Educ Psychol Rev (2011) 23:221–234
DOI 10.1007/s10648-011-9161-z

ESSAY

A Model of Cognition: The Missing Cornerstone of Assessment

Nathaniel J. S. Brown · Mark Wilson

Published online: 15 May 2011

Abstract When we rely upon gains on some measure to support statements of prescription, we have the obligation to ensure that those measures are valid. Nearly 10 years after an influential National Research Council (2001) report on educational assessment identified an explicit model of cognition as one of three necessary components of a valid assessment system, we note that most measures still lack this fundamental cornerstone. In this paper, we draw attention to the *construct modeling* approach to assessment that strives for coherence and consistency with a model of cognition in which student proficiency varies along a continuum of competence. This approach is illustrated in the context of an assessment of conceptual understanding of certain scientific phenomena given to undergraduates at a large public university (National Research Council 2001).

Keywords Cognition · Assessment · Validity · Construct modeling

When we rely upon gains on some measure to support statements of prescription, we have the obligation to ensure that those measures are valid. The National Research Council (2001) report *Knowing What Students Know* described three necessary components of a valid assessment system: "a model of student *cognition* and learning in the domain, a set of beliefs about the kinds of *observations* that will provide evidence of students' competencies, and an *interpretation* process for making sense of the evidence" (p. 44). We note that 10 years later, most measures still lack the first leg of the stool: an explicit model of cognition.

N. J. S. Brown (✉)
Learning Sciences, School of Education, Indiana University, 201 N. Rose Ave. Rm. 4020, Bloomington, IN 47405-1005, USA
e-mail: njsbrown@indiana.edu

M. Wilson
Berkeley Evaluation and Assessment Research (BEAR) Center, Graduate School of Education, University of California, Berkeley 94720, USA
e-mail: MarkW@berkeley.edu

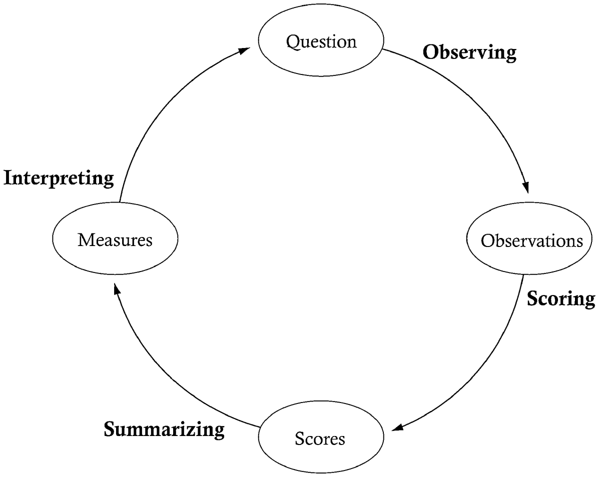
Cornerstones of Assessment

Expanding upon the National Research Council (NRC 2001) model, we characterize the practice of assessment as a cyclical process involving four steps (Fig. 1). Any cycle of assessment starts with a *question*, the answer to which will be a ranking (as in norm referencing), characterization (as in criterion referencing), or decision (as in establishing cut points). Each of these answers requires us to *measure* something—to determine the values of one or more latent variables representing, for example, abilities, attitudes, aptitudes, activations, or effectivities. The four steps in the cycle are (1) *observing*: eliciting performances assumed to depend upon the latent variable(s), leading to a set of *observations*; (2) *scoring*: categorizing different observed performances and assigning them relative value, or *scores*; (3) *summarizing*: combining the values of the individual performances to yield *measures* of each latent variable, and (4) *interpreting*: using the measures of the latent variable(s) to answer the question.

In order to carry out the practice of assessment, tools are required to mediate and accomplish each of these four steps (Fig. 2). Observations are elicited by designed *tasks*, scores are assigned with the aid of a set of *scoring guides*, summary measures are the result of applying a *measurement model* to the individual scores, and interpretation of the measures depends upon a *model of cognition*. Together, these four are the cornerstones of any assessment instrument (Fig. 3).

This cornerstone model of assessment practice is intended to be sufficiently general to accommodate a wide range of assessment purposes and models of cognition. The purpose of assessment may be formative or summative and may provide information for immediate, close, proximal, distal, or remote stakeholders (Hickey *et al.* 2006; Ruiz-Primo *et al.* 2002). A particular model of cognition may be unidimensional or multidimensional, involve variables that are continuous or discrete and bounded or unbounded, locate cognition within individuals, groups, or activity systems, and represent cognitive processes ranging from simple recall to ill-defined real-world problem solving.

