Psicología Educativa 20 (2014) 117-123

Psicología Educativa

www.elsevier.es/psed

Cognitive frameworks for assessment, teaching, and learning: A validity perspective

Marcos cognitivos para la evaluación, la enseñanza y el aprendizaje: una perspectiva centrada en la validez

Michael T. Kane* and Isaac I. Bejar

Educational Testing Service, Princeton, New Jersey, U.S.A.

The papers in this issue address the general question of how to add value to educational assessments, particularly in terms of student growth in academic disciplines. In addressing this goal, the papers focus on several recent and emerging model-based methodologies: in particular, learning progressions and cognitive models of learning (Pellegrino, this issue; de la Torre & Minchen, this issue), evidence-centered design (ECD) as a framework for assessment design and development (Zieky, this issue), and cognitively based assessment *of*, *for*, and *as* learning (Deane & Song, this issue; van Rijn et al., this issue).

These model-based methodologies involve major developments in how we interpret assessment results and, therefore, they have strong implications for how we evaluate the psychometric quality of the assessments. The model-based interpretations of each student's assessment results involve relatively complex descriptions of each student's achievement emphasizing the student's overall level of sophistication as specified by a list of skills mastered and not mastered (de la Torre & Minchen, this issue), or by a level in a learning progression (Pellegrino, this issue), rather than the student's standing on a unidimensional scale (or on several scales). The goal is to develop assessments that promote learning by providing information that is useful in teaching and learning, and to generate evidence that supports the proposed interpretation and usefulness of the assessment results.

Our main point in this paper is that while grounding assessment design in cognitive theories and model-based methodologies is highly desirable, rigorous evaluation of the resulting scores is still necessary. Specifically, the basic definition of validity in terms of the extent to which the interpretation and use of test scores is supported by appropriate evidence and analysis does not need to change. However, as discussed in more detail later, the structure of the arguments used to support the proposed interpretations and uses of the scores and the evidence needed to evaluate these arguments will need to be adapted to fit the proposed interpretations and uses of the test results. Similarly, the analyses of the precision, or reliability, of the results will need to be reconsidered; for example, to the extent that the focus is on placement in a learning progression rather than on a score on a continuous scale, analyses of precision would focus on consistency of placement (in the progression), rather than on traditional reliability or generalizability coefficients.

Learning Progressions and Cognitive Models

Cognitive models for learning seek to explain and predict student performance on assessment tasks in terms of profiles of student skills and corresponding task requirements. If a student has mastered all of the attributes required by the task, we would expect the student to be consistently successful in performing the task; if the student has not mastered all of the attributes required by the task, we would expect the student to be less successful, or completely unsuccessful in performing the task, or to perform at some chance level, depending on the assumptions built into the model (de la Torre & Minchen, this issue). An assessment involving a number of tasks with different attribute requirements can then be used to identify the attributes that have been mastered by the student and those that have not been mastered. It is easy to see how this kind of information could be useful to teachers and students.

Identifying the person and task attributes that are most relevant to a discipline is potentially a labor intensive activity, as is the development of an appropriate statistical model for specifying the relationship between the attributes mastered by a student, the attributes required by a task, and the expected performance of the student on the task. There are also questions about how large the domain being modeled should be (the domain size) and how general or specific the attributes should be (the attributes' grain size). As de la Torre and Minchen (this issue) point out, given that we cannot have more than five to ten attributes in the statistical model without running into serious problems in estimation, there is a tradeoff between the domain size and the grain size but there is also a need to insure that the attributes being assessed are the most relevant given the purpose of an assessment. The attributes in a cognitive model are not necessarily ordered or hierarchical, but they can be; that is, the assumption that mastery of one attribute is a prerequisite for another attribute can be built into the model as a constraint.

Model-based assessments can provide relatively detailed information on the attributes (e.g., skills and conceptual understandings) that each student has mastered and not mastered, and with a small grain size, this information can be quite detailed. Such specific indications of the weaknesses in a student's mastery of a topic can be used to target instruction on those soft spots. With a larger grain size more general guidance can be obtained. But there is no such thing as a free lunch. In order to realize these benefits to a substantial degree, it is necessary that the model fit the data and that it provide a coherent and instructionally relevant explanation

* Correspondence concerning this article should be addressed to Michael Kane. 6973 Apprentice Pl. Middleton WI, 53562. E-mail: MKane@ets.org

DOI: http://dx.doi.org/10.1016/j.pse.2014.11.006

of student performance, and that the assessment be built in a way that supports accurate estimation of model parameters. In order to meet these requirements, the assessment may need to be quite lengthy or the number of parameters estimated may need to be quite limited.

