Language Teaching (2025), 1–13
doi:10.1017/S0261444825000096

CAMBRIDGE UNIVERSITY PRESS

FIRST PERSON SINGULAR

Assessing approaches to writing assessment: Considering claims about fairness

Kristen Di Gennaro and Meaghan Brewer

Pace University, NY, USA
Corresponding author: Kristen Di Gennaro; Email: kdigennaro@pace.edu

(Received 21 February 2025; accepted 25 February 2025)

1. Introduction

Despite the prevalence of assessment and reflections on (un)fairness in everyday life, writing teachers often fail to question if their assessments of students' writing are fair. More specifically, some fail to identify and reduce potential threats to fairness. Recently, several writing scholars (situating their work mainly in L1 contexts) have proposed assessment approaches that, they claim, promote fairness and/or justice. These scholars, however, seem unaware or even dismissive of the ongoing conversations about fairness that have long taken place in related, adjacent disciplines, such as educational measurement, language teaching and assessment, and second language writing. Likewise, while the concept of fairness has been theorized in depth by academics in educational measurement and language testing over the last several decades, such discussions tend to focus on large-scale language assessments rather than everyday classroom teaching. As such, scholars in those disciplines often omit practical suggestions toward fair methods of writing assessment that classroom teachers can apply without requiring complete program revision or waiting for ideological changes in society at large. In this position paper, we assess current writing assessment approaches and then translate relevant theory into practice by providing an operationalization of fairness applicable to classroom-based writing assessment contexts.

This paper was motivated, in part, by our experiences as writing program administrators encountering a surprising amount of acrimony toward assessment from faculty at our university, a sentiment which has extended, as we will discuss later, to first language (L1) writing scholarship more broadly. We thus begin by noting how we are surrounded by assessment and questions of fairness in many areas of everyday life, not just in our classrooms. We use these examples to illustrate how the ubiquity of assessment outside teaching contexts often goes unnoticed, while frequent attention to fairness, or more likely, unfairness, in everyday contexts suggests that we know what it is without having to define it. But claims that a particular writing assessment procedure is fair must be founded on clear understandings of both assessment and fairness. To this end, we define these and related terms, such as VALIDITY and JUSTICE, noting, in particular, distinctions between FAIRNESS and JUSTICE, as these terms are often used interchangeably by laypersons and scholars alike with the unproductive consequence that fairness is only minimally addressed. We then describe current approaches to writing assessment presented as promoting fairness, including ungrading, labor-based grading, and specifications grading, noting how fairness is addressed (or not) in each approach.

Our aim here is not to introduce a novel approach to fairness in writing assessment but rather to critically evaluate approaches claiming to address fairness. As we discuss later, because these

© The Author(s), 2025. Published by Cambridge University Press.

https://doi.org/10.1017/S0261444825000096 Published online by Cambridge University Press



Figure 1. Image of packaging depicting a checklist for assessing paper towels.

approaches come from L1 writing contexts, we caution second language (L2) writing instructors who encounter ungrading and labor-based grading strategies to consider the shaky foundations on which these practices are based. Closer scrutiny of these practices reveals their similarity to pre-1970s indirect writing assessments, which are widely criticized by writing assessment scholars today. For this and other reasons that we will describe below, we question the fairness of such approaches, despite their claims. While we do not address fairness in standardized writing assessment, based on the distinction between fairness and justice that we adopt here, we note that the (mis)uses and consequences of standardized writing assessments are best discussed through the lens of justice. We conclude by proposing that well-designed rubrics in general, and specifications grading in particular, provide theory-grounded methods that can be tailored to address fairness.

2. What is assessment?

While language testing and assessment researchers tend to use the terms ASSESSMENT and GRADING descriptively, our experiences as writing program administrators have shown us that these terms often evoke negative interpretations among writing instructors and academics more broadly. For example, one of us was recently in a meeting of our university's curriculum committee where it was announced that a newly added program was due for an assessment, which seemed to us like a common, periodic practice. The prospect of having to undergo an assessment was met with such outrage from faculty leading the program that the meeting ended without resolution. Another colleague shared with us that at her university, the program formerly known as the Office of Institutional Research and Assessment had recently replaced the word ASSESSMENT with EFFECTIVENESS.

While we find extreme negative connotations for ASSESSMENT unfortunate, it is also true that viewing assessment as inherently neutral and objective fails to consider the potential consequences of assessments on individuals. Regardless of positive, negative, or neutral associations with the word itself, assessment practices are ubiquitous and unavoidable, not just in educational settings but in everyday life. For example, even when choosing which brand to purchase for something as mundane as paper towels, we assess our options. Product manufacturers present us with features to consider in our decision. In addition to objective, quantitative features such as price, size, and total amount, in the case of paper towels, we might also consider somewhat subjective, quality features, such as absorbency, durability, and texture. The manufacturer acknowledges these additional considerations by providing a checklist on the packaging, highlighting features they believe will lead consumers to favor their product, as depicted in the image in Figure 1.

In the language of assessment, the checklist provided in Figure 1 includes criteria (similar to a grading scheme or rubric) that customers might use to evaluate the quality of the product:

https://doi.org/10.1017/S0261444825000096 Published online by Cambridge University Press

- Thick and absorbent
- Strong when wet
- Additional softness and absorbency.

That is, the packaging offers a visible, proposed definition of quality for this product. In short, it proposes a theory of quality for paper towels. The same can be said for criteria used in writing assessment. As writing teachers, we recognize that judging written language includes subjectivity, and thus, our assessments of students' writing cannot rely on universal, objective criteria applicable in all contexts. Our goal in assessing writing, however, is not to deny these subjective elements, but rather to identify which elements matter for a specific assessment context and then focus on those features.

