Title
Three Interpretative Frameworks: Assessment of English Language Arts-Writing in the Common Core State Standards Initiative

Journal
Journal of Writing Assessment, 8(1)

Authors
Elliot, Norbert
Rupp, Andre A.
Williamson, David M.

Publication Date
2015

Peer reviewed

Three Interpretative Frameworks: Assessment of English Language Arts-Writing in the Common Core State Standards Initiative

by Norbert Elliot, Andre A. Rupp, David M. Williamson

We present three interpretative frameworks by which stakeholders can analyze curricular and assessment decisions related to the Common Core State Standards Initiative in English Language Arts-Writing (CCSSI ELA-W). We pay special attention to the assessment efforts of the Smarter Balanced Assessment Consortium (Smarter Balanced) and the Partnership for Assessment of Readiness for College and Careers (PARCC). Informed by recent work in educational measurement and writing assessment communities, the first framework is a multidisciplinary conceptual analysis of the targeted constructs in the CCSSI ELA-W and their potential measurement. The second framework is provided by the Standards for Educational and Psychological Testing (2014) with a primary focus on foundational principles of validity, reliability/precision, and fairness. The third framework is evidence-centered design (ECD), a principled design approach that supports coherent evidentiary assessment arguments. We first illustrate how Standards-based validity arguments and ECD practices have been integrated into assessment work for the CCSSI ELA-W using Smarter Balanced and PARCC assessment reports. We then demonstrate how all three frameworks provide complementary perspectives that can help stakeholders ask principled questions of score interpretation and use.

Three Interpretative Frameworks:

Assessment of English Language Arts in the Common Core State Standards Initiative

By the end of the nineteenth century in the United States, demand for universal public education had become equated with assurance of participatory democracy. In 1869-1870, 7.48 million students enrolled in kindergarten and grades one through eight. By 1899-1900, that number had risen to 14.98 million. This increase was accompanied by a dramatic rise in high school enrollment as advanced education became necessary for better paying jobs. In 1869-1870, 80,000 students were enrolled in grades nine through twelve. In 1899-1900, that number had risen to 519,000 (Snyder, 1993, p. 34, Table 8).

Accompanying this new influx of students were those who believed they knew best how to shape the curriculum. Archetypal responses—the humanism of Charles W. Eliot (1892), the developmentalism of G. Stanley Hall (1883), the social efficiency of Joseph Mayer Rice (1893), and the social meliorism of Lester Frank Ward (1883)—were to continue throughout the twentieth century (Kliebard, 2004). Today, one may identify these enduring themes in the calls for equity by Diane Ravitch (2010), the cognitive modeling of Howard Gardner (2006), the emphasis on effective teaching by Bill and Melinda Gates (2015), and the progressivist agenda of Arne Duncan (2015).

With enrollment projections for the school year 2015-2016 estimated at 49.8 million public elementary and secondary school students (Snyder & Dillow, 2015, p. 86, Table 203.10), these and other voices emerge to give council on how best to spend a projected education budget of no less than $669 billion (Snyder & Dillow, 2015, p. 58, Table 106.10). There is a loud roar of voices accompanying initiatives associated with the term “educational reform,” which has become nearly deafening as the national debate has turned to the Common Core State Standards Initiative (CCSSI) and associated state-led curricular guidelines for a national school curriculum assessed by two consortia: the Smarter Balanced Assessment Consortium (Smarter Balanced) and the Partnership for Assessment of Readiness for College and Careers (PARCC).

As the most comprehensive effort in American history to leverage uniform goal-based instruction, the CCSSI is designed to ensure that high school graduates are prepared to take credit-bearing courses in two- or four-year college programs or enter the workforce. At the present writing, forty-two states, the District of Columbia, four territories, and the Department of Defense Education Activity have adopted the CCSSI. Assessments in English language arts and mathematics have taken place in the 2014-2015 school year, and preliminary results are being released at the time of this writing.

The development of the CCSSI and its assessment has been accompanied by three categories of criticism: warnings of the dangers of neoliberalism; concerns over the constraint of the writing construct; and fears that the achievement of equity continues to elude educational reform. From their creation (in order to enhance global competitiveness and workplace success) to their solicitation (in order to encourage proposals for next generation assessment systems), the CCSSI have been informed by “a form of cultural politics and a set of economic principles, policies, and practices devoted to handing over as much of social life as possible to private interests” (Gallagher, 2011, p. 453).

Referencing this depiction of neoliberalism, Wilson has been critical of the ways that such framing has diminished teacher agency (Shannon, Whitney, & Wilson, 2014). In interacting with students and teachers, she argued, “you see what matters, and you realize that these grand plans that Bill Gates has for how it is that we’re going to improve education just don’t make any sense” (p. 299). In similar fashion, Addison and McGee (2015) warned that the role of the Gates Foundation compromises local efforts such as those sponsored by the National Writing Project, “to gain compliance” with the CCSSI (p. 215). Concentrating on the limits of construct

representation following from the neoliberal policy climate, Kristine Johnson (2015) found curricula based on the CCSSI "would focus almost exclusively on expository/informational and fact-based argumentative writing, with some narrative descriptive writing"—a "narrowing effect" that diminishes coverage of the writing construct (p. 520). Applebee (2013) has also identified this narrowing effect in his identification of four areas—separate emphasis on foundational skills, grade-by-grade standards, absence of a developmental writing model, and implementation issues—with "equal potential to distort curriculum" (p. 28).

While public debate swirls around societal impact, often absent are voices of stakeholder groups directly involved with students: parents and guardians; teachers and administrators; legislators; and workforce leaders. It is our aim in this paper to suggest directions of inquiry for those stakeholder groups. Specifically, we seek to empower these stakeholder groups by discussing how a deeper understanding of the traditions, terminologies, and best practices of educational measurement and writing assessment provide an excellent way to ask critical questions about new curriculum and assessment initiatives.

Such strategies are needed to navigate a maze of complex debates in which everything and its opposite both appear to be true. As researchers in writing assessment (Elliot), cognitively-grounded diagnostic measurement (Rupp), as well as automated scoring and modern psychometrics (Williamson), we are positioned to enter the controversial roar in a very precise way.

While we acknowledge and honor the ontological and axiological force of voices interested in the social dimension of assessment, we focus in this paper on structuring discussions around significant technical issues in assessment design and use for the CCSSI in English Language Arts-Writing (CCSSI ELA-W). These issues are discussed through the lens of three interpretative frameworks that provide complementary perspectives and ways of thinking about key issues for stakeholders: multidisciplinary research on writing (e.g., Elliot & Perelman, 2012), the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014), and evidence-centered design (ECD) (e.g., Mislevy, Steinberg, Almond, & Lukas, 2006).

While we are certainly encouraging readers to consider the different perspectives we present through these frameworks, our discourse modes are primarily expositive, descriptive, and narrative. That is, we do not seek to criticize the CCSSI or the work of Smarter Balanced and PARCC in any absolute or relative terms; rather, we want to illustrate how the three interpretative frameworks provide conceptual scaffolds for asking critical questions that lead to enriched discussions among stakeholders. We believe that such discussions—and the associated heightened awareness of the complexities of many curricular and assessment design decisions—can help the diverse communities affected by the CCSSI gain a stronger appreciation for the relative strengths and weaknesses of various political, instructional, and assessment efforts.

Interpretative Framework 1: Multidisciplinary Research on Writing

Part of the discipline of education, the field of educational measurement finds its origin in 1892 with the founding of the American Psychological Association (Fernberger, 1932) and the subsequent 1945 designation of Division 5, Evaluation and Measurement (Benjamin, 1997). Part of the discipline of English language and literature, the field of writing assessment finds its origin with the founding of the National Council of Teachers of English in 1911 (Lindemann, 2010) and the 2010 designation of Rhetoric and Composition/Writing Studies as its own specialized field (Phelps & Ackerman, 2010).

Recent multidisciplinary research between educational measurement and writing assessment has addressed the present landscape of writing assessment, as well as methodology, consequence, and future directions for the field (Elliot & Perelman, 2012). Clearly, the two fields have begun to influence each other; the acknowledgement of mutually beneficial research agendas, for instance, has resulted in recommendations for next-generation assessments to focus on social and rhetorical knowledge, domain knowledge and conceptual strategies, writing processes, and knowledge of conventions (Sparks, Song, Brantley, & Liu, 2014). Such a multidisciplinary perspective provides a way to frame the CCSSI assessment of ELA-W in terms of reflective attention to definitions and measurement of the writing construct.