Fig. 1 The four steps in a cycle of assessment: observing, scoring, summarizing, and interpreting



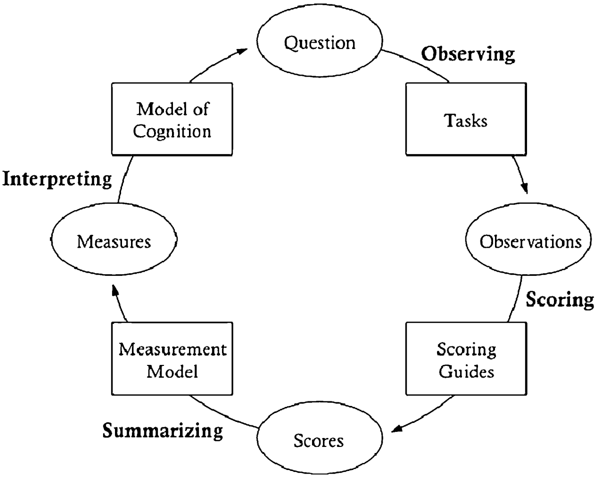


Fig. 2 The four cornerstones of assessment—a model of cognition, tasks, scoring guides, and a measurement model—each mediating one of the four steps in an assessment cycle. Image source: Brown, N. J. S., Nagashima, S. O., Fu, A., Timms, M., & Wilson, M. (2010). A framework for analyzing scientific reasoning in assessments. *Educational Assessment, 15*, 142–174. Reprinted by permission of the publisher, Taylor & Francis Ltd

Instrument Development and Validity

Instruments are generally developed following the path shown in Fig. 4. In this path, items are developed first with the guidance of content experts. After the items are written, scoring guides are developed to score possible responses. The items are then pilot tested with a sample from the intended population. Their raw scores are subsequently analyzed to produce scale scores. At this point, items are analyzed using statistical methods to examine their measurement properties. Assuming the item statistics are satisfactory, the instrument is considered finished. When time and money permit, or when poor results demand, a second development cycle begins using the results of the calibration to suggest revisions to the items. Note that throughout this process, the model of cognition is marginalized or ignored.

While cyclical, this model of a development cycle is fundamentally orderly in that the development of each successive cornerstone depends primarily, if not solely, upon the cornerstone that immediately proceeds it in the cycle. Likewise, current conceptions of validity that focus on gathering evidence to support inferences within an interpretive

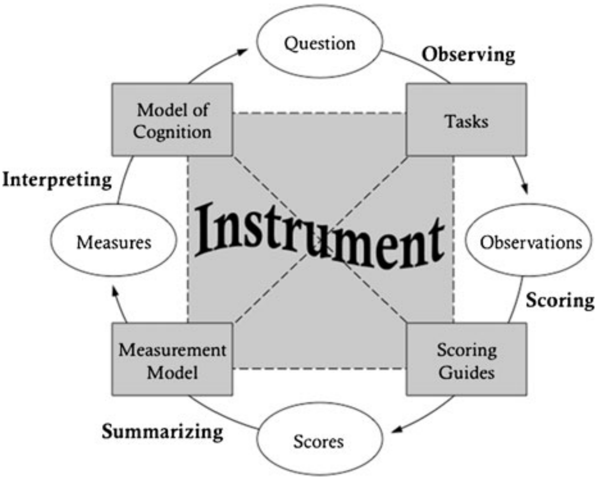


Fig. 3 A coherent measurement instrument comprised of the four cornerstones of assessment

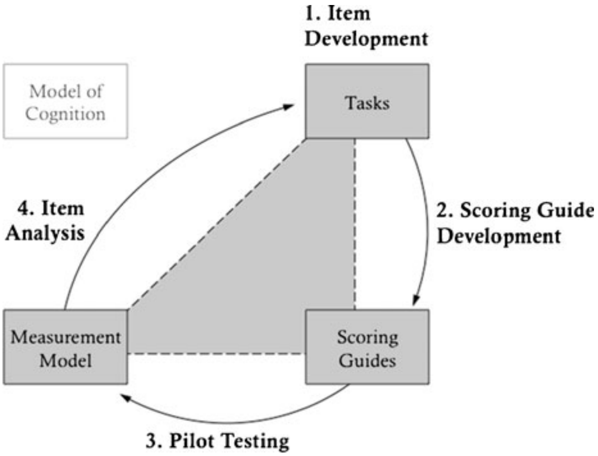


Fig. 4 The four steps of traditional instrument development and validation, consisting of an initial item development followed by a repeating cycle of scoring guide development, pilot testing, and item analysis

argument (American Educational Research Association *et al*. 1999; Cronbach 1988; Kane 2001; Messick 1989; Mislevy 1996) implicitly associate forms of evidence with single cornerstones and the assessment stages they mediate. For example, consider the development of a typical assessment consisting of a set of multiple-choice items. The scoring guide for each item is the list of possible choices—the correct answer and the distracters—presented to the student. This list is typically generated with the items, and only the items, in mind, drawing upon a task decomposition provided by content experts and/or actual student responses to the items. Consideration is given to whether the scoring guide accurately represents the sorts of things that students are likely to say and whether those responses accurately indicate the presumed underlying cognitive processes. In the most recent *Standards for educational and psychological testing* (American Educational Research Association *et al*. 1999), this is referred to as evidence based on response processes. In the eight-stage validation model of Crooks *et al*. (1996), this is referred to as the scoring link (see also Kane 2006). In Messick's (1995) construct framework, this is referred to as the substantive aspect. During this stage in the development and validation cycle, the connection between the tasks and the scoring guides is foregrounded, while the measurement model and the model of cognition are backgrounded.

