*Journal of Educational Measurement*
*Winter 2007, Vol. 44, No. 4, pp. 285–291*

# Guest Editors' Introduction and Overview: IRT-Based Cognitive Diagnostic Models and Related Methods

**Louis V. DiBello**
*Learning Sciences Research Institute, University of Illinois at Chicago*
**William Stout**
*University of Illinois at Urbana-Champaign*

## Introduction and Overview

Much of the 20th century psychometric focus on testing has been on the unidimensional continuous scaling of examinees in major subject or cognitive areas, such as verbal reasoning, mathematics ability, etc., often using item response theory (IRT) modeling, especially 1PL, 2PL, 3PL modeling. In fact, many of the well-known, currently nationally administered, standardized tests in America are supported wholly or in part by such continuous unidimensional IRT models. Such psychometric modeling is very useful and psychometrically reliable for summative assessment of students in broad-based subject or cognitive areas. Most measurement specialists, including the co-editors of this special issue, would say this work has been both an intellectual and practical triumph, in spite of various cautionary issues about large-scale standardized tests being raised currently. The design and IRT model-based calibration of such tests as well as the statistical-analysis-based scaling of test-taking populations of examinees have been greatly facilitated by these profound advances in psychometrics. Many summative decisions concerning admission, placement, scholarships and fellowships, as well as ranking examinees on broadly defined proficiencies, are based in whole or in part on such unidimensionally modeled tests.

In the last few decades, a confluence of recent scientific, political, and educational developments has been strongly encouraging the development and use of psychometrically grounded tests with a fundamentally different purpose than the dominantly unidimensional summative tests that are well modeled by the logistic IRT models. There is rapidly emerging a powerful need, and demand, for tests designed to formatively assess an appropriately chosen moderate number of relatively fine-grained chunks of knowledge in major subject or important cognitively defined areas. By "formative" is meant that the results of the assessment are used to directly support teaching and learning, as contrasted with summative testing, which evaluates the student after the instruction is over. The development of tests specifically designed for the purpose of providing information about student understanding that can directly be used to guide teaching and learning at the individual student level is now being addressed. Such informative assessments may be applied at the individual classroom level and may be used in a benchmarking role throughout the course being taught. In the simplest case of skills modeled dichotomously, students are assigned multidimensional skills profiles by assigning mastery versus nonmastery for each skill or attribute (these two terms are used generically and interchangeably throughout the Special Issue to describe the "knowledge chunks" referred to above).

This article focuses primarily on psychometric model-grounded approaches to skills diagnosis with an emphasis on the various classes of measurement models that can be appropriately used. The formative assessment challenge requires a multistage implementation process of which the important step of selecting an appropriate psychometric model is only one component. So it is important to embed the emphasis on measurement models in this Special Issue into the larger context of what is necessary for a successful formative assessment. As pointed out in Pellegrino et al.'s *What Students Know* National Research Council (NRC) report (2001), assessments can be described as evidentiary systems that consist of the so-called "assessment triangle" of cognition, observation, interpretation. This implies that the assessment development process must be based not only on content and instructional considerations but must also consider the cognitive processing of the students being assessed. The desired skills diagnostic interpretation of the assessment data is facilitated by a statistical analysis of an appropriate and often parametrically complex skills-based IRT model. Moreover, these inferences must be communicated in ways that are maximally useful to the key stakeholders (especially students, teachers, parents, and relevant school administrators).

An adaptation of the evidence-centered design (ECD) paradigm developed by Mislevy, Almond, and colleagues provides a convenient organizational structure for delineating the major components required for a successful formative assessment. The basic principle of ECD is that the items should be carefully designed to provide as efficiently and completely as possible the necessary evidence required to satisfy the assessment purpose, which in this Special Issue is focused on inference about multiple skills in a particular domain. The necessary components to meet this fundamental ECD principle are (1) describing the assessment purpose, (2) modeling/selecting the latent skills space appropriately, (3) development of the assessment's items (possibly complex, and, as such, sometimes called "tasks," a more inclusive term including items), (4) selecting an appropriate IRT-based cognitive diagnostic model (ICDM) to link examinee skills to examinee performance on the selected items/tasks, (5) selection of statistical and computational methods that are accurate and practicable, and (6) developing assessment reports that effectively serve all the key stakeholders. Much of the research and development work needed for skills-level formative assessment to truly flower remains to be carried out. We like to refer to this body of needed work as the engineering science required for doing effective formative assessments.