Construct Definition

A construct such as writing, which is the core focus of the definition and empirical representation of models of student competence for CCSSI ELA-W assessment, is generally defined rather broadly. Its description, however, should be as concrete, comprehensive, and systemic as possible to be useful for instructional guidance and assessment development. The operationalization of the way the construct is measured through assessment tasks and their associated scoring rules is a great leverage point for obtaining clarity about the boundaries of the construct definition as targeted in an assessment.

Beginning with the protocol analyses of Flower and Hayes (1981), writing has been understood as a complex process in which readers and writers construct meaning through detailed, often internal, cognitive iterations concerning variables such as discourse conventions, social context, language, purpose, and knowledge. In negotiating meaning, writers create "webs of intention, carrying out complex, individual, and socially bounded purposes, shaped by attitudes and feelings, and other people" (Flower, 1994, p. 54). In recent iterations of the model, attention has been drawn to the importance of source-based investigation, the design of visual content, and management of attention and motivation (Hayes, 2012; Leijten, Van Waes, Schriver, & Hayes, 2014). As evidence of

their enduring presence, Beringer (2012) has documented the origin, traditions, and future directions of cognitive perspectives on writing research. Based on construct models derived from these perspectives, Dean and his colleagues (2015) have recently developed a key practice framework linking ECD, scenario-based assessment, and cognitively-based assessment in order to create English Language Arts task sequences that support both instruction and assessment. Social cognitive models are understood to yield high quality, specific information about both the writing construct and its boundaries.

Informed by models of social cognition, the CCSSI ELA-W is designed to specify performance-level objectives—knowledge descriptions that can be mapped to grade levels. By these strategies, the CCSSI ELA-W models writing from kindergarten through grade 12. That is, in the CCSSI ELA-W, the construct is defined in actionable terms: "Students should demonstrate increasing sophistication in all aspects of language use, from vocabulary and syntax to the development and organization of ideas, and they should address increasingly demanding content and sources" (CCSSI, 2015c). By extension, writing is also viewed as part of the broader construct of ELA:

> The Common Core asks students to read stories and literature, as well as more complex texts that provide facts and background knowledge in areas such as science and social studies. Students will be challenged and asked questions that push them to refer back to what they've read. This stresses critical-thinking, problem-solving, and analytical skills that are required for success in college, career, and life. (CCSSI, 2015a)
>
> As a blend of both reading and writing, this definition advances a conception of language arts that envisions writing and reading as integrated constructs.

In turn, this blended, integrated construct is then rendered specific within grade levels across kindergarten through grade 12. For example, the standards for grades 11 and 12 are further defined in terms of the following conceptual anchors: text types and purposes (to "write arguments to support claims in an analysis of substantive topics or texts, using valid reasoning and relevant and sufficient evidence"); production and distribution of writing (to "produce clear and coherent writing in which the development, organization, and style are appropriate to task, purpose, and audience"); research to build on present knowledge (to "conduct short as well as more sustained research projects to answer a question [including a self-generated question] or solve a problem; narrow or broaden the inquiry when appropriate; synthesize multiple sources on the subject, demonstrating understanding of the subject under investigation"); and range of writing (to "write routinely over extended time frames [time for research, reflection, and revision] and shorter time frames [a single sitting or a day or two] for a range of tasks, purposes, and audiences") (CCSSI, 2015b).

Construct Measurement

While the CCSSI ELA-W is research-based, it is important to understand that the conceptual model—the way the elements of writing are understood in their relationship to each other within the given construct—was based on consensus opinion. Distinct from construct definitions based on evidence from reflective latent variable models (Graham, McKeown, Kiuhara, & Harris, 2012; Graham & Perin, 2007; Hillocks, 1986; Rogers & Graham, 2008), this consensus definition is, in reality, a "stew" of elements that might or might not be empirically related to each other (National Research Council, 2012). Put differently, as a consensus model, the development and instantiation of the CCSSI has, so far, been a state-led effort based on adoption, not on data collection. The means of assessing students and the information resulting from that assessment are left to the discretion of the states as an activity distinct from the CCSSI—a very complex task for individual states and collections of states.

The era of modern assessment has arguably been characterized by a focus on creating writing tasks that are closely aligned with modern views of writing from expert communities. In fact, without this involvement of the writing community it would be difficult to imagine how this new generation of assessment would be different than the print-born bubble and booklet tests of the past. This involvement has led to the use of digitally-delivered stand-alone writing tasks and the embedding of writing activities in domain or profession-specific complex performance tasks (Tucker, 2009). Designed to capture blended constructs, integrated tasks incorporating content from source materials offer benefits such as providing realistic, challenging activities, engaging students in writing responsible to specific content, obviating practice effects associated with conventional item types, evaluating language abilities consistent with integrated models of literacy, and offering diagnostic value for instruction or self-assessment.

Challenges nevertheless remain. Cumming (2013) has noted that integrated writing tasks have associated risks. These include confounding measurement of writing ability with abilities to comprehend source materials, merging assessment and diagnostic information together in ineffective ways, and invoking genres that are emerging and therefore difficult to score. As we discuss below, navigating the complex system of tradeoffs when designing individual assessments and systems of assessments over time for CCSSI ELA-W can be substantially facilitated, integrated, and scrutinized using the *Standards* and ECD frameworks as guidance.

Interpretative Framework 2: *Standards for Educational and Psychological Testing*

Recently revised, the *Standards* and their adaptations by testing companies (e.g., Educational Testing Service, 2015) can be seen

as cohesive interpretative frameworks that lend focus to assessment design. Use of standards-based reasoning results in logical approaches to evidence in light of desired arguments about individual test-takers, test-taker groups, and the assessments themselves.

A consensus statement of its own, the *Standards* (2014) are intended "to provide criteria for the development and evaluation of tests and testing practices and to provide guidelines for assessing the validity of interpretations of test scores for the intended test uses" (p. 1). A consensus statement of its own, the *Standards* (2014) are intended "to provide criteria for the development and evaluation of tests and testing practices and to provide guidelines for assessing the validity of interpretations of test scores for the intended test uses" (p.1). However, while *Standards* are designed for raising awareness and guiding decision-making about assessment systems. However, while *Standards* are designed for raising awareness and guiding decision-making about assessment systems at a high conceptual level, the document is not designed to be step-by-step instructions of how to do the necessary work on a day-to-day basis. That role falls to principled assessment design frameworks like ECD, which we discuss in the next section.

Calls for increased assessment literacy such as those found in the *Standards* (see pp. 192-193) are not incidental to our purpose in this paper. Any fixed set of curricular approaches or assessment methods yields particular kind of interpretation and any such methodological exclusivity is inappropriate when dealing with complex assessments such as the CCSSI-ELA-W. In fact, assessment of the CCSSI-ELA-W is designed to generate the kinds of evidence needed to validate multiple proposed interpretations and uses.

While the present version of the *Standards* is our concern here, the 4th revision (1999) was the common referential point for both the Smarter Balanced and PARCC consortia. Indeed, the five sources of validity evidence identified by Sireci (2012) in his report of the Smarter Balanced research agenda—a report to which we will turn later in order to establish the informed view of validity used to support score interpretation and use (Kane, 2013, 2015) in the design of the CCSSI ELA-W assessment—are taken directly from the 1999 version. The *Standards* have played, and will continue to play, a significant role in the development of assessments related to the CCSSI.

In their present form, the *Standards* are divided into three sections: foundations, operations, and applications. By far, the foundations section is the most significant in terms of assessment of the CCSSI ELA-W. It is here we find extended discussion of the three overarching principles of validity, reliability/precision, and fairness. Because these foundational concepts deeply inform Smarter Balanced and PARCC assessment designs, a brief definition and discussion of each is warranted. Nevertheless, the concepts are not intended to be separated; rather, validity, reliability/precision, and fairness are intended to be used in support of proposed interpretation and use of scores associated with the CCSSI ELA-W assessment.

