Writing Assessment Validity: Adapting Kane's Argument-Based Validation Approach to the Assessment of Writing in the Post-Process Era

by Josh Lederman, Brandeis University

This article examines the translatability of Kane's (1992, 2006, 2013, 2016) argument-based validation approach to the aims and needs of college writing assessment, particularly given (a) the elusive ways in which writing is theorized in the post-process era, and (b) the composition/writing assessment community's concern for matters of social justice. As part of this process, I review current conceptualizations of writing in the composition literature; I then discuss some overarching implications of assessment as a decision-making process; and then I review the underlying principles of Kane's approach. From there, I discuss which elements of Kane's theory apply more to testing and test development and less to writing assessment as theorized and practiced today; and then I offer criticism of Kane's theory and call for adaptations that could be made to help forge a theory of assessment validity more suited to the goals of writing assessment and composition studies at large.

Introduction

Validity is “the most fundamental consideration in developing and evaluating tests” (American Education Research Association, American Psychological Association, & the National Council on Measurement in Education, 2014, p. 11), yet the educational measurement/testing community has struggled mightily to employ consistent and thorough applications of contemporary validity theory (see Cizek, Bowen, & Church, 2010; Cizek, Rosenberg, & Koons, 2008). In composition studies, since scholars like Brian Huot and Michael M. Williamson began importing validity theory into writing assessment contexts (e.g., Huot, 1996; Williamson, 1994; Williamson & Huot, 1993), the writing assessment community has similarly struggled to apply validity theory in appropriate and consistent ways. The reasons for these struggles vary and even include a history of infighting about what the word *validity* itself means. But there is perhaps a positive reason for these difficulties: Educational assessments try to do very complex, difficult things; they seek to gather finite samples of students' knowledge or abilities and to base life-level decisions (e.g., passing a course, entering a profession, admission to a university) upon inferences drawn from these samples. These decisions have tremendous impact not only upon individuals but also upon social-historical groups—as Lee J. Cronbach (1988) reminded us, “Psychological and educational tests influence who gets what in society” (p. 5)—so perhaps the people responsible for ensuring the quality (validity) of assessments *ought* to find the task to be a consistent struggle.

Cronbach's (1988) above statement is powerful, but I suggest we could extend it further: Aside from influencing who gets what in society, these tests also influence who *gives* what *to* society (e.g., in the form of professional licensure or admission to higher education, and therefore increased positions of power and influence), often upholding systems in which people in positions of power and authority tend to come primarily from historically privileged social groups. We might also expand Cronbach's relationship between test scores and social goods because “who gets what in society” also has great influence on who gets which test scores. Due to histories of social advantages (socioeconomic status, parental education level, access to coaching, home and community English dialect, etc.), students from certain mainstream social groups tend to see their historical advantage manifest in higher test scores (which can then lead to more social goods, and so on); meanwhile, members of historically marginalized social-historical groups, who lack this type of access and advantage, tend to receive lower scores on average (on nearly any large-scale standardized test), and then these lower scores can reinforce continued lack of access and advantage, and so on. What's more, social-psychological concepts like *stereotype threat* (e.g., Steele & Aronson, 1995) help reveal the impact of past *generations'* scores upon the scores of individual test-takers from similar marginalized social-historical groups, in order to further complicate the connection between testing/assessment and who gets what in society.

The point here is that assessment plays a major role in the distribution of social goods and in the reproduction of social power and marginalization, and, as such, the process of *validation*—the process of evaluating assessments—should not be treated as a simplistic, paint-by-numbers concept. Indeed, the very notion of what makes a given assessment valid, what determines “the soundness of [its] interpretations, decisions, or actions” (Moss, Girard, & Haniford, 2006, p. 109), should be diligently theorized, given the stakes of assessment as a social enterprise. This is where validity theory, the topic of this article*,* is crucial to any assessment practice.

The present article regards the adaptation of validity theory, as published in the educational measurement/testing scholarship, to the specific aims and needs of writing assessment—particularly, the types of locally-controlled, site-based assessment (Huot, 1996, 2002) that most compositionists/writing program administrators (WPAs) perform, primarily on college campuses (as opposed to larger-scale writing assessments such as state-wide test, the TOEFL, or other non-local, published and purchased assessments). As the authors of *Writing Assessment, Social Justice, and Advancement of Opportunity* (Poe, Inoue, & Elliot, 2018) collectively argued, “Writing assessment remains under-theorized and, as such, often follows theories that may not align with humanistic dedication to value each student. Writing assessment theories are needed that advance dual aims of justice and opportunity” (p. 381). I see the current project as taking up that call.

Granted, writing assessment scholarship has seen a number of forays (if somewhat sporadically) into this type of work over the years. Assessment scholars like Schendel and O'Neill (1999), Perry (2008), Inoue (2009), and West-Puckett (2016) have all employed critical theory to explore the interconnections between validity and the social force of writing assessment, and authors like Poe, Elliot, Cogan, and Nurudeen (2014) and Slomp, Corrigan, and Sugimoto (2014) have used various quantitative and qualitative methodologies to explore and extend the ways in which measurement/test validity theory could inform writing assessment. Yet, the majority of current writing assessment scholarship (from 2010 to the present), when it does engage validity theory, either seeks to correctly apply existing measurement/test validity theory to writing assessment, or it clings to historical conceptualizations of validity that are no longer current in the measurement literature. What we have not seen is a sustained, systematic approach of the type mentioned by Poe et al. (2018) above, exploring how current measurement/test validity theory could be translated to writing assessment, particularly in ways that both engage our current theories of writing and that "advance [our] dual aims of justice and opportunity" (p. 381).