We take issue with this orderly approach, arguing instead that all four cornerstones should be foregrounded at each stage of development and validation. Instead of focusing on the current and immediately prior cornerstone, developers should consider the coherence of the entire instrument as they develop cornerstones that are explicitly tied to the others. In our experience, the development and pilot testing of each cornerstone reveal potential problems and sources of incoherence that require tweaking to most, if not all, the cornerstones. Consistent with the recommendations of the NRC (2001) report, we believe the *relationships* between the cornerstones, illustrated in Fig. 3 by the dashed lines, are the true foundations of internal validity.[1] Of particular importance is the consistency

[1] By using the term *internal validity*, we mean to exclude for the moment discussion of evidence based on relations to other variables and evidence based on consequences of testing (American Educational Research Association *et al*. 1999).

between the tasks, scoring guides, and measurement model and an explicit model of cognition.

The Missing Cornerstone: A Model of Cognition

The model of cognition "refers to a theory or set of beliefs about how students represent knowledge and develop competence in a subject domain" (National Research Council 2001, p. 44). Describing what he calls *standard test theory*, Mislevy (1996) argues that most educational measurement lacks such a model of cognition. Instead, standard uses of classical test theory, generalizability theory, item response theory, and factor analysis or structural equation modeling are bound up in a paradigm of behavioral psychology in which the target of interest is "a summary of a behavioral tendency in a class of stimulus situations—an overall proficiency in the prescribed domain of tasks" (Mislevy 1996, pp. 386–387). As a consequence, the valid interpretations of test scores are limited:

> Through the use of standard test theory, evidence can be characterized and brought to bear on inferences about students' overall proficiency in behavioral domains, for determining students' levels of proficiency, comparing them to other students or to a standard, or gauging changes from one point in time to another. Conjectures about the nature of this proficiency or how it develops falls largely outside the mental measurement paradigm's universe of discourse (Mislevy 1996, p. 388).

In educational assessment, and in classroom assessment particularly, the nature of the proficiency and how it develops is of critical importance. Having a model of how students represent knowledge and develop competence strongly benefits curriculum, instruction, and evaluation (National Research Council 2001). In terms of curriculum, a model of cognition supports the development of more effective sequences of content delivery and classroom activities, consistent with how material is more naturally, more accessibly, or more efficiently learned. In terms of instruction, a model of cognition permits the meaningful interpretation of assessment results in terms of what students know, guiding choices about what needs to be reviewed and what instruction should target next. In terms of evaluation, a model of cognition ensures that the competencies we value go beyond behavioral tendencies to include constructs such as higher-order cognitive processes, new literacies, and twenty-first century skills (Jenkins 2009).

In this paper, we draw attention to an assessment practice—*construct modeling* (Wilson 2005)—in which the model of cognition is defined by one or more proficiencies represented by unbounded continuous latent variables. This is a relatively simple yet flexible and powerful model of cognition. For each proficiency, values vary along a continuum from intuition (large negative values of proficiency) to expertise (large positive values of proficiency). In this model, focus is shifted away from disconnected and individually acquired knowledge and skills. Instead, learning is conceived as progress toward higher levels of sophistication and competence as new knowledge is linked to existing knowledge and deeper understandings are developed. The growing importance of this perspective is evident behind the recent interest in learning progressions in science education (National Research Council 2006, National Research Council 2007; Smith *et al.* 2006; Songer *et al.* 2009; Wilson 2009).

Construct modeling involves the explicit design, implementation, and evaluation of the four cornerstones described above. Each of these cornerstones, or *building blocks* (Wilson 2005), is designed to embody and be consistent with the model of cognition described above (Fig. 5).

This approach is demonstrated below in the next few sections. The context is an assessment of conceptual understanding of certain scientific phenomena involving dynamic equilibrium, such as chemical phase equilibrium and solubility. This assessment, called the Multidimensional Measure of Conceptual Complexity (MMCC), was given to 103 undergraduates at a large public university taking one of three chemistry courses: general chemistry for non-majors, general chemistry for majors, and organic chemistry (Brown 2005).

Construct maps

A core principle of construct modeling is that assessment design should begin not with item creation, but with the development of a *construct map*. Construct maps are representations of models of cognition by which the results of the assessment can be interpreted. Without construct maps in place from the beginning, assessment designers are handicapped with a largely implicit model of cognition and no clear guidance for item developers on how to create assessments with construct validity. A key insight here is that the *content* experts to whom we turn are often not *construct* experts. Cognitive research has shown that experts are expert in part because they see past the surface features of their domain that are most salient for novices (Chase and Simon 1973; Chi *et al.* 1981). Consequently, while possessing content knowledge, experts often lack sufficient pedagogical content knowledge about how their domain is understood and best learned by novices (Shulman 1986, 1987). This makes it difficult for them to design easier items that more effectively measure the proficiency of beginning students. As such, a construct map developed with the help of teachers, educational psychologists, professionals in related areas, and measurement experts, in addition to content experts, is a key means for establishing the construct validity of the assessment.

Having undergone several iterations during development, one of the construct maps for the MMCC is shown in Fig. 6. Note the bidirectional arrow representing greater and fewer amounts of the construct in question: conceptual depth.