A different kind of cognitive model, learning progressions (Pellegrino, this issue), has been proposed as a way of making assessment results more meaningful and more useful to teachers, policy makers, and others involved in education in cases where learning can be modeled in terms of a progression of increasingly sophisticated levels of mastery of a domain or discipline. The general idea is to define the levels in the progression in terms of mastery of the core principles and methods in the discipline. Whereas the cognitive models tend to describe performance in terms of the constituent processes involved in the performance and therefore can be quite detailed, learning progressions tend to model performance in terms of increasingly sophisticated tasks and evaluative criteria for the performances.

Learning progressions are tied to the idea that learning a discipline (or a major part of a discipline) is a complex, long-term activity in which students gradually master the core components (methods, techniques, generalizations) of the discipline and develop the capability to function effectively in the discipline. The levels are defined in terms of qualitatively different levels of skill, understanding, and mastery, and they focus on a relatively small number of "higher-order cognitive skills" (Pellegrino, this issue), or general competencies in the discipline (e.g., the general principles and strategies defining the discipline), rather than specifying a large number of specific objectives (e.g., factual knowledge, specific skills). Assessments based on learning progressions tend to be associated with programs of instruction and curricula that are designed to develop the kinds of skill, understanding, and competency defining the associated learning progression. Learning is seen as a process of student growth, involving the assimilation of frameworks, schemas, and understanding of core conceptions and methods, rather than as a movement along a unidimensional scale.

The learning progression has an end point *(*what Pellegrino, this issue, calls "target performances" or "learning goals"), and intermediate levels that can be considered steps that characterize progress toward mastery of the domain. As Heritage (2008) put it:

> Another idea represented in these definitions of learning progressions is progression, that is, there is a sequence along which students can move incrementally from novice to more expert performance. Implicit in progression is the notion of coherence and continuity. Learning is not viewed as a series of discrete events, but rather as a trajectory of development that connects knowledge, concepts and skills within a domain. (Heritage, 2008, p. 4)

The number and "spacing" of the levels in a learning progression will depend on the length of instruction and learning and the "grain size" required for a particular application. The learning progression for K-12 curriculum in mathematics would have a large number of broadly defined levels. The learning progression for a third-grade unit on computing areas would probably have a much smaller number of more closely spaced levels. But in all cases, the levels are to reflect meaningful differences in terms of the development of competence in the discipline and to provide meaningful intermediate goals, or stages, that can be used to guide instruction.

It is the nature of the levels, rather than their number that has major implication for the analyses of the psychometric properties of the assessments. As explicated by Pellegrino (this issue) the levels are not simply points on a unidimensional scale. Each level is likely to be defined in terms of a cluster of related skills, understandings, and competencies that are learned and practiced together and that can be used together to perform various kinds of tasks within the discipline. The highest level of a learning progression is shaped by program expectations and requirements and is defined in terms of the level of competence expected of those who successfully complete the program. The entry level is defined in terms of the expectations for new students, or novices.

Between the entry level and the highest level of the learning progression, we have some number of levels or stages through which students are expected to pass as they go through the instructional program. The intermediate levels of sophistication are referred to as *levels of achievement* (Pellegrino, this issue). The levels of achievement are defined, in part, in terms of achievement on clusters of related *progress variable*s that reflect a set of core competencies, which are being developed within an instructional program.

It is expected that most students at a level of achievement would be at the prescribed levels on all or most of the progress variables. Assuming that this is the case and that students are distributed across a number of achievement levels, the measures of the progress variables should be positively correlated with each other; students with high scores on one progress variable would be expected to have relatively high scores on the other progress variables.

The structure of a learning progression tends to depend on the structure of the discipline and the criteria for effective mastery of the discipline. The structure of the instructional program (the sequencing of instructional topics and activities, or the curriculum) should also depend on the structure of the discipline and expectations about how students develop increasing competence and sophistication in the discipline, and the instructional program may, in fact, be based on the learning progression. As a result, the learning progression and the instructional program should be closely related.

It is expected that information about a student's achievement level in the learning progression will be helpful in guiding instruction and learning, because the learning progression reflects natural stages in mastering the discipline, and because these stages are reflected in the instructional program. A student who has demonstrated the competencies defining one achievement level in the learning progression can be encouraged to work on activities associated with the next level, independent of where other students are in the learning progression. If a student is struggling with the activities at an achievement level, information about their performance on the progress variables or about their ability to integrate these specific competencies into effective performance at the level of achievement might suggest strategies for helping the student. For example if the student is deficient in a particular progress variable, it may be helpful to have the student work on the specific skills or kinds of conceptual understanding defining that variable.

In short, learning progressions can be viewed as an attempt to combine the psychometric tradition in which achievement is represented by unidimensional scales and a rich qualitative tradition that describes achievement in terms of extended, holistic descriptions of performance and change. The levels of the learning progression are defined holistically, in terms of a general level of performance that may require many specific skills, but is not simply the sum of these skills, and thereby, they provide meaningful goals for instruction and learning. In addition, the levels are ordered and, therefore, provide a natural basis for at least an ordinal scale of measurement that could be used to describe achievement and growth. Working out how such a scale can be integrated into current psychometric theory and into educational practice will require the development of new and richer assessments (Dean and Song, this issue; Zieky, this issue), new psychometric models (Pellegrino, this issue; van Rijn et al., this issue), and new interpretive and use models. To the extent that the levels are clearly defined and distinct and have an appropriate grain size, they can provide an assessment framework that meets educational needs.