Below, I present an argument for adapting Michael T. Kane's argument-based validation approach (e.g., Kane, 1992, 2006, 2013), which represents the current state of the art in measurement/test validity theory, to best fit writing assessment theory and practice. I will argue that (a) measurement/test validity theory does contain a strong foundation for a theory of writing assessment, including the concern for promoting social justice and disrupting social inequities, but that (b) it cannot be adopted wholesale to writing assessment if it is to achieve these ends. It must be tailored, adapted, and translated from one context (educational measurement/testing) into another context (writing assessment), which has similar but distinct aims and needs. As such, I argue against the drive to reject measurement/testing concepts, specifically the word *validity*, as some assessment scholars have suggested (e.g., Lynne, 2004; Petruzzi, 2011; Royer & Gilles, 1998, hint at a similar argument), because, as others have pointed out (e.g., Huot, O'Neill, & Moore, 2010; Poe et al., 2014; Slomp et al., 2014), measurement/test validity theory has seen key developments over the years—it has become less mechanistic and more humanistic (e.g., Messick, 1989), less psychometric and more rhetorical (e.g., Kane, 2006, 2013)—and so current validity theory aligns far better with composition sensibilities than its older instantiations. But, because measurement/test validity is theorized primarily by and for those who develop and research *tests* as commonly conceived (which I distinguish from *writing assessment* below), this work remains often only tangentially relevant to the types of assessment practices that most college compositionists develop and use. Thus, in what follows, I explore how Kane's approach to validation can be adapted to help craft a theory of writing assessment validity that would support both our social/ethical aims and our current conceptualizations of teaching, learning, and writing/literacies.

What are We Assessing When We Assess 'Writing'?

Whether readers are fond of or resistant to the term *post-process*, today's conceptualizations of writing (what it is, how it is learned and taught) are quite different from those of the 1970s through early 1990s, that is, the process era. Perhaps the most fundamentally altered assumption about *what* writing is concerns the idea of *where* writing is. Process-era wisdom considers the writing process to be located within the individual writer's mind, but newer theories of writing tend to complicate, or even reverse, this positioning of writer and process—viewing writing process(es) as a larger set of practices that writers engage *in*, as opposed to the processes happening within a writer's mind. According to Kristopher M. Lotier (2016):

> [W]hereas the majority of process-era inventional schemes presupposed cognitive "internalism," the idea that one's mind is separate from other minds and from the world in which those minds exist, many (and perhaps all) post-process approaches… assume an 'externalist' viewpoint: that no cognitive action can occur without the contribution of human or nonhuman others, including languages and various technological artifacts. (p. 362)

I refer back to these concepts of internalism and externalism throughout what follows.

Composition studies experienced a social turn (Trimbur, 1994) sometime in the late 1980s/early 1990s, as newer conceptualizations of writing—and of writer/self/mind/identity—began to problematize the notion of the writer as a stable, asocial entity, separate and separable from writing. Gary A. Olson (2002), for one, critiqued the process-era "[assumption] that writing can be untethered from specific contexts, that somehow we can describe writing detached from specific acts of writing, specific attempts to communicate particular messages to particular audiences for particular reasons" (p. 425). Earlier, Patricia Bizzell asserted that "an anti-foundationalist understanding of discourse would see the student's way of thinking and interacting with the world, the student's very self, as fundamentally altered by participation in any new discourse" (as cited in Breuch, 2002, p. 138-139). In that same year, Marilyn Cooper (1986) published "The Ecology of Writing," where she argued "the time ha[d] come for some assessment of the benefits and limitations of thinking of writing as essentially—and simply—a cognitive process" (p. 364), claiming that such a notion "obscures many aspects of writing we have come to see as not peripheral" (p. 365).

Cooper's (1986) insight now seems prescient, given the intervening proliferation of digital technologies, the ramping up of "hyper-circulatory networks of contemporary information exchange" (Dobrin, 2010, p. 268) that all but define the 21st century. As Rebecca Tarsa (2015) added, "The literacy opportunities these [digital] spaces offer emphasize 'the public nature of writing' (Sabatino 42), with users rapidly delivering and receiving feedback through digitally produced and mediated text" (p. 12); this *digital landscape* "shifts the purpose of literate activity 'from individual expression to community involvement'" (p. 12). Thus, what Cooper had argued as a theoretical stance, Tarsa (2015) can now describe concretely. "Events," said Rau�l Sa�nchez (2012), "have caught up with theory" (p. 235).

It should be noted that the overall ethos of these philosophies—the externalization of writing and even mind/cognition/thinking, the reconceptualization of the writer-writing binary into a larger ecosystem—while not accompanied by pre-scripted pedagogical methods (and necessarily so), does have great potential for impact on classroom practice. Whether we consciously subscribe to ecological, networked, post-process philosophies of writing, teaching, and learning, many composition instructors find that as we encourage students' engagement in the network/ecology of existing and always moving, growing ideas, information, actions, and contribution, their writing reads less like classroom-based exercises/simulations *of* writing, and more like "real" writers doing real writing, engaged in authentic conversation with other scholars/writers/thinkers, not mere passers-by—something that many of them have not experienced in their previous schooling.

