INTERNATIONAL JOURNAL OF TESTING, *7*(2), 95–125

The Answer is in the Question: A Guide for Describing and Investigating the Conceptual Foundations and Statistical Properties of Cognitive Psychometric Models

André A. Rupp
Institut zur Qualitätsentwicklung im Bildungswesen (IQB)
Humboldt-Universität zu Berlin
Berlin, Germany

One of the most revolutionary advances in psychometric research during the last decades has been the systematic development of statistical models that allow for cognitive psychometric research (CPR) to be conducted. Many of the models currently available for such purposes are extensions of basic latent variable models in item response theory (IRT). While the requirements of basic IRT models in terms of data collection designs, measurement scales of variables, and sample sizes for learners and items are relatively well understood, the added requirements that the cognitive components of modeling processes for CPR bring along with them are understood to a much lesser degree. Therefore, this article contains a guide in the form of a series of questions, which support measurement and substantive specialists in describing and investigating the conceptual foundations and statistical properties of models for CPR. An application of the guide to the rule-space methodology (e.g., Tatsuoka, 1983, 1985, 1995) is presented, which extends the conceptual-logical discussion of Gierl, Leighton, and Hunka (2000) by focusing more heavily on cognitive and statistical considerations.

Key words: item response theory (IRT), cognitive modeling, rule-space methodology

Correspondence should be addressed to André A. Rupp, Institut zur Qualitätsentwicklung im Bildungswesen (IQB), Humboldt-Universität zu Berlin, Unter den Linden 6, Sitz: Jägerstrasse 10 / 11, 10099 Berlin, Germany. E-mail: andre.rupp@iqb.hu-berlin.de

With the integration of information from cognitive psychology into psychometric modeling processes, either through explanatory post-hoc analyses or through explicit operationalizations of construct components through model parameters, a new realm of psychometric research has emerged at the end of the last century. Specifically, researchers nowadays are increasingly looking to develop assessments that transcend a mere norm-referenced, rank-ordering functionality and include criterion-referenced diagnostic capacities so that stakeholders have detailed information about learner characteristics and assessment properties at their disposal (see, e.g., de Boeck & Wilson, 2004; Embretson, 1998; Junker 1999; Mislevy, 1996; Nichols, Chipman, & Brennan, 1995; Snow & Lohman, 1989). Developments of such assessments have been facilitated, on the one hand, by research findings in cognitive psychology that form the foundations of such assessment systems and, on the other hand, by advances in psychometric research fueled by the rapidly increasing power of computer systems in the 1990s (for an overview see, e.g., Rupp & Mislevy, in press).

But with the conceptual and computational power comes the danger that researchers may be overwhelmed by the complexity of the statistical models, the complexity of the psychological theories, or the complexity of the linkages between these two contributing components unless they are quite knowledgeable or even experts in multiple research areas. To support researchers in tackling these conjoint complexities, this article presents a guide for the systematic description and investigation of the conceptual foundations and statistical properties of psychometric models useful for scoring assessments whose structure and design are explicitly based on theories about response processes from cognitive psychology. The guide consists of a set of questions that—if carefully answered either through conceptual analysis and expert consensus or through empirical investigations—facilitate a deeper understanding of the strength and limitations of the statistical models at the core of such assessments and show areas of research where unanswered questions await empirical study.

Its primary aim is, thus, to stimulate and facilitate constructive discussions about the features of such assessments and their constituent statistical models that encourage measurement and substantive specialists to make overt the many subtle assumptions that often implicitly inform their practice. Concurrently, its aim is to aid those researchers in the development of designs for either generating simulated or collecting real data that are needed in order to understand the proper functionality and scope of applicability of the statistical models across domains. Put differently, this article is of a conceptual rather than empirical nature, because it is believed that many specialists would benefit from a guide that allows them to discuss meaningfully the strength and limitations of different statistical models for assessments with a diagnostic purpose grounded in cognitive theories of response processes.

To accomplish these goals, this article is divided into two main parts. In the first part, the development of this guide is situated within the context of other recent relevant developments in psychometric modeling theory. In the second part, the guide itself is presented and rationales for the selection of the individual constituent questions are provided. Moreover, the application of the guide is illustrated concurrently with the *rule-space methodology* (e.g., Tatsuoka, 1983, 1985, 1995), a probabilistic pattern classification approach for diagnosing learners' competencies in various skill domains.

THEORETICAL BACKGROUND

In this section of the article, relevant terminology will be introduced, important statistical developments and cognitive considerations for model development will be presented, and influential related conceptual frameworks will be discussed.

Terminology

In his description of a methodological framework for performing psychometric research informed by cognitive psychology, Nichols (1994) coined the phrase *cognitively diagnostic assessment*, which also became the title of a subsequent seminal book (Nichols et al., 1995). While this phrase accurately captures the integration of cognitive psychology into the design and interpretation of assessments—predominantly from education, sociology, and the health sciences—it also seems to suggest that their main purpose is the *diagnosis* of learner characteristics and subsequent profiling of learners. This is limiting, however, because several strands of research that draw on information from cognitive psychology and involve such assessments are only marginally concerned with diagnosis.

For example, the area of *automated item generation* (e.g., Diehl, 2002; Embretson, 1999b; Gorin, 2005; Irvine & Kyllonen, 2002) is concerned predominantly with the validation of the constructs that the psychometric instruments are supposed to be tapping and the subsequent collection of empirical evidence for the relationships of these constructs within a nomological network. This line of research includes the explicit modeling of model parameters like item difficulty so that novel items with known calibration values and desired psychometric properties can be automatically generated and banked. Therefore, providing diagnostic information to learners is typically not the primary goal of these studies even though it may be a feasible byproduct of the cognitive enterprise. To circumvent the semantic restriction imposed by the term "cognitively diagnostic assessment," the terms *cognitive psychometrics (CP)*, *cognitive psychometric model (CPM)*, and *cognitive psychometric research (CPR)* will be used in this article.