We should note that we are not arguing for a return to the indirect and so-called objective tests of writing dismissed decades ago, in which right/wrong dichotomous test items (e.g., multiple-choice or true-false statements) were used to measure students' writing ability in place of actual writing. Rather, our position recognizes that since writing is context-specific, our assessments should be too. In fact, we will discuss later how currently popular writing assessment approaches ironically return us to indirect assessment, often in the name of fairness. Since fairness intersects with validity, or in some interpretations is a part of validity, in the next section we define fairness as it relates to validity.

3. Validity and fairness

Most educators are familiar with a rather simplistic definition of validity, such as validity is the degree to which an assessment measures what it purports to measure. Along these lines, claims are made as to whether a specific test or assessment 'has validity' or 'is valid.' In educational measurement and language assessment, however, 'validity is not a property of the test or assessment as such, but rather the meaning of the test' in context, which includes the test questions, conditions, test-takers, evaluators, and results (Messick, 1995, p. 741). Most importantly, validity rests in the interpretation of an assessment as well as in the consequences stemming from that interpretation. Viewed in this light, an assessment (or test) is not inherently valid or invalid but can be argued to have greater or less validity depending on how well its results reflect meaningful interpretations within a specific context and are free from unintended consequences. For results to be meaningful, they must reflect the construct being assessed within a given context; that is, they ask us to attend to the expected features of the writing task.

Assessments that only partially reflect a construct suffer from construct under-representation, while those that (inadvertently) introduce influences unrelated to the construct as it has been defined in the assessment context suffer from construct irrelevance. Ideally, we aim to maximize construct representation and minimize construct irrelevance. Attention to both construct representation and construct (ir)relevance is critical for ensuring fairness in assessments, with the latter often the cause of unseen threats to fairness.

Significantly for our discussion, definitions of the same construct can vary depending on the uses of the assessment. For example, a writing assessment intended for placement purposes will define the construct in reference to a particular program's curriculum, while an assessment of students' summary writing in response to a specific task following classroom instruction will define the construct based on features addressed during instruction. In fact, in our writing program administration roles, one of our most common reminders to faculty is that if they are assessing a feature of writing, then they should be teaching it. When an overly broad or narrow construct definition is used to assess students' writing in response to specific classroom tasks, such as when writing instructors use the same generic rubric or set of grading criteria for all assessments, they introduce construct (under)representation and/or irrelevance as potential threats to validity. Similarly, when writing assessments attend mainly to spelling and surface-level grammar rules, construct representation is threatened.

https://doi.org/10.1017/S0261444825000096 Published online by Cambridge University Press

The most egregious threats to construct (ir)relevance and (under)representation occur when the construct is left undefined or taken for granted, without a rubric or scoring scheme to aid in the assessment procedure. Such a scenario is frequently the case in L1 (namely, US) college writing classes where, as we noted earlier, instructors often view assessment unfavorably, seeing rubrics and grading procedures either as limited to primary or secondary schooling (and thus too elementary for their context), or as tools from standardized testing (and thus too commercial, inauthentic, or dehumanizing). In fact, when one of us recently reviewed a syllabus for a writing intensive course and asked to see the rubrics being used for assessment (a requirement for these courses), we received in response a screed against rubrics.

Validity also includes acknowledging the consequences of assessments, which brings us back to fairness. Similar to assessment, attention to fairness is ubiquitous in everyday activities. We sense fairness when we stand in line and are received in order and unfairness when someone is permitted to skip the line because they are friends with the manager. These and other common scenarios demonstrate our intuitive sense (or folk definition) of what fairness (or unfairness) is in everyday situations. Yet this intuition often fails us, or is inadequate, in writing assessment contexts. The inclusion of consequences as part of validity requires understanding what constitutes fairness in writing assessment.

A lack of attention to fairness in writing assessment could be explained by a division in expertise within writing assessment, where one branch specializes in assessment and another in pedagogy, and with the latter viewing the former with skepticism. As L1 writing studies scholar Yancey (1999) admits, 'there still continues reluctance at best, and aversion at worst, to writing assessment' (p. 495). This 'aversion' has, unfortunately, led writing scholars in L1 contexts to believe they can have expertise in teaching writing without a basic understanding of assessment, or worse, that teaching can take place without assessment. Yet teaching without assessment can hardly be described as teaching, unless we limit teaching to the simple delivery of information or self-help coaching. If teaching is intended to induce learning, then it also requires assessment, however informal or low-stakes it might be (Crusan, 2010, 2015; Weigle, 2002).

This misunderstanding of the purpose of assessment has also allowed many writing scholars to remain unenlightened about how fairness has been theorized in language assessment and its relevance for writing pedagogy. Furthermore, ongoing skepticism toward educational measurement, most notably by L1 scholars in the US, has produced scholarship that attempts to center fairness yet fails to theorize or define the concept in operational terms that would benefit pedagogy and assessment practices. For example, in a recently released edited volume on 'emerging theoretical and pedagogical practices' in writing assessment, half of the volume is dedicated to issues of fairness (Kelly-Riley et al., 2024). A look at the chapters, however, namely those with the terms FAIRNESS and/or VALIDITY in their titles, reveals a disappointing continued isolation from long-standing assessment and measurement scholarship.

For discussions about fairness to be productive, they should at least attempt to define the term. In the following section, thus, we summarize theoretical discussions of fairness in the language assessment literature. Since language assessment scholars straddle both language and assessment, their discussions of fairness link back to definitions of validity from educational measurement.