Items design

An items design is an explicit framework for designing tasks that most effectively elicit evidence of the latent proficiency, as represented by the construct maps. The items design is

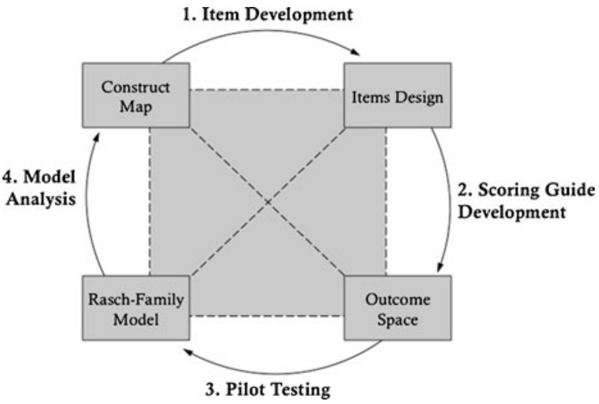
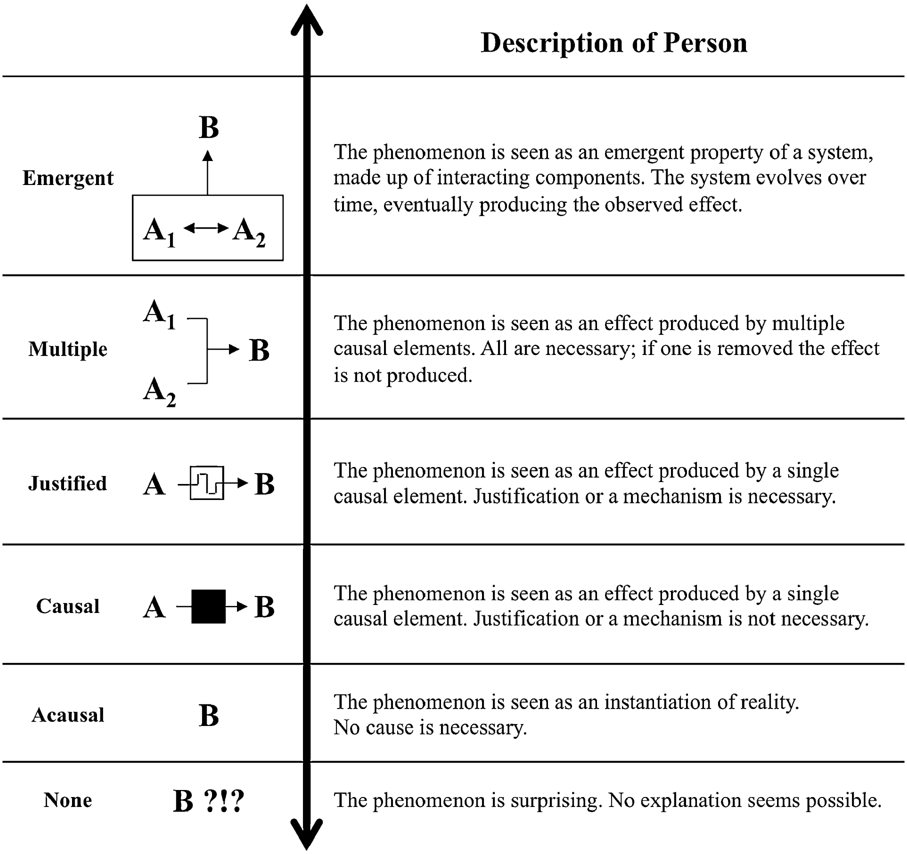


Fig. 5 The four steps of assessment development in the construct modeling approach, consisting of a repeating cycle of item development, scoring guide development, pilot testing, and model analysis



		Description of Person
Emergent	B $A_1 \leftrightarrow A_2$	The phenomenon is seen as an emergent property of a system, made up of interacting components. The system evolves over time, eventually producing the observed effect.
Multiple	A_1, A_2 → B	The phenomenon is seen as an effect produced by multiple causal elements. All are necessary; if one is removed the effect is not produced.
Justified	A → B	The phenomenon is seen as an effect produced by a single causal element. Justification or a mechanism is necessary.
Causal	A ■→ B	The phenomenon is seen as an effect produced by a single causal element. Justification or a mechanism is not necessary.
Acausal	B	The phenomenon is seen as an instantiation of reality. No cause is necessary.
None	B ?!?	The phenomenon is surprising. No explanation seems possible.

Fig. 6 A construct map illustrating the model of cognition associated with the construct conceptual depth, comprised of six ordered, qualitatively distinct kinds of cognition

a tool that structures observing so that the observations are most likely to be useful, informative, and easily interpretable with respect to the latent proficiency.