ECD and Learning Progressions

As just noted, learning progressions serve multiple purposes ranging from informing curriculum to the design of the assessment. However, incorporating learning progressions into assessments is a far more complex process than traditional test development. Assessment designers have traditionally relied on content standards that serve to delineate the scope and depth of the content coverage of an assessment (Schmeiser & Welch, 2006). The outcome of this process is typically represented as a table indicating the content areas covered by the test, their weight, and a classification of the content coverage along a relevant dimension such as "cognitive level", based, for example, on Bloom's taxonomy. This table, in turn, can be used to create a test blueprint, a more detailed version of the table, from which test forms could be produced.

In an educational context, interpreting the scores from a test developed in that fashion has typically involved an ordering of test takers along some continuum (i.e., a norm-referenced interpretation), or it has required a process known as standard setting (Cizek & Bunch, 2007; Hambleton & Pitoniak, 2006), whereby regions of the score scale are referenced to descriptions of what students know and can do. The boundaries between levels and the levels themselves are also known as "achievement levels" in the standard setting literature, although they do not correspond necessarily or typically to levels in a learning progression, which are intended to be a more concrete and actionable approach to reporting assessment results. In the United States, the National Assessment of Educational Progress (NAEP) and many state tests have used this kind of standard setting to make distinctions such as Basic, Proficient, and Advanced, which however, lack the developmental implications of learning progressions.

Specifically, since the standard setting process for each grade is typically carried out independently by grade, the potential for incomparable standards (Lissitz & Wei, 2008) is substantial, in that the rigor in the definition of the different levels is not necessarily consistent across grades. Several approaches to address the problem have been suggested (Ferrara et al., 2007; Lewis & Haug, 2005). Among them, is incorporating learning progressions in the design of an assessment, such that the vertical articulation takes place explicitly, naturally, and by design, rather than retrospectively (Bejar, Braun, & Tannenbaum, 2007; Kannan, in press). To achieve that goal, the assessment needs to be developed with learning progressions as an input to the assessment design process following a methodology such as evidence-centered design.

Zieky (this issue) outlined the basic elements of evidence-centered design as consisting of several layers (Mislevy & Haertel, 2006). The first layer of ECD, domain analysis, reviews relevant research on student learning, including a high level view of the sort of situations that would serve to elicit evidence about students' learning and specifically the level at which they performed. The levels in the learning progression provide milestones against which student progress can be evaluated. Unlike more traditional achievement levels such a Basic, Proficient and Advanced, the levels in a learning progression are associated with well articulated performances, as markers, and for that reason, they can be used to guide the development of an assessment that reflects the progression.

The next ECD layer is domain modeling where the assessment argument is formulated. That is, the intended inference or claim, for example, that the student's performance is characteristic of students at a given level of a learning progression, is justified by enumerating the warrants for score interpretation or use and corresponding supporting information. Another outcome of domain modeling is one or more design patterns as a way to begin specifying tangible design components. For an assessment that is informed by a learning progression, the design pattern would include distinctions between levels in the learning progression (e.g., Mislevy et al., 2014). In addition, the design pattern would include the type of performance evidence that would be characteristic of students at different levels of the learning progression. Finally, the outlines of situations that could serve to elicit such evidence would complete the design pattern.

The Conceptual Assessment Framework (CAF), the third layer, recasts the design up to this point into a set of models represented as a set of variables. The student model describes the student, typically as the "ability" parameter(s) in a suitable psychometric model. Task models include the features of task and guidelines for task development, for example task templates (Riconscente, Mislevy, & Hamel, 2005) to produce actual tasks. The evidence model describes the scoring of the student performance and consists of two parts, scoring rules, and the psychometric model to update the student model based on student performance.

Increasingly, assessments are being delivered on computer and therefore ECD can take advantage of technological supports (Mislevy, Bejar, Bennett, Haertel, & Winters, 2010) for their delivery and the processing of the interaction of the student with the assessment, which would include the scoring of such interactions. Recasting of evidence rules as an automated scoring process (Williamson, Mislevy, & Bejar, 2006) is natural in that context, for example. For assessments informed by a learning progression, the distinctions among the levels of the learning progression can serve as the scoring rubric for a task. That is, responses can be classified as being characteristic of the student at a given level of the progression. Ordered multiple choice (OMC) (Briggs & Alonzo, 2012) items are an example of that idea for the multiple choice case. A similar idea can be implemented in the open-ended case by designing the automated scoring process to classify responses with respect to a learning progression.

