

Comments on Lissitz and Samuelsen

Validity by Design

by Robert J. Mislevy

Lissitz and Samuelsen (2007) argue that the unitary conception of validity for educational assessments is too broad to guide applied work. They call for attention to considerations and procedures that focus on “test development and analysis of the test itself” and propose that those activities be collectively termed *content validity*. The author of this article describes work that makes more explicit the underlying principles of assessment design, thereby providing conceptual foundations for familiar practices and supporting the development of new ones. By structuring design activities around assessment arguments, the test developer accrues evidence in passing for what Embretson (1983) calls “construct representation” argumentation for validity.

Keywords: assessment design; construct validity; content validity

Robert W. Lissitz and Karen Samuelsen, in their article “A Suggested Change in Terminology and Emphasis Regarding Validity and Education” (this issue of *Educational Researcher*, pp. 437–448), argue that the unitary conception of validity for educational assessments is too broad to guide the applied work of testing professionals. This view is shared by Kane (2006), McNamara (2006), Wiley (1991), Borsboom, Mellenbergh, and van Heerden (2004), and others. To guide practical work, Lissitz and Samuelsen call for attention to considerations and procedures that focus on “test development and analysis of the test itself” (p. 437), to be grouped under the appellation *content validity*. I resonate with much that they say about the need to provide support at this layer, as this is where my own interests lie. For example, Lissitz and Samuelsen reason as follows:

A third example of a systematic approach to the theory underlying the performance within a test, which we characterize as a type of content validity, is the work of Mislevy [., Steinberg, & Almond] (2003), although the authors would certainly disagree with our labeling of their approach. Our basis for this characterization of their approach is that their conceptual assessment framework specifies variables that characterize students and schemas for getting evidence about those students. Our discussion above should clarify that we would put the work of Mislevy et al. in the camp of content validity and not in construct validity, because the focus of much of their work is essentially inside the test of interest. We see much of their work as being in the spirit of Rulon (1946), Lindquist (1951), and Lennon (1956) and their emphasis on the

operations and person response processes. The area of the cognitive analysis of a test is one of the most productive and promising areas in psychometric application today. (p. 445)

Lissitz and Samuelsen are right in saying that cognitive analysis is productive and promising, although I will emphasize a constructive stance even more strongly. It is not “the cognitive analysis of a test” that is productive and promising but the design of assessments from the very beginning through a cognitive frame.¹ Because a cognitive perspective is not sufficient to determine design choices, however, considerations about test use, even if implicit, are integral to design. Lissitz and Samuelsen are also right that I do not think about my work in terms of content validity,² although I can see why they do, given the way they define the term.

In this commentary I say more about this work on assessment design. Its origins can in fact be traced to some of the concerns that motivate Lissitz and Samuelsen. My colleagues and I were working on new forms of assessment, we wanted them to support valid inferences, and we did not find the support we needed in validity theory per se. I sketch some of the sources and representations that we and others have developed to serve these needs, and I relate them to validity terminology. In the conclusion I avoid the temptation to propose new types of validity or redefine existing ones.

Origins of Evidence-Centered Assessment Design

Our work on a framework for evidence-centered assessment design (ECD) began in the 1990s at the Educational Testing Service (ETS). This was the time that the assessment community was poring over Messick’s (1989) definitive chapter on validity. It not only solidified validity as a unitary concept but extended consideration beyond inferences based on test scores to the consequences of the use of the scores. The scope of validation responsibilities thus outstripped the best practices that had developed at ETS and elsewhere over the years: These practices included some procedures falling under predictive validity, to gauge the value of test scores for predicting criterion measures; some procedures falling under construct validity (in the narrower sense that Cronbach and Meehl [1955] gave the term), to examine correlations with other theoretically linked tests through factor analysis and multitrait-multimethod studies; and—most germane to our day-to-day work—other procedures falling under content validity, to carry out domain analyses, build test specifications, and work with substantive experts to write good items to fill in specifications tables.