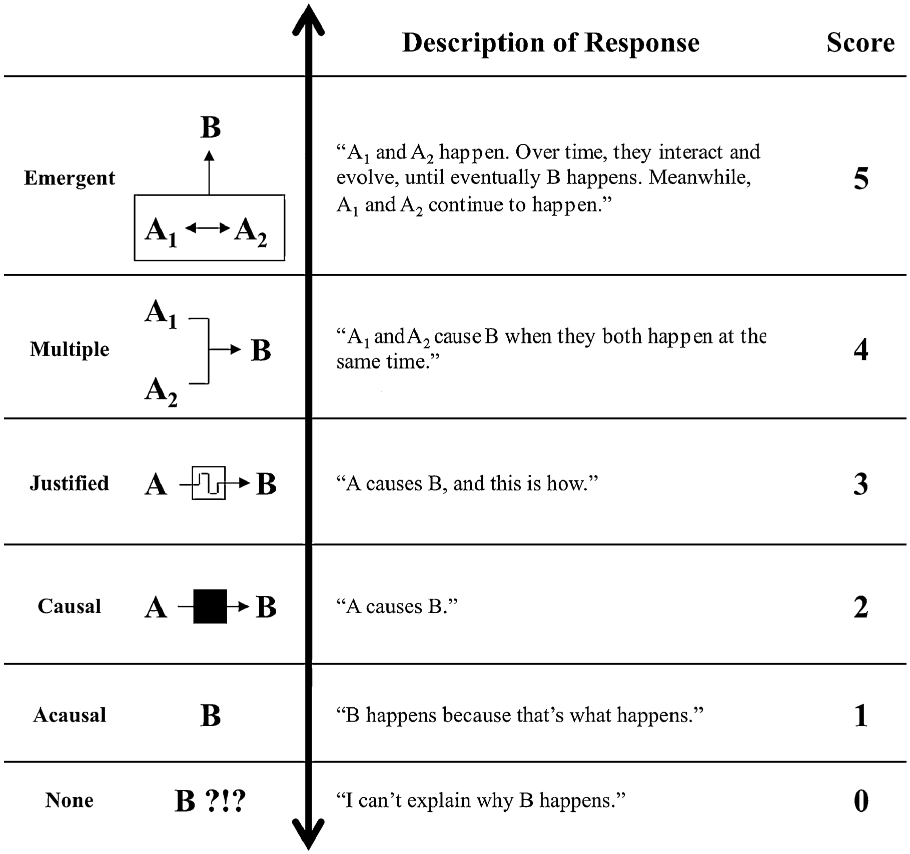
The MMCC used open-ended items that were designed to elicit responses at all levels of the construct. Open-ended items are not a requirement of the construct modeling approach, however. Ordered multiple-choice items are an attempt to explicitly connect scoring guides to the model of cognition by designing distracters that map to specific regions of the construct (Briggs *et al.* 2006). As Briggs *et al.* have demonstrated, this coherence has several benefits: such items typically provide more information than regular multiple-choice items and exhibit greater validity and reliability.

Outcome spaces

Outcome spaces describe in detail the qualitatively different kinds of student response elicited by the items and map these categories of response to the levels of the construct maps. In doing so, outcome spaces are tools that facilitate the process of “scoring”—i.e., categorizing and valuing—student responses, ensuring that scores are meaningfully related to the latent proficiency. For the conceptual depth construct in the MMCC, an outcome space was developed using an iterative process in which the levels of the construct map (Fig. 6) were reconciled with a phenomenographic analysis (Marton 1986) of actual student responses to the items (Fig. 7).

Measurement model

We choose as our measurement approach the Rasch model and its associated family of models because they have features that we believe to be particularly suitable for use in the construct mapping formulation. The Rasch model family parameterizes the persons and items in a way that is directly interpretable in terms of the construct map and, hence, the model of cognition described above. Although more complex item response models such as the 2PL and 3PL models allow a somewhat more detailed modeling of the items (while having the same model for the person), their results cannot be readily expressed using the visual representation of the empirical construct map we call a Wright map. The Wright map, as we will illustrate below, is central to the interpretation and ease of use of the assessment results. The cost of using the Rasch model is that more items need to be developed in the first case, as fewer will be found to fit the simpler (and hence, more rigorous) Rasch model.



	Description of Response	Score
Emergent	"A_1 and A_2 happen. Over time, they interact and evolve, until eventually B happens. Meanwhile, A_1 and A_2 continue to happen."	**5**
Multiple	"A_1 and A_2 cause B when they both happen at the same time."	**4**
Justified	"A causes B, and this is how."	**3**
Causal	"A causes B."	**2**
Acausal	"B happens because that's what happens."	**1**
None	"I can't explain why B happens."	**0**

Fig. 7 An outcome space for the construct conceptual depth, comprised of six ordered, qualitatively distinct types of response scored from 0 to 5

However, we have found this to be a relatively small burden in development, as many items are modified or discarded during development for a multitude of reasons beyond this issue of model fit.

The polytomous response data from the MMCC were modeled using Masters' (1982) partial credit model (PCM). The PCM is a submodel of the multidimensional random coefficients multinomial logit model, which accommodates a wide range of Rasch family models for which item response probabilities can be modeled by a set of random variables usually associated with person proficiencies and fixed parameters usually associated with item difficulties (Adams *et al*. 1997).

The proficiency of each person (n=1, …, 103) was modeled by a random variable, θ_n. A set of I=9 items (i=1, …, I) each permits the completion of K_i steps (k=1, …, K_i) between response category k–1 and the next highest response category k. The maximum possible K_i for each item, K, is K=5; the actual value of K_i is less for items for which not all possible response categories were observed in the data. For each item, the difficulty of each step was modeled by a fixed parameter, δ_{ik}.