The question is, how do we assess "writing" conceived this way, when the instability/fluidity of both writer and writing seem incompatible with traditional concepts of educational measurement? Taking Kane's lead, I offer that such questions are answerable only after we articulate *why* we want to assess it.

Why We Assess Writing: Assessment as Decision-Making

The question of why we assess is foundational for any assessment following Kane's argument-based approach, but it also seems logical that theorizing any practice ought to begin at the beginning—with an exploration of why we are doing it at all. In the measurement literature, we see a common theme emerge in response to this question: W. James Popham (2010), for example, answered, "We don't test for the sheer joy of testing or because 'it is interesting.' Instead, we assess students in order to make better *decisions* about the curricular ends we should be pursuing" (p. 5, emphasis original). And, Kane (1992) added, "If the test scores were not relevant to any decision, it is not clear why the test would be given" (p. 530). This notion of assessment-as-decision-making is vital for a theory of writing assessment validity. If validity refers to the overall quality of the assessment—the extent to which it achieves its stated purposes (see below)—then we might be validating very different things depending upon what those purposes are. If the purpose of an assessment *were* for the sheer joy of measurement, then we might call the procedure valid if it yielded accurate measurements (readers may recognize this as an old, though still sometimes referenced, definition of validity), for that would be its end goal. However, if the purpose is educational decision-making—whether we seek to place students into writing courses, to evaluate whether they should be granted exit from our program, or whether we seek information about how we might best make curricular revisions—then the validity of the assessment must extend beyond the accuracy of the test's measurements because no amount of test-score accuracy can ensure, by itself, high quality decisions (Kane, 2006, 2013). As Kane (2006) stated:

> Evidence for the accuracy of the information is certainly relevant to the evaluation of a decision procedure but mainly because more accurate information is expected to lead to better decisions . . . . However, even indisputable evidence for accuracy does not justify a decision procedure. (p. 54)

This matter of focusing on, and working backwards from, the purpose of assessment could have rippling effects upon how we theorize the assessment enterprise itself, including how we search for and make claims of validity for any given assessment practice. Traditional, internalist conceptualizations of assessment and validity do seem to require that writing (and mind) must be conceptualized as stable, internal, and isolatable, facilitating accurate measurement. But, focusing on assessment-as-decision-making, rooted in Kane's larger theory, can problematize these assumptions, setting the stage for a similar type of externalist, ecological turn in assessment validity theory as experienced in composition studies over the past decades (see Zumbo et al., 2015, for a review and application of ecologically based test validation).

Adapting Validity Theory From Educational Measurement Contexts to Writing Assessment Contexts

A major goal of this article is to problematize two common-sense beliefs about writing assessment:

1. In order to assess with validity, we need to ground our practices in traditional, internalist educational measurement concepts (which may run counter to the very conceptualizations of writing at the center of our programs).
2. The accuracy of our assessment scores (however defined) should take precedent over unwanted, unintended consequences (including patterned consequences that affect social-historical groups and group-members) that arise from the decisions we base upon those scores.

Neither of these beliefs is grounded in current measurement/test validity theory, as most writing assessment specialists already realize. However, from my experience in a number of college writing assessment contexts, most non-assessment-specialist WPAs, deans, and other relevant audiences do tend to believe that both of these statements are fundamentally, perhaps unassailably, true. Below, I hope to show specifically where Kane refutes the first belief; then, I argue for certain adaptations to Kane's approach—from a measurement/testing context to a writing assessment context—that could help ground a theory of writing assessment in a rejection of the second belief.

Testing Versus Writing Assessment

As mentioned above, most validity theory is published within educational measurement books and journals, a community which has some diverging interests from the composition community, with our emphases on locally-controlled, site-based writing assessment practices. Here it becomes critical to delineate some differences between *testing* and *writing assessment* as they relate to validity theory and its translation from one context to the other.

4.1.1 Differences in purposes. Within the measurement/testing community, those who research and develop tests tend to be a different population from those who *use* those tests for decision-making purposes, and the distinction between *test developers* and *test users* is quite sharp in the measurement literature. The College Board, for example, develops the SAT not for their own use but so that others can use the scores to facilitate (local) educational decisions. Measurement scholars in university settings similarly work on matters that apply the study of testing and validation in general, not to use the assessments for their own decision-making purposes. But, compositionists who develop writing assessments do tend to use them for local decision-making, often solely for that purpose. This difference in purpose (i.e., developing assessments for local use) will have some key implications for which elements of measurement/test validity theory are most and least relevant to a theory of writing assessment.

Differences in methodology. Methodologically, testing differs from most current writing assessment practices in some key ways. Tests seek to standardize elements of the sample collection (e.g., time conditions, response choices, prompts/questions) in ways that make little sense to current writing assessment. Older (though still widely used) methods of writing assessment, such as timed-essay scoring, sought to blend the standardization of testing conditions with the authenticity/representativeness of directly observing the students' writing. As our assessment theories and practices have evolved, though, our desire to standardize both content (i.e., every student responding to the same prompt) and contexts (i.e., time and place constraints) has been overrun by a concern for *representativeness*—the ways in which the samples of student writing approximate the actual writing practices they do, and will, engage in.