4. Fairness in (language) assessment

As we noted above, for the past several decades, validity has been theorized in language assessment as comprising the construct, contexts, and consequences pertaining to an assessment, yet there remains a lack of agreement as to the scope of fairness and its relationship to the concept of justice. For example, leaving aside (mis)conceptions of fairness as reliability (cf. White, 1993), fairness has been defined in language assessment very narrowly as the absence of bias in results for certain individuals or groups (Kane, 2010), but also very broadly as encompassing validity and justice (Kunnan, 2000;

https://doi.org/10.1017/S0261444825000096 Published online by Cambridge University Press

Randall et al., 2024). It is also described as a prerequisite for validity (Xi, 2010), as distinct from yet a requirement for justice (Kunnan, 2014), and as not a quality but a process (Walters, 2021).

For writing assessment purposes, McNamara and Ryan's (2011) definition of fairness is the most useful, as it is sufficiently yet not overwhelmingly technical, avoids philosophical and performative digressions about justice (e.g., Inoue, 2022), and, most importantly, lends itself to operationalization in teaching contexts. Fairness, in their terms, refers to properties internal to an assessment, making it something that we, as writing teachers, can control, at least partially, with our own assessment procedures. The values associated with an assessment, by contrast, are questions of justice, external to classroom-based assessment and beyond the control of individual writing teachers. Many writing scholars, unfortunately, conflate the two concepts, resulting in a type of 'folk fairness' (Walters, 2021, p. 567) with limited usefulness for writing assessment purposes. By decoupling fairness from justice, and understanding the scope of fairness to encompass qualities inherent to an assessment while leaving qualities external to the assessment as issues of justice, we can begin to operationalize fairness to address it in our classroom writing assessments.

5. Fairness in current approaches to writing assessment

In the next section, we describe approaches to writing assessment that have seen recent gains in popularity based on claims that they counter unfairness associated with traditional assessments. We will argue, however, that not all approaches hold up under scrutiny, especially with regard to fairness.

5.1. Ungrading

Ungrading is an approach to writing assessment in which instructors are discouraged from assigning any grades to students' written work. Based on the work of L1 writing studies scholar Peter Elbow (1993, 1997) and education critic Alfie Kohn (2011), ungrading advocates argue that grades are overly simplistic representations of student achievement and thus should not exist at all (Stommel, 2017, 2020). Scholarship on ungrading has an unapologetic activist bent to it, depicting grades as products of hierarchical capitalist systems. Ungrading proponents propose abandoning grading systems completely, replacing them with qualitative descriptions of students' work. One of the most outspoken promoters of ungrading, Jesse Stommel, describes ungrading as separating assessment from grading, with assessment viewed favorably as a qualitative description of a writer's strengths and weaknesses (often in the form of peer reviews and instructor feedback on individual tasks) and grading depicted as a reductive quantitative measure (often as a letter or numerical score). For ungraders, their descriptive assessments of students' writing are valid in that they aim to help students grow as writers, while quantitative feedback in the form of grades is simply inherently unfair. As Elbow argued, 'Let's do as little ranking and grading as we can. They are never fair and they undermine learning and teaching' (Elbow, 1993, p. 127). Echoing this sentiment, Stommel (2020) states bluntly, 'grades aren't fair. They will never be fair' (p. 28).

In descriptions of ungrading, the inherent unfairness of grades is taken for granted without interrogation, as is the preference for qualitative over quantitative assessment. Moreover, Stommel and others create a false binary by pitting grades and fairness against compassion. This positing of fairness and compassion as opposing constructs reveals Stommel's limited understanding of both fairness and assessment. His failure to define fairness suggests that Stommel relies on some universal intuition of what fairness is and uses this folk definition to dismiss all grading as unfair.

In terms of validity, the pitfalls of ungrading show up when instructors determine grades. Although ungrading advocates argue that we should work towards a system without grades, in our current institutions, most of us are required to record grades for students at the end of a course (and, notably, instructors with contingent positions may feel the least power to resist traditional grading structures). This, of course, calls up issues of validity: what are these grades based on? Proponents

https://doi.org/10.1017/S0261444825000096 Published online by Cambridge University Press

of ungrading often describe basing students' final grades on student reflections, which the instructor uses to gauge what learning took place, and/or on a grade that is negotiated between student and instructor, with the student proposing a grade and the instructor having some authority to negotiate or override it (Perman, 2024; Stommel, 2020). But are grades determined by reflections and student self-assessment necessarily fair? Is it possible that certain student groups will (at least appear to) argue more effectively for higher grades, or that already advantaged students who have received good feedback in prior assessments will assume that they deserve higher grades, and less confident students will humbly assign themselves lower grades? Notably, ungrading advocate Tony Perman (2024) states that assigning their own grades is 'a lot for some students to handle,' so he allows them to 'opt out and request that … [he] grade them' instead, a sentiment echoed in other ungrading literature. This of course begs the question of which students opt out of determining their grades and why, and by what criteria are grades then being determined?

More important to the current discussion, what is the role of fairness in this procedure and eventual outcomes? Even more concerning are reports that ungrading simply replicates the biases against disadvantaged student groups that were already present in traditional grading. For example, ungrading could work against disadvantaged students as it removes information, structure, or signposting that helps them understand how well they are doing in a given class (see Crusan, 2024), a concern also raised, as we discuss in the next section, about labor-based grading. In other words, even if grades contain some degree of unfairness, they still provide critical information for students, and removing them could make the system less fair. Moreover, if we consider issues of validity, rather than dismiss grades altogether, time might be better spent working to improve the validity (and thus fairness) of grading processes – something ungrading fails to address entirely.