At the same time, we at ETS, along with the wider assessment community, were encountering challenges that lay beyond the procedures that had evolved for standardized tests. Developments

in technology and psychology were expanding the ranges of assessment contexts, ways of gathering data, and conceptions of knowledge and skill about which inferences were desired. For example, task-based language testing moved beyond sets of discrete skills to a capability to use language in real-life situations, interactively, with other people, to accomplish meaningful goals (Brindley, 1994). Computer simulations of environments such as those for medical diagnosis (Melnick, 1996) and electronics troubleshooting (Lesgold, Lajoie, Bunzo, & Eggan, 1992) enabled us to observe people making things and solving problems interactively and iteratively, in richer environments. The value of capturing evidence of teacher candidates' interactions with their students in situ, through video, direct observation, and instructional artifacts, was recognized. But how should we make sense of the complex performances we could now capture? How could we design environments and procedures effectively to elicit, capture traces of, and then interpret evidence of the kinds of knowledge and capabilities that were at issue?

Neither conceptual discussions of the ever-more-encompassing nature of validity nor established procedures for familiar tests offered much guidance for the practical work we faced. These projects relied on the insights of domain experts and talented test developers and often took a great deal of time and effort in trial and error, starts and restarts. We needed scaffolding specifically for the phase of assessment design. The design layer at which we were now working was populated by standards and practices for familiar assessments but was sparse on procedures, let alone a conceptual framework, for the new assessments we were charged with developing.

Of course much important work had been done along these lines for us to build on. Embretson (1983) had distinguished after-the-fact nomological construct validity arguments from what she called *construct representation* arguments, and in *Test Design: Developments in Psychology and Psychometrics* (1985) she made a case for, and began to illustrate, an integration of cognitive theory, task design, and psychometric models. Roid and Haladyna's (1982) *A Technology for Test-Item Writing* brought together work on more principled methodologies of task development. Wiley's (1991) "Test Validity and Invalidity Reconsidered," which I draw upon later in this commentary, provided a conceptual framework for unpacking complexes of skills and tasks in ways that supported domain analysis, task construction, and validity argumentation at this level. Messick (1994) himself provided constructive advice for designing complex performance tasks with regard to construct-relevant and -irrelevant sources of variation, as seen in the interplay among task features, evaluation procedures, and intended inferences.

Reading David Schum (1987, 1994) on inference under uncertainty was the watershed in the way that my colleagues and I thought about assessment (Mislevy, 1994, 2003, 2006; Mislevy, Almond, & Steinberg, 2002; Mislevy, Steinberg, & Almond, 2003; Mislevy, Steinberg, Breyer, Johnson, & Almond, 2002). His investigations span philosophy, science, statistics, literature, jurisprudence, intelligence analysis, and cognitive psychology. They provide a unifying foundation for the kind of reasoning that is needed to design assessments from first principles—to flesh out a framework for practical assessment design work, in light of intended inferences, and to understand and coordinate an interplay among ideas arriving from psychology, technology, statistics, and learning domains.

The Structure of Assessment Arguments

In ECD we can distinguish layers at which different kinds of thinking and different kinds of work take place in assessment. In domain analysis, for example, we study the nature of competencies of interest, how they are acquired, and how they are used. In the conceptual assessment framework that Lissitz and Samuelsen referred to, we craft specifications for the machinery through which an assessment becomes operational: task schemas, rubrics, psychometric models, presentation procedures, and the like. But the layer that organizes information about the domain for the purpose of assessment, that gives meaning to the operational elements, we call *domain modeling*. It is here that one builds assessment arguments to guide practical work and at the same time lay out what will become an essential strand of the validity argument.

Schum's work concerns what is often called *informal reasoning*, in contrast to the formal logic that implicitly grounds the positivist stance of Cronbach and Meehl (1955). This is because we are always reasoning about "particular objects or situations, . . . substantive, timely, local, [and] situation-dependent" (Toulmin, 2001, p. 24), even in cases that do revolve around scientific laws and operational definitions.