The requirements for a suitable psychometric model, the second component of the evidence model, are also formulated as part of the CAF. As noted earlier, learning progressions are not constrained by the convenient assumptions, such as unidimensionality, made by off-the-shelf psychometric models, and will require far more flexible approaches (Wilson, 2012). Since learning progressions are ordered, polytomous item response models (van Rijn et al., this issue) are necessary. Modeling need not be limited to item response models, and Bayesian networks (West et al., 2012) could also be suitable for fitting to data from such an assessment.

As noted by Zieky the last two layers of the design process are implementation and delivery. Implementation is concerned with the authoring of tasks or algorithms for the production of tasks and actual scoring, fitting response models, and similar implementation details. Delivery is concerned with the orchestration of test administration, including the delivery of the items. It interacts with the student model, especially in the case of adaptive testing, for choosing the next item. Accommodating students with special needs is also handled at this stage provided the design process took that requirement into account, which is entirely possible within ECD (Hansen & Mislevy, 2008).

The pay off of gearing domain analysis to formulate a learning progression is illustrated in the context of the CBAL project, where the design led to the idea of a scenario-based assessment (Sheehan & O'Reilly, 2012). Deane and Song (this issue) describe the process of developing a learning progression for argumentation skills. They noted that in light of the complexity of the argumentation construct, ECD was an ideal approach to design the assessment. The domain analysis they conducted suggested that argumentation can be seen as involving five stages: understanding the stakes, exploring the subject, considering positions, creating and evaluating arguments, and organizing and presenting arguments. The fine granularity of the analysis is appropriate for a school context where the goal is not simply to assess learning but also to promote and scaffold the development of skills through actionable feedback on student performance. By contrast, an admission test, such as the Graduate

Record Examination (GRE®), also addresses argumentation skills but focuses on the last stage identified by Deane and Song, namely presenting an argument, which is appropriate given the purpose of the GRE.

Deane and Song review the developmental literature to identify the possible levels of a progression and begin by postulating a set of skills (KSAs) that underlie the different phases of mastering argumentation from appeal building to building a case. The explicitly developmental goal is to identify how proficiency in a domain like argumentation develops, so that assessments can serve the multiple functions that CBAL aspires to, namely not just assess the students' current standing but also serve to promote learning. CBAL provides a working prototype for the use of cognitive models for assessment design, implementation, and analysis.

In short, ECD has been useful in designing a complex assessment involving a learning progression. Taking into account multiple considerations *during* the design of the development process is more likely to result in an assessment that yields valid scores. However, even when following a disciplined approach to design much can go wrong and for that reason the process of validation is still necessary, as we discuss next.

Validity

To validate an interpretation or use of assessment results is to evaluate whether the proposed interpretation and use of the results is adequately supported by appropriate evidence. The validation can be facilitated by first stating the proposed interpretation and use in some detail, in terms of an interpretation/use argument (or IUA) that lays out the inferences and assumptions inherent in the interpretation and use, and the interpretation and use can then be validated by evaluating the completeness and coherence of the IUA and by evaluating the plausibility of the inferences and assumptions in the IUA (Kane, 2013). (When ECD has been used to design an assessment, the proposed interpretation and use and the argument to support that interpretation exists, at least in part, as a byproduct of the design process.)

Learning progressions provide an interpretation based on a developmental model of performance in a discipline. Rather than reporting results in terms of a continuous score scale, a student's assessment performance is reported and interpreted in terms of the student's standing in the learning progression, where the achievement levels are intended to represent qualitatively different levels of sophistication in the discipline. Alternately, an interpretation based on a cognitive model might describe a student's current state of mastery of a topic or domain in terms of their mastery or nonmastery of each of a set of binary attributes (skills, understandings) specified in a cognitive model.

An assessment designed to identify students' levels in a learning progression would need to involve tasks that require the kinds of performances associated with the different levels in the learning progression. An assessment task or a part of an assessment task associated with a particular achievement level would require the kind of performance that students at that level of achievement should be capable of performing.

An assessment designed to provide estimates of each student's mastery or nonmastery of the attributes in a cognitive diagnostic model would need to involve assessment tasks that require different subsets of the attributes and would need to include a sufficient number and variety of such tasks to identify the particular attributes that each student has mastered and those that the student has not mastered.

The IUA and the Validity Argument

As noted, a learning progression is an ordered set of levels defined with respect to a developmental or curricular model. What are relevant criteria for evaluating an assessment based on a learning progression? That question becomes all the more important in light of the shift toward assessments that are used across jurisdictions, such as countries in the case of international assessments, or states in the case of the U. S., where consortia are developing assessments intended to be used across states. In international assessments the potential for country-by-item interactions has been noted when different languages are involved (Ercikan, 2002). Similarly, the potential for jurisdiction-by-item interactions could be relevant if consequential inferences are to be drawn regarding the relative performance of the different jurisdictions.

The IUA for assessments based on a learning progression would start with the student performances on the assessment tasks and would end with conclusions about the student (e.g., where the student is in the learning progression), and in applied settings, with suggestions about what to do next.