We now briefly discuss these ECD-grounded components of a cognitively diagnostic assessment, likely intended to be used formatively. To carry out these components, as part of an emerging engineering science of skills diagnostic assessment, an appropriate cross-disciplinary team of experts is required. This team will cooperate on all aspects of the design and development of the assessment, including selecting the skills, constructing the assessment instrument, and reporting assessment feedback to users. First, the assessment purpose must be clearly delineated within its full cognitive, curricular, and instructional context. This delineation will then have a strong influence on the remaining components, especially on the number and type of skills chosen as the skills space, the construction of the assessment items/tasks, and the class of measurement models chosen, for example, IRT latent class (discrete)

modeling for classification purposes and IRT latent trait (continuous) modeling for continuous scaling purposes. The assessment purpose will dictate the type of score reporting needed.

A vital part of the formative assessment development process is that a reasonably sized set of skills be carefully and expertly chosen to be formatively assessed. Choosing the skills for such a formative assessment requires cooperative interdisciplinary expertise in subject/curricular matters, instructional issues, cognitive aspects of learning, skills-based psychometric theory, etc., to be blended appropriately in selecting a manageable number of important skills to be assessed. These skills should capture the core aspects or components of the subject or cognitive area being tested, or, if administered as an interim benchmark, the core aspects taught to date. In particular, the set of skills must be psychometrically, instructionally; and cognitively appropriate. Hence, a multidisciplinary collaboration of expertise will produce a more valid and more psychometrically informative set of skills.

As is well understood, state and district subject standards do not lend themselves for direct use as the set of skills to be formatively assessed. A specific effort is required for converting standards to skills whenever the assessment is required to be "standards based," a challenging and largely unexplored research topic of considerable importance.

The items, possibly including complex items or, even more generally, open-ended tasks, must be designed from a cross-disciplinary evidentiary perspective to effectively measure well all of the specified skills. An open research topic is whether fewer complex items or greater numbers of simple (perhaps multiple choice) items are most informative for a given assessment application, both from the validity and the "reliability" perspectives. The construction of items for carrying out effective formative assessments is an emerging challenge that will benefit from further foundational research, with the goal of elucidating general ECD-grounded principles on how to build skills-level formative assessments.

For each skill, examinees could be scaled continuously or placed in a discrete ordered set of categories of size of two or moderately greater, such as: being seriously below standard, being below standard, meeting the standard, or exceeding the standard—a four-level ordered set, for example. This scaling decision influences the choice of the psychometric model. The latent class modeling paper in this Special Issue mostly focuses on dichotomously graded skills (mastery vs. nonmastery). However, as noted in the Bayes net Special Issue paper, dealing with so-called ordered polytomous skills is of obvious applications importance and may be required in many settings. Many cognitive diagnostic IRT models used in practice include a skill by item 0/1 incidence $Q$ matrix that specifies for each item which of the chosen skills is required to solve the item. From this perspective, one challenge of item construction/selection is to produce a test with a resulting $Q$ matrix that satisfies various criteria, such as the test being balanced with respect to the skills space in the sense that each skill is measured by roughly the same number of items and that the number of items per skill is large enough to provide sufficient skills measurement accuracy.

As stated, the major focus of this Special Issue is to provide an appropriate sampling of various types of ICDM models that show promise from the viewpoint of facilitating all the components of the ECD-inspired formative assessment

implementation framework. This requires sufficient ICDM modeling parametric complexity such that most of the relevant information provided by the assessment can be extracted by estimation of informative parameters in the model. However, a balance is required because excessive parametric complexity can lead to nonidentifiability in extreme cases and to lack of sufficient information in the data to reliably calibrate the model in less extreme cases. Clearly, parametrically complex IRT modeling is called for, but with the minimal amount of complexity needed to satisfy the assessment's purposes. That is, the chosen model should display as much modeling parsimony as possible within the bounds of the diagnostic purpose. Although achieving good model fit is important, we see the central issue as selecting measures of model—fit that are strongly related to the diagnostic purpose, and to work hard to satisfy such diagnostically relevant fit measures with models that are as parametrically simple as possible, but still capture the assessment's diagnostic purpose. More precisely, the model simultaneously must be complex enough to provide sufficient skills information to meet user needs and also must be parametrically parsimonious enough and must display enough appropriate model fit so that the data provide sufficient skills information to meet user needs.