The item response probability model, which models the probability of person n achieving score $X_{in} = k$ on item i, is given by:

$$P(X_{in} = k) = \frac{\exp\left(k\theta_n - \sum\limits_{j=0}^{k} \delta_{ij}\right)}{\sum\limits_{h=0}^{K_i} \exp\left(h\theta_n - \sum\limits_{j=0}^{h} \delta_{ij}\right)},$$

where $\delta_{i0}\equiv 0$. The sum of the average of the step difficulty parameters for each item,

$$\sum_{i=1}^{I} \frac{1}{K_i} \sum_{j=1}^{K_i} \delta_{ij},$$

was defined to be zero to identify a unique set of estimated item difficulties.

Estimation

Step parameters were estimated using a marginal maximum likelihood procedure (Adams *et al*. 1997). The distribution of the random variables representing person proficiencies was assumed to be normal with mean μ and variance σ^2. Subsequently, weighted least-mean-square estimates of θ_n were calculated (Adams *et al*. 1997). These estimates and their standard errors were used to calculate a person separation reliability of r=0.7 (Wright and Masters 1982). Weighted mean-square fit statistics and associated t values were calculated for all estimates (Adams and Wu 2011; Wu 1997). The software ConQuest (Wu *et al*. 2007) was used for all calculations.

Results

When items, scoring guides, and a measurement model are designed and chosen that are consistent with the proficiency model of cognition, it becomes possible to illustrate all of the cornerstones within a single diagram, called a Wright map (Fig. 8). A Wright map plots person proficiency and step difficulty estimates on their common interval scale. Estimates of person proficiencies on the scale are represented by the histogram on the left side of the

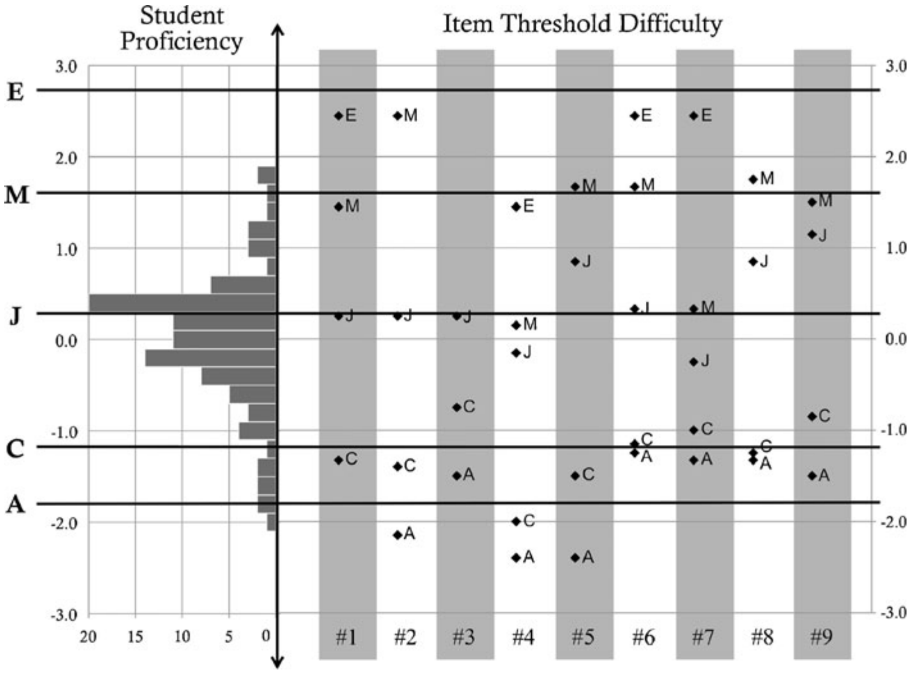


Fig. 8 A Wright map displaying the results of analyzing data from the Multidimensional Measure of Conceptual Complexity for the construct conceptual depth. The unidimensional, scaled construct runs vertically through the figure represented by the *double-headed arrow* and the units (in logits) running from −3.0 to 3.0. Person proficiencies are displayed in the histogram to the *left of the scale*. Item difficulties are displayed to the *right of the scale*, with the Thurstonian thresholds for each of the nine items displayed in a separate column. *Capital letters* correspond to the levels of the construct map (and outcome space): *A* marks the threshold between the None and Acausal levels; *C* marks the threshold between the Acausal and Causal levels; *J* marks the threshold between the Causal and Justified levels; *M* marks the threshold between the Justified and Multiple levels; and *E* marks the threshold between the Multiple and Emergent levels. The *thick horizontal lines* are weighted averages of each threshold across all items, illustrating the overall relation to the construct map

figure. For ease of interpretation, step difficulty estimates have been converted into Thurstonian thresholds (Wu *et al.* 2007). Each threshold marks the proficiency above which a person is more likely to give a response in category *k* or higher ($k, \ldots, K_i$) than a response in any lower category (0, …, *k*–1). For each item, Thurstonian thresholds appear in a separate column on the right side of the figure. Thresholds are labeled with codes corresponding to the level names in the outcome space (Fig. 7). In general, people near the bottom of the scale are likely to give poor responses on all the items, while people near the top are likely to give proficient responses on all the items. From the item perspective, response categories represented by thresholds near the bottom of the scale are likely to be achieved by nearly all the people, while response categories represented by thresholds near the top are achieved relatively infrequently.

