Routledge
Taylor & Francis Group

Theory Into Practice

ISSN: 0040-5841 (Print) 1543-0421 (Online) Journal homepage: http://www.tandfonline.com/loi/htip20

Assessment of Complex Cognition: Commentary on the Design and Validation of Assessments

James W. Pellegrino & Mark Wilson

To cite this article: James W. Pellegrino & Mark Wilson (2015) Assessment of Complex Cognition: Commentary on the Design and Validation of Assessments, Theory Into Practice, 54:3, 263-273, DOI: 10.1080/00405841.2015.1044377

To link to this article: http://dx.doi.org/10.1080/00405841.2015.1044377

Accepted online: 21 May 2015.

Theory Into Practice, 54:263–273, 2015

ISSN: 0040-5841 print/1543-0421 online
DOI: 10.1080/00405841.2015.1044377

Routledge
Taylor & Francis Group

James W. Pellegrino
Mark Wilson

Assessment of Complex Cognition: Commentary on the Design and Validation of Assessments

THE SEVEN articles in this special issue are concerned with the challenges of assessing complex aspects of cognition in the domains of mathematics, reading, history, and science. Each describes the design of assessments and their interpretive use, with a particular focus on assessments closely tied to classroom instruction. Individually and collectively, they make valuable contributions, highlighting many conceptual and practical considerations that need to be addressed in designing and validating assessments of key aspects of mathematical, literary, scientific, and historical reasoning.

James W. Pellegrino is the Distinguished Professor of Liberal Arts and Sciences, Psychology, and Education at the University of Illinois at Chicago and Co-director of UIC's Learning Sciences Research Institute and Mark Wilson is Professor of Education at the University of California, Berkeley.

Correspondence should be addressed to Professor James W. Pellegrino, Learning Sciences Research Institute, University of Illinois at Chicago, 1240 West Harrison (M/C 057), Chicago, IL 60607. E-mail: pellegjw@uic.edu.

Our discussion is divided into three parts. Part 1 presents three conceptual frames regarding the nature of assessment and assessment design, providing an interpretive language for discussing the seven articles. Part 2 applies these frames to the articles as a way to interpret the specifics of each case. Part 3 highlights challenges that remain in operationalizing and validating assessments of complex cognition.

Part 1: Three Conceptual Frames

The C-I-A: Curriculum, Instruction, and Assessment

Assessment does not and should not stand alone in the educational system. Rather, it is one of three coordinated components—curriculum, instruction, and assessment. *Curriculum* refers to knowledge and skills in subject matter areas that teachers teach and students are supposed to learn. It generally consists of a scope of content in a given subject area—such as mathematics, history, or science—and a sequence for learning.

Instruction refers to methods of teaching and the learning activities used to help students master the content and objectives specified by a curriculum. *Assessment* is the means used to measure the outcomes of education and students' achievements with regard to important competencies. Ideally, an assessment should measure what students are actually being taught, and what is taught should parallel the curriculum one wants students to master. Aligning the three components is often a challenge; each article in this section addresses this challenge, albeit in different ways.

Assessment as Evidentiary Reasoning

Assessment enables educators to learn about what students know and can do, but cannot offer a direct window into a student's mind. An assessment is a tool designed to observe students' behavior, to produce data that can be used to draw reasonable inferences about what students know. In the process of generating and interpreting evidence to support inferences about what students know, all assessment procedures operate from a chain of reasoning about learning. This is true for classroom quizzes, standardized achievement tests, computerized tutoring programs, and even the conversation between a student and teacher as they work through a problem together, or discuss the meaning of a historical text or a scientific diagram.

This process of reasoning from evidence has been portrayed as a triad of three interconnected elements: the *assessment triangle* (Pellegrino, Chudowsky, & Glaser, 2001). The vertices represent three key elements underlying any assessment: a model of student *cognition* and learning in the domain of the assessment; a set of assumptions and principles about the kinds of *observations* that will provide evidence of students' competencies; and an *interpretation* process for making sense of the evidence. For effective and valid assessment, the three elements must be in synchrony.

The assessment triangle provides a useful framework for analyzing the underpinnings of assessments to determine how well they accomplish intended goals, for designing assessments, and for establishing their validity (e.g., Marion & Pellegrino, 2006). Each of the elements of the triangle must make sense on its own, and must connect meaningfully to each of the other two, to lead to an effective assessment and sound inferences. Central to this process are theoretically grounded and empirically supported understandings of how students learn, what students know as they develop competence, and how students' performances reflect these competencies. Such considerations are reflected differently in each article in this section.