5.2. Labor-based grading

Labor-based grading (LBG) has recently enjoyed even more popularity than ungrading. It can be considered a reiteration, or perhaps rebranding, of contract grading, which has been around since at least the 1920s in various disciplines in both secondary and post-secondary contexts (Cowan, 2020). Simply put, in contract grading, the instructor articulates some set of requirements or objectives that students must achieve to receive a specified grade; for example, a student might have to achieve all objectives for a grade of 'C' before moving on to higher grades. An early feature of contract grading was a conversation between the instructor and students about the contract before students agreed to it. Cases where students were informed but had little say about the contracts were later termed 'unilateral' contracts; cases where students assisted in articulating the requirements for each grade were called 'negotiated' contracts (see Cowan, 2020; Danielewicz & Elbow, 2009). Despite their popularity and long history in L1 writing contexts, as Cowan (2020) points out, strikingly few empirical studies of their efficacy for learning exist (p. 5).

LBG in writing assessment is usually associated with L1 scholar Asao Inoue (2022), and again Elbow (1993, 1997) was one of its earliest proponents. Many critical literacy theorists also argued for contract grading as a key component of liberatory pedagogies aimed at combating oppressive educational structures. As such, they implicitly invoke fairness and/or justice; in fact, since the 1990s, contract grading is explicitly connected to social justice (see Cowan, 2020, for a full history).

In arguing for LBG, Inoue frequently refers to 'the writing assessment ecology of [his] classroom' as performing 'social justice work' (Inoue, 2022, p. 13, 61). He contends that 'using labor as the only way to grade my students allows my classroom assessment ecologies to engage in larger social justice projects … ones that interrogate and attempt to dismantle white language supremacy in schools and society' (Inoue, 2022, pp. 3–4). Inoue's reliance on Rawls' interpretation of 'justice as fairness' (Inoue, 2022, p. 66) explicitly equates the two concepts, a simplification that we have argued above is unproductive for effectively addressing fairness in classroom-based writing assessment. While we agree with Inoue's interrogation of language-based (in)justices (again, not new in language assessment),

https://doi.org/10.1017/S0261444825000096 Published online by Cambridge University Press

and we describe elsewhere a relevant pedagogy drawing on critical language awareness in the writing classroom (Di Gennaro et al., 2023), simply repeating statements about the existence of linguistic injustice does not in itself create fair writing assessments.

We also question the extent to which our construct definition of writing should rely on the amount of labor (usually measured in time) that students dedicate to each task. Is labor by itself sufficient for assessing students' writing, essentially serving as a proxy for the construct? If time on task is considered all or part of the writing construct, validity arguments in support of LBG must account for this by showing how it is relevant, something proponents potentially could but fail to do.

As with supporters of ungrading, LBG theorists are comfortable ignoring issues of validity addressed toward their own models while they vehemently challenge the validity of assessment models based largely on standard language ideologies. In all their (arguably justified) ranting against standard language ideologies, mainly in L1 contexts, they seem to think their position is new, failing as they do to acknowledge that most L2 writing assessment scholars, following research in sociolinguistics (Labov, 1970; Lippi-Green, 2012; Trudgill, 1999), have long argued against overvaluing standard languages. Perhaps the root of the problem is not that grades are inherently unfair or that certain students are (dis)advantaged based on their preexisting proximity to a standard variety of English, but rather L1 writing scholars' lack of familiarity with relevant linguistic, educational measurement, language teaching, and L2 writing assessment theories and practices (Crusan, 2024; Di Gennaro & Brewer, 2024). As Crusan (2024) also argues, while L2 writing instructors likely agree on many of the problems that ungrading and LBG proponents attempt to address, many of the practices illustrated in their approaches (e.g., self-assessment, increased student agency, decreased centrality of Standard Written English [SWE]) are either practices that L2 writing instructors have already been doing, or are based on limited foundations or research.

By ignoring significant problems with construct representation and relevance, Inoue and other LBG proponents seem to believe that writing assessments based exclusively on student labor – an arbitrary and quantitative measurement of time – automatically result in increased fairness. This view fails to consider that time, as with proximity to SWE, is a commodity to which students have unequal access (see Carillo, 2021; Wolfe, 2024). In other words, LGB simply replaces one potential for bias with another while still failing to address fairness (or validity) in writing assessment procedures.

Inoue (2022) falsely gives the impression of addressing validity by citing criticisms of standardized assessments of writing competency and not, we should note, by exploring models of classroom-based assessments. His superficial reading of writing assessment literature leads him to erroneously draw circular conclusions that 'using ANY scale based on our own judgments of writing is suspect in terms of its validity to make judgments on students and their writing, and the reliability of our judgments themselves' (p. 190, emphasis ours). Inoue seems to think that, because construct definitions of writing are based on theoretical models of what 'good writing' is, by using these constructs to create grading criteria, we are 'reify[ing]' something that 'is not real and does not exist' (pp. 191–192). Aside from overlooking that construct definitions of writing can and often do rely on both theoretical models and empirical research findings, Inoue suggests that construct validity is itself a myth, based as it is, in his view, simply on a random collection of abstract or imaginary features. Dare we note how closely this argument resembles that of William Jenning Bryan's questioning of the theory of evolution during the Scopes trial, where he posited that (counter to scientific evidence) 'Evolution is not truth, it is merely an hypothesis – it is millions of guesses strung together' (Time Staff, 1925)?

Unfortunately, by failing to explore language assessment and L2 writing literature more substantively in his book-length descriptions of LBG, Inoue fails to notice LBG's lack of attention to either validity or fairness in writing assessment. Despite using the term CRITERIA or CRITERION at least 32 times in his book (2022), and articulating specific criteria for assessing students' labor, Inoue disregards the potential for well-designed, criterion-based writing assessment focused on more objective than subjective criteria to address fairness. Even a slightly deeper understanding of language and

https://doi.org/10.1017/S0261444825000096 Published online by Cambridge University Press

writing assessment found in introductory chapters, encyclopedia entries, and books written for non-technical audiences (Crusan & Matsuda, 2018; Elsheikh, 2018; Di Gennaro, 2017, 2018; Greve et al., 2018) would have uncovered the usefulness of meaningful, task-specific rubrics for writing assessment, both for combating the standard language ideology that Inoue objects to as well as the fairness that LBG proponents claim to care about. A cynical reader might conclude that perhaps the point of LBG is not actually to address fairness in writing assessment, but rather to promote its promoters (Schmenk et al., 2018).