A further model selection issue is whether the basic modeling approach should be conjunctive (which assumes mastery of all skills required by the item is necessary for solving the item correctly and, as such, often seems more congruent with the cognitive perspective), or fully compensatory (which assumes a low level of mastery on one required skill can be compensated for by sufficiently high mastery on one or more other skills required by the item).

Appropriate statistical approaches and computational methods are required so that estimation accuracy is achievable and the needed data-analytic computations are possible in sufficient brief time to meet the assessment's reporting timeline. The Special Issue briefly discusses estimation and statistical computation in its various articles but does not treat the topic in detail. The ever-increasing speed and storage capacity and the ever-decreasing cost of computing enables the use of the complex IRT modeling needed to accurately underpin such skills-based testing even if such modeling requires computationally complex estimation algorithms. It is noted that modern developments in statistics, including the development of fast, efficient, and accurate estimation algorithms for analyzing large data sets modeled by parametrically complex models (often Bayes or empirical Bayes), also permit accurate complex IRT modeling and accurate statistical analysis of cognitive diagnostic tests.

The assessment score reporting issue is central. The score reports must provide sufficient information in a manner that communicates to the stakeholders, especially students, teachers, parents, but also possibly school administrators, curriculum specialists, state and city departments of education, etc. A prerequisite for good score reporting is of course that the assessment process yield appropriate and accurate statistical inferences concerning information relevant to the formative assessment purpose. In this regard, one essential task is to decide on the specific model-based statistical skills mastery/nonmastery estimators that underpin the scoring reports. For example, in reporting examinee skill masteries, the statistical issue arises whether it suffices to use highly intuitive (model-based) subscores with cutpoints used to

assign mastery levels, rather than using more complex (and highly nonintuitive from the user's perspective) likelihood-based mastery assignments that take full advantage of the model. The Special Issue only touches on the issue of presenting adequate and appropriate assessment-derived information for use by the various stakeholders. However, it does provide a paper on ICDM-based subscoring used to assess skill mastery levels, which may provide more face validity for some users when they require that the skill classification methods that lie behind the reported skills diagnostic masteries have some intuitive clarity.

To illustrate some of the components of an effective ECD-grounded assessment, envision an algebra test that has as its assessment purpose to formatively assess individual students concerning their mastery of the major components of algebra competence. Then, using the combined input from various experts, a moderately sized set of fundamental algebra skills may be chosen, perhaps including linear equations solving, rules of exponents, quadratic formula, algebra word problem solving, factoring of polynomials, rational functions, understanding and solving linear and quadratic inequalities, and perhaps including some skills that are more cognitive in flavor. Then a test could be designed in which each skill appears on at least four items and on six items on average, say. Suppose via some sort of psychometric analysis this is judged as likely to produce adequate "reliability" for the relatively "low-stakes" assessment purpose of determining for classroom usage what topics each individual student needs to study more. After consideration, perhaps the Fusion Model will be chosen to model the item response functions, and the Arpeggio calibration and classification system will be chosen to do the statistical analysis needed. Then each student's reported profile would indicate mastery versus nonmastery of each of the proposed algebra skills, information that if reliable and well presented should be very useful for students and teachers alike. Because the teacher finds sum-scoring more intuitively reasonable and easier to grasp, perhaps it is decided to use the calibrated Fusion Model to produce model-based sum-scores with model-specified cutpoints to assign mastery or nonmastery on each skill for each student (see the Henson et al.'s paper in this Special Issue). Then it is perhaps necessary to develop a narrative scoring report (as an example, see the College Board PSAT Score Report Plus reporting system used operationally and suggesting for each PSAT examinee certain skills that may need further study).

The No Child Left Behind (NCLB) legislation, in spite of the controversy surrounding it, stands as perhaps the best known example of the new accountability emphasis in education that tends to call for skills diagnosis. Among other things, the NCLB guidelines explicity call for formative assessment information to be communicated so as to be useful for students, parents, teachers, principals, school districts, state departments of education, etc. This NCLB mandate of course strengthens the need for skills-based formative assessment.