Scoring

Given the structure, interpretation, and expected uses of assessment results in terms of learning progressions or cognitive diagnostic models, the scoring system would be designed to assign each student to a particular level in the progression or to an attribute profile, based on the requirements built into the model. The assignment might also include some differentiation within levels to distinguish, for example, between students who have clearly mastered a level, students who seem to be at the level but are somewhat inconsistent, and students who have mastered the previous level and are beginning to develop the skills of this level.

For the scoring procedures to make sense, they must be consistent with the assumptions built into the model and with the structure and content of the assessment; we have to collect appropriate data for the estimation of the attributes used to characterize each student's achievement. A careful analysis of the performance domain and of the model being adopted (e.g., using ECD) can make a strong preliminary case for the fit between the performance domain, the theoretical model, and the data collection procedures (de la Torre & Minchen, this issue).

In addition, the observed relationships within the data should be consistent with the assumptions built into the model and any empirical predictions that can be derived from the model (van Rijn et al., this issue). For example, the achievement levels in learning progressions are typically strongly hierarchical in the sense that a student who is assigned to a level in the progression should generally be able to meet the requirements for lower levels, and should generally not be able to meet the requirements for higher levels. There may be some exceptions and slippage, especially for adjacent levels, but the hierarchical structure of the learning progression should generally hold.

Van Rijn et al. (this issue) propose two criteria for evaluation. One is whether the leaning progressions can be 'recovered" from test data; the second criterion is whether tasks that are built based on a learning progression and intended to be parallel to each other, in fact, behave in that manner.

Most students classified as being at a certain level in the progression should also be more or less at the levels of the progress variables associated with that level of the progression, and this pattern should hold across major subgroups of students (e.g., defined by gender, race) as well as across jurisdictions. It will generally not be possible to evaluate all such relations across all groups (e.g., because of small sample sizes), but where possible, the differences should be evaluated, to ensure that the model-based interpretations are invariant across relevant groups.

Similarly, the fit of cognitive diagnostic models should be invariant across relevant groupings of students, as well as across jurisdictions, where the potential exists that the match between the

local curricula and the modeling assumptions is not uniform across jurisdictions.

Generalization

As with any assessment, the interpretation of the results of these model-based assessments assumes that the results would not change much if the assessment were replicated (e.g., using a different set of equally appropriate tasks) at about the same time. If a student were evaluated and found to have a particular profile of attributes, we would expect the student to have a similar profile on the replication. Similarly, if a student were found to be at a particular level in a learning progression, we would expect the student to be at the same level or, perhaps, an adjacent level on the replication. If this kind of consistency or generalizability were not found (i.e., if the results varied), it would be difficult to interpret or use the assessment results in any coherent way.

In this context, standard reliability indices, which focus on how precisely the assessment scores place students on a score continuum, would be of marginal relevance at best. As de la Torre and Minchen (this issue) state, the cognitive diagnostic models do not provide estimates of true scores on a continuum, but provide information about profiles of discrete binary attributes. The learning progressions do focus on the ordering of students in the progression, but in this case also, the focus is on categorizing the student's level of achievement in the learning progression.

The progress variables in a learning progression can have standard psychometric interpretations (as scores on a unidimensional and more-or-less continuous scale), and therefore the question of generalizability over occasions, tasks, raters, and so on could be handled in standard ways (van Rijn et al., this issue). A generalizability, or reliability coefficient and/or their associated standard errors of measurement, would provide an indication of how closely we could expect the scores to agree with each other across replications of the assessment (e.g., using different tasks from the same domain, different scorers).

However, more generally, it would probably be appropriate to address the issue of generalizability by examining the extent to which the results reported to the teachers and other users of the results (e.g., the achievement levels in a learning progression or the ability profiles assigned to the students) remain the same or change over replications of the assessment. In this context, the variability across parallel tasks, and more generally across tasks, can be particularly critical (van Rijn et al., this issue). If the assessment results are interpreted in terms of achievement levels in a learning progression, then the main concern about generalizability is whether a student is consistently assigned to the same level, and sources of variability that do not have much impact on the achievement level are not serious. Similarly, for cognitive diagnostic models, variability in student performance over replications becomes substantial when it leads to substantial differences in the ability profiles reported for students.

Extrapolation

The fact that a model (learning progression or cognitive diagnostic) provides a useful framework for interpreting and using assessment results in the context in which it is developed does not necessarily imply that it provides an equally useful framework in other contexts, involving different settings, different curricula, different teachers, or different students. The match between the curriculum and the model is likely to be particularly important. For example, reporting assessment results in terms of a learning progression is likely to be especially effective in contexts where the structuring and sequencing of instruction are consistent with the learning progression. Unlike most of the rest of the world, in the U. S. curricula are not national. Therefore, there is a potential for what might be called "curricular differential item functioning". That is, students exposed to different curricula perform differentially on an assessment.