If the SAT were to similarly de-standardize testing conditions, if examinees could take the test at home, were allowed as much time as they wanted, and could even seek help if they wished, the test scores would be almost useless for their intended purposes; the standardization of testing conditions is vital for the interpretation and use of SAT scores to make admissions decisions. But, contemporary writing assessment sensibilities wouldn't consider the fact that one student spends more time on a paper to be a sign that he or she has lower writing abilities—in fact, we might view more time spent on a writing project as a sign that certain course goals (e.g., revision, careful reading) had been met. In fact, I suggest that standardizing the conditions under which students are allowed to write would tell me far *less* about their writing abilities/processes/development that I could otherwise learn, and therefore that whatever decisions I based upon such an assessment would have greater likelihood of being unsound, less valid.

Differences regarding internalism and externalism. Traditional theories of testing and validity are rooted in the measurement of psychological constructs, which can be understood as internalist/psychological attributes that cause a test-taker to receive a certain score (e.g., Borsboom et al., 2004). Within this paradigm, a test-score helps us make inferences about the examinee's possession of a theorized construct, and the possession of that construct is what leads to the assessment-based decision. So, rather than the test/task performance itself justifying a decision, the performance serves as an indicator of an internal construct, and the possession of the construct justifies the decision.

For many, the notion of defining and measuring psychological constructs is foundational to educational assessment and therefore validity, which would render moving beyond such methodologies all but impossible. But Zumbo (2009) pointed out that this approach is not a fundamental given; it was introduced at a specific time for a particular reason. He recounts that, when Cronbach and Meehl (1955) introduced the notion of the construct into validity theory, they did so as a response to 50 years of “psychology's focus on observed behavior and theories of learning, as well as its relatively recent break from psychoanalytic and introspective methods” (Zumbo, 2009, p. 68). The introduction of *construct validity* made it “safe and respectable, again, to talk in the language of unobservables (e.g., constructs)” (Zumbo, 2009, p. 68). This theory of assessment and validity would work nicely with process-era conceptualizations of writing: The *writing process* would be that unobservable, internalist construct, which would explain the writer's performance on a given writing assessment task, and which would justify assessment-based decisions. But what happens when writing is no longer understood as a psychological, internalist attribute of an individual writer? In what follows, I describe how Kane's approach requires no internalist, psychological construct (e.g., Kane, 1992, p. 530; Kane, 2013, p. 21; Kane, 2016, p. 65; see also Newton & Shaw, 2015, pp. 135-141), and I show what he offers instead. Then, I offer specific adaptations to Kane's approach that could more effectively guide the theory and practice of writing assessment—giving matters of social justice a more central role.

Kane's Argument-Based Approach

In Kane's (2013) words, “The argument-based approach is straightforward: state the claims being made and then evaluate the plausibility of these claims” (p. 16). Argument-based validation consists of a two-phase process: The *interpretation/use argument* (IUA) lays out the purposes of the assessment, making an argument for what it plans to achieve and what rationales support those goals; then, the *validity argument* evaluates the IUA, examining “the coherence and completeness of [the IUA] and of the plausibility of its inferences and assumptions” (Kane, 2013, p. 9). Kane's approach is highly adaptable because its rigor comes not from a universal set of criteria to follow, but rather from its one fundamental premise, namely, that all matters of validity and validation stem from the explicitly stated purposes/goals of the assessment at hand, the IUA. For Kane, outside of the IUA, there is simply nothing to validate:

> If someone showed us an unlabeled test booklet and a set of test administration guidelines and asked us to validate the test, what would we do? The first thing I would do would be to ask how the test scores are to be interpreted and used, the population from which the test takers will come, and the contexts in which all this is to happen. With a proposed interpretation and use spelled out, the claims being made can be evaluated. (Kane, 2013, p. 25)

Once the IUA is articulated,

> It is clear where to begin (by specifying the IUA), how to proceed (by evaluating the coherence and completeness of the IUA and the plausibility of its inferences and assumptions), and when to stop (when the inferences and assumptions have been evaluated). (Kane, 2013, p. 9).

As mentioned above, Kane's model can work within a traditional, internalist measurement paradigm, but it doesn't require one. When Kane (1992) first published “An Argument-Based Approach to Validity,” he stated: “The term *argument-based approach* to validity has been used here instead *of construct validity . . .* applying as it does to theoretical constructs as well as to attributes defined in terms of specific content or performance domains” (p. 530), and he has made similar statements in subsequent works (e.g., Kane, 2013, p. 21; Kane, 2016, p. 65). Kane's (2013) discussion of *observable attributes* and their assessment is where he explores this point in the most detail, and where I see the strongest connection between measurement/test validity theory and post-process era, externalist conceptualizations of writing.

Writing as an observable (externalist) attribute. Kane (2013) defined observable attributes (OAs) as “tendencies to perform or behave in some way (e.g., a tendency to act aggressively in some contexts or to solve algebraic equations correctly) under some circumstances” (p. 21), which works well with post-process era conceptualizations of writing in several ways. First, Kane avoided overly reified language, calling OAs *tendencies* and clarifying that they are based on circumstances, similar to ideas of context and situatedness, which are fundamental to current composition sensibilities. Also, Kane (2013) explicitly delinked OAs from underlying, explanatory attributes/constructs, stating, “There is no need to assume that the performances are associated with a single trait or with a specific theory that accounts for, explains, or causes differences in performance” (pp. 21-22). I believe this point lays the groundwork for a theory of assessment that can operate outside of a paradigm of internalism, a theory that could work with current conceptualizations of writing.