Construct-Centered Design

The design of an actual assessment is a challenging endeavor that must be guided by theory and research about cognition in context, as well as practical prescriptions regarding the processes that lead to productive and potentially valid assessments for particular contexts of use. Design is always a complex process that applies theory and research to achieve near-optimal solutions under multiple constraints, some of which are outside the realm of science. Assessment design is influenced in important ways by variables such as its purpose (e.g., to assist learning, to measure individual attainment, or to evaluate a program), the context in which it will be used (e.g., classroom, district or international-comparative), and practical constraints (e.g., resources and time).

The logic embedded in the assessment triangle is exemplified by the work of two groups of researchers that have generated frameworks for developing assessments: (a) the evidence-centered design (ECD) approach, developed by Mislevy and colleagues (see, e.g., Mislevy & Haertel, 2006; Mislevy & Riconscente, 2006), and the construct-modeling approach, developed by Wilson and his colleagues (see, e.g., Wilson, 2004a; Wilson & Draney, 2004; Wilson & Sloane, 2000). They both use a construct-centered approach to task development, and both closely follow the assessment triangle's emphasis on the logic of evidentiary reasoning.

Traditional approaches to assessment design tend to focus primarily on surface features of tasks, such as how they are presented to students, or the format in which students are asked to respond. In a construct-centered approach, the selection and development of assessment tasks, as well as the scoring rubrics and criteria, and the modes and style of reporting, are guided by the construct to be assessed and the best ways of eliciting evidence about a student's proficiency with that construct.

In a construct-centered approach, the process of assessment design and development is characterized by the following developmental steps, which are common to both ECD and construct modeling:

- Analyzing the cognitive domain that is the target of an assessment;
- Specifying the constructs to be assessed in language detailed enough to guide task design;
- Identifying the inferences that the assessment should support;
- Laying out the type of evidence needed to support those inferences;
- Designing tasks to collect that evidence, modeling how the evidence can be assembled and used to reach valid conclusions; and
- Iterating through the previous stages to refine the process, especially as new evidence becomes available (Pellegrino et al., 2001, p. xx).

The articles in this section differ in the ways they balance these criteria, and in their approaches to incorporating evidence into assessment design.

Part 2: Consideration of the Seven Articles

The C-I-A Frame: Addressing the Alignment of Curriculum, Instruction, and Assessment

Although assessments are currently used for many purposes in the educational system, a premise of the *Knowing What Students Know* report (Pellegrino et al., 2001) is that their effectiveness and utility must ultimately be judged by the extent to which they promote student learning. The aim of assessment should be "*to educate and improve* student performance, not merely to *audit* it" (Wiggins, 1998, p. 7). Thus, beyond the immediate use of assessments to provide teachers with student measures of a construct, (a) the results can also be used to improve classroom instruction, to evaluate educational institutions and educational programs; and (b) the assessments themselves can also have impacts as signifiers of what is educationally important, which might be called the "signification" effect of the assessments (Wilson, 2004b). Because assessments are developed for specific purposes, the nature of their design is very much constrained by their intended use. Although a dichotomy is often made between *internal* classroom assessments, administered by instructors, and *external* tests, administered by districts, states, or nations or other agencies, such a dichotomy is an oversimplification of a continuum that reflects the proximity of an assessment to the enactment of specific instructional and learning activities. Moreover, many assessments are used in multiple ways that are at different points on this continuum (albeit not always with sound validity evidence for each one). Ruiz-Primo, Shavelson, Hamilton, and Klein (2002) defined five discrete points on a continuum of assessment distance: *immediate* (e.g., observations or artifacts from the enactment of a specific instructional activity), *close* (e.g., embedded assessments and semiformal quizzes of learning from one or more activities), *proximal* (e.g., formal classroom exams of learning from a specific curriculum), *distal* (e.g., criterion-referenced achievement tests such as required by the federal No Child Left Behind legislation), and *remote* (broader outcomes measured over time, including norm-referenced achievement tests and some national and international achievement measures). Different assessments should be understood as different points on this continuum if they are to be effectively aligned with each other and with curriculum and instruction. In essence, an assessment is a test of transfer and it can be near or far transfer depending on where the assessment falls along the continuum noted earlier. The level

at which an assessment is intended to function, which involves varying distance in space and time from the enactment of instruction and learning, has implications for how and how well it can fulfill various functions of assessment, be they formative, summative, or program evaluation (National Research Council, 2003).