Guiding Instruction and Learning

Ultimately, the added benefit associated with these model-based assessments would derive from their anticipated usefulness in instruction and learning, and therefore, the model-based assessment programs would be evaluated in terms of their effectiveness in promoting learning. To the extent that they provide effective guidance for teachers and students, they can add substantial value, over and above the kind of summative assessment that simply assigns each student to a point on a unidimensional scale or to points on a small number of such scales. If they do not provide effective guidance for teachers and students, then claims about added value would be questionable.

To the extent that the interpretation of the assessment results can be validly interpreted in terms of levels in a learning progression or ability profiles derived from a cognitive model, the assessment can be said to provide information that is not otherwise available, but this does not in itself justify a claim that the use of the assessment will promote effective instruction. To be helpful, the assessment-based information will need to be relevant to the instructional context, be meaningful to teachers, be timely, be at about the right grain size, and not be otherwise available. For example if the curriculum and or the teachers organize instruction on a topic in a way that is different from the schemas incorporated in the model-based assessment, or in a different order, it may be difficult for teachers to make effective use of the additional information. The authors of these papers clearly recognize the potential for this kind of lack of alignment and point out the need to link instruction and assessment closely to each other and to the cognitive models. Either the instructional program can be designed to fit the assessment's cognitive model, or the cognitive model can be designed to reflect the instructional program, or better still, the two components can be developed and implemented together. In any case, professional development for teachers will probably be needed for the system to work well.

Claims for instructional effectiveness could be evaluated using several kinds of data. First, the claims should be supported by a theory of action (Bennett, 2010) that indicated how the information provided by the assessment is to be used by teachers, students and, possibly, others. Evidence (e.g., interview or observational data) indicating that teachers find the data helpful and are using it in the ways anticipated in the theory of action associated with the assessment program would indicate the theory of action is being implemented. Statistical studies (longitudinal or comparative) indicating how teacher behavior and student achievement change with implementation of the cognitively-based program could be used to evaluate claims that the implementation of the assessment/instructional program is effective. An optimal approach to evaluating efficacy claims, at least initially, would probably involve qualitative, descriptive studies of how the program is functioning, its impact on what's happening in classrooms, and small-scale outcome studies.

It is important to keep in mind that good teachers generally have some kind of cognitive model in mind when they work with students, and they update these student models more or less continuously as they interact with the students. They know that some students are operating at a relatively advanced level, while others are struggling with more basic tasks. They also have some idea of the specific competencies each student has mastered. So the value added by the model-based assessment may be diminished by the fact that some of the information being supplied by the assessment may be largely redundant. The model-based assessment results may be much more helpful to teachers with less well developed student models than

they are to teachers with sophisticated student models; in such cases, the model-based assessment programs may be particularly helpful as a form of professional development for teachers who do not make much use of student models or have poorly developed student models.

Concluding Remarks

Large-scale assessments developed along traditional psychometric lines are designed to produce scores that reflect each student's position on some continuum reflecting overall achievement in some domain. Such assessments are useful for some purposes, but they are not directly helpful in instruction and learning. The papers in this issue highlight several approaches, based on cognitive models that can provide assessments that are more directly useful for instruction and learning.

These methodologies emphasize more complex, model-based descriptions of student achievement in terms of levels of achievement in a learning progression (Pellegrino, this issue) or profiles of concepts and skills that the student has mastered (de la Torre & Minchen, this issue). The goal is to provide model-based feedback that is useful for teaching and learning. The evaluation of such assessments in terms of their psychometric properties and their educational utility raises new questions, some of which can probably be answered using existing methods, but a full realization of these new approaches to assessment will undoubtedly require substantial rethinking of our existing methodology, and may also require some radically new methods.

Resumen

Los artículos de este número especial abordan la cuestión general de cómo proporcionar valor añadido a la evaluación educativa, sobre todo en términos del progreso de los estudiantes en las distintas disciplinas académicas. Para ello, los trabajos presentados se centran en varias metodologías basadas en modelos que han surgido recientemente, en particular las progresiones de aprendizaje y los modelos cognitivos de aprendizaje (véase los artículos de Pellegrino y de la Torre y Minchen, respectivamente), el diseño centrado en la evidencia (véase Zieky) y la evaluación cognitiva de/por/para el aprendizaje (véase los trabajos de Deane y Song, y de van Rijn, Graf y Deane). Se trata, en definitiva, de diseñar evaluaciones que promuevan el aprendizaje proporcionando información útil para el proceso de enseñanza-aprendizaje y de obtener evidencia que garantice que se pueden interpretar y utilizar de la forma deseada los resultados de la evaluación.

Estas metodologías implican un cambio considerable en la forma de interpretar las puntuaciones de los tests que tiene, a su vez, una importante repercusión a la hora de evaluar su calidad métrica.