To summarize, ungrading and LBG are based on similar claims that most grading systems are unfair, either due to their very nature (ungrading) or to subjective judgments often associated with white language supremacy (LBG). Both rely on misunderstandings of reliability and validity in writing assessment to support their claims of unfairness, as neither is based on research in educational measurement, language assessment, or even writing assessment. Despite their claims to address fairness, both overlook extensive discussions about fairness (and justice) in these adjacent disciplines. As a result, both dismiss the existence of alternative fair (or valid) assessments of writing, and thus substitute direct assessments of writing with proxies: students' reflections on their own writing (ungrading) or the amount of time they put into writing tasks (LBG). In short, ungrading and LBG are modern-day examples of indirect measures of writing ability arguably no better than the multiple-choice and short answer tasks that writing scholars and teachers objected to decades ago.

5.3. Specifications or criterion-based grading

Though we take a critical stance toward ungrading and LBG, and we urge writing teachers seeking fair methods for L1 and L2 writing assessment to look elsewhere, we do not disagree that writing assessments based on unstated, vague, or primarily subjective criteria are unfair. In fact, we hope that these same concerns motivate proponents of ungrading and LBG to broaden their perspectives to consider writing assessment methods in which criteria are transparent, concrete, and as objective as possible. Furthermore, by linking fairness with validity, namely construct representation and (ir)relevance, and limiting fairness to outcomes within the immediate context of a writing assessment (i.e., classroom-based assessments) and excluding larger concerns of social justice (which are better addressed as program, policy, and political matters), we believe fair writing assessment is simpler and easier to achieve than ungrading and LBG promoters would have teachers believe through their conflation of problems associated with large-scale standardized testing and classroom-based writing assessments. We, too, disagree with the misuse of such tests (see Di Gennaro, 2017).

For classroom-based writing assessment, we argue that if we take the time to identify features of our writing assignments that are representative of the construct within a given task, articulate clearly and share these features with students, and then assess students on these features without introducing new or irrelevant criteria later in the assessment process, then we are moving toward fair writing assessment. To put it simply, we find specifications grading the best model for fair writing assessment.

Many writing scholars are familiar with distinctions between norm-referenced and criterion-referenced assessments, where the former is used to rank and compare test-takers with one another (often with major consequences such as acceptance or rejection to a specific university) and the latter is used to determine if a test-taker's performance meets specific, task-related requirements (as in certifying exams). Implementing criterion-referenced assessment relies on establishing representative and relevant criteria, often referred to a rubric or grading scheme. Multiple criteria on the rubric might be consolidated into a single level (as in holistic scoring rubrics) or each criterion might be evaluated separately and/or to varying degrees (as in analytic scoring rubrics).

Specifications grading is a type of criterion-based assessment with an analytic rubric but differs in that each criterion is limited to only two levels of assessment: satisfactory or not (Nilson, 2015). While specifications grading is often described as 'a new approach to grading,' its roots can be traced

https://doi.org/10.1017/S0261444825000096 Published online by Cambridge University Press

to earlier, competency-based education and other pass/fail or contract grading methods (Nilson, 2015, p. 14). The lack of attention to specifications grading that we have noticed in many L1 writing assessment contexts, however, is perhaps a consequence of its association with rubrics. While the use of rubrics or scoring schemes is common practice in many (namely L2) teaching contexts (Crusan, 2010; Weigle, 2002), many L1 writing scholars avoid them, claiming, as we noted earlier, that they are reductive measures promoted by commercial testing companies or only appropriate for primary or secondary schooling. We suspect, however, that such objections would decrease if more writing teachers received adequate training in assessment procedures (Crusan, 2015; Dempsey et al., 2009).

As this discussion illustrates, a rubric is an attempt to make a theoretical construct more concrete and transparent for assessment purposes. Our earlier example of the paper towels packaging shows how rubrics (or criterion-based assessments) are ubiquitous in everyday life because they serve a real purpose. For everything we judge or assess, we apply a set of criteria, or mental model, of what we expect and consider important. Our mental models may not even be visible to ourselves unless we are asked to articulate them. Besides aiding in our decisions for which products to purchase, just by articulating the features we value, we uncover our values and present them for others to consider, adopt, or question. Recognizing how we use criteria to evaluate and assess real-life concerns, even without our full awareness, helps put assessment in its place, not as an inherently damaging or evil activity designed to harm individuals or groups, but simply as a common, decision-making activity. Assessments in educational or writing course contexts are not inherently bad (or good) but rather a basic part of instruction in that assessments show us what an instructor has taught, and (hopefully) what students have learned. It follows, then, that a fair and well-designed assessment relies on a fair, well-defined, and articulated set of representative and relevant criteria aligned with a particular class or curriculum.

Despite this, criticisms of rubrics are omnipresent in L1 writing instruction contexts. Kohn (2020), whose work has formed the basis of the ungrading movement, depicts rubrics as inherently unfair, requiring 'detoxifying' because they 'are all about evaluation' (p. xvii). Without rubrics, however, the 'what' of what students are doing is unclear, making them more rather than less anxious about their performance as they try to meet a certain target but one that is invisible. That is, the lack of transparency in assessments is the real cause for concern for many students (Crusan, 2024; Elbow, 1993; Inman & Powell, 2018). Moreover, Inoue's (2022) and Elbow's (1993) frequent references to 'criteria' in their work show that regardless of what we think we are doing in our assessments, we are always evaluating students in reference to some set of criteria, articulated or not. And while it is true that some rubrics lack fairness by containing construct-irrelevant criteria, by making our criteria visible in the form of rubrics, we can work to eliminate construct-irrelevant criteria and communicate evaluation procedures clearly and transparently to students, which increases fairness.