The seven articles in this special issue all involve assessments that are tightly coupled to curriculum and instruction, although they vary on the space and time continuum and the explicitness of their connection to specific curricula or instructional practices.

For example, the article by Schoenfeld gives a brief overview of developments regarding standards in mathematics education over the last 25 years, leading up to recently released items from the California Standards Test (CST), the Smarter Balanced Assessment Consortium (SBAC), and then focuses on contributions from the Mathematics Assessment Project (MAP). He then discusses sample items from the projects. The first two projects are quite different in their purposes from the third, with the CST and SBAC items being intended for a context of state standardized assessments, and the MAP items intended to be used directly by teachers in a formative assessment process. Hence, the example items that are discussed are of very different natures, with the CST and SBAC ones being *remote* in the Ruiz-Primo et al. (2002) classification, and the MAP ones being *close*. Perforce, this difference in purpose results in considerable differences between the two item types, and Schoenfeld makes use of this to illustrate the different sorts of interpretations that can be made from the results of each. He advocates strongly for development and deployment of items more like the MAP ones in standardized assessments.

The article by Graf and Arieli-Attali is based on work from the Cognitively Based Assessment of, for, and as Learning (CBAL) project (Bennett, 2010; Bennett & Gitomer, 2009). The domain model for CBAL Mathematics (which they refer to as a competency model) includes processes and content areas, and, consistent with Common Core State Standards for Mathematics (CCSSM), these are seen as crosscutting. The title of the project itself implies a potentially very wide range of the assessments on Ruiz-Primo's classification, from *close* to *remote*. The assessment that is shown as an example appears to be towards the remoter end, however.

The article by Afflerbach, Cho, and Kim focuses on assessment of higher-order thinking in reading. Rather than giving an account of the products of a particular assessment project, their contribution is to explicate a framework they have developed for understanding higher-order thinking in reading. The framework itself is quite catholic in its applicability to Ruiz-Primo's classification—items could be developed that were at any level, although none are actually shown in the article.

The article by Lee and Goldman addresses the assessment of literary reasoning, seen as an example of the assessment of literacy in substantive fields, explicitly including also history and science literacy. As with the Afflerbach et al. article, this article is describing a generalized approach, rather than a specific product, although it is indeed produced by a particular project—Project READI. And, just as with the Afflerbach et al. article, this generalized approach could be applied across the range of Ruiz-Primo et al.'s (2002) classification—items could be developed that were at any level, although none are actually shown in the article.

The assessment development work discussed in the Pearson, Knight, Cannady, Henderson, and McNeill article seems to encompass the first three levels in the Ruiz Primo et al. (2002) continuum—*immediate, close*, and *proximal*—although it is not altogether clear exactly when a particular form or type of assessment is actually enacted and how. Furthermore, their assessments are deeply embedded in a particular curriculum and instructional program—*Seeds of Science/ Roots of Reading*—with the goal of providing information about student competence that can be used by the teacher for both formative and summative purposes.

The Ryoo and Linn article describes assessments that are also closely tied to key science learning standards and classroom instructional practices and seem to best fit the characteristics of

proximal assessments that can be used to evaluate the progress of student learning with regard to key conceptual understanding of complex relationships among elements of biological systems. The assessments are embedded in technology-based inquiry units using the WISE system platform, and they appear to be designed in such a way that the evidence obtained from student performance could be used for both formative and summative purposes. Clearly, the design was based on a collaborative partnership involving teachers and the validity of the assessments rests in part on teachers' perceptions of the relevance of the tasks vis-à-vis the curricular and instructional goals of the various units in concert with the forms of evidence of student thinking made visible via the assessment tasks.