Este tipo de interpretaciones supone trabajar con descripciones relativamente complejas del rendimiento en las que se pone el acento en el nivel general de sofisticación alcanzado por el estudiante, tal como es especificado mediante una lista de atributos o destrezas que éste domina o no (de la Torre y Minchen, en este número) o mediante un nivel de competencia en una progresión de aprendizaje (Pellegrino, en este número), más que mediante su ubicación en una escala unidimensional (o en varias). El aprendizaje se considera como un proceso donde el estudiante avanza asimilando los marcos, esquemas y conocimientos relativos a los métodos y conceptos clave de una disciplina, en lugar de moviéndose a lo largo de una escala unidimensional. Los niveles de una progresión de aprendizaje no son puntos que se ubican en una escala unidimensional sino que están definidos holísticamente como un conjunto relacionado de conocimientos, destrezas y habilidades que se aprenden y practican a la vez y que se pueden utilizar conjuntamente para realizar distintos tipos de tareas dentro de una disciplina. Cómo integrar una escala de estas características en la actual psicometría y en la práctica educativa requiere el desarrollo de nuevos tipos de evaluación (y evaluaciones más ricas y elaboradas), nuevos (y más flexibles) modelos psicométricos y nuevos modelos de uso e interpretación de las puntuaciones.

Para empezar, incorporar en una evaluación las progresiones de aprendizaje supone introducir una complejidad en el proceso de construcción y diseño del test que no resulta fácil de acomodar en el marco tradicional de diseño de las pruebas de evaluación. El trabajo de Deane y Song en este número muestra lo útil que resulta para ello la metodología del diseño centrado en la evidencia, metodología que también es utilizada por de la Torre y Minchen (en este número) para ilustrar cómo diseñar una evaluación para el diagnóstico cognitivo.

En el trabajo presentado en este número, van Rijn et al. utilizan modelos politómicos de teoría de respuesta al ítem para asignar estudiantes de secundaria a los niveles de las progresiones de aprendizaje formuladas para la capacidad de argumentar en el trabajo de Deane y Song (en este número). Ahora bien, como señalan West et al. (2012), también se podría recurrir a redes bayesianas para modelar los datos de una evaluación basada en progresiones de aprendizaje.

El punto central del presente trabajo tiene que ver con la interpretación y uso de los resultados de la evaluación: la validación sigue siendo necesaria, aun cuando la prueba haya sido cuidadosamente diseñada. Si bien es claramente deseable que el diseño del test se realice a partir de metodologías basadas en modelos, eso no exime en modo alguno de una evaluación rigurosa de los resultados obtenidos.

La validación se puede facilitar formulando con cierto detalle una determinada interpretación o utilización de las puntuaciones del test, esto es, desarrollando un Argumento de Interpretación/Uso (AIU) que establezca las correspondientes inferencias y supuestos implícitos en esa interpretación/uso; seguidamente hay que validar dicha interpretación/uso evaluando la coherencia y grado de compleción del AIU, así como la plausibilidad de las inferencias y supuestos de dicho argumento (Kane, 2013).

Una vez validada una determinada interpretación (bien en términos de un nivel de competencia en una progresión de aprendizaje basada en un modelo de desarrollo, bien en términos de un perfil de atributos o destrezas derivado de un modelo cognitivo), no se puede considerar sin más que ese programa de evaluación contribuye a mejorar el proceso de enseñanza-aprendizaje, no se puede dar por hecho su valor o utilidad formativa, sino que una vez más es necesario obtener evidencia que permita concluir su eficacia en la instrucción y, de este modo, confirmar el valor añadido de estas evaluaciones basadas en modelos. Se proponen distintas vías para ello y se discuten también aspectos tan importantes en la validación como la asignación de puntuaciones, su generalización y extrapolación.

Conflict of Interest

The authors of this article declare no conflict of interest.

References

Bejar, I. I., Braun, H., & Tannenbaum, R. (2007). A prospective, predictive and progressive approach to standard setting. In R. W. Lissitz (Ed.), *Assessing and modeling cognitive development in school: Intellectual growth and standard setting* (pp. 31-63). Maple Grove, MN: JAM Press.

Bennett, R. E. (2010). Cognitively Based Assessment of, for, and as Learning (CBAL): A preliminary theory of action for summative and formative assessment. *Measurement: Interdisciplinary Research & Perspective, 8*, 70-91.

Briggs, D. C., & Alonzo, A. C. (2012). The psychometric modeling of ordered multiple-choice item responses for diagnostic assessment with a learning progression. In A. C. Alonzo & A. Wenk Gotwals (Eds.), *Learning progressions in science* (pp. 293-316). Boston, MA: Sense Publishers.

Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: a guide to establishing and evaluating performance standards on tests*. Thousand Oaks, CA.: Sage Publications.

