair and valid assessment of linguistic minorities. *Educational Research and Evaluation, 19*, 245–263. https://doi.org/10.1080/13803611.2013.767632.
Solano-Flores, G., Backhoff, E., Contreras-Niño, L. A., & Vázquez-Muñoz, M. (2015). Language shift and the inclusion of indigenous populations in large-scale assessment programs. *International Journal of Testing, 15*, 136–152. https://doi.org/10.1080/15305058.2014.947649.
Spalding, E., & Cummins, G. (1998). It was the best of times. It was a waste of time: University of Kentucky students' views of writing under KERA. *Assessing Writing, 5*, 167–199.
Spearman, C. (1904). "General Intelligence," objectively determined and measured. *The American Journal of Psychology, 15*, 201–292.
Sternglass, M. S. (1997). *Time to know them: A longitudinal study of writing and learning at the college level.* Mahwah, NJ: Lawrence Erlbaum.
Strauss, A., & Corbin, J. J. (1998). Grounded theory methodology: An overview. In N. K. Denzin, & Y. S. Lincoln (Eds.). *Strategies of qualitative inquiry* (pp. 158–183). Thousand Oaks, CA: Sage.
Sudweeks, R. R., Reeve, S., & Bradshaw, W. S. (2005). A comparison of generalizability theory and many-facet Rasch measurement in an analysis of college sophomore writing. *Assessing Writing, 9*, 239–261.
Taylor, D. (1983). *Family literacy: Young children learning to read and write.* Portsmouth, NH: Heinemann.

U.S. Const. amend. XIX. Retrieved https://www.govinfo.gov/content/pkg/GPO-CONAN-2002/pdf/GPO-CONAN-2002-9-15.pdf.
Veira, K. (2016). *American by paper: How documents matter in immigrant literacy*. Minneapolis, MN: University of Minnesota Press.
Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes.* Cambridge, MA: Harvard University Press.
White, E. M., & Thomas, L. L. (1981). Racial minorities and writing skills assessment in the California State University and Colleges. *College English, 43*, 276–283.
White, H. (2014). *The practical past.* Evanston, IL: Northwestern University Press.
Williamson, M. (1994). The worship of efficiency: Untangling theoretical and practical considerations in writing assessment. *Assessing Writing, 1*, 147–173. https://doi.org/10.1016/1075-2935(95)90021-7.
Wind, S. A., & Engelhard, G. (2013). How invariant and accurate are domain ratings in writing assessment. *Assessing Writing, 18*, 278–299.
Wood, S., & Elliot, N. (2019). Commemorating community: Forty years of writing assessment in *WPA: Writing Program Administration. WPA: Writing Program Administration, 42*, 28–35.
Yen, W. M., & Fitzpatrick, A. R. (2006). Item response theory. In R. L. Brennan (Ed.). *Educational measurement* (pp. 111–153). (4th ed.). Westport, CT: American Council on Education/Praeger.

Mya Poe is Associate Professor of English and Director of the Writing Program at Northeastern University. Her research focuses on writing assessment and writing development with particular attention to equity and fairness. She is the co-author of *Learning to Communicate in Science and Engineering* (CCCC Advancement of Knowledge Award, 2012), co-editor of *Race and Writing Assessment* (CCCC Outstanding Book of the Year, 2014), and co-editor of *Writing, Assessment, Social Justice, and Opportunity to Learn* (2019). She has also guest-edited special issues of *Research in the Teaching of English* and *College English* dedicated to issues of social justice, diversity, and writing assessment. She is series co-editor of the *Oxford Brief Guides to Writing in the Disciplines.* In 2015–2016, she won the College of Social Sciences and Humanities Outstanding Teaching Award and the Northeastern University Teaching Excellence Award.

Norbert Elliot is Professor Emeritus of English at New Jersey Institute of Technology. In 2016, he was appointed Research Professor in the Department of English at University of South Florida. He presently serves on the editorial boards of *Assessing Writing*, *IEEE Transactions in Professional Communication*, and *Research in the Teaching of English*. From 2017 to 2019, he served as Editor-in-Chief of *The Journal of Writing Analytics*. Most recently, he is co-author with Richard Haswell of *Early Holistic Scoring of Writing: A Theory, A History, A Reflection* (Utah State University Press, 2019). With Diane Kelly-Riley, he is co-editor of *Improving Outcomes: Disciplinary Writing, Local Assessment, and the Aim of Fairness* (forthcoming, Modern Language Association).