Validity

In the *Standards*, validity is defined as the "degree to which accumulated evidence and theory support a specific interpretation of test scores for a given use of a test. If multiple interpretations of a test score for different uses are intended, validity evidence of each interpretation is needed" (p. 225). Although still considered by many as an "up-or-down vote" or a simple "stamp of approval," the 2014 edition is clear on the imprecision of such summary judgment: "Statements about validity should refer to particular interpretations and consequent uses. It is incorrect to use the unqualified phrase 'the validity of the test'" (p. 23).

While the origin of this characterization of validity may be found in the 1985 edition of the *Standards*, it is important to reflect on just how enduring the work of Messick (1989) has become in his characterization of validity as "an integrated evaluative judgement of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of *inferences* and actions based on test scores or other modes of assessment" (p. 13, emphasis in original). Equally important is the work of Kane (2013) and his call for evidence-based interpretation and use arguments: "To validate an interpretation or use of test scores is to evaluate the plausibility of the claims based on the test scores" (p.1).

Validation therefore requires a clear statement of the claims inherent in the proposed interpretations and uses of the test scores. "Public claims require public justification" (Kane, 2013, p. 1). Influential in the development of the *Standards* and their manifestation in the assessment of the CCSSI ELA-W, Kane (2015) has offered a two-step approach to validation:

> First, the interpretation and use is specified as an interpretation/use argument, which specifies the network of inferences and assumptions leading from test performances to conclusions and decisions based on the test scores. Second, the interpretation/use argument is critically evaluated by a *validity argument*. (p. 4, emphasis in original). As a result of this orientation, validity becomes a property of score interpretations—not as a property of the assessment: "Once we adopt an interpretation, it can make sense to talk about 'the validity of a test', but the 'validity' is relative to that interpretation" (Kane, 2015, p. 2).

This “flexible framework for validation,” as Kane terms it, is important in that it allows for—indeed, encourages—multiple interpretations that may arise from multiple groups. As Kane concludes, “[T]o restrict our conception of validity to one kind of interpretation seems unnecessary and would greatly limit our ability to respond to the varied applications of test scores” (2015, p. 3).

Reliability/Precision

Reliability/precision is defined as

> the degree to which test scores of a group of test takers are consistent over repeated applications of a measurement procedure and hence are inferred to be dependable and consistent for an individual test taker; the degree to which scores are free of random errors of measurement for a given group. (AERA, APA, & NCME, 2014, pp. 222-223)

In other words, the empirical quantification of reliability requires the existence of replication of assessment across conditions that are comparable (e.g., test forms, administration conditions, subsets of items, and sets of raters).

Once seen strictly as quantifiable by the familiar statistical coefficient of classical test theory, reliability was re-conceptualized by Lord (1980) through a more complex mathematical model for the relationships among test item performance, item characteristics, and test taker proficiency with respect to the construct(s) under examination. This framework is known in the educational measurement literature as item response theory (IRT) (e.g., de Ayala, 2009; de Boeck & Wilson, 2004) and is the most commonly applied framework for large-scale assessment apart from classical test theory. IRT can accommodate reporting on single and multiple dimensions, the existence of nested data structures (e.g., students nested in schools nested in districts), and the inclusion of variables to explain performance differences for test-takers and tasks. It can be effectively used to create large banks of tasks that can be used for adaptive assessment systems and the efficient delivery of comparable assessments with varying composition for international, national, and state-wide survey purposes.

As is the case with validity, misunderstanding about reliability abounds. For example, still considered by many as the equivalent of the railroad standard gauge, the value of 0.7 for a single reliability coefficient such as internal consistency, inter-reader agreement, or cross-administration score correlation often appears to be the sole level of attainment in the hearts and minds of many. However, with frameworks like IRT the notion of precision of measurement can be assessed more finely at different points of the reporting scale, which is important for optimizing pass-fail decisions or test assembly in high-volume testing contexts.

Consequently, the authors of the *Standards* do their best to dispel such reductionism and offer general guidelines that allow for the proper use of modern measurement approaches for capturing evidence about reliability/precision, validity, and fairness. To this end, the authors of the *Standards* also underscore that reliability and validity must be considered in conjunction with fairness considerations. For example, while the need for precision at some points of the scale increases as the consequence of score use increase, the authors acknowledge that the sacrifices in reliability/precision that may result from using performance-based writing tasks instead of multiple choice items may, in fact, be acceptable. Despite being more costly to score, these tasks may reduce construct-irrelevant variance (difference in scores attributable to elements extraneous to the test) and/or diminish construct underrepresentation (failure to tap significant aspects of the construct that the assessment is designed to measure), which lessen the validity of the intended interpretation/use argument and its critical evaluation by the validity argument.

Fairness

In the *Standards* fairness is defined as

> the validity of test score interpretations for intended use(s) for individuals from all relevant subgroups. A test that is fair minimizes construct-irrelevant variance associated with individual characteristics and testing contexts that otherwise would compromise the validity of scores for some individuals. (p. 219).

This section of the *Standards* has been expanded substantially over previous revisions, with emphasis given to fairness for all examinees. Again, we see the presence of Messick (1989) who linked forms of validity with consequences related to score use—an emphasis that has been maintained by Kane (2006, 2013).

Significantly, special attention is given in the *Standards* to the opportunity to learn—“the extent to which individuals have had exposure to instruction or knowledge that affords them the opportunity to learn the content and skills targeted by the test” (p. 56). In an analysis consistent with this emphasis on exposure, Pullin (2008) has highlighted connections among assessment, equity, and opportunity to learn, as both a reflection of the learning environment and a concept demanding articulated connections between the assessment and the instructional environment. Such characterizations afford identification and removal of barriers to valid score interpretation for the widest possible range of individuals and subgroups, interpretative validity for examined populations, and the development of suitable testing accommodations and safeguards to protect fair score usage.

Equally associated with fairness—and of special interest in terms of equity to all stakeholders—is adherence to the principles of universal design.An approach to assessment that strives to minimize construct distortion and maximize fairness through uniform access for all intended examinees, universal design has been identified in the *Standards* (2014) as a way to leverage fairness for all examinees (p. 63). As Ketterlin-Geller (2008) has established, when student characteristics are considered during the conceptualization, design, and implementation phase of test development under principles of universal design (e.g., specifying content and cognitive complexity in the test blueprint, as well as information about the target and access skills), test performance of students with special needs is more likely to reflect their construct knowledge. Furthermore, Mislevy et al. (2013) has demonstrated that a combination of ECD and universal design results in an increased sense of fairness as construct-irrelevant barriers to student success are proactively removed in comprehensive efforts to provide all students with an opportunity to perform at their best during assessment episodes.

Operations and Applications

As the authors of the *Standards* wrote, test design "begins with considerations of expected interpretations for intended uses of the scores to be generated by the test" and therefore "test design and development procedures must support the validity of the interpretations of test scores for their intended uses" (p. 75). The influence of Kane is again palpable. Issues related to validity, reliability/precision, and fairness are thus interwoven into the development process from the creation of test specifications to the copyright responsibilities of test users; as we will see, this perspective is embodied by the ECD framework that we discuss in the next section.

While the foundations discussed in the first three sections of the *Standards* are essential for understanding and navigating the complex decision-making space surrounding assessments, additional guidance is needed to put these articulated principles into practice. In the assessment operations section of the *Standards*, chapters are devoted to test design and development processes that lead to reported scores, scales, and norms as well as processes for score linking (processes used to facilitate score comparisons) and cut score setting (processes used to divide scores in order to act upon them). The authors also included chapters on test administration; scoring, reporting, and interpretation; supporting documentation for tests; the rights and responsibilities of test takers; and the rights and responsibilities of test users. The final section of the *Standards* is devoted to testing applications. Attention is given to psychological, workplace, and educational assessment, as well as the role of tests in program evaluation, policy studies, and accountability.