Toulmin's (1958) argument schema is a starting point for assessment arguments at the level of design, as it also is for validity arguments about inferences from scores (Bachman, in press; Kane, 1992, 2006; Messick, 1989). We wish to support a claim with data. A warrant is a generalization that justifies the inference from the particular data to the particular claim. Theory and experience back the warrant. Alternative explanations can weaken the argument, which in turn may be supported or weakened by rebuttal evidence. Validity accrues as the warrant better fits the circumstances at hand, as its backing is stronger and comes from different sources, and as more alternative explanations can be countered more convincingly.

This schema suggests the structure of an argument but not its contents. The character of data, claims, warrants, and alternative explanations arises from an understanding of the substance of the argument. In assessment, this is a conception of the capabilities of interest—how they are acquired, the situations in which people might (or might not) bring them to bear, and how we might recognize them when we see them.

Assessment arguments are more complicated than the basic Toulmin diagram. Schum points us to the allied system that Wigmore (1937) developed to chart evidence in terms of recurring patterns and principles and to guide thinking across the multiple steps of reasoning, the widely varied particulars, and the dependencies among claims and various pieces of data that appear in legal contexts. Figure 1 adapts ideas from Toulmin and Wigmore to key relationships in assessment arguments. The bottom half of the figure is a design argument, and the top half is the extension to assessment use (after Bachman, 2003). This diagram is simplified in many respects. I will mention below some additional key points that it does not depict, including the synthesis of information across multiple observations; the role of probability-based measurement models; and the evaluation of performances that arise from evolving, interactive, and sometimes multiperson situations.

Note that the claim of the design argument—some summary statement concerning the examinee's actions in the assessment setting—constitutes data for the use argument and that the backing