The Ercikan and Seixas article is probably best understood as an example of a *distal* assessment in the sense that it is not tied directly to any specific curriculum and instructional program and the primary use of the results of the assessment is likely for a summative judgment about key aspects of students' historical reasoning skill. Nevertheless, it is not devoid of a connection to issues of curriculum and instruction, because the valid use of such an assessment and any inferences to be drawn from student performance rely on assumptions about the nature of the instruction that has previously ensued. It is designed to serve as a test of the capacity of students to transfer specific aspects of historical reasoning that are the focus of the assessment task to a new situation using unfamiliar specific content. Thus, their assessment is designed as a test of far, rather than near, transfer, as is the case with both the Pearson et al. and Ryoo and Linn assessment examples.

Taken together, these articles illustrate that issues of assessment design and validation depend on the coupling of those assessments to curriculum and instruction. Such coupling can vary substantially in terms of the space and time continuum articulated by Ruiz Primo et al. (2002), and it has implications for assessment design and interpretive use. The seven articles illustrate variations in that coupling and some of the design issues and challenges associated with the closeness of that coupling. Despite their differences, they all point out the need to consider the educational value of the results of a given assessment relative to critical issues of teaching and learning, especially when the constructs of interest are forms of knowledge and reasoning that go well beyond recall of mathematical procedures or historical or scientific facts, as well as the signification-power of the assessments themselves.

The Evidentiary Reasoning Frame: Addressing the Three Components of the Assessment Triangle

The *assessment triangle* hinges on an explicit model of domain cognition, i.e., key constructs in the domain of mathematical, literary, historical, or scientific knowledge and reasoning. One question is whether such a domain model has been articulated, laying out the forms of knowledge and the reasoning practices that define the target domain for each of the developed assessments. A second question is whether and how that cognitive model guides the design and selection of the tasks and activities presented to students. A third question is what elements of student work are the focus for the interpretation of performance relative to the underlying cognitive model.

In terms of evidentiary reasoning, the Schoenfeld article leaves the domain model unspecified—he does note that the SBAC items are intended to be based upon the CCSSM, and he does discuss these in general, but the specific standards for example items are not given. He does note the general domains that SBAC is intending to assess, though the way that is attained through specific items is not laid out (in his defense, he was working with SBAC-released items, so the shortcoming is theirs, not his). In terms of MAP, he gives an example of how the results from an item in a formative assessment lesson (FAL) can be related very directly to a common issue that students have, and thence to sample suggested questions and prompts that the teacher might use in the classroom.

CBAL Mathematics structures the content and processes in a learning progression (Smith, Wises, Anderson, & Krajcik, 2006), which they

see as being a useful way to embody connections among content areas and processes. They give an example that focuses on linear (and nonlinear) functions, showing five successive levels of the progression, each embodying a particular combination of processes and contents that is keyed to the CCSSM framework. They give a summary of how the progression itself, along with associated documentation, can help a teacher understand the learning that a student needs to accomplish as they progress up the progression. The example they show is a simple linear order, but, of course, this need not be the case—for example, many different shapes of progression have been discussed by Wilson (2009, 2014).

The Afflerbach et al. framework is based on a modification of Kratwohl's (2002) taxonomy of cognitive processes, itself a revision of Bloom's (1956) redoubtable taxonomy. In their Table 1, they show how it can be seen as relating to both the National Assessment of Educational Progress (NAEP) Framework, and the Common Core State Standards (CCSS). They give no actual items as examples, although they do list some representative types of tasks relating to different levels of the framework in the same Table 1.

The Lee and Goldman article abounds in structures of various sorts—there are the cognitive domains:

> Dimensions of Knowledge, Skills, and Practices, with the categories Epistemology, Inquiry Strategies, Key Concepts, Types of Texts, and Discourse and Ways of Using Language.

Furthermore, there are two types of assessment structure included

> Text Complexity with the categories Theme, Character and Structure, and Question Types, with the categories Basic Stated Information, Key Details, Stated Relationships, Simple Implied Relationships, Complex Implicit Relationships, Author Generalizations, and Structural Generalizations.

Note that each of these categories may well be multidimensional in any given context. Pointing out the complexity of these structures is not a criticism—the phenomena that assessments are dealing with are inherently complex, and people seldom acknowledge this in the very full way that it is displayed in this article. Lee and Goldman do not see their approach as necessarily matching to the CCSS Reading structure, and, in fact, use their structure to critique what they see as arbitrary divisions created by CCSS Reading.