Assessing OAs: Performance testing and operational definitions. Kane (2013) dedicated 12 pages (pp. 35-46) to the assessment of OAs. This section is particularly valuable for a theory of writing assessment validity, as distinct from measurement/test validity, though it is rarely discussed in either the writing assessment scholarship or the measurement literature. When assessing OAs, Kane (2013) described two approaches—one which looks more like current writing assessment practices, and one which looks more like traditional testing. One approach, *performance testing*, seeks to observe the OA in a “slice of life” (Kane, Crooks, & Cohen, 1999, p. 12) setting, in which the sample observed is as representative of the real-world act as possible, covering as much of that domain as possible. In short, if you want to assess someone's piano playing, observe their piano playing; if you want to assess students' writing, read their writing. Similar to portfolio assessment, performance testing seeks larger, more representative samples of what students/writers can do in various writing contexts.

The second approach entails operationalizing the OA, breaking it down into pre-specified criteria, “defin[ing] the attribute in terms of a measurement or testing procedure” (p. 24). This approach would operationalize what *writing* means and then test the student on standardized components of that definition—similar to the 1950s and 60s era multiple-choice testing as writing assessment, as well as to many direct writing assessment methodologies (see Huot, 1990) of the succeeding decades.

There are trade-offs for either decision, just as there are trade-offs associated with assessing writing via standardized tests, holistically scored timed-essays, portfolios, or other newer approaches. For one, operational definitions lend themselves to much higher levels of reliability between raters and/or generalizability over replications of the test. But, while performance testing may yield lower reliability/generalizability, it has greater potential for sound *extrapolations* about the test-taker's real-world tendencies, which are vital for assessment-as-decision-making purposes. Kane (2013) referred to this trade-off as the reliability/validity paradox (pp. 30-31), saying:

> If the test observations are highly standardized and very similar in content and format, generalizability is likely to be secure but extrapolation to a broadly defined target domain [“real-world” tasks and performances in non-test contexts (p. 22)] may be relatively shaky. To the extent that we make the test more representative of the target domain, generalizability might decrease but the extrapolation inference would be easier to justify. (p. 36)

My experience working with non-assessment-specialists at various locations (including countless conversations and listserv posts) suggests a prevailing belief that standardization and generalizability/reliability are the hallmarks of valid assessment, and that sound real-world decisions about examinees will simply follow once these matters are in place. But Kane's model does not require prioritizing reliability/generalizability over extrapolations. Instead, because the validity of assessment is centered around the IUA, the purpose of the assessment at hand dictates the methodological approach. An assessment/IUA centered around local decision-making may need to place its focus upon extrapolations to real-world tendencies, and may need to de-emphasize more traditional psychometric concerns for generalizability/reliability if they interfere with those goals. These matters depend upon the specifics of the IUA, not upon prescriptive assumptions about what makes for good assessment.

Thus, a theory of writing assessment validity rooted in Kane's philosophy could help justify methodological approaches that work with compositionists' current understandings of writing, teaching, and learning—and it could help refute assumptions that we are obligated to follow traditional, common-sense assessment principles that work against these sensibilities. What's more, if pushed in certain directions, I believe this approach could allow for concerns about social justice, fairness, equity, and the advancement of opportunity, to become inseparable from the validity of any writing assessment practice.

Pushing Kane's Model Further

Coherence and completeness. Kane (2013) stated some version of this sentence no fewer than six times:

> A proposed interpretation or use can be considered valid to the extent that the IUA is coherent and complete (in the sense that it fully represents the proposed interpretation or use) and its assumptions are either highly plausible *a priori* or are adequately supported by evidence. (Kane, 2013, pp. 2-3)

But, I argue that coherence and completeness should not be enough to validate an IUA; too many lurking unforeseen possible negative consequences might arise from attempts to assess the writing of various populations of students/writers.

The relationship between the consequences of test use and validity is perhaps the most contentious topic in the measurement literature. Many do argue that the consequences of test use should be a fundamental part of validity and validation (e.g., Cronbach, 1988; Haertel, 2013a; Linn, 1997; Messick, 1989; Moss, 1996, 2016; Shepard, 1997; Sireci, 2016), but others believe that validity should refer only to the quality of score-based inferences/interpretations (e.g., Cizek, 2012, 2016; Markus, 1998; Mehrens, 1997; Popham, 1997; Reckase, 1998). But, again, most measurement/test validity theory is written by and for those who develop tests, not those who use tests for decision-making in local contexts. It makes some sense that the measurement community would argue over who bears responsibility for the consequences of test use—the developer or the user—as these are distinct identities in the testing industry. This is precisely why I argue for a theory of writing assessment validity, as distinct from measurement/test validity. Because compositionists both develop and use assessments to make decisions within our specific contexts, narrowing our definition of validity to exclude the consequences of these decisions seems hard to justify. Without investigating questions, like *Did the assessment achieve its desired impact? Did it lead to problematic, unforeseen consequences?*, it would seem impossible to claim our assessments (assessment-based decisions) as valid.