An important characteristic of specifications grading as described by Nilson (2015) is linking scores, or grades, to outcomes, whether they be course- or assignment-specific. As Nilson argues, 'outcomes achievement is not a matter of degree; a student either achieves an outcome or does not achieve it in any given assessment' (p. 25). Essentially, students are assessed on a pass/fail or satisfactory/unsatisfactory basis for each specific criterion. This is in contrast to rubrics containing multiple gradations for each criterion, a format that introduces greater subjectivity into assessments. Incidentally, such multi-level rubrics might be what Kohn (2020) envisions (and objects to) with his reference to 'gradations' (p. xvii). While not arguing for specifications grading by name, in his frequently-cited article in both the ungrading and LBG literature, Elbow (1993) admits, 'I sometimes do a bit of ranking even on individual papers, using two "bottom-line" grades: H and U for "Honors" and "Unsatisfactory"' (p. 193). Elbow also supports fewer gradations in grading, acknowledging that, 'we can get along with fewer occasions for assessment but also with fewer gradations in scoring' (p. 194). We contend that, by removing the subjectivity and vague distinctions between levels of

https://doi.org/10.1017/S0261444825000096 Published online by Cambridge University Press

Table 1. Reviewer feedback questions for college composition journal

Please consider the following criteria as you write your review:
Significance
• Does the scholarship in this piece contribute something new and/or innovative to the body of research and scholarship in Rhetoric and Composition studies? Is the significance of the contribution signaled in the piece itself? • Will this piece appeal to the audience of *CCC*, which includes a wide range of educators from a variety of institutional backgrounds?
Theory/Methodology/Methods/Sources
• Is the piece contextualized relative to scholarship in Rhetoric and Composition studies, and to writing studies broadly defined? • Does the piece have an appropriate balance of the writer's and of other's work? Of primary and secondary sources (if applicable)? • Are any crucial sources overlooked? • Are the theories/methodologies/methods used appropriate for this topic? • Does this piece present concepts such as 'students or teachers or classrooms' in nuanced, specific, diverse ways?
Organization and Style
• Does the piece present clear arguments with well-substantiated ideas? • Is the focus of the piece well defined and articulated?

achievement found in both holistic and analytic rubrics, our assessment procedures promote greater fairness than either ungrading or LBG.

Some might argue that specifications grading is itself a kind of LBG, since students are graded not on how (subjectively) well they do but rather on their completion of a set of tasks or the inclusion of certain features in their writing. Indeed, specifications grading and LBG are not mutually exclusive, as rubrics can (and should) give attention to labor in that students are assessed for in completing writing tasks or attending to specified writing elements (e.g., using examples or citing a specified number of outside sources). In contrast with LBG, however, labor in specifications grading is not measured by the simple, quantitative construct of time, but is translated into evidence that the student worked toward the writing task. In other words, in specifications grading, a rubric or list of criteria specify which labors students should undertake. Interestingly, the reviewer feedback questions from a major journal for teachers of L1 college composition, *College Composition and Communication*, take the form of a specifications grading rubric (though they do not label it as such) in that each criterion is addressed through a simple Yes/No question, leading to just two levels of achievement for each (CCC editorial office, personal communication, November 29, 2023: see Table 1).

Finally, and perhaps most importantly, drawing on and developing writing assessment rubrics forces us to articulate, and make visible for others, our theory of writing for a given context or task. Such transparency has obvious connections to fairness: stakeholders know how they are being assessed, instructors are disinclined to add criteria that could privilege individual students who impress them on unforeseen and potentially construct-irrelevant criteria later in the grading process, and criteria can be questioned. Transparency also connects our practices with theory-building, as we discover the values we hold through our construct definitions and assignment descriptions. In this sense, rubrics also encourage a justice-oriented approach to writing assessment in that when visible, values can also be challenged and modified as programs, policies, and politics draw our attention to societal injustices that potentially can be addressed through our methods of assessment.

6. Conclusion

Table 2 summarizes how each of the three approaches to assessment described above addresses fairness.

https://doi.org/10.1017/S0261444825000096 Published online by Cambridge University Press

Table 2. Summary of fairness in three approaches to writing assessment

Assessment method	How fairness is defined and addressed
Ungrading	• Fairness is left undefined but all grades are determined to be unfair. • Qualitative (formative) assessment is preferred, while quantitative (summative) assessment is dismissed as inherently unfair. • Criteria for determining grades are left unarticulated until students receive feedback (lack of fairness). • Assessments rely on indirect measures of writing, such as student reflections and self-assessments.
Labor-based grading	• Fairness is defined in terms of reliability; since reliability of judgments is questioned, so are assessments based on judgments. • Construct definitions of writing are described as imaginary or unfounded, leading to the complete inattention to construct representation and relevance. • Fairness is equated with justice in that writing assessments are criticized to the extent that they rely on SWE. • Labor, defined in terms of time (another indirect measure of writing), is used as a proxy for writing.
Specifications grading	• Fairness exists in the rubric itself to the extent that criteria are aligned with teaching and construct-relevant writing tasks. • Full transparency exists to the extent that criteria are shared with students before, during, and after a writing task, and when instructors apply the criteria as stated. • Fewer gradations related to quality reduce potential for bias in assessment results.