The Pearson et al. article is focused on the domain of constructing and critiquing scientific arguments based on text. They talk about adopting a learning progression frame of reference with respect to the construct(s) of interest and mapping it to the three modalities of reading, writing, and talking. One of the many challenges they faced in their assessment design process was the limited base of theory and research regarding the progression of student cognition for scientific argumentation. Nevertheless, they articulate a set of constructs related to the structure and dialogic process of scientific argumentation and briefly discuss the translation into assumptions about levels in their learning progressions for each of the three areas of performance—reading, writing, and listening. Their cognitive model appears to have significantly influenced the observation component of their assessment development work. The tasks and activities that they have chosen to present to students and the ways in which evidence is seemingly extracted from those performances for purposes of mapping back to the levels in the learning progression appears to represent a coordinated and coherent system as regards the cognition-observation-interpretation components of the assessment triangle. Space limitations precluded providing many of the relevant details regarding their final assessments and the validation process and evidence. They do, however, mention exploring various measurement models and psychometric approaches as part of that validation process.

When viewed through the analytic lens of the assessment triangle, the Ryoo and Linn work is more challenging with respect to identifying how all three components come together and are reflected in their Energy Stories and MySystem

assessment tasks. Especially unclear is the cognitive model that undergirds the assessment development process and the design of the particular tasks and activities. The cognitive construct that seems to guide their work is that of integrated knowledge in which various elements of conceptual understanding become interconnected as part of an increasingly sophisticated understanding of energy storage and transformation in biological systems. The tasks they have designed and the data they extract from those tasks appear to be guided by a focus on students' ability to describe and illustrate how the various elements of such a system are interconnected and the ability to apply that knowledge to explanations of situations and phenomena. One might argue that there is a tight coupling among the three elements of the assessment triangle—cognition-observation-interpretation—even though the nature of key aspects of that coupling requires further explication.

The Ercikan and Seixas article is the most explicit with regard to the nature of the model of cognition and learning that has guided their work in developing an assessment of historical reasoning. They point out that there are differing views of the nature of student cognition in this domain with major variations in European versus North American conceptions of the key constructs of interest that should be the focus of teaching, learning, and assessment. The cognitive model they chose to apply focuses on three key constructs: use of evidence, taking a historical perspective, and the ethical dimension of historical interpretations. Given this cognitive model, they then proceeded to design tasks and materials to elicit these aspects of historical thinking and reasoning in addition to specifying the relevant forms of evidence for interpreting student proficiency with respect to the constructs of interest.

Of the seven articles in this issue, Ercikan and Seixas most clearly articulate the *assessment as reasoning from evidence* perspective and the role of the three components of the assessment triangle in connecting those components together as part of a coherent assessment development process. That said, the *reasoning from evidence* perspective seems to be at the core of the work of the others, as well. All of the efforts in developing assessments of complex thinking and reasoning in the domains of mathematics, literature, science, and history are connected to contemporary conceptions of complex cognition in the domain of interest and how such cognition might be expected to change as a function of curriculum and instruction designed to foster development of the critical competencies.

The Construct-Centered Design Frame: Addressing the Purposes, Contexts of Use, and Practical Constraints Shaping Assessment Design

The adoption of a construct-centered framework highlights design decisions as valuable opportunities to make the evidentiary logic of an assessment clearly visible. As such, several important design challenges are described in these articles, and the authors leverage them differently to highlight key issues related to development and validation of their assessments. Turning first to the two articles exploring issues in mathematics education, we see some common themes arising.

In terms of the design of assessment materials and activities, the Schoenfeld article (this issue) discusses how two aspects of the SBAC items will limit their usefulness in attaining the CCSSM aims—the exclusive use of computer administration, and the use of adaptive testing, both of which he judges to be problematic. The former is problematic in that it forces students to express themselves in ways that are (at least at present state of technology in schools) quite limited, forcing students to write their open-ended responses, and making sketching of mathematical representations quite awkward. The latter has a potential to narrow the assessments that are given to any particular student. In contrast, he notes that the MAP-style item

> supports student engagement in a number of fundamentally important aspects of learning: dealing with conceptually rich mathematics, being given the opportunity to engage (and be supported in engaging with) challenging

> problems, and to discuss and present their own ideas. (Schoenfeld, p. 191)

He also sees the possibility, if the items from the SBAC consortium (and the Partnership for Assessment of Readiness for College and Careers [PARCC] consortium) changed to be like the MAP items, then that would lead to considerable changes for the better in mathematics education.