A Wright map allows teachers and researchers to easily make valid inferences about students' proficiency. In the example, the majority of students are located above the Causal threshold and span the Justified threshold. This means that they are able to give causal explanations for the scientific phenomena on the assessment and are struggling to justify those explanations with more than one step. The most proficient students are struggling to

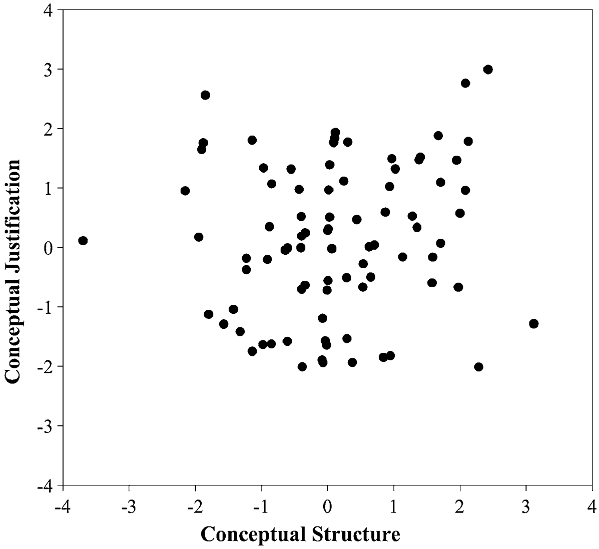


Fig. 9 Person proficiency estimates for the two-dimensional space defined by the conceptual structure and conceptual justification constructs (r=0.11)

give explanations with multiple causal elements and, as represented by their position below the Emergent threshold, have difficulty giving system-related explanations. There are exceptions, and these too are noted on the Wright map. For example, almost half of the students were able to construct an explanation with multiple causal elements for Item #4, and the most proficient students are able to give emergent explanations. This item dealt with the evaporation of water, a canonical system in undergraduate chemistry—consequently, it should be no surprise that students were more familiar with this system and more able to delve deeper with their explanations.

Improving Validity

In the traditional model of instrument development and validation (Fig. 4), the fourth step focuses on the relationship between the measurement model and the tasks. Because of this focus, this process is usually called item analysis and positive results (e.g., adequate point-biserial correlations) are often considered sufficient evidence of validity based on internal structure. When all four cornerstones are present and are designed to be coherent,

Table 1 Measurement models of the Rasch family appropriate for various models of cognition

Aspects of a model of cognition	Rasch family measurement model
Binary choices	Simple logistic (Rasch 1980)
Ratings	Rating scale (Andrich 1978a, b)
Multiple steps	Partial credit (Masters 1982)
Multiple strategies	Ordered partition model (Wilson 1992)
Multiple sub-tasks	Many facets (Fischer 1983; Linacre 1990)
Dependencies between tasks	Item bundles (Wilson and Adams 1995)
Developmental stages	Saltus (Wilson 1989)

additional opportunities to strengthen the evidence for the validity of the system become available. For example, the results for the measurement model should be consistent with the model of cognition. In the construct modeling approach, this means, among other things, that the item parameter estimates should fall in the order predicted by the construct map (and have satisfactory fit statistics too). In the example above, the item fit statistics were satisfactory. However, approximately 20% of the respondents were fitting poorly. Examination of their response patterns revealed that these students were of average proficiency yet were not exhibiting any Justified responses, only Causal, Multiple, and even Emergent. This conflict with the model of cognition led to a refinement of the cognitive model, splitting apart the proficiencies of conceptual composition (the original depth construct minus the Justified level) and conceptual justification (involving the number of steps in the causal chain).

To model this multidimensional set of proficiencies, a confirmatory multidimensional item response model was used (Adams *et al*. 1997). The results (Fig. 9) illustrated that the two proficiencies—conceptual composition and conceptual justification—are nearly uncorrelated (r=0.11), a striking result and one that would not have been discovered had the data not been analyzed with an explicit model of cognition in mind, one that can be compared with the other cornerstones of assessment and revised if necessary.

Discussion

Another notable effort to better synchronize tasks, scoring guides, and measurement models with explicit models of cognition is evidence-centered assessment design (Mislevy and Riconscente 2006; Mislevy *et al*. 2002a). While the evidence-centered approach is more often associated with task analysis and the identification of item facets and/or production rules (Mislevy 1996), we note in passing that this approach is compatible with a construct modeling formulation (Mislevy *et al*. 2002b). Indeed, one of the benefits of construct modeling as an assessment practice is its flexibility. In addition to the multidimensional models described above, the construct modeling framework is comprehensive and allows for specification of a variety of models given varied task formats and scoring assignments. For example, the models in Table 1 are available, representing a wide range of models of cognition. Importantly, any of these models can be combined to create hybrid models under a generalized Rasch modeling framework (Adams *et al*. 1997; Embretson and Reise 2000; De Boeck and Wilson 2004). This gives instrument designers the freedom to design an appropriate model of cognition, based on research on how students represent knowledge and develop competence. Whichever model of cognition one chooses, the search for coherence between the four cornerstones of assessment is the key to maximizing sources of evidence that support valid statements of prescription.