Interpretative Framework 3: Evidence-Centered Design (ECD)

As the discussions in the previous *Standards* section have made abundantly clear, to build an evidentiary argument for assessment scores so that intended interpretations and decisions comply with the *Standards* is a complex process. This complex process is exemplified in the CCSSI assessment aim as it is identified by Smarter Balanced: "The assessment system being developed by the Consortium is designed to provide comprehensive information about student achievement that can be used to improve instruction and provide extensive professional development for teachers" (Sireci, 2012, p. 4). As such, "the assessment system focuses on the need to strongly align curriculum, instruction, and assessment, in a way that provides valuable information to support educational accountability initiatives" (p. 4).To help facilitate the construction of arguments supporting such aims and to imbue the assessment ecosystem with appropriate characteristics that support intended interpretations and decisions, a principled design framework for practice such as ECD is needed. Proposed to make explicit the evidentiary reasoning process of assessment interpretation and decision-making, ECD helps organize assessment practices in ways that yield cohesive integrated thinking about assessment aims, delivery capability, and justification of score use. As such, ECD can be viewed as providing the "evidentiary grammar" for evidence-based assessment arguments.

At its best, ECD is a powerful professional development tool that can help interdisciplinary teams of experts (e.g., assessment developers, statisticians, information technology specialists, policy-makers, and other stakeholders) develop common language, mental models, design artifacts, and best practices. In addition, it can help such teams utilize these capacities to develop targeted artifacts that move the assessment process forward in ways that best capture the connected thinking underlying the design process. These goals are always laudable and important, of course, but become especially important as the assessments become more performance-oriented, more reliant on models of social cognition, more responsive to correlates such as engagement or motivation, and more situated within community practices. In short, ECD is highly relevant for task-based CCSSI assessments of ELA-W.

Mislevy, Sternberg, and Almond (2003) identified five core structural/conceptual elements for ECD and arrange them in what they term the conceptual assessment framework: student models that characterize knowledge and skill; task models that provide constructed response test items to elicit student knowledge and skills; evidence models that provide a chain of inferential reasoning from student test performance to knowledge and skill, with emphasis on scores and their measurement; assembly models that specify how individual tasks are combined to produce the final assessment; and presentation models that specify how individual tasks are administered to students. In practice, spelling out these different models means creating artifacts such as databases, spreadsheets, and text files to document the key decisions that underlie the reasoning process.

Thus, a second layer in the day-to-day practice of assessment development is putting the decisions captured in these artifacts into practice by setting up a delivery, scoring, and reporting architecture, which Mislevy, Sternberg, and Almond described as a four-process model of activity selection (the process of selecting and sequencing assessment tasks), presentation (the process of presenting the assessment task to the student), response processing (the process that evaluates the essential features of the student response to the task), and summary scoring (the process that produces inferences about student ability based on evidence accumulated across the task). Each of these processes emanates from an understanding of the domain that inferences are tied to and the processes of analyzing and modeling the domain tasks for assessment development purposes (Almond, Steinberg, & Mislevy, 2002).

As noted above, ECD is a framework or mechanism for making explicit the evidence-based reasoning practices of interdisciplinary teams charged with assessment design, delivery, scoring, and reporting. At a fine-grained technical level the decomposition of the argumentation is based on Toulmin's argument schema (1958/2003), which is well known to the writing assessment community (White, Elliot, & Peckham, 2015, Figure 3.5) and the educational measurement community (Mislevy, 2007, Figure 1). Moreover, Bachman (2005) extended the Toulmin diagram/argument from assessment interpretations to assessment decisions. Recent scholarship has elaborated on the Toulmin model as a way to formalize three credentials of an evidential datum—relevance, credibility, and inferential force—that must be established in analyzing its relationship to a hypothesis (Anderson, Schum, & Twining, 2005). As the Toulmin model reveals, evidence, warrants, claims, and qualifications are important in establishing the two aspects of overarching validation arguments proposed by Kane (2006, 2013, 2015) noted above: an interpretive argument, which documents the network of inferences and assumptions leading from the performance to the conclusions and decisions on use; and the validity argument, which serves as a check on the interpretative argument by evaluating its plausibility. As Mislevy (2007) has observed, the Toulmin model serves an important function, which is to render the validity argument "public, sharable, and reusable" (p. 437).

For CCSSI ELA-W assessments, the validity argument is used as a vehicle to articulate the characteristics and boundaries of a designated construct. In the next section we describe how the *Standards* and the ECD framework have been instrumental in the development of curricular and assessment efforts surrounding the CCSSI ELA-W.

***Standards*-based Validity Arguments and ECD Practices:**

Integration into CCSSI ELA-W Assessment

In this section we use three Smarter Balanced and PARCC assessment reports to illustrate how *Standards*-based validity arguments and ECD practices have been integrated into assessment work for the CCSSI ELA-W.

Consider first the report entitled "Smarter Balanced Assessment Consortium: Comprehensive Research Agenda" (Sireci, 2012). The author's detailed validity argument is intended to "put potential misperceptions to rest" that the Consortium has adopted a research agenda that has unfortunately resulted in fragmentation (p. 63). To counterbalance these claims, Sireci advanced seven principles, or claims, of the assessments: that they are grounded in a standards-based curriculum and are part of an integrated system; that they produce evidence of student performance; that they are part of a state-led effort with a transparent and inclusive governance structure; that they are structured to continuously improve teaching and learning; that they provide useful information on multiple measures educative for all stakeholders; that their implementation strategies adhere to established professional standards; and that teachers have been integrally involved in the development and scoring of the assessments.

The claims are then followed by two tables: one providing the details of 55 studies proposed by the Consortium; and the other providing a way to map the studies to the five sources of evidence—validity based on test content, internal structure, response processes, relationships to other variables, and consequence—identified by the consortium. Explicitly and by name, the report utilizes ECD as a way to evaluate the degree to which the assessment specifications represent the CCSSI and the degree to which the constructed response items themselves capture the assessment specifications (p. 25).

Second, consider the "Memorandum on Instructional Sensitivity Considerations for the PARCC Assessments" (Way, 2014). The author uses a validity argument to map a research agenda of the instructional sensitivity of the assessments, defined as the extent to which a test item is sensitive to instruction. Rather than viewing instructional sensitivity as an isolated concept, Way proposed that it is "tied up with related concepts governing what is supposed to be taught in the classroom, what is actually taught in the classroom, and how well tests and items align with what is taught" (p. 3).

Way noted that while the PARCC assessments are designed to measure integrated stills (such as those that require evaluation, synthesis, analysis, reflective thought, and research), this particular type of integration might not be taught in a given school year. As such, the assessments could possibly become tests in search of a curriculum. To address this dilemma, Way proposes the use of IRT plots as predictors based on ability level, as well as classroom observations and teacher reports of classroom content. Framing a research agenda in anticipation of validity argument used to establish assessment and curricular connections suggests the centrality of evidentiary reasoning throughout the CCSSI design process.

Finally, consider the PARCC report "Evidence and Design Implications Required to Support Comparability Claims" (Luecht & Camara, 2011). In it, the authors have paid close attention to score use—to the ways to compare student performance across schools, districts and states, to measure growth across grade levels, and to evaluate year-to-year changes. Because of the importance of such comparisons and goal setting, the authors emphasized the need for "well-articulated, cognitively-based constructs" based on the CCSSI, which should be developed in order to establish the ordered claims and evidence requirements by grade level.

Luecht and Camara noted that the ECD approach "may offer some advantages over conventional item design and test specifications because such new design approaches prioritize more explicit connections between items from task models which are directly derived from evidence" (p. 15). Task models resulting from ECD, as the report acknowledges, allow designers to control for content through an emphasis on cognitive demand and yield greater efficiency in development of the assessment over time.

As these three examples demonstrate, strategic use of *Standards*-based and ECD frameworks at the planning stage yields a validity agenda and evidentiary processes. In the next section, we provide some guiding questions for stakeholder networks that can help to raise awareness about what it means to translate the different concepts in the *Standards* and ECD into thoughtful assessment practice that supports meaningful interpretations and decisions.

Guiding Questions for Stakeholders

In this section we turn to four key stakeholder groups—students and guardians, teachers and administrators, legislators, and workforce leaders—and provide questions intended to empower each to grapple with the decisions that must be made as a result of information issuing from the three interpretative frameworks discussed above. It is our belief that these stakeholders would be well served by raising a series of such very specific questions that can lead to informed judgments regarding score use stemming from the assessment of the CCSSI ELA-W by Smarter Balanced and PARCC. Made on a state-by-state basis this judgment will, we argue, be best made if informed by the perspective gained when key stakeholders think along the same lines.