Deane, P., & Song, Y. (2014). A case study in principled assessment design: Designing assessments to measure and support the development of argumentative reading and writing skills. *Psicología Educativa 20*, 99-108.

de la Torre, J., & Minchen, N. (2014). Cognitively diagnostic assessments and the cognitive diagnosis model framework. *Psicología Educativa 20*, 89-97.

Ercikan, K. (2002). Disentangling sources of differential item functioning in multi-language assessments. *International Journal of Testing, 2*, 199-215.

Ferrara, S., Phillips, G. W., Williams, P. L., Leinwand, S., Mahoney, S., & Ahadi, S. (2007). Vertically articulated performance standards: An exploratory of inferences about achievement and growth. In R. W. Lissitz (Ed.), *Assessing and modeling cognitive development in school: Intellectual growth and standard setting*. Maple Grove: MN: Jam Press.

Hambleton, R. K., & Pitoniak, M. (2006). Setting performance standards. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 433-470). Westport, CT: Praeger.

Hansen, E. G., & Mislevy, R. J. (2008). *Design patterns for improving accessibility for test takers with disabilities* (Research Report 08-49). Princeton, NJ: Educational Testing Service.

Heritage, M. (2008). *Learning progressions: supporting instruction and formative assessment*. National Center for Research on Evaluation, Standards, and Student Testing (CRESST). Retrieved from http://www.k12.wa.us/assessment/ClassroomAssessmentIntegration/pubdocs/FASTLearningProgressions.pdf

Kane, M. T. (2013). Validation as a pragmatic, scientific activity. *Journal of Educational Measurement, 50*, 115-122.

Kannan, P. (in press). *Content and performance standard articulation practices across the states: Report summarizing the results from a survey of the state departments of education* (Research Report). Princeton, NJ: Educational Testing Service.

Lewis, D. M., & Haug, C. (2005). Aligning policy and methodology to achieve consistent across-grade performance standards. *Applied Measurement in Education 8*, 11-34.

Lissitz, R. W., & Wei, H. (2008). Consistency of standard setting in an augmented state testing system. *Educational Measurement: Issues and Practice, 27*(2), 46-55.

Mislevy, R. J., Bejar, I. I., Bennett, R. E., Haertel, G. D., & Winters, F. I. (2010). Technology supports for assessment design. In G. McGaw, E. Baker & P. Peterson (Eds.), *International Encyclopedia of Education* (3rd ed., vol. 8, pp. 56-65). Amsterdam, The Netherlands: Elsevier.

Mislevy, R. J., & Haertel, G. D. (2006). Implications of evidence-centered design for educational testing. *Educational Measurement: Issues and Practice, 25*(4), 6-20.

Mislevy, R. J., Oranje, A., Bauer, M. I., von Davier, A., Hao, J., Corrigan, S., … John, M. (2014). *Psychometrics considerations in game-based assessment*. GlassLab Research, Institute of Play. Retrieved from http://www.instituteofplay.org/wp-content/uploads/2014/02/GlassLab_GBA1_WhitePaperFull.pdf

Pellegrino, J. (2014). Assessment as a positive influence on 21st century teaching and learning: A systems approach to progress. *Psicología Educativa, 20*, 65-77.

Riconscente, M. M., Mislevy, R. J., & Hamel, L. (2005). *An introduction to PADI task templates* (Technical Report 3). Menlo Park, CA: SRI International. Retrieved from http://padi.sri.com/downloads/TR3_Templates.pdf

Schmeiser, C. B., & Welch, C. J. (2006). Test development. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 307-353). Westport, CT: Praeger Publishers.

Sheehan, K. M., & O'Reilly, T. (2012). The case for scenario-based assessments of reading competency. In J. Sabatini, E. Albro & T. O'Reilly (Eds.), *Reaching an understanding: Innovations in how we view reading assessment* (pp. 19-33). Lanham, MD: Rowman & Littlefield Education.

Van Rijn, P., Graf, E. A., & Deane, P. (2014). Empirical recovery of argumentation learning progressions in scenario-based assessments of english language arts. *Psicología Educativa 20*, 109-115.

West, P., Wise Rustein, D., Mislevy, R. J., Liu, J., Levy, R., Dicerbo, K., & Beherens, J. T. (2012). A Bayesian appraoch to modeling learning progessions. In A. C. Alonzo & A. Wenk Gotwals (Eds.), Learning progressions in science (pp. 257-292). Boston, MA: Sense Publishers.

Williamson, D. M., Mislevy, R. J., & Bejar, I. I. (2006). *Automated scoring of complex tasks in computer-based testing*. Mahwah, NJ: Erlbaum.

Wilson, M. (2012). Responding to the challenge that learning progressions pose to measurement practice: Hypothesised links between dimensions of the outcome progression. In A. C. Alonzo & A. Wenk Gotwals (Eds.), *Learning progressions in science* (pp. 317-344). Boston, MA: Sense Publishers.

Zieky, M. J. (2014). An introduction to the use of evidence-centered design in test development. *Psicología Educativa 20*, 79-87.