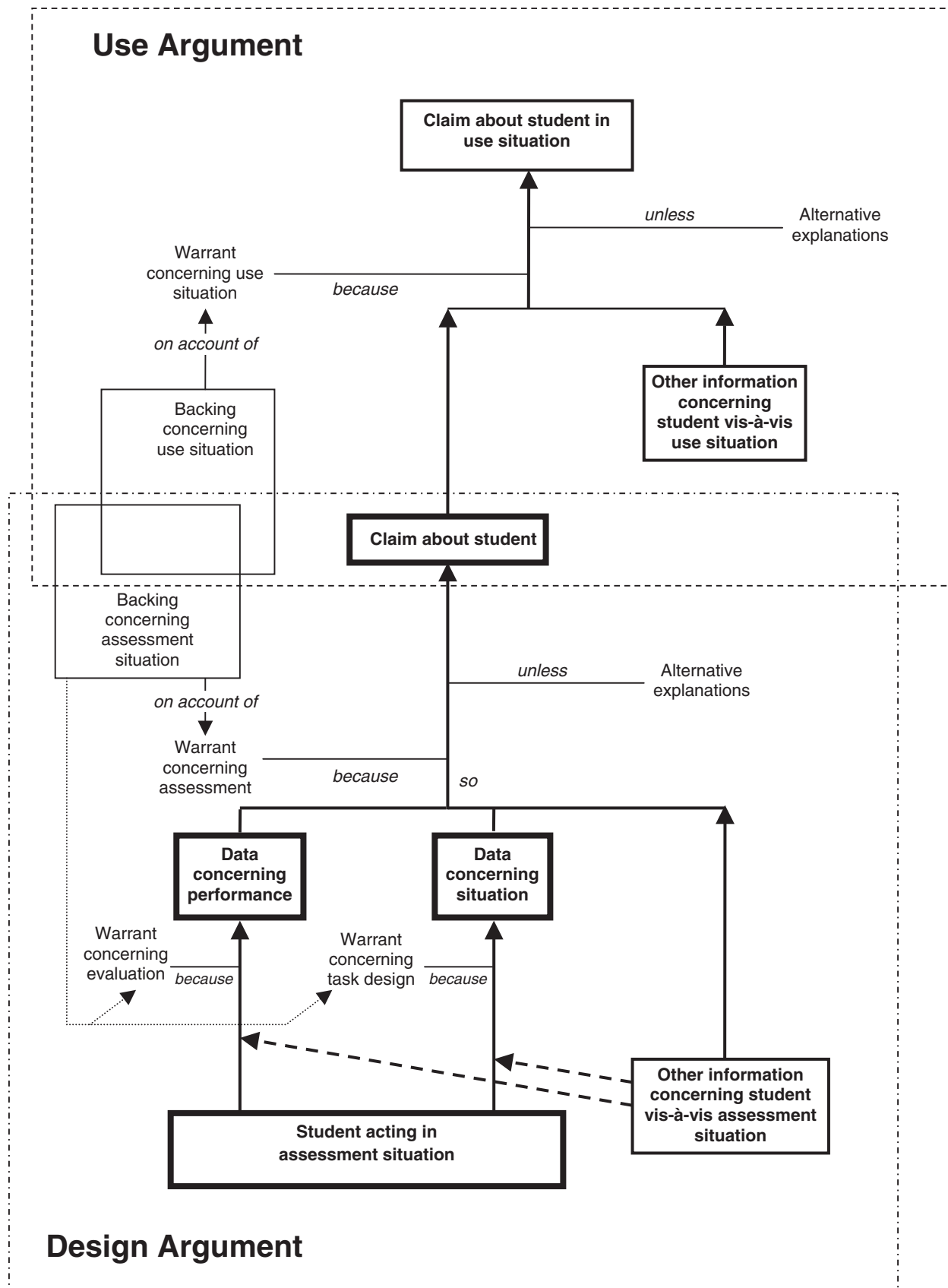


FIGURE 1. *Assessment design and use arguments.* Adapted from *Educational Measurement (4th ed., p. 206)*, Robert L. Brennan (Ed). Copyright © 2006 by American Council on Education and Praeger Publishers. Reproduced with permission of Greenwood Publishing Group, Inc., Westport, CT.

for the two arguments overlaps. Questions about which aspects of capabilities are important in the first place—and with regard to what kinds of situations, through what psychological lenses, and toward what social ends—are always present, even when these capabilities are not the primary focus of various necessary test construction activities. Figure 1 is embedded within, and for a given assessment project can only be fully detailed within, a social context. These are the questions addressed in validity arguments that concern use, values, and consequences (e.g., see Frederiksen & Collins's [1989] "systemic validity"). These questions condition design choices, although they do not determine in and of themselves what the design space is for a given assessment project, nor do they provide conclusive rationales for making all design choices or supply machinery to implement those choices. The ways we build and use assessments affect the ways we organize schooling and instruction: A test says as much to students (and teachers, and parents, and policy makers) about the nature of knowledge and learning as students' scores on it say about them.

At the bottom of Figure 1 is a unique human event, namely, a student's actions in an assessment situation: The student says, does, or makes something, possibly extending over time, possibly interacting with others. Interpretations of the student's actions, rather than the actions themselves, constitute data. The evaluation procedures embody subarguments that also require warrants and backing. These warrants, and the evaluations they ground, are cast in terms of the psychological perspective and the substantive grounding of the application, in concert with the nature of the claim they are meant to support (Messick, 1994). In task-based language testing, for example, for different purposes claims can be organized around evidence of linguistic proficiencies, tendencies to perform in situations with particular features, or targeted capabilities to achieve particular goals in particular kinds of settings (Chapelle, 1998).