More broadly, the perspective offered by these questions is commensurate with comprehensive validation arguments and coherent evidentiary reasoning practices embodied in the *Standards* and ECD, respectively. It is therefore appropriate to think of the questions raised in Tables 1 and 2 as applicable to any large-scale assessment of ELA-W that has been created under the contemporary evidentiary reasoning practices presented in this paper. As evidence of the force of multidisciplinary research, we note that our perspective is congruent with the emphasis on networks and their logic proposed by Gallagher (2011); that is, the questions we provide are intended to provide "analytic tools for understanding how actors exercise power by virtue of their *locations* and *relations*" (p. 466, emphasis in original).

Heuristics and Bias

We have informed our questions by the heuristics and biases research of Amos Tversky and Daniel Kahneman. Together, these scholars in the field of decision-science advanced a program of research since the early 1970s that revolutionized our understanding of human judgment (Kahneman, 1973; Tversky & Kahneman, 1973). Their system is too complex for discussion save its core concept: attention to the heuristics that we use to ask questions and the cognitive biases that result in tangled reasoning. Defined as "a simple procedure that helps find adequate, though often imperfect, answers to difficult questions," Kahneman (2011, p. 98) had found that heuristics are a consequence of intuition (termed System 1 thinking) and strategy (the corrective System 2). While we think associatively, metaphorically, and causally with some ease and accuracy as a result of intuition, he noted, even the most educated have trouble thinking about more abstract concepts like probabilities and uncertainties to make appropriate strategic inferences.

Complexities that arise from the overestimation of what we know and the underestimation of chance are potentially important for two reasons in educational assessment and measurement. First, as we have demonstrated in our three interpretative frameworks, modern assessment requires that we embrace evaluative techniques as complex as the humans we seek to learn about. In this process, meaningful and informed questions are of paramount importance lest we underestimate the demands of assessment. Just below the surface, foundational concepts are associated with probabilities, and the nuanced nature of the evidence produced from modern assessment systems requires acknowledgement of contingency. Second, while we are experientially familiar with the forms of logic that assessment designers use in test design, we know less about the forms of logic that the stakeholders use to make interpretations and decisions based on assessment scores. The more we can learn about the logic of stakeholder networks, the better we will be able to communicate our evidentiary processes.

In the absence of such information, the questions in Table 1 and Table 2 are intended to help networks of non-specialists structure conversations that may, in turn, help specialists learn more about the cares and concerns of all stakeholders. The guiding questions are designed to help uncover implicit assumptions, potential biases in reasoning, and connections between various design decisions within the teaching and assessment ecosystem. We deeply believe that it is of value to connect the logic of educational

measurement and writing studies research with the logic of heuristics and biases research, if only to remind everyone that complex ventures obligate us to think in complex ways.

In each table, we have used the *Standards* to generate a series of broad foundational and operational questions that, in turn, are made specific by focusing on specific facets of measurement. Because our focus is on an educational assessment, we have integrated that application into the foundational and operational question and, hence, no additional table is provided for that section of the *Standards*.

Table 1. *Foundational Questions for Stakeholder Groups in English Language Arts-Writing.*

Standard	Students and Guardians	Teachers and Administrators	Legislators	Workforce Leaders
Validity: "Clear articulation of each intended test score should be set forth, and appropriate validity evidence in support of each intended interpretation should be provided" (p. 23).	• How will scores be used? ➢ Will scores be used to draw conclusions about an individual student's present writing ability? ➢ Will scores be used to make decisions about an individual student's ability to perform in subsequent courses?	• Has validity evidence been provided that will allow interpretation of test scores for a specified use? ➢ Has the sample of test takers been defined from which scores have been drawn? ➢ How does this sample represent the population of interest in terms of socio-demographic or developmental characteristics?	• What evidence has been provided that the assessment has positive consequences for stakeholders? ➢ If unintended consequences have occurred, have investigations been made of both categories of validity evidence and factors external to the assessment?	• What evidence has been provided that the assessment captures a construct that is relevant in the workplace? ➢ If the scores are to be used for credentialing, how will they be distributed and what interpretative materials will be provided?
Reliability/Precision: "Appropriate evidence of reliability/precision should be provided for the interpretation for each intended score use" (p. 42).	• Have estimates of reliability/precision of scores been provided so that scores use can be justified? ➢ Have estimates of reliability/precision been provided for each relevant student subgroup so that comparisons can be made between individual and group performance?	• How do the methods for estimating sub-scores contribute to the interpretation and justification of score use? ➢ In the case of automated scoring of essay items, have descriptions of the scoring algorithms and scores associated with those algorithms been made available?	• What evidence has been provided that administrative conditions of the assessment have remained stable? ➢ What evidence has been provided of reliability/precision to justify score interpretation and use?	• When compared to a meaningful workplace criterion variable, what evidence has been provided that the assessment reliably predicts workplace performance? ➢ Is workplace performance reliably predicted for subgroups of employees?
Fairness: "All steps in the testing process, including test design, validation, development, administration, and scoring procedures, should be designed in such a manner as to minimize construct-irrelevant variance and to promote valid score interpretations for the indented used of all examinees in the intended population" (p. 63).	• What evidence has been provided that scores contribute to equality of opportunity and opportunity to learn for individual students? ➢ Has each student been provided with the opportunity to learn the construct as it is being assessed?	• What evidence has been provided that principles of universal design have been followed in creating the assessment? ➢ Have barriers been identified and mitigated that impede access to the construct as it is being assessed?	➢ Have safeguards been developed to discourage inappropriate score interpretations and score use? ➢ If value added methods have been considered in determining school or teacher performance based on test scores, does evidence justify a fixed weight in decision-making?	➢ What evidence is available that the scores have the same meaning for all individuals? ➢ If meanings differ for different individuals or groups, how will evidence be provided to justify score interpretation and use?

Table 2. *Operational Questions for Stakeholder Groups in English Language Arts-Writing.*

Standard	Students and Guardians	Teachers and Administrators	Legislators	Workforce Leaders
Test Design and Development: "Tests and testing programs should be designed and developed in a way that supports the validity of interpretations of the test scores for their intended uses. Test developers and publishers should document steps taken during the design and development process to provide evidence of fairness, reliability, and validity for intended uses for individuals in the intended examinee population" (p. 85).	• What is the relationship among the following: the curriculum at the individual student's school, the curricular goals, and the assessment? ➢ How have the steps of the assessment processes been documented and communicated by those responsible for developing the assessment?	• How have assessment specifications been provided regarding the construct under examination, the examinee populations, and the proposed interpretations of scores and their use? ➢ How have the assessment developers communicated the standards for item review, the administration and scoring procedures, and the basis for revision of the assessment?	• How have the assessment developers demonstrated that they have designed their assessments in ways to support the validity, reliability/precision, and fairness associated with their intended use? ➢ What processes have been established, and what funds have been designated, to revise the assessment based on new information resulting from the present administration?	• How have the assessment developers demonstrated that their test development and design process have taken into consideration important workplace needs associated with construct competency? ➢ How have rationales been developed that justify linkages between test design and development processes and workplace needs for credentialing, selection, placement, and promotion?
Scores, Scales, Norms, Score	• If decisions regarding	• If cut scores have been	• How have the assessment	• How have the assessment