We began this article by noting how some faculty and administrators have negative associations with assessment, believing it to be either time-consuming and pointless or, worse, a threat to their courses or programs. These simplistic, under-informed understandings of assessment, illustrated here in ungrading and LBG, also fail to recognize that teaching without assessment is hardly teaching. While adherents of ungrading and LBG argue that their approaches to writing assessment are more fair than other, perhaps more traditional methods, by overlooking construct representation and (ir)relevance, ungrading and LBG lack validity arguments and, consequently, lack evidence of fairness. In fact, by favoring indirect measures of writing (self-reflections, self-assessment, and time on task), they signal a return to pre-1970s indirect writing assessment methods even if students are required to write.

Despite the amount of personalized feedback instructors might offer students, ungrading and LBG as assessment practices are far from fair in that the criteria for assessment remain undefined until students receive feedback. In many cases, instructors may not even know what they are looking for until they read students' writing. Some may argue that this lack of specificity provides flexibility in the feedback process, allowing for students to be assessed on different criteria as these emerge in their writing. We contend, conversely, that this lack of transparency promotes unfair, covert assessment practices. For writing assessments to be fair, grading criteria must be made transparent for all stakeholders, especially students. This does not mean that only one set of criteria is considered fair, as contexts for writing vary, and thus, assessment criteria will also vary according to the writing task and context (Greve et al., 2018). Nor are we suggesting that a perfect set of criteria exists for a given task or context. As we argued earlier, assessment criteria represent a mental model of the quality or ability being assessed, and such models can and do change.

While we leave it to readers to determine the most relevant criteria for their own writing contexts, we offer the following considerations to check for fairness in their classroom writing assessments:

1. On what criteria will you grade students' work? How have you selected these criteria (i.e., from a widely accepted understanding of writing ability)?

https://doi.org/10.1017/S0261444825000096 Published online by Cambridge University Press

2. How do your criteria relate to the specific task, class instruction, and time constraints?
3. Have you given students the criteria you will use to assess their work?
4. Are your writing prompts clear? How have you verified that students understand what they are expected to do?
5. What hidden criteria pose a threat to your assessment? What might influence your assessment, either negatively or positively?

Developing fair assessments using rubrics requires some time and thought when creating an assignment, but saves time and confusion later, both for instructors and students. In essence, instead of responding instinctively to student's writing, more time is spent teaching within an integrated learning and assessment process. In a sense, fair assessment relies not on feedback but on 'feedforward,' where instruction and clarity about expectations occur BEFORE students submit their work rather than after.

Competing interests. The author(s) declare none.

References

Carillo, E. C. (2021). *The hidden inequities of labor-based contract grading*. Utah State University Press.

Cowan, M. (2020). A legacy of grading contracts for composition. *The Journal of Writing Assessment*, *13*(2), 1–16. https://escholarship.org/uc/item/0j28w67h

Crusan, D. (2010). *Assessment in the second language writing classroom*. University of Michigan Press. doi: 10.3998/mpub.770334

Crusan, D. (2015). Dance, ten; looks, three: Why rubrics matter. *Assessing Writing*, *26*, 1–4. doi: 10.1016/j.asw.2015.08.002

Crusan, D. (2024). Ungrading: Revolution or evolution. *Journal of Second Language Writing*, *66*, 1–7. doi: 10.1016/j.jslw.2024.101149

Crusan, D., & Matsuda, P. K. (2018). Classroom writing assessment. In J. I. Liontas (Ed.), *The TESOL Encyclopedia of English language teaching* (online). Wiley 1–7.

Danielewicz, J., & Elbow, P. (2009). A unilateral grading contract to improve learning and teaching. *College Composition and Communication*, *61*(2), 244–268. https://www.jstor.org/stable/40593442

Dempsey, M. S., PytlikZillig, L. M., & Bruning, R. H. (2009). Helping preservice teachers learn to assess writing: Practice and feedback in a Web-based environment. *Assessing Writing*, *14*(1), 38–61. doi: 10.1016/j.asw.2008.12.003

Di Gennaro, K. (2017). SAT scores are useful for placing students in writing courses. In D. Loewe, & C. Ball (Eds.), *Bad ideas about writing* (pp. 294–298). Digital Publishing Institute. https://textbooks.lib.wvu.edu/badideas/badideasaboutwriting-book.pdf.

Di Gennaro, K. (2018). Subjectively-scored formats. In J. I. Liontas (Ed.), *The TESOL Encyclopedia of English language teaching* (online). Wiley 1–6.

Di Gennaro, K., & Brewer, M. (2024). Linguistic currents in Writing Studies scholarship: Describing variation in how linguistic terms have been borrowed and (re-)interpreted in Writing Studies. *Across the Disciplines*, *21*(2–3), 82–101. doi: 10.37514/ATD-J.2024.21.2-3.02

Di Gennaro, K., Choong, K. W., & Brewer, M. (2023). Uniting CLA with WAW via SLA: Learning about written language as a model for college writing courses. *Journal of Second Language Writing*, *60*, 100967. doi: 10.1016/j.jslw.2023.100967

Elbow, P. (1993). Ranking, evaluating, and liking: Sorting out three forms of judgment. *College English*, *55*(2), 187–206. doi: 10.2307/378503

Elbow, P. (1997). Grading student writing: Making it simpler, fairer, clearer. *New Directions for Teaching and Learning*, *69*, 127–140. doi: 10.1002/tl.6911

Elsheikh, A. (2018). Rubrics. In J. I. Liontas (Ed.), *The TESOL encyclopedia of English language teaching* (online). Wiley 1–8.

Greve, C., Morris, W., & Huot, B. (2018). Use and misuse of writing rubrics. In J. I. Liontas (Ed.), *The TESOL encyclopedia of English language teaching* (online). Wiley 1–8.