According to the Graf and Arieli-Attali article, CBAL Mathematics is explicitly constructed following the ECD procedures, and the authors show screens from a specific computer-based item and they explain how it is related to a broader task-model that has been developed, and which explicitly links features of the items to the learning progression. They explain how this task-model is useful as a template for developing multiple variants of items, and can also be important as a validation tool. The same criticism by Schoenfeld, mentioned previously, can be made of this particular computerized item, although there is no limitation of the general ideas of CBAL in this.

The aims of the two articles in the area of reading are somewhat different from the rest: Given its aim to describe a framework for assessment, rather than focusing on an actual assessment, the Afflerbach et al. article does not address the questions raised in this section. The situation for the Lee and Goldman article is similar, though they do use an extended example relating to Alice Walker's story *The Flower* to illustrate many of their points.

Turning now to the three articles that address science and history assessments, one can see that all three describe aspects of the envisioned contexts of use for each assessment that constrain and shape the design: the amount of time and effort required of students and/or teachers (e.g., Ercikan & Seixas, this issue; Pearson et al., this issue), alignment with content and practices covered in the curriculum (e.g., Ercikan & Seixas; Ryoo & Linn), and considerations of text difficulty in the selection of materials (e.g., Ercikan & Seixas; Pearson et al.). These and many other issues contributed to design decisions and design revisions as the developers of these assessments created, piloted, revised, and implemented their various tasks and methods of scoring and interpretation. These last three articles articulate some of the many challenges they faced in designing assessment materials and activities that would elicit key forms of evidence to substantiate claims about student proficiency driven by their explicit or implicit model of domain cognition. Each of the three projects engaged in an iterative process of articulating the nature of the evidence that would be most relevant and useful for substantiating a claim about student proficiency and designing tasks and activities that had features to effectively elicit that evidence. The choices about task types and modes of responding vary across the three articles, in part reflecting the aspects of scientific and historical reasoning that were the focus of the work, and in part reflecting pragmatic constraints of time, cost, feasibility, scoring, not to mention serious attention to the reduction of construct irrelevant variance as a contributor to performance. The Pearson et al. article is the most explicit about the iterative nature of the design and validation process although it is likely that such was also the case in the work of Ryoo and Linn for scientific reasoning, and Ercikan and Seixas for historical reasoning. Finally, Ercikan and Seixas most explicitly discuss the application of ECD in their work and illustrate via a table showing the connections among key elements resulting from that design process.

Part 3: Challenges in Connecting Theory and Research to Assessment Practice

The articles in this Special Issue reveal several challenges that exist in connecting theory and research on assessment development to the practice of designing assessments across a range of subject areas, from mathematics and reading to historical and scientific thinking and reasoning intended for use in everyday educational contexts. What follows is a very brief discussion of two of the most important of those challenges.

Content Versus Thinking Processes

Generally, the articles agree on the need to move away from conceptions of school subjects as static bodies of facts and towards the importance of a citizenry versed in content knowledge as well as subject-relevant thinking practices. Yet, all recognize that engaging in (and assessing) such complex thinking and reasoning necessarily involves some subject-relevant content. Ercikan and Seixas argue that familiarity with a historical topic may change the competencies that students make evident in an assessment. On the other hand, when one is interested in the integration of knowledge regarding complex biological systems, familiarity with the elements of that system is essential. The resolution of exactly how to incorporate content into assessments depends, in part, on the closeness of the assessment to ongoing teaching and learning, the purposes of the assessment, and the claims for which evidence is sought. The issue of studied-versus-unknown content in assessments of scientific or historical thinking needs to be explicitly considered in interpreting the observed performances to support claims about what students know and can do. These articles offer useful examples of variation in this important aspect of assessment design.

Literacy Demands of the Assessments

The literacy demands of an assessment are a necessary set of considerations in the design of assessment tasks. This applies to comprehension as well as production. One needs to ask whether students have been provided with opportunities to learn to write like scientists or historians, as well as to read like scientists or historians. For example, using an ECD process, the literacy demands would be addressed in the design of the assessment task models (Mislevy & Riconscente, 2006).