Moreover, we note that an assessment argument encompasses three kinds of data:

- aspects of the person's actions in the situation
- aspects of the situation in which the person is acting
- other information about the person's history or relationship to the observational situation

The common view that assessment data are simply item scores is too narrow on several counts. It overlooks the inextricable roles of the other kinds of data in making sense of a performance. It hides the argument for just what aspects of a performance or a product are to be recognized and why—an argument that must be made even when the mechanism to effect the evaluation is just identifying whether a multiple-choice response is correct. It narrows the designer's focus to scoring items rather than scrutinizing performances for clues about what students know or can do in what kinds of situations—which sometimes boils down to scoring items but at other times may require integrative evaluation through theories of expertise as revealed in situated actions (Moss, 1994, 1996). The challenge of evaluating complex performances in an extended patient management problem or making sense of a mass of disparate material in an art portfolio draws this issue to the fore. It quickly forces the assessment designer to think not in terms of a self-contained scoring problem but in terms of the broader design

argument, always with an eye toward particular or potential uses. Bennett and Bejar (1998) wrote, for example, that "a comprehensive discussion of validity and automated scoring includes the interplay among construct definition and test and task design; examinee interface; tutorial; test development tools; automated scoring; and reporting—for in the development process these components affect one another" (p. 9).

A performance arises from the interaction between a person and a situation, and any conception of capability ultimately concerns potential interactions between persons and situations of various kinds. Characterizing assessment situations and use situations through the same lens permits a test developer to distinguish essential features of assessment task and targeted use situations. Building tasks around them at once offers practical guidance and constitutes construct representation validity evidence. Furthermore, the differences between test situations and use situations, and the capabilities entailed by one but not the other, raise theoretically motivated, alternative explanations to be explored empirically in use situations. Such theory-grounded backing for task design, and hence for construct-representation validity arguments, can draw upon cognitive studies variously from the information-processing, expertise research, situative psychology, and sociocultural literatures. Embretson (1998) illustrates this approach with psychological ability tests, for example. Bachman and Palmer (1996) provide practical guidance for task-based language tests that is grounded in psycholinguistic and sociocultural research. Baker (1997) and her colleagues structure tasks around "big ideas" in learning domains. In the Principled Assessment Design for Inquiry project (PADI; see Mislavy & Haertel, 2006; Mislavy & Riconscente, 2006), we developed design patterns for building tasks around key aspects of reasoning in science (Mislavy, Hamel, et al., 2003).

The last of the three kinds of data that appear in assessment design arguments concerns information that is further required to interpret the person's action in the situation, to interpret the situation as it applies to this particular person, or to interpret the aforementioned kinds of data as they pertain to the claim. As a simple example, the reading guidelines of the American Council on the Teaching of Foreign Languages (ACTFL, 1989) contrast intermediate readers' competence with texts "about which the reader has personal interest or knowledge" with advanced readers' comprehension of "texts which treat unfamiliar topics and situations." The same performance would constitute different evidence for a student we knew was familiar with the topic of a text, for another we knew was not familiar, and for yet another whose familiarity was unknown to us. Familiarity affects the weight of the evidence and the validity of the inference, even though it does not appear in test specifications or measurement models. The quality of inferences from assessments depends on both contextual features and the knowledge states of test users, upon which inference is necessarily conditioned.

Information about the intended use, or potential range of supportable uses, of a test influences, in myriad ways, how we should characterize the performance and the situation and what alternative explanations must be considered. These considerations are usually dealt with implicitly, as we build tests to use in classrooms in light of what we know our students to have been studying, what we want to learn about them, and what we will do with the information. In the same way, recommendations for how standardized tests

should be used in light of the conceptions under which they were designed are just as important to the validity of an inference as are the testing materials and procedures. We become aware of these considerations only when something goes wrong, such as differential item functioning, needs for accommodations for students with disabilities, or obviously incorrect outcomes when tests are used in situations other than those for which they were designed. Should we call these additional information considerations internal or external? Do they concern content validation, as I expect Lissitz and Samuelsen would propose; or test validation, as in Wiley's (1991) framework; or construct representation strands of construct validation, in Embretson's (1983) terms; or simply construct validation, because one might interpret all lines of evidence as construct validation? It does not much matter, as long as we do it, and if we have the words, representational forms, and examples to support us, we are likely to do it.

Regarding multiple observations and measurement models: In most assessments we obtain several performances or multiple aspects of complex performances, sometimes all similar to one another, sometimes diverse. (Diversity in tasks allows one to head off different alternative explanations and to observe variation across contexts and characterize it if it is integral to the claim; Chalhoub-Deville, 2003.) All of the observations require warrants as to their character and rationale for their determination. Multiple observable variables, perhaps conditionally dependent, result, all capturing nuggets of evidence to support the claim.