Thus, while Kane (2013) says that validation equals “evaluating the coherence and completeness of the IUA and the plausibility of its inferences and assumptions” (p. 9), if we have every opportunity to check these matters *a posteriori*, in our local context, there seems to be no reason (ethical or methodological) to end validation before we perform such inquiry. No matter how plausible a set of placement decisions (for example) may seem, no matter how complete and coherent the IUA that facilitated the decisions, given that the students are here in our program, we need to investigate whether our decisions are actually yielding the specific profitability (O'Neill, Moore, & Huot, 2009, p. 89) we intended if we are to validate the assessment. For this purpose, I offer an adjustment to Kane's two-phase model, in which the two phases are thought of as a *potentiality argument* and then an *actualization argument.*

The IUA as *potentiality argument.* While I believe that Kane's approach holds the potential to forefront matters of social justice and equity, its current form holds too much room for turning a blind eye toward these matters. But, while nothing explicitly requires concern for disrupting assessment-based social injustices, nothing foundational to Kane's model excludes it. Based on Kane's approach, a theory of writing assessment validity (and a model of assessment validation) could hold that the IUA should include not only the intended consequences of the decisions to be made, but also a statement of potential unintended negative consequences the assessment would seek to actively disrupt (Lederman & Warwick, 2018). If spelled out this way, the validation of that IUA would necessarily entail the empirical, *a posteriori* investigation of both those positive intended consequences and a more active search for negative unintended consequences. I reframe the IUA as a potentiality argument because the concept of potential implies a necessary second step—actualization.

So, the first adaptation I propose is that the IUA/potentiality argument still specifies the goals of the assessment (the decisions to be made and their intended positive consequences), but it would include arguments for the potentiality of the assessment to disrupt cycles of negative social consequences, such as differential impact on social-historical groups and group members, or reproduced systems of access based on such group membership. The second adaptation I propose is reframing Kane's second phase (the validity argument) as an *actualization* argument, which would require empirical evaluation of these matters through multi-method inquiry—of the type that Messick (1989) so strongly called for and that Moss (1996, 1998) so strongly reiterated but which few have ever engaged. Placing concern for these consequences in the IUA/potentiality argument could be a powerful move toward insisting that these less visible consequences of assessment-based decisions be taken seriously in the validation of writing (or any educational) assessment.

As a brief example, consider a program that is implementing a directed self-placement (DSP) method to place students into first-year writing courses. Kane's approach would entail an argument (IUA) for why this method would likely be effective for placing students. This might include reviewing studies from other programs using similar methods, exploring theoretical principles suggesting that DSP should promote learner agency, and so on. The IUA/potentiality argument should be explicit enough to spell out the specific goals this assessment system would seek to achieve, along with the rationales for why it has the potential to do so. My amendment here would be that one consider potentiality of unintended negative consequences in the very IUA/potentiality argument, which, if found, would jeopardize the validity of the assessment. Then, while Kane's approach would end with an evaluation of the completeness and coherence of this IUA and the plausibility of its inferences and assumptions, because all of these assessment-based decisions take place in this local context, I suggest we need an actualization argument to round out the potentiality argument, to see if that potential is being realized. Such inquiry would need to be multi-method; grades and retention rates could tell part of the story, but clearly not the whole thing. Poe et al.'s (2014) use of disparate impact analysis serves as an example of quantitative inquiry that goes well beyond grades and retention rates and into matters of social justice and equity. And, Slomp et al.'s (2014) use of instrumental case study could be a model for a qualitative component, peering into the lived experience of the assessed students. I would also argue that critical inquiry—exploring the deeper problematic social forces that may or may not be at play and/or the various social relations that may be reproduced even through this assessment methodology—would be a necessary component of the actualization argument (see Perry, 2008; Schendel & O'Neill, 1999 for examples of such inquiry).

I don't need to articulate here the power of critical, qualitative inquiry for the purpose of investigating the lived experience of individuals and social-historical group members and/or examining social power arrangements. Many compositionists are already experts in feminist, queer, postcolonial, antiracist, and many other critical, qualitative traditions that delve into these matters. In fact, it is measurement scholars like Moss (1996) and Haertel (2013b) who charged their colleagues with finding collaborators in other disciplines who specialize in these types of methodologies. My goal here is to simply articulate how such methodological approaches could be a fundamental part of validating local writing assessment practices. The proposition of using such multi-method inquiry as a mainstay of validation inquiry is not new; theorists like Messick (1989) and Moss (1996, 1998) have been arguing this for decades. But, despite their pleas, nothing has seemed to keep these matters anywhere but in the distant background.

The way to avoid relegating such inquiry as optional would be to include it in the IUA/potentiality argument, to actively include goals such as disrupting “relations of forces that arranged people in unequal and unfair ways” (Inoue, 2009, p. 103), for that is what would render socially just writing assessment valid or invalid. Once they have been laid out in the IUA/potentiality argument, we have committed ourselves to empirically investigating these matters. These are ambitious goals, to be sure, but a concern for writing assessment and social justice, opportunity, fairness, equity would require such ambition.