Linking, and Cut Scores: "Test scores should be derived in a way that supports the interpretation of test scores for the proposed uses of tests. Test developers and users should document evidence of fairness, reliability, and validity of test scores for their proposed use" (p. 102).	placement and progression are to be made from the assessment, have cut scores been established for categories of student performance? ➢ If cut scores have been established, has the procedure been documented and communicated in terms of both technical specifications and policy decisions?	established, are these scores to be used for descriptive or decision-making purposes? ➢ How have assurances been made that the establishment of cut scores does not undermine the validity of score interpretations?	developers demonstrated that scores have been normed with student populations similar to those found at individual schools or school districts? ➢ How have differentiated norms been established for different gender, race/ethnicity, language, disability, economically disadvantages, grade, and age groups?	developers demonstrated that the norms and cut scores established are congruent with workforce populations and employment needs? ➢ How have interpretations been established to help employers interpret and use the established norms and cut scores?
Test Administration, Scoring, Reporting, and Interpretation: "To support useful interpretations of score results, assessment instruments should have established procedures for test administration, scoring, reporting, and interpretation. Those responsible for administering, scoring, reporting, and interpreting should have sufficient training and supports to help them follow the established procedures. Adherence to the established procedures should be monitored, and any material errors should be documented and, if possible, corrected" (p. 114).	• How have the assessment developers designed the digital administration so that technical disruptions do not contribute to construct-irrelevant variance? ➢ Have distinctions been made between accommodations for test takers based on need and accommodations based on misalignment between the digitally-based assessment and the print-based curriculum?	• Because different stakeholder groups may administer, score, report, and interpret the assessment, how have procedures been established to ensure that score interpretation and use are not compromised by failure of standardization? ➢ How have assessment developers demonstrated that standardization will ensure that students have the same ability to demonstrate their competency?	• How have resources been leveraged to ensure that the diverse stakeholder groups needed to administer, score, report, and interpret the assessment have the competency and resources necessary to ensure standardization? ➢ In cases of students with disabilities or different language backgrounds, how have nonstandard models been established that will allow these students to demonstrate competence?	• How have test administration, scoring, reporting, and interpretation processes been designed so that scores can be used to establish connections with workplace needs? ➢ How have standardization processes resulted in the anticipation and removal of construct-irrelevant variance so that scores from the assessment can be used on a long-time basis?
Supporting Documentation for Tests: "Information relating to tests should be clearly documented so that those who use tests can make informed decisions regarding which test to use for a specific purpose, how to administer the chosen test, and how to interpret test scores" (p. 125).	• When scores are released, how have interpretations appropriate for both students and their guardian been communicated? ➢ When technical information on development and scoring is released to students and guardians, has this information been adequately explained so that score interpretation is informed?	• How have documents been prepared so that teachers and administrators can understand and communicate to students and their guardians the development process, administration and scoring, and appropriate use of scores associated with the assessment? ➢ What milestones have been established so that these supporting documents are made available to teachers and administrators in a timely manner?	• How have resources been allocated so that supporting documentation has been examined for its intended audiences? ➢ Based on knowledge about aim, genre, and discourse communities, have supporting documents been prepared so that they will discourage score misuse and contribute to justified score interpretation?	• How has supporting documentation been prepared so that workplace users of the assessment will be able to receive additional interpretative support when summaries of technical information are needed to interpret scores? ➢ In cases where the workplace is international in nature, have supporting materials been prepared in digital form and translated into languages users will need to interpret assessment scores?
Rights and Responsibilities of Test Takers: "Test takers have the right to adequate information to help them prepare for a test so that the test results accurately reflect their standing on the construct being assessed and lead to fair and accurate score interpretations. They also have the right to protection of their personally identified score results from unauthorized access, use, or disclosure. Further, test takers have the responsibility to present themselves accurately in the testing process and to respect copyright in test materials" (p. 133).	• How has the student been provided with accurate, free information about the assessment? ➢ As a means of reducing construct-irrelevant variance, how has the student been provided with practice access to the digital environment in which the test will be administered?	• How has the instructor provided students with information about the assessment, intended score use, scoring criteria, administrative policy, available of accommodations, and confidentiality? ➢ How have the students been informed of their rights and the rights of their parents to access assessment results and be protected from unauthorized use of results?	• In order to protect students from potentially adverse consequences, how has the legislative process been used to delay justified score use? ➢ If the legislative process has been used to delay score use, how have specific determinations been made regarding a range of decisions and a timeline for justified score use?	• If assessment scores are to be used to determine workplace competency, how have assurances be established to assure that students have information about how employers are using scores? ➢ If assessment scores are to be transferred to employers, how have the data systems be designed to assure confidentiality?
Rights and Responsibilities of Test Users: "Test users are responsible for knowing the validity evidence in support of the intended interpretations of scores on tests that they use, from test selection through the use of scores, as well as common positive and negative consequences of test use. Test users also have a legal and ethical	• What assurances exist that those who use assessment scores have the training and credentials necessary for responsible score interpretation and use? ➢ How have those individuals been prepared to deliver consistent and timely interpretations of scores and their use?	• How has a clear and distinct role been established for instructors in the communication of assessment results? ➢ If teachers and administrators disagree with justified interpretation and use, have processes been designed to allow warranted	• In order to protect students from potential misinterpretations of scores, how have legislators minimized these foreseeable misrepresentations? ➢ What processes have legislators put in place to prevent score misrepresentations?	• How have workplace leaders been educated to interpret and use scores in ways leading to the advancement of equity and opportunity to learn? ➢ How have workplace leaders been educated about anticipating negative consequences of score use?

responsibility to protect the security of test content and the privacy of test takers and should provide pertinent and timely information to test takers and other test users with whom they share test scores" (p. 142).		disagreement while maintaining a stance that will not compromise student motivation or parental interest?		

A Town Hall Thought Experiment

To envision how the questions in Table 1 and Table 2 might be used together, we propose a thought experiment: a series of town hall meetings in which local stakeholders are brought together to address assessment issues associated with the CCSSI ELA-W. If frequently asked questions arising from these tables were prepared and distributed in advance, fact finding could occur before the meeting and the participants could then focus on establishing common ground.

Imagine that town hall meeting were to occur in the beginning of the 2015 school year, a time at which many questions of proper score interpretation and use remain unanswered. Using questions from Table 2 in order to establish the relationships among validity, reliability/precision, and the operational obligations of assessment developers, curriculum developers, and teachers, students and their guardians might justifiably ask how scores have been established for categories of student performance and if those scores will, in turn, lead to decisions regarding promotion and placement.

During the imagined town meeting, attention might be drawn to the Smarter Balanced Consortium (2014b) document entitled "Interpretation and Use of Scores and Achievement Levels" that we discussed in the previous section. Recall that scale scores and achievement level descriptors are identified in alignment with the *Standards* in the document. Using the validity questions from Table 1, teachers and administrators might focus on discussing the relationship between test results and the curriculum in their classrooms, schools, and districts. Choices in test design, administration, and reporting become critical as questions are raised regarding the constructive alignment—the integrated instructional and assessment systems and efforts used to map learning activities to outcomes (Biggs & Tang, 2011)—that must be established among the individual student's school, the CCSSI ELA-W, and Smarter Balanced and PARCC assessments. Critically discussing the implications of various decisions based on questions around constructive alignment would help establish a common understanding of the extent to which the scores are faithful demonstrations of individual student ability.

Similarly, in using the questions to investigate sources of evidence related to reliability, teachers and administrators would benefit by paying attention to the concept of measurement precision and not just an overly simplistic single descriptive statistic (Sireci, 2012). Estimates of score reliability (internal consistency) and those based on examining students more than once (parallel forms) thus become important sources of information to consider when determining appropriate and less appropriate interpretations of scores.

For students, guardians, teachers, and administrators, questions of what constitutes appropriate score interpretation and use would be especially relevant in light of the disaggregated information about student performance obtained from the Smarter Balanced field test that was administered between March and June 2014 (Smarter Balanced, 2014a). The test revealed clear performance differences among key student subgroups that allow for a critical discussion of how these differences are related to potential differences in opportunities to learn.

Specifically, at the Grade 11 level, 40.9 percent of total students examined (n = 31,018) met the cut score of Level 3 (or above) in achievement levels ranging from Level 1 (novice) to Level 4 (advanced). Among American Indian/Alaskan Native students (n = 777), 26.6 percent passed; Asian students (n = 2,334) passed at 54.1 percent; Black/African American students (n = 2,552) passed at 21.2 percent; Hispanic/Latino students (n = 10,041) passed at 32.4 percent; Native Hawaiian/Other Pacific Islander students (n = 195) passed at 32.8 percent; White/Caucasian students (n = 16,020) passed at 46.2 percent; Multi-ethnic/Multi-racial students (n = 889) passed at 45.1 percent. Among those enrolled in an Individualized Education Program (n = 2,084), 9.0 percent passed; among those classified as Limited English Proficient/English language learners (n = 1,767), 5.7 percent passed; among those classified under special program enrollment preventing discrimination based on disability (n = 366), 36.1 percent passed; among those classified as Economically Disadvantaged students (n = 13,962), 32.6 percent passed (Smarter Balanced, 2014a, p. 12).