Measurement models can then be used to synthesize the information in terms of variables that indicate the nature and the strength of a claim, expressed in terms of a probability distribution over variables that characterize the target aspects of proficiency in accordance with the grain size, nature, and psychological perspective that is tuned to the assessment's purpose. The familiar notion of adding up item scores to get a test score is again too narrow (Behrens, Mislevy, Bauer, Williamson, & Levy, 2004). Even though in some circumstances this turns out to be exactly the right thing to do from an evidentiary-reasoning perspective (Rasch, 1977), standard practices that have evolved to suit familiar assessments can fall short for more ambitious projects. Probability models need to be designed in coordination with task situations and evaluation procedures from the start to embody an assessment argument that is both internally coherent and tuned to intended uses. Measurement models are invaluable in assessment, but the measurement metaphor alone is too constraining. The models are part of a web of argument, the particular task of which is to characterize support for claims in a context, through a point of view, for a purpose (see Schum, 1994, on the role of probability-based reasoning in evidentiary arguments).

What Should We Call It?

It may seem that I have gone on at some length about assessment design without saying much about validity. But working through a structured design argument and embodying it in the pieces of machinery that constitute an operational assessment accomplishes two goals. The first goal, a practical one, is making the elements of both the argument and its instantiation public, sharable, and reusable. The second is producing explicit construct representation validity argumentation. Validity emerges from design activities for new forms of tests and new development procedures, just as it did with familiar tests and familiar procedures. The

difference, and the contribution of this line of assessment research, is an emerging theory of assessment design: a unified framework, terminology, representations, data structures, and procedures, through which we draw upon advances in contributing fields, to better understand existing tests and create new ones (National Research Council, 2001).

Just 2 years after Cronbach and Meehl (1955) explicated construct validity in terms of nomothetic nets, Loevinger (1957) promoted it to the status of a container: "Since predictive, concurrent, and content validities are essentially *ad hoc*, construct validity is the whole of validity from a scientific point of view" (p. 636; cited in Messick, 1989, p. 17). Continually expanding in scope, it becomes more profound and less useful. Ongoing work by researchers such as Bachman, Baker, Embretson, Gorin (2005), Leighton and Gierl (2007), Luecht (2002), Wiley, Wilson (2005), my colleagues and me, and many others is transforming assessment design into a discipline—which, like many other disciplines, arises in large part from trying to understand the principles, the intuitions, and the work of the best practitioners, in this case test developers.

I find it useful to be able to call attention to this layer of the assessment enterprise, as indeed Figure 1 does visually. Like Lissitz and Samuelsen, I think it is convenient to have a name for these activities as they contribute to validity argumentation. Wiley's (1991) "test validation" operates at about the right level, as do the steps that Lissitz and Samuelsen describe as internal test evaluation procedures. These authors nevertheless do not emphasize the constructive nature of the process as much as I would like; we as test creators are not carrying out validation activities but carrying out design activities structured in such a way that validity evidence emerges. When I need to use a term, I use Embretson's (1983) phrase "construct representation argumentation for construct validity." I am not persuaded by Lissitz and Samuelsen's proposal to promote the term *content validity* to the status of a container to encompass these developments. It has had a century to become understood as something much less.

NOTES

This work is supported by the National Science Foundation under Grant REC-0129331. The opinions expressed are those of the author and not necessarily of the foundation. I am grateful to Geneva Haertel for conversations on which this commentary is based.

¹I use the term *cognitive* broadly, to include insights from what might be called information-processing, situative, developmental, and socio-cultural perspectives (Mislevy, 2006).

²I have been privileged to collaborate over the years with Linda Steinberg and Russell Almond in work such as the article that Lissitz and Samuelsen cited, and with John Behrens, Geneva Haertel, and many others. The views expressed here, however, may not represent theirs.