Conclusion

The present article explored Kane's argument-based approach to validation as a baseline approach for translating measurement/test validity theory to writing assessment—particularly given our field's conceptualizations of *writing* and interconnections between social justice and writing assessment. I have argued that Kane's (2013) discussion of observable attributes offers an approach toward writing assessment and validation that would allow our current, externalist theories of writing to exist, without being compromised for the sake of assessment.

As many compositionists worry, with good reason, that traditional assessment and validation models show too little concern for ethics, fairness, or social justice, I have argued that Kane's argument-based model contains the flexibility to address these matters.

But only with certain adaptations. Specifically, after following Kane's baseline approach of (a) specifying in argument form (the IUA) what goals a writing assessment seeks to achieve and (b) evaluating the coherence and plausibility of that argument, I have argued that we need to go beyond the matters of evaluating the coherence and plausibility of the IUA, that we must empirically check on the actual impact of the assessment-based decisions in order to find out whether the potentiality of the IUA is being achieved.

I am therefore arguing that even a cost-benefit approach to validation, in which unintended negative social consequences of an assessment operate as a threat against an otherwise valid assessment (e.g., Crooks, Kane, & Cohen, 1996), is itself problematic. Instead, I suggest we need to make the disruption of negative consequence cycles one of the *intended* positive consequences embedded within the IUA; in that way, the failure to disrupt these cycles would not threaten an otherwise valid assessment—it would reveal an assessment that had not met its intended goals, that had not achieved the actualization of its IUA, and which could therefore not yet be considered valid.

Author Note: Josh Lederman teaches teaches first-year writing courses, along with graduate courses in composition pedagogy, at Brandeis University

Acknowledgments: I would like to thank Norbert Elliot, who worked with me through several drafts of this article. I'd also like to thank Mya Poe and Chris Gallagher, who provided tremendous feedback on earlier versions of what would become this article. Many thanks to Michael M. Williamson, as well, who introduced me to validity theory, and encouraged me to continue my work in this direction, over a decade ago.

References

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

Borsboom, D., Cramer, A., Kievit, R., Zand Scholten, A., & Franic, S. (2009). The end of construct validity. In R. Lissitz (Ed.), *The concept of validity* (pp. 135–170). Charlotte, NC: Information Age Publishers.

Borsboom, D., Mellenbergh, G., & van Heerden, J. (2004). The concept of validity. *Psychological Review*, *111*, 1061–1071.

Breuch, L.-A. M. K. (2002). Post-process "pedagogy": A philosophical exercise. *JAC, 22*(1), 119-150.

Cizek, G. J. (2012). Defining and distinguishing validity: Interpretations of score meaning and justifications of test use. *Psychological Methods, 17*(1), 31-43.

Cizek, G. J. (2016). Validating test score meaning and defending test score use: Different aims, different meth. *Assessment in Education: Principles, Policy & Practice, 23*(2), 212-225.

Cizek, G. J., Bowen, D., & Church, K. (2010). Sources of validity evidence for educational and psychological tests: A follow-up study. *Educational and Psychological Measurement, 70*(5), 732-743.

Cizek, G. J., Rosenberg, S., L., & Koons, H. H. (2008). Sources of validity evidence for educational and psychological test. *Educational and Psychological Measurement, 68*, 397-412.

Cooper, M. M. (1986). The ecology of writing. *College English, 48*(4), 364-375.

Cronbach, L. J. (1988). Five perspectives on validity. In H. Wainer & H. I. Braun (Eds.), *Test validity* . Hillsdale, NJ: Erlbaum.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281-302.

Crooks, T. J., Kane, M. T., & Cohen, A. S. (1996). Threats to the valid use of assessment. *Assessment in Education: Principles, Policy, and Practice, 3*(3), 265-286.

Dobrin, S. I. (2010). Through green eyes: Complex visual culture and post��literacy. *Environmental Education Research, 16*(3-4), 265–278.

Haertel, E. (2013a). How is testing supposed to improve schooling? *Measurement: Interdisciplinary Research and Perspectives,* *11*(1-2), 1-18.

Haertel, E. (2013b). Getting the help we need. *Journal of Educational Measurement, 50*(1), 84-90.

Huot, B. (1990). The literature of direct writing assessment: Major concerns and prevailing trends. *The Review of Educational Research, 60*(2), 237-263.

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

Huot, B. (2002). *(Re)articulating writing assessment*. Logan, UT: Utah State Press.

Huot, B., O'Neill, P., & Moore, C. (2010). A usable past for writing assessment. *College English, 72*(5), 495.

Inoue, A. B. (2009). The technology of writing assessment and racial validity. In C. Schreiner (Ed.), *Handbook of research on assessment technologies, methods, and applications in higher education* (pp. 97-120). Hershey, PA: Information Science Reference.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*, 527-535. doi:10.1037/0033-2909.112.3.527

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

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

Kane, M. (2016). Validation strategies: Delineating and validating proposed interpretations and uses of test scores. In S. Lane, M. R. Raymond, & T. M. Haladyna (Eds.), *Handbook of test development* (2nd ed., pp. 64-80). New York: Routledge.

Kane, M. T., Crooks, T., & Cohen, A. (1999). Validating measures of performance. *Educational Measurement: Issues & Practice, 18*(2), 5-17.