A few words on the term "statistical model" are useful at this point as well. As authors such as Zumbo and MacMillan (1999) or Rupp (2002) remind us, and as is exemplified in heated discussions about model equivalency [e.g., Fischer, 1993, and Wilson, 1993], such a definition is not easy to develop. Nonetheless, how one defines "statistical model" has significant implications for the types of statistical structures that are going to be included in any subsequent discourse. In this article, "statistical models" is meant to stand for mathematical equations that formalize probabilistic processes and whose parameters can be estimated with appropriate routines for real and simulated data-sets under certain required assumptions.

Specifically, CPMs include parameters that operationalize components of response processes or mental faculties, whose existence can be justified through theories grounded in cognitive psychology. The parameters are provided by specialists and are typically collected in so-called *Q-matrices*. A Q-matrix is an item × attribute matrix that lists which cognitive attributes (e.g., aspects of abilities, aspects of complex cognitive processes) are required to solve each item to what degree. Its entries may be binary (i.e., indicating the absence or presence of a component) as well as ordinal or even continuous (i.e., indicating the degree to which a component is present). The entries for the Q-matrix are always provided by experts from the domains under study and are subject to empirical verification and rational consensus. In fact, since the Q-matrix is a fundamental building block of these models, it is essential that the choice of entries can be justified comprehensively to avoid a "garbage-in, garbage-out" phenomenon where interpretations of model parameters become questionable because the cognitive information used to build the model was weak.

In many CPMs such as the *linear logistic test model* (LLTM) [e.g., Fischer, 1973, 1997], the *DINA* and *NIDA models* (e.g., de la Torre & Douglas, 2004; Junker & Sijtsma, 2001), or the *rule-space methodology* (e.g., Tatsuoka, 1983, 1995), the expert ratings in the Q-matrix remain fixed in the calibration stage, whereas in some models such as the *reparametrized unified model/fusion model* (e.g., Bolt & Fu, 2004; diBello, Stout, & Roussos, 1995; Hartz, 2002; Hartz, Roussos, & Stout, 2002; Templin, Roussos, & Stout, 2003) the entries are subject to empirical updating. Since most CPMs seek to classify learners into a number of unobservable classes whose number is predetermined by the number of cognitive components and the number of levels each can take on, CPMs are typically *multiple classification* or *restricted latent class models* (e.g., Haertel, 1989).

Not only the number of cognitive components represented through CPMs has to be provided by specialists, but also the structure of the statistical model should be chosen to match the way in which learners combine the cognitive components in responding to items (see Maris, 1995). For example, in componential IRT models, the response processes are typically modeled in a noncompensatory fashion meaning that all required cognitive components need to be activated by learners to a sufficient degree to respond successfully to an item. This has important practical consequences as data on both subtasks and the composite tasks may be required (e.g., Hoskens & de Boeck, 1995, 2001; Janssen & de Boeck, 1996a, 1996b; see also Embretson, 1983). As a result of these considerations, thinking about "statistical models" for CPR also implies thinking quite profoundly about "cognitive process models" at some level of detail so that the statistical model and the psychological process become intricately intertwined.

Important Statistical Developments for CPR

One of the deciding features of psychometric research in the 1990s was a strong push toward a unification and integration of conceptual and statistical approaches to assessment where it has become clear that many approaches that had previously been discussed separately in the literature share common roots (e.g., Goldstein & Wood, 1989; McDonald, 1999; Mellenbergh, 1994; Muthén, 2002; Takane & de Leeuw, 1987). This unification and integration process has been facilitated through the refinement of Bayesian estimation methods, which have significantly broadened the classes of estimable psychometric models and the classes of complex data structures that these can accommodate (for overviews see, e.g., Gelman, Carlin, Stern, & Rubin, 1995; Patz & Junker 1999a, 1999b; Rupp, Dey, & Zumbo, 2004). At the same time that these psychometric foundations were brought together, the dominant discourse surrounding inferential validity advocated a unitary and integrative conceptualization of construct validity, which takes into account notions of causality and the consequences that inferences from assessment processes have for all agents involved in those processes and the disciplines they are embedded in (e.g., Borsboom, Mellebergh, & van Heerden, 2004; Messick, 1989, 1995).

The most important statistical developments for modeling data from designs that foster CPR have been in the area of parametric latent variable modeling, most notably *item response theory (IRT)* [e.g., Embretson & Reise, 2000; van der Linden & Hambleton, 1997]. In basic IRT models, unobserved latent variables are created as reflexive indicators of the constructs that the assessment instrument is tapping (Borsboom, Mellenbergh, & van Heerden, 2003) meaning that the latent variables are constructed to represent unobservable characteristics of the learners such as proficiencies, aptitudes, or dispositions. These models allow for the joint calibration of item parameters (e.g., difficulty, discrimination, pseudo-guessing) and learner parameters (e.g., proficiency, aptitude, disposition) on a common scale thereby allowing for scale equating, item banking, and analyses of parameter stability across learner groups, time points, and measurement conditions. Despite their statistical flexibility, however, the latent scales represent rather *undifferentiated continua* and, thus, provide only global information about latent learner characteristics akin to other latent variable models. To increase the interpretative and practical value of basic IRT models, they have recently been augmented to allow for the incorporation of external information about model parameters that often have origins in response process models from cognitive psychology (e.g., Junker, 1999; Embretson, 1999a; National Research Council, 2001); therefore, CPMs are also called *structured IRT models* by some researchers (e.g., Rupp & Mislevy, in press).