REFERENCES

- American Council on the Teaching of Foreign Languages. (1989). *ACTFL proficiency guidelines*. Yonkers, NY: Author.
- Bachman, L. F. (2003). Building and supporting a case for test use. *Language Assessment Quarterly*, 2, 1–34.
- Bachman, L. F. (in press). What is the construct? The dialectic of abilities and contexts in defining constructs in language assessment. In J. Fox, M. Wesche, & D. Bayliss (Eds.), *What are we measuring? Language testing reconsidered*. Ottawa, Canada: University of Ottawa Press.
- Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford, UK: Oxford University Press.

- Baker, E. L. (1997). Model-based performance assessment. *Theory Into Practice*, 36, 247–254.
- Behrens, J. T., Mislevy, R. J., Bauer, M., Williamson, D. M., & Levy, R. (2004). Introduction to evidence centered design and lessons learned from its application in a global e-learning program. *International Journal of Testing*, 4, 295–301.
- Bennett, R. E., & Bejar, I. I. (1998). Validity and automated scoring: It's not only the scoring. *Educational Measurement: Issues and Practice*, 17(4), 9–17.
- Borsboom, D., Mellenbergh, G., & van Heerden, J. (2004). The concept of validity. *Psychological Review*, 111, 1061–1071.
- Brindley, G. (1994). Task-centred assessment in language learning: The promise and the challenge. In N. Bird, P. Falvey, A. Tsui, D. Allison, & A. McNeill (Eds.), *Language and learning: Papers presented at the Annual International Language in Education Conference, Hong Kong, 1993* (pp. 73–94). Hong Kong, China: Hong Kong Education Department.
- Chalhoub-Deville, M. (2003). Second language interaction: Current perspectives and future trends. *Language Testing*, 20, 369–383.
- Chapelle, C. A. (1998). Construct definition and validity inquiry in SLA research. In L. F. Bachman & A. D. Cohen (Eds.), *Interfaces between second language acquisition and language testing research* (pp. 32–70). New York: Cambridge University Press.
- Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52, 281–302.
- Embretson, S. E. (Whitley) (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin*, 93(1), 179–197.
- Embretson, S. E. (Ed.). (1985). *Test design: Developments in psychology and psychometrics*. Orlando, FL: Academic Press.
- Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods*, 3, 380–396.
- Frederiksen, J. R., & Collins, A. (1989). A systems approach to educational testing. *Educational Researcher*, 18(9), 27–32.
- Gorin, J. S. (2005). Manipulating processing difficulty of reading comprehension questions: The feasibility of verbal item generation. *Journal of Educational Measurement*, 42, 351–373.
- Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin*, 112, 527–535.
- Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: American Council on Education/Praeger.
- Leighton, J. P., & Gierl, M. J. (Eds.). (2007). *Cognitive diagnostic assessment for education: Theories and applications*. Cambridge, UK: Cambridge University Press.
- Lennon, R. T. (1956). Assumptions underlying the use of content validity. *Educational and Psychological Measurement*, 16, 294–304.
- Lesgold, A. M., Lajoie, S. P., Bunzo, M., & Eggen, G. (1992). Sherlock: A coached practice environment for an electronics troubleshooting job. In J. H. Larkin & R. W. Chabay (Eds.), *Computer-assisted instruction and intelligent tutoring systems* (pp. 202–274). Hillsdale, NJ: Lawrence Erlbaum.
- Lindquist, E. F. (1951). Preliminary considerations in objective test construction. In E. F. Lindquist (Ed.), *Educational measurement* (pp. 119–158). Washington, DC: American Council on Education.
- Lissitz, R. W., & Samuelsen, K. (2007). A suggested change in terminology and emphasis regarding validity and education. *Educational Researcher*, 36, 437–448.
- Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports*, 3(Monograph Supplement 9), 635–694.
- Luecht, R. M. (2002). *From design to delivery: Engineering the mass production of complex performance assessments*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.