There is a stronger form of formative assessment called embedded assessment, which postulates that formative assessments should be seamlessly and periodically embedded in the curriculum for the purpose of improving teaching and learning. One noted example of classroom-embedded assessment is the Mark Wilson–directed University of California at Barkeley SEPUP (Science Education for Public

Understanding Program) embedded assessment program, which is briefly discussed in the continuous latent trait paper in the special issue as a good example of what may be possible.

Another point worth brief mention is that those political forces that express varying degrees of opposition to large-scale standardized testing, as practiced these days, are usually neutral toward or in favor of properly carried out formative assessment. Thus formative assessment promises to be an inherently harmonizing factor in the politics of standardized testing in education. Instead of the test possibly controlling the teaching as is often claimed for traditional summative standardized tests, the teaching indeed controls the test in formative assessment, thus focusing the test particularly on those skills the teacher wants to be teaching well.

As stated, the main purpose of this Special Issue is to survey a collection of IRT-based ICDM approaches that seem to provide promising ways to carry out skills diagnostic modeling. These ICDMs are broken down by paper into broad areas: latent class, namely discrete latent space IRT modeling (all knowledge states are discrete, here dichotomous for lack of space to cover the important ordered polytomous skills levels case), continuous latent trait IRT models (all knowledge components are continuous), Bayesian networks–based cognitive diagnostic modeling (in which the basic IRT model appears as an acyclic-directed graph with discrete skill-level nodes: allowing very complex settings to be modeled, an advantage), the nonparametric artificial intelligence grounded IRT approach as embodied by the Tatsuoka Rule Space Approach, and IRT model–based sum-scoring (allows the assignment of mastery vs. nonmastery for each skill to be based on an intuitive model-based subscore of items).

The ICDM modeling approaches as discussed in this Special Issue occur within a larger context, which we refer to as the engineering science of doing effective formative assessments. In simple terms, this broader perspective takes an evidentiary perspective and combines theory and method with express attention paid to practical matters necessary for successful implementation of these methods and ideas in real educational settings. This approach draws in numerous other pragmatic aspects, including skills selection, item construction, development of instructional materials for teachers using formative assessments, psychometric issues of statistical calibration and examinee assessment approaches and the recasting of reliability/validity as appropriate for this new skills-focused type of psychometric modeling, the large-scale delivery of such tests, the linking of formative assessments to state and district standards, the interface between cognitive psychology and skills-level educational testing, working with teachers to facilitate the embedding of periodic formative assessment, cost containment, the blending of psychometric, cognitive, curricular, and instructional perspectives in developing this engineering science of informative assessment, evaluation of the effectiveness of various formative assessment efforts, exploring various environments in which such testing/assessment seems appropriate, developing the instruction of education majors concerning their future classroom use of formative assessment at the skills level, methods of item scoring including those sensitive to diagnostic distracters in multiple-choice items, and sophisticated (e.g., hierarchical) modeling of the examinee test-taking population. With the limited space available, most of these topics are omitted or merely touched upon lightly. It is the editors' belief that this engineering science of skills diagnosis, of which

the psychometric modeling component is a foundational and necessary piece, is essential for the flowering of skills diagnostic formative assessment and widespread implementation throughout actual education and training contexts. Hence its various aspects must be developed if skills diagnostic testing is to have a major impact on education. This includes both initial pilot projects and subsequent large-scale projects.

## Authors

LOUIS V. DIBELLO is a Research Professor, Learning Sciences Research Research Institute, University of Illinois at Chicago, 1006 West Harrison Street (Mail Code 057), Chicago, IL 60607-7137; ldibello@uic.edu. His primary research interests include psychometrics, diagnostic assessment, and informative assessment.

WILLIAM STOUT is Professor Emertus, University of Illinois, Department of Statistics, 101 Illini Hall, 725 S. Wright Street, Champaign, IL 61820; stout@stat.uiuc.edu. He is also a Research Professor, Learning Sciences Research Institute, University of Illinois at Chicago. His primary research interests include psychometrics in general and latent dimensionality analysis, DIF, and cognitive diagnostic assessment in particular.