References

Adams, R., & Wu, M. (2011). The construction and implementation of user-defined fit tests for use with marginal maximum likelihood estimation and generalised item response models. In N. J. S. Brown, B. Duckor, K. Draney, & M. Wilson (Eds.), *Advances in Rasch measurement* (Vol. 2). Maple Grove: JAM Press.

Adams, R., Wilson, M., & Wang, W. (1997). The multidimensional random coefficients multinomial logit model. *Appl Psychol Meas, 21*, 1–23.

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.

Andrich, D. (1978a). Application of a psychometric rating model to ordered categories which are scored with successive integers. *Appl Psychol Meas, 2*, 581–594.

Andrich, D. (1978b). A rating formulation for ordered response categories. *Psychometrika, 43*, 561–573.

Briggs, D. C., Alonzo, A. C., Schwab, C., & Wilson, M. (2006). Diagnostic assessment with ordered multiple-choice items. *Educ Assess, 11*, 33–63.

Brown, N. J. S. (2005). *The multidimensional measure of conceptual complexity.* Berkeley: University of California.

Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive Psychology, 4*, 55–81.

Chi, M. T. H., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science, 5*, 121–152.

Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 3–17). Hillsdale: Lawrence Erlbaum.

Crooks, T. J., Kane, M. T., & Cohen, A. S. (1996). Threats to the valid use of assessments. *Assessment in Education: Principles, Policy & Practice, 3*, 265–285.

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

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: LEA.

Fischer, G. (1983). Logistic latent trait models with linear constraints. *Psychometrika, 48*, 3–26.

Hickey, D. T., Zuiker, S. J., Taasoobshirazi, G., Schafer, N. J., & Michael, M. A. (2006). Balancing varied assessment functions to attain systemic validity: Three is the magic number. *Studies in Educational Evaluation, 32*, 180–201.

Jenkins, H. (2009). *Confronting the challenges of participatory culture: Media education for the 21st century.* Cambridge: Massachusetts Institute of Technology.

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement, 38*, 319–342.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Lanham: Rowman & Littlefield.

Linacre, J. (1990). *Many-facet Rasch measurement.* Chicago: MESA Press.

Marton, F. (1986). Phenomenography: A research approach to investigating different understandings of reality. *Journal of Thought, 21*, 29–49.

Masters, G. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149–174.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher, 18*, 5–11.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *The American Psychologist, 50*, 741–749.

Mislevy, R. J. (1996). Test theory reconceived. *Journal of Educational Measurement, 33*, 379–416.

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

Mislevy, R. J., Steinberg, L. S., Breyer, F. J., Almond, R. G., & Johnson, L. (2002a). Making sense of data from complex assessments. *Applied Measurement in Education, 15*, 363–389.

Mislevy, R. J., Wilson, M. R., Ercikan, K., & Chudowsky, N. (2002b). Psychometric principles in student assessment. CSE Technical Report. Los Angeles, CA: Center for the Study of Evaluation, University of California.

National Research Council. (2001). *Knowing what students know: The science and design of educational assessment*. Committee on the Foundations of Assessment. J. Pellegrino, N. Chudowsky, & R. Glaser (Eds). Washington, DC: National Academies Press.

National Research Council. (2006). *Systems for state science assessment.* Committee on Test Design for K-12 Science Achievement. M. Wilson, & M. Bertenthal (Eds.). Washington, D.C.: National Academies Press.

National Research Council. (2007). *Taking Science to School: Learning and Teaching Science in Grades K-8*. Committee on Science Learning, Kindergarten through Eighth Grade. R. A. Duschl, H. A. Schweingruber, & A. W. Shouse (Eds.). Washington, D.C.: National Academies Press.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. Chicago: University of Chicago Press. Original work published in 1960.

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.

Shulman, L. S. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher, 15*, 4–14.

Shulman, L. S. (1987). Knowledge and teaching: Foundations of the new reform. *Harvard Educational Review, 57*, 1–22.

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. *Measurement, 14*, 1–98.

Songer, N. B., Kelcey, B., & Gotwals, A. (2009). How and when does complex reasoning occur? Empirically driven development of a learning progression focused on complex reasoning about biodiversity. *Journal of Research in Science Teaching, 46*, 610–631.

Wilson, M. (1989). Saltus: A psychometric model of discontinuity in cognitive development. *Psychological Bulletin, 105*, 276–289.

Wilson, M. (1992). The ordered partition model: An extension of the partial credit model. *Applied Psychological Measurement, 16*, 309–325.

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah: Lawrence Erlbaum.

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

Wilson, M., & Adams, R. (1995). Rasch models for item bundles. *Psychometrika, 60*, 181–198.

Wright, B., & Masters, G. (1982). *Rating scale analysis: Rasch measurement*. Chicago: MESA Press.

Wu, M. (1997). *The development and application of a fit test for use with marginal maximum estimation and generalized item response models*. Master's thesis, Victoria, Australia: University of Melbourne

Wu, M., Adams, R., Wilson, M., & Haldane, S. (2007). *ACER ConQuest version 2.0: Generalised item response modelling software [Computer software and manual]*. Camberwell: ACER Press.