- McNamara, T. (2006). Validity in language testing: The challenge of Sam Messick's legacy. *Language Assessment Quarterly*, 3, 31–51.
- Melnick, D. (1996). The experience of the National Board of Medical Examiners. In E. L. Mancall, P. G. Vashook, & J. L. Dockery (Eds.), *Computer-based examinations for board certification* (pp. 111–120). Evanston, IL: American Board of Medical Specialties.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: American Council on Education/Macmillan.
- Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, 23(2), 13–23.
- Mislevy, R. J. (1994). Evidence and inference in educational assessment. *Psychometrika*, 59, 439–483.
- Mislevy, R. J. (2003). Substance and structure in assessment arguments. *Law, Probability, and Risk*, 2, 237–258.
- Mislevy, R. J. (2006). Cognitive psychology and educational assessment. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 257–305). Westport, CT: American Council on Education/Praeger.
- Mislevy, R. J., Almond, R. G., & Steinberg, L. S. (2002). On the roles of task model variables in assessment design. In S. Irvine & P. Kyllonen (Eds.), *Generating items for cognitive tests: Theory and practice* (pp. 97–128). Hillsdale, NJ: Lawrence Erlbaum.
- Mislevy, R. J., & Haertel, G. (2006). Implications for evidence-centered design for educational assessment. *Educational Measurement: Issues and Practice*, 25, 6–20.
- Mislevy, R., Hamel, L., Fried, R. G., Gaffney, T., Haertel, G., Hafter, A., et al. (2003). *Design patterns for assessing science inquiry* (PADI Technical Report 1). Menlo Park, CA: SRI International.
- 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: Lawrence Erlbaum.
- Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, 1, 3–67.
- Mislevy, R. J., Steinberg, L. S., Breyer, F. J., Johnson, L., & Almond, R. G. (2002). Making sense of data from complex assessments. *Applied Measurement in Education*, 15, 363–378.
- Moss, P. A. (1994). Can there be validity without reliability? *Educational Researcher*, 23(2), 5–12.
- Moss, P. A. (1996). Enlarging the dialogue in educational measurement: Voices from interpretive research traditions. *Educational Researcher*, 25(1), 20–28, 43.
- National Research Council. (2001). *Knowing what students know: The science and design of educational assessment*. Committee on the Foundations of Assessment, J. Pellegrino, R. Glaser, & N. Chudowsky (Eds.). Washington DC: National Academy Press.
- Rasch, G. (1977). On specific objectivity: An attempt at formalizing the request for generality and validity of scientific statements. *Danish Yearbook of Philosophy*, 14, 58–94.
- Roid, G. H., & Haladyna, T. M. (1982). *A technology for test-item writing*. New York: Academic Press.
- Rulon P. J. (1946). On the validity of educational tests. *Harvard Educational Review*, 16, 290–296.
- Schum, D. A. (1987). *Evidence and inference for the intelligence analyst*. Lanham, MD: University Press of America.
- Schum, D. A. (1994). *The evidential foundations of probabilistic reasoning*. New York: John Wiley.
- Toulmin, S. E. (1958). *The uses of argument*. Cambridge, UK: Cambridge University Press.

- Toulmin, S. E. (2001). *Return to reason*. Cambridge, MA: Harvard University Press.
- Wigmore, J. H. (1937). *The science of judicial proof* (3rd ed.). Boston: Little, Brown.
- Wiley, D. E. (1991). Test validity and invalidity reconsidered. In R. Snow & D. E. Wiley (Eds.), *Improving inquiry in social science* (pp. 75–107). Hillsdale, NJ: Lawrence Erlbaum.
- Wilson, M. R. (2005). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Lawrence Erlbaum.

AUTHOR

ROBERT J. MISLEVY is a professor at the University of Maryland, College Park, Department of Measurement, Statistics and Evaluation, 1230-C Benjamin Building, College Park, MD 20742; rmislevy@umd.edu. In his research he seeks to apply developments in statistics, psychology, and technology to practical problems in educational assessment.

Manuscript received August 23, 2007

Accepted September 17, 2007