Lederman, J., & Warwick, N. (2018). The violence of assessment: Writing assessment, social (in)justice, and the role of validation. In M. Poe, A. B. Inoue, & N. Elliot (Eds.), *Writing assessment, social justice, and the advancement of opportunity* . Boulder, CO: University Press of Colorado; Fort Collins, CO: WAC Clearinghouse.

Linn, R. L. (1997). Evaluating the validity of assessments: The consequences of use. *Educational Measurement: Issues and Practice, 16*(2), 14-16. doi:10.1111/j.1745-3992.1997.tb00587.x

Lotier, K. M. (2016). Around 1986: The externalization of cognition and the emergence of postprocess invention. *College Composition and Communication, 67*(3), 360.

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

Markus, K. A. (1998). Science, measurement, and validity: Is completion of Samuel Messick's synthesis possible? *Social Indicators Research, 45*(1/3), 5-34.

Mehrens, W. A. (1997). The consequences of consequential validity. *Educational Measurement: Issues and Practice, 16*(2), 16-18. doi:10.1111/j.1745-3992.1997.tb00588.x

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: Macmillan.

Moss, P. A. (1996). Enlarging the dialogue in educational measurement: Voices from interpretive research traditions. *Educational Researcher, 25*(1), 20-43. doi:10.3102/0013189X025001020

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

Moss, P. A. (2016). Shifting the focus of validity for test use. *Assessment in Education: Principles, Policy & Practice, 23*(2), 236-

251. doi:10.1080/0969594X.2015.1072085

Moss, P. A., Girard, B. J., & Haniford, L. C. (2006). Chapter 4: Validity in educational assessment. *Review of Research in Education, 30*(1), 109-162. doi:10.3102/0091732x030001109

Olson, G. A. (2002). Why distrust the very goals with which you began? *JAC, 22*(2), 423-428.

O'Neill, P., Moore, C., & Huot, B. (2009). *A guide to college writing assessment*. Logan, UT: Utah State University Press.

Perry, J. (2008). *Institutional cunning: Writing assessment as social reproduction* (Unpublished doctoral dissertation). Kent State University, United States -- Ohio.

Petruzzi, A. (2011). Convalescence from modernity: Writing assessment in the epoch of scientism. *Canadian Journal for Studies in Discourse and Writing, 23*(1), 1-25.

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 & Communication, 65*(4), 588-611.

Poe, M., Inoue, A. B., & Elliot, N. (Eds.). (2018). *Writing assessment, social justice, and the advancement of opportunity.* Boulder, CO: University Press of Colorado; Fort Collins, CO: WAC Clearinghouse.

Popham, W. J. (1997). Consequential validity: Right concern-wrong concept. *Educational Measurement: Issues and Practice, 16*(2), 9-13. doi:10.1111/j.1745-3992.1997.tb00586.x

Popham, W. J. (2010). *Everything school leaders need to know about assessment*. Thousand Oaks, CA: Corwin.

Reckase, M. D. (1998). Consequential validity from the test developer's perspective. *Educational Measurement: Issues and Practice, 17*(2), 13-16. doi:10.1111/j.1745-3992.1998.tb00827.x

Royer, D. J., & Gilles, R. (1998). Directed self-placement: An attitude of orientation. *College Composition and Communication, 50*(1), 54-70.

Sánchez, R. (2012). Outside the text: Retheorizing empiricism and identity. *College English, 74*(3), 234-246.

Schendel, E., & O'Neill, P. (1999). Exploring the theories and consequences of self-assessment through ethical inquiry. *Assessing Writing, 6*(2), 199-227.

Shepard, L. A. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice, 16*(2), 5-24. doi:10.1111/j.1745-3992.1997.tb00585.x

Sireci, S. G. (2016). On the validity of useless tests. *Assessment in Education, 23*(2), 226-235. doi:10.1080/0969594X.2015.1072084

Slomp, D., Corrigan, J. A., & Sugimoto, T. (2014). A framework for using consequential validity evidence in evaluating large-scale writing assessments: A Canadian study. *Research in the Teaching of English, 48*(3), 276-302.

Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology, 69*(5), 797-811. doi:10.1037/0022-3514.69.5.797

Tarsa, R. (2015). Upvoting the exordium: Literacy practices of the digital interface. *College English, 78*(1), 12-33.

Trimbur, J. (1994). Taking the social turn: Teaching writing post-process. *College Composition & Communication, 45*(1), 108-118.

West-Puckett, S. (2016). Making classroom writing assessment more visible, equitable, and portable through digital badging. *College English, 79*(2), 127-151.

White, E. M., Elliot, N., & Peckham, I. (2015). *Very like a whale: The assessment of writing programs*. Boulder: University Press of Colorado.

Williamson, M. (1994). The worship of efficiency: Untangling theoretical and practical considerations in writing assessment. *Assessing Writing, 1*(2), 147-173.

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

Zumbo, B. D. (2009). Validity as contextualized and pragmatic explanation, and its implications for validation practice. In R. W. Lissitz (Ed.), *The concept of validity: Revisions, new directions, and applications* (pp. 65-82). Charlotte, NC: Information Age Publishing.

Zumbo, B. D., Liu, Y., Wu, A. D., Shear, B. R., Astivia, O, L. O., & Ark, T. K. (2015). A methodology for Zumbo's third generation DIF analyses and the ecology of item responding. *Language Assessment Quarterly, 12*(1), 136–151.