The literature associated with opportunity to learn is a particularly rich framework for advancing instructional equity among student groups (Moss, Pullin, Gee, Haertel, & Young, 2008). In terms of the fairness questions raised in Table 1, using scores as a way to promote opportunity to learn can help in identification of barriers to success and creation opportunities to foster educational advancement. Making *Standards*-based conceptual and empirical connections among issues around validity, reliability/precision, and fairness through the lens of opportunity to learn is, we believe, an especially powerful logic that can be used to guide discussion of assessment results.

Because the continuum among school, college, and workplace writing appears to exhibit more disjuncture than congruence

(Burstein, Elliot, & Molloy, in press; Melzer, 2014), Table 2 might be used to call attention to the especially difficult generalization inference between academic and workplace writing established by the CCSSI ELA-W. Because the CCSSI specifically identifies both academic and workplace readiness, it is reasonable for post-secondary academic and workplace leaders to ask questions that allow them to obtain more clarity on critical assessment design, delivery, and scoring decision. Moreover, it is important that the ensuing discussions are used to elucidate any remaining ambiguities around how performance certification decisions should be informed by scores from CCSSI ELA-W assessments. In terms of the report "Interpretation and Use of Scores and Achievement Levels" that we discussed in the previous section, questions of score use become especially important in light of the fact that parallel operational definitions and frameworks are still under development for career readiness (Smarter Balanced Consortium, 2014b, p. 2). Present at the imagined town meeting, academic and workplace leaders could certainly highlight issues regarding the learning continuum.

Legislators will want to attend to both the intended and unintended consequence of the CCSS ELA-W in terms of validity evidence and factors external to the assessment. Determination of score use is especially important in the case of value-added methods used to make inferences about teacher performance, especially when current research reveals that the scores resulting from such procedures may be systematically biased in favor of some instructors and against others (Haertel, 2013). In anticipating legal issues associate with CCSSI ELA-W assessment, stakeholders will find the empirical techniques associated with quantifying disparate impact equally useful (Poe, Elliot, Cogan, & Nurudeen, 2014) so that they can meaningfully help to advance opportunities to learn.

Conclusion

As these examples from our town hall thought experiment illustrate, while the questions in Table1 and Table 2 are not meant to be exhaustive, they might prove useful for three reasons. First, because their phrasing is informed by the program of research begun by Tversky and Kahneman (2011), it is possible that such questions might act as a bridge between the kinds of evidence-based, argumentative logic that assessment designers employ in ECD (Mislevy, Steinberg, & Almond, 2003) and the availability, representativeness, and adjustment involved in heuristic reasoning that other assessment stakeholders may use in decision-making (Gilovich & Griffin, 2002). Bridging the logic of the assessment developer and the logic of the assessment user is a worthy goal that might be served by attention to decision-making under uncertainty. Tables 1 and 2 contribute to our desire to help stakeholders ask principled questions about assessment design, score use, and consequences. Second, attention to diverse reasoning processes is inherent in the social cognitive view of writing that informs the CCSSI ELW-W and its assessment. As Gilovich & Griffin (2002) have observed, the heuristic reasoning program fits well with our present understanding of how the mind works. Third, the imagined town meeting as the forum for deliberative discussion suggests the need for the development of what Rawls (2001) has referred to as overlapping consensus. The aim of reasonable pluralism is a worthy goal that may be achieved if common referential frames are established of the kinds we have suggested here.

The concepts we have presented in this paper are complex, and the challenges we have identified are real and must be addressed. We believe that our collective logic can be guided by interpretative frameworks such as the three presented here that speak to core issues associated with advancement of opportunity to learn. As present curricular and assessment innovations merge to produce information about student performance, many questions nevertheless remain. Especially notable are questions regarding the relationship between assessment and opportunity structure. Future work must turn to questions left unanswered here.

Acknowledgements

The authors would like to thank Jesse R. Sparks, Robert J. Mislevy, and Donald E. Powers for their careful review and suggestions for revision. The authors would also like to thank the anonymous reviewers for their detailed notes. Throughout the process, Diane Kelly-Riley and Carl Whithaus encouraged and supported our work.

References

Addison, J., & McGee, S. J. (2015) To the core: College composition classrooms in the age of accountability, standardized testing, and Common Core State Standards. *Rhetoric Review*, *34*, 200-218.

Almond, R. G., Steinberg, L. S., & Mislevy, R. G. (2002). A four process architecture for assessment delivery, with connections to assessment design. Princeton: Educational Testing Service. Retrieved from https://www.education.umd.edu/EDMS/mislevy/papers/ProcessDesign.pdf

American Educational Research Association, American Psychological Association, & National Council on Measurement in

Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

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

Anderson, T., Scum, D., & Twining, W. (2005). *Analysis of evidence* (2nd ed.). Cambridge, UK: Cambridge University Press.

Applebee, A. N. (2013). Common Core State Standards: The promise and the peril in a national palimpsest. *English Journal*, *103*, 25-33.

Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly, 2*, 1-34.

Benjamin, L. T. (1997). The origin of psychological species. *American Psychologist*, *57*, 725-732.

Berninger, V. W. (2012). (Ed.) *Past, present, and future contributions of cognitive writing research to cognitive psychology*. New York, NY: Tailor and Frances.

Biggs, J., & Tang, C. (2011). *Teaching for quality learning at university* (4th ed.). New York, NY: McGraw-Hill.

Burstein, J., Elliot, N., & Molloy, H. (in press). Informing automated writing evaluation using the lens of genre: Two studies. *Calico Journal*.

Common Core State Standards Initiative. (2015a). English language arts standards. Retrieved from http://www.corestandards.org/ELA-Literacy/

Common Core State Standards Initiative. (2015b). English language arts standards, writing, grade 11-12. Retrieved from http://www.corestandards.org/ELA-Literacy/W/11-12/

Common Core State Standards Initiative. (2015c). English language arts standards, writing, introduction 6-12. Retrieved from http://www.corestandards.org/ELA-Literacy/W/introduction-for-6-12/

Cumming, A. (2013). Assessing integrated writing tasks for academic purposes: Promises and perils. *Language Assessment Quarterly*, *10*, 1-8.

de Ayala, R. J. (2008). *The theory and practice of item response theory*. New York, NY: Guilford Press.

De Boeck, P., & Wilson, M. (Eds.). (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. New York, NY: Springer.

Duncan, A. (2015, April 9). Remarks by U.S. Secretary of Education Arne Duncan on the 50th anniversary of Congress passing the Elementary and Secondary Education Act. Retrieved from http://www.ed.gov/news/speeches/remarks-us-secretary-education-arne-duncan-50th-anniversary-congress-passing-elementary-and-secondary-education-act

Educational Testing Service (2015). *ETS standards for quality and fairness, 2104*. Princeton, NJ: Educational Testing Service. Retrieved from https://www.ets.org/s/about/pdf/standards.pdf

Eliot, C. W. (1892). Wherein popular education has failed. *The Forum*, *14*, 411-428.

Elliot, N., & Perelman, L. (Eds.). (2012). *Writing assessment in the 21st century: Essays in honor of Edward M. White*. New York, NY: Hampton Press.

Fernberger, S. W. (1932). The American Psychological Association: A historical summary, 1892-1930. *Psychological Bulletin*, *29*, 1-89.

Flower, L. (1994). *The construction of negotiated meaning: A social cognitive theory of writing*. Carbondale and Edwardsville, IL: Southern Illinois University Press.

Flower, L., & Hayes, J. R. (1981). A cognitive process theory of writing. *College Composition and Communication*, *32*, 365–87.

Gallagher, C. W. (2011). Being there: (Re)making the assessment scene. *College Composition and Communication*, *63*, 450-476.

Gates, B., & Gates, M. (2015). College-ready education. Retrieved from http://www.gatesfoundation.org/What-We-Do/US-Program/College-Ready-Education

Gardner, H. (2006). *Five minds for the future: Leadership for the common good*. Cambridge, MA: Harvard Business School Press.

Gilovich T. & Griffin, D. (2002). Introduction—Heuristics and biases: Then and now. In T. Gilovich, D. Griffin, & D, Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 1-18). Cambridge, UK: Cambridge University Press.

Graham, S. (2006). Writing. In P. Alexander & P. Winne (Eds.), *Handbook of educational psychology* (pp. 457-478). Mahwah, NJ: Erlbaum.

Graham, S., McKeown, D., Kiuhara, S. A., Harris, K. R. (2012). A meta-analysis of writing instruction for students in the elementary grades. *Journal of Educational Psychology*, *104*, 879-896.

Graham, S., & Perin, D. (2007). A meta-analysis of writing instruction for adolescent students. *Journal of Educational Psychology, 99,* 445-476.

Hall, G. S. (1883). The contents of children's minds. *Princeton Review*, *11*, 249-272.

Haertel, E. H. (2013). Reliability and validity of inferences about teachers based on student test scores. William H. Angoff 14th memorial lecture, National Press Club, Washington, DC. Retrieved from https://www.ets.org/Media/Research/pdf/PICANG14.pdf

Hayes, J. R. (2012). Modeling and remodeling writing. *Written Communication*, *29*, 369-88.

Hillocks, G. (1986). *Research on written composition: New directions for teaching*. Urbana, IL: National Council of Teachers of English.

Johnson, K. (2013). Beyond standards: Disciplinary and national perspectives on habits of mind, *College Composition and Communication*, *64*, 517-541.

Kahneman, D. (1973). *Attention and effort*. Englewood Cliffs, NJ: Prentice-Hall.

Kahneman, D. (2011). *Thinking fast and slow*. New York, NY: Farrar, Straus and Giroux.

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

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

Kane, M. T. (2015). Explicating validity. *Assessment in Education: Principles, Policy & Practice*, *22*, 1-14.

Ketterlin-Geller, L. R. (2008). Testing student with special needs: A model for understanding the interaction between assessment and student characteristics in a universally designed environment. *Educational Measurement: Issues and Practice*, *27*, 3-16.

Kliebard, H. M. (2004). *The struggle for the American curriculum, 1893-1958* (3rd ed.). New York, NY: Routledge.

Leijten, M., Van Waes L., Schriver, K., & Hayes, J.R. (2014). Writing in the workplace:

Constructing documents using multiple digital sources. *Journal of Writing Research*, *5*, 285-337.

Lindemann, E. (Ed.). (2010). *Reading the past, writing the future: A century of American literacy education and the National Council of Teachers of English*. Urbana, IL: NCTE.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillside, NJ: Erlbaum.

Luecht, R. M., & Camara, W. J. (2011). Evidence and design implications required to support compatibility claims. Retrieved from http://www.parcconline.org/sites/parcc/files/PARCCWhitePaperRLuechtWCamara.pdf

Melzer, D. (2014). *Assignments across the curriculum: A national study of college writing*. Logan, Utah: Utah State University Press.

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

Mislevy, R. J. (2007). Validity by design. *Educational Researcher, 36*, 463-69.

Mislevy, R. J. (2008). How cognitive science challenges the educational measurement tradition. Retrieved from http://umdperg.pbworks.com/f/CommentaryHaig_Mislevy.pdf

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2004). A brief introduction to evidence-centered design (ETS Research Report-03-16). Princeton, NJ: Educational Testing Service. Retrieved from https://www.ets.org/Media/Research/pdf/RR-03-16.pdf

Mislevy, R.J., Haertel, G., Cheng, B., Ructtinger, L., DeBarger, A., Murray, E., …Vendlinski. T. (2013). A "conditional" sense of fairness in assessment. *Educational Research and Evaluation: An International Journal on Theory and Practice*, *19*, 121-140.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessment (with discussion). *Measurement: Interdisciplinary Research and Perspective*, *1*, 3-62.

Moss, P. A., Pullin, D. C. , Gee, J. P., Haertel, E. H. & Young, L. J. (Eds.), *Assessment, equity, and opportunity to learn.* Cambridge, UK: Cambridge University Press.

National Research Council. (2012). *Education for life and work: Developing transferable knowledge and skills in the 21st century*. Committee on Defining Deeper Learning and 21st Century Skills, J.W. Pellegrino & M.L. Hilton (Eds.). Board on Testing and Assessment and Board on Science Education, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

Phelps, L. W., & Ackerman, J. M. (2010). Making the case for disciplinarity in rhetoric, composition, and writing studies: The visibility project. *College Composition and Communication*, *18*, 180-215.

Poe, M., Elliot, N., Cogan, J. A., & Nurudeen, T. G. (2014). The legal and the local: Using disparate impact analysis to understand the consequences of writing assessment. *College Composition and Communication*, *65*, 588-611.

Pullin, D. C. (2008). Assessment, equity, and opportunity to learn. In P. A. Moss, D. C. Pullin, J. P. Gee, E. H. Haertel, & L. J. Young (Eds.), *Assessment, equity, and opportunity to learn* (pp. 333- 351). Cambridge, UK: Cambridge University Press.

Ravitch, D. (2010). *The death and life of the great American school system: How testing and choice are undermining education*. New York, NY: Basic Books.

Rawls (2001). *Justice as fairness: A restatement*. Cambridge, MA: Harvard University Press.

Rice, J. M. (1893). *The public-school system of the United States.* New York, NY: Century.

Rogers, L., & Graham, S. (2008). A meta-analysis of single subject design writing intervention research. *Journal of Educational Psychology, 100*, 879-906.

Shannon, P., Whitney, A. E., & Wilson, M. (2014). The framing of the Common Core State Standards. *Language Arts*, *91*, 295-302.

Sireci, S. G. (2012). Smarter Balanced Assessment Consortium: Comprehensive research agenda. Retrieved http://www.SmarterBalanced.org/wordpress/wp-content/uploads/2014/08/Smarter-Balanced-Research-Agenda_Recommendations-2012-12-31.pdf

Smarter Balanced Assessment Consortium. (2014a). Disaggregated data from the Smarter Balanced field test. Retrieved from

http://www.Smarter Balanced.org/wordpress/wp-content/uploads/2014/12/Disaggregated-FieldTestDataFINAL.pdf

Smarter Balanced Consortium (2014b). Interpretation and use of scores and achievement levels. Retrieved from http://www.Smarter Balanced.org/wordpress/wp-content/uploads/2014/11/Interpretation-and-Use-of-Scores.pdf

Snyder, T. D. (1993). *120 years of American education: A statistical portrait*. Washington, DC: National Center for Education Statistics. Retrieved http://nces.ed.gov/pubs93/93442.pdf

Snyder, T. D., & Dillow, S. A. (2015). *Digest of education statistics 2013* (NCES 2015-011). Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education.

Sparks, J. R., Song, Y., Brantley, W., & and Liu, O. L. (2014). Assessing written communication in higher education: Review and recommendations for next-generation assessment (ETS Research Report No. RR-14-37). Princeton, NJ: ETS.

Toulmin, S. E. (1958/2003). *The uses of argument*. Cambridge, UK: Cambridge University Press.

Tucker, B. (2009). The next generation of testing. *Educational Leadership*, *67*, 48-53.

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, *5*, 202-32.

Ward, L. F. (1883). *Dynamic sociology, or applied social science as based upon statistical sociology and the less complex sciences*. New York: Appleton.

Way, W. (2014). Memorandum on instructional sensitivity considerations for the PARCC assessments. Retrieved from http://www.parcconline.org/sites/parcc/files/Instructional_sensitivity_memo_final%2808%2015%2014%29.pdf

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

Author Biographies

Norbert Elliot is Professor Emeritus at New Jersey Institute of Technology. With Edward M. White and Irvin Peckham, he is author, most recently, of *Very Like a Whale: The Assessment of Writing Programs* (Logan, UT: Utah State University Press, 2015).

André A. Rupp is Research Director for work focused on best practices for evidence identification, accumulation, and alignment for automated scoring systems as well as standard setting and alignment work situated within the assessment capabilities division at the Educational Testing Service. With Jonathan Templin and Robert A. Henson, he is co-author of *Diagnostic Measurement: Theory, Methods, and Applications* (New York: Guilford Press, 2010).

David M. Williamson is Vice President for New Product Development at Educational Testing Service. With Norbert Elliot, he co-edited a special issue of *Assessing Writing* (2013, 18.1) on automated writing evaluation.