Comprehension: Authentic documents versus adapted materials. Not unsurprisingly, adapting source texts has been a serious bone of contention for curriculum developers and assessment developers. On the one hand, students need to be able to access the content to reason with it. If documents are too complex or contain high proportions of esoteric vocabulary or complex syntax, students will simply not read the material and it will not be possible to obtain observations of their scientific or historical thinking. Accordingly, designers may justify the use of extracts or adapted materials that do the translations for the students. On the other hand, if students are never confronted with actual scientific or historical documents that they have to struggle to make sense of, they will develop neither strategies (cognitive and interpersonal) for dealing with complex and challenging materials, nor the confidence to tackle them. Students' development as thinkers will be dependent on the presence of document translators. This issue of comprehension is explicitly addressed by Pearson et al. and by Ercikan and Seixas. It is an area that needs a great deal more attention, especially regarding ways in which instructional supports can assist students in tackling challenging texts when reading like a literary critic, scientist, or historian (cf. Goldman, 2012; Goldman & Snow, in press; Reisman, 2012; Schoenbach, Greenleaf, & Murphy, 2012).

Production: Constructed versus selected response. The assessment tasks that were featured in these articles varied from those with high production demands (e.g., essay writing) to those with lower production demands (e.g., selected response items). Several issues rest on these choices. First, essays and construction tasks such as those used by Ryoo and Linn (this issue) can be difficult to score, especially if the assessment developers have no exemplar responses. Rubric development is time consuming, and reliability across scorers is often difficult to obtain without clear criteria. The scoring issue is less complicated for short-answer constructed responses, but unless the criteria for reasoning are clear, reliability issues are just as problematic. Also, short-answer responses may not afford students rich opportunities to demonstrate mathematical, literary, scientific or historical reasoning. Students are also often able to demonstrate

more sophisticated reasoning orally than in written form. Thus, the literacy demands of written production may mask the thinking and reasoning that students can do. The work of Ryoo and Linn, in which the products that students construct to demonstrate their integrated knowledge are embedded in technology, offers the prospect of automated scoring. In fact, it is hard to imagine that we will be able to assess many aspects of complex thinking in most subject-matter domains and make the results of those assessments useful for educators in a timely fashion without employing intelligent automated scoring engines to assist in the interpretive process.

Conclusion

Looking over the seven articles, one thing that stands out about the accounts is something that is, with two exceptions, entirely missing—a discussion of the empirical support for the cognitive structures that are discussed in the articles. (The exceptions are the Ryoo and Linn and Pearson et al. articles, where validity evidence is indeed discussed, though the testing of the constructs, even there, is not explicit.) This is somewhat odd, as it would seem to be a needed piece of evidence before one went ahead and used the assessments (e.g., its referred to as "internal structure validity evidence" in the "Testing Standards"—AERA, APA, NCME, 2014) and, in addition, the construct-centered approaches described previously both posit an empirical approach for the testing of the hypothesized construct structure as essential components of the iterative process of instrument development. Perhaps one part of an explanation lies in the nature of some of the articles, which are structured more as *domain analyses*; perhaps another part lies in the nature of this journal, which emphasizes Theory (although one would think that even theory with a capital *T* would resort to empirical support). Perhaps another explanation lies in the state of the art of assessment development, which is often more art than science; or perhaps another lies in the requested relative brevity of the articles. One hopes future such reviews of the state-of-the art will include treatment of empirical results as part of the validity argument.

Assessment development in any disciplinary domain is a challenging endeavor. We applaud the work of these authors to specify key aspects of their respective domains for purposes of assessment, and their articulation of design models and cases. Although they share a common goal, it is clear that there is considerable variation in how they have gone about the task of assessing student competence. These cases are instructive, and offer the opportunity for further dialogue about how to meet both conceptual and practical challenges in the assessment of complex reasoning and thinking in the domains of mathematics, literature, science, and history.

Acknowledgements

The preparation of this commentary was supported, in part, by Project READI, a multidisciplinary, multiinstitution collaboration aimed at research and development to improve complex comprehension of multiple forms of text in literature, history, and science. The first author's thinking on matters of assessment of historical, scientific and literary reasoning have benefitted from discussions with his READI colleagues. Project READI is supported by the Institute of Education Sciences, US Department of Education, through Grant R305F100007 to University of Illinois at Chicago. The opinions expressed are those of the authors and do not represent views of the Institute or the US Department of Education.

References

AERA, APA, NCME. (2014). The standards for educational and psychological testing. Washington, DC: AERA.

Bennett, R. E. (2010). Cognitively based assessment of, for, and as learning: A preliminary theory of action for summative and formative assessment.