Inman, J. O., & Powell, R. A. (2018). In the absence of grades: Dissonance and desire in course-contract classrooms. *College Composition and Communication*, *70*(1), 30–56. https://www.jstor.org/stable/26772544.

Inoue, A. (2022). *Labor-based grading contracts: Building equity and inclusion in the compassionate writing classroom* (2nd ed.). University Press of Colorado. doi: 10.37514/PER-B.2022.1824

Kane, M. (2010). Validity and fairness. *Language Testing*, *27*(2), 177–182. doi: 10.1177/0265532209349467

Kelly-Riley, D., Macklin, T., & Whithaus, C. (Eds.). (2024). *Considering students, teachers, and writing assessment, Volume 2: Emerging theoretical and pedagogical practices*. The WAC Clearinghouse; University Press of Colorado. doi: 10.37514/PER-B.2024.2326

https://doi.org/10.1017/S0261444825000096 Published online by Cambridge University Press

Kohn, A. (2011). The case against grades. *Educational Leadership*. https://www.alfiekohn.org/article/case-grades/
Kohn, A. (2020). Foreword. In S. D. Blum (Ed.), *Ungrading: Why rating students undermines learning (and what to do instead)* (pp. xiii–xx). West Virginia University Press.
Kunnan, A. J. (2000). Fairness and justice for all. In A. J. Kunnan (Ed.), *Fairness and validation in language assessment* (pp. 1–13). Cambridge University Press.
Kunnan, A. J. (2014) Fairness and justice in language assessment. In A. J. Kunnan (Ed.), *The companion to language assessment* (1st ed., vol. 3, pp. 1–17). John Wiley & Sons.
Labov, W. (1970). The logic of non-standard English. In F. Williams (Ed.), *Language and poverty* (pp. 153–189). Rand McNally.
Lippi-Green, R. (2012). *English with an accent: Language, ideology and discrimination in the United States*. Routledge. doi: 10.4324/9780203348802
McNamara, T., & Ryan, K. (2011). Fairness versus justice in language testing: The place of english literacy in the Australian citizenship test. *Language Assessment Quarterly*, *8*(2), 161–178. doi: 10.1080/15434303.2011.565438
Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*(9), 741–749. doi: 10.1037/0003-066X.50.9.741
Nilson, L. B. (2015). *Specifications grading: Restoring rigor, motivating students, and saving faculty time*: Routledge. Stylus.
Perman, T. (2024, July 16). Ungrading for hope. *Inside Higher Ed*. https://www.insidehighered.com/opinion/career-advice/teaching/2024/07/16/four-powerful-benefits-ungrading-opinion
Randall, J., Poe, M., Slomp, D., & Oliveri, M. E. (2024). Our validity looks like justice. Does yours? *Language Testing*, *41*(1), 203–219. doi: 10.1177/02655322231202947
Schmenk, B., Breidbach, S., & Küster, L. (2018). *Sloganization in language education discourse: Conceptual thinking in the age of academic marketization*. Multilingual Matters. doi: 10.21832/9781788921879-002
Stommel, J. (2017, October 26). Why I don't grade. https://www.jessestommel.com/why-i-dont-grade/
Stommel, J. (2020). How to ungrade. In S. D. Blum (Ed.), *Ungrading: Why rating students undermines learning (and what to do instead)* (pp. 25–41). West Virginia University Press.
Time Staff. (1925, August). Education: Dixit. *Time Magazine*. https://time.com/3602563/education-dixit/
Trudgill, P. (1999). Chapter 5: Standard English: What it isn't. In T. Bex, & R. J. Watts (Eds.), *Standard English: The widening debate* (pp. 117–128). Taylor & Francis.
Walters, F. S. (2021). Ethics and fairness. In G. Fulcher, & S. Harding Eds., *The Routledge handbook of language testing* (2nd ed., pp. 563–577). Routledge.
Weigle, S. (2002). *Assessing writing*. Cambridge University Press. doi: 10.1017/CBO9780511732997
White, E. (1993). Holistic scoring: Past triumphs and future challenges. In M. M. Williamson, and B. Huot (Eds.), *Validating holistic scoring for writing assessment: Theoretical and empirical foundations*. Hampton (pp. 79–108).
Wolfe, J. (2024). What educational psychology can teach us about providing feedback to black students: A critique of Asao Inoue's antiracist assessment practices and an agenda for future research. *College Composition and Communication*, *75*(4), 759–788. doi: 10.58680/ccc2024754759
Xi, X. (2010). How do we go about investigating test fairness? *Language Testing*, *27*(2), 147–170. doi: 10.1177/0265532209349465
Yancey, K. B. (1999). Looking back as we look forward: Historicizing writing assessment. *College Composition and Communication*, *50*(3), 483–503. doi: 10.2307/358862

Kristen di Gennaro is an associate professor of English at Pace University. After teaching a range of English courses in Italy, the UK, and the US, she currently teaches undergraduate writing and linguistics courses. Her scholarly interests include writing pedagogy, assessment of writing, and issues in sociolinguistics. Her work has appeared in various journals including *Language Teaching, Assessing Writing, Applied Linguistics Review, Journal of Second Language Writing*, and *ELT Journal*.

Meaghan Brewer is an associate professor of English at Pace University, where she teaches courses in language, literacy, and writing. Her current research examines disciplinarity, writing pedagogy and assessment, and linguistic phenomena like compliments and microaggressions. Her work has recently appeared in *Across the Disciplines, Journal of Second Language Writing, College Composition and Communication*, and *Applied Linguistics Review*.

Cite this article: Di Gennaro, K and Brewer, M. (2025). Assessing approaches to writing assessment: Considering claims about fairness. *Language Teaching*, 1–13. https://doi.org/10.1017/S0261444825000096

https://doi.org/10.1017/S0261444825000096 Published online by Cambridge University Press