Measurement: Interdisciplinary Research and Perspectives, *8*, 70–91.

Bennett, R. E., & Gitomer, D. H. (2009). Transforming K–12 assessment: Integrating accountability testing, formative assessment, and professional support. In C. Wyatt-Smith & J. Cumming (Eds.), *Educational assessment in the 21st century* (pp. 43–61). New York, NY: Springer.

Bloom, B. (1956). *Taxonomy of educational objectives: Cognitive and affective domains* (1st ed.). New York, NY: David McKay.

Goldman, S. R. (2012). Adolescent literacy: Learning and understanding content. *Future of Children*, *22*, 89–116.

Goldman, S. R., & Snow, C. (in press). Adolescent literacy: Development and instruction. In A. Pollatsek & R. Treiman (Eds.), *The Oxford handbook of reading*. New York, NY: Oxford University Press.

Krathwohl, D. (2002). A revision of Bloom's Taxonomy: An overview. *Theory into Practice*, *41*, 121–218.

Marion, S. F., & Pellegrino, J. W. (2006). A validity framework for evaluating the technical quality of alternate assessments. *Educational Measurement: Issues and Practice*, *25*, 47–57.

Mislevy, R. J., & Haertel, G. D. (2006). Implications of evidence-centered design for educational testing. *Educational Measurement: Issues and Practice*, *25*, 6–20.

Mislevy, R. J., & Riconscente, M. M. (2006). Evidence-centered assessment design: Layers, concepts, and terminology. In S. Downing & T. Haladyna (Eds.), *Handbook of test development* (pp. 61–90). Mahwah, NJ: Erlbaum.

National Research Council. (2003). *Assessment in support of learning and instruction: Bridging the gap between large-scale and classroom assessment*. Washington, DC: National Academies Press.

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (Eds.). (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academies Press.

Reisman, A. (2012). Reading like a historian: A document-based history curriculum intervention in urban high schools. *Cognition and Instruction*, *30*, 86–112.

Ruiz-Primo, M. A., Shavelson, R. J., Hamilton, L., & Klein, S. (2002). On the evaluation of systemic science education reform: Searching for instructional sensitivity. *Journal of Research in Science Teaching*, *39*, 369–393.

Schoenbach, R., Greenleaf, C., & Murphy, L. (2012). *Reading for understanding: How reading apprenticeship improves disciplinary learning in secondary and college classrooms*, 2nd Edition. San Francisco: Jossey-Bass.

Smith, C. L., Wiser, M., Anderson, C. W., & Krajcik, J. (2006). Implications of research on children's learning for standards and assessment: A proposed learning progression for matter and the atomic molecular theory. Focus article. *Measurement: Interdisciplinary Research and Perspectives*, *14*, 1–98.

Walker, A. (1994). *The Complete Stories*. London: The Women's Press. (p. 107).

Wiggins, G. (1998). *Educative assessment: Designing assessments to inform and improve student performance*. San Francisco, CA: Jossey-Bass.

Wilson, M. (2004a). *Constructing Measures: An Item Response Modeling Approach*. Mahwah, NJ: Erlbaum.

Wilson, M. (2004b). A perspective on current trends in assessment and accountability: Degrees of coherence. In M. Wilson (Ed.), *Towards coherence between classroom assessment and accountability. 103rd Yearbook of the National Society for the Study of Education, Part II* (pp. 272–283). Chicago, IL: University of Chicago Press.

Wilson, M. (2009). Measuring progressions: Assessment structures underlying a learning progression. *Journal for Research in Science Teaching*, *46*, 716–730.

Wilson, M. (2014). Considerations for measuring learning progressions where the target learning is represented as a cycle. *Pensamiento Educativo. Revista De Investigación Educacional Latinoamericana*, *51*, 156–174.

Wilson, M., & Draney, K. (2013). A strategy for assessment of competencies in higher education: The BEAR assessment system. In S. Blomeke, O. Zlatkin-Troitschanskaia, C. Kuhn, & J. Fege (Eds.), *Modeling and measuring competencies in higher education: Tasks and challenges* (pp. 61–80). Rotterdam, The Netherlands: Sense Publishers.

Wilson, M., & Sloane, K. (2000). From principles to practice: An embedded assessment system. *Applied Measurement in Education*, *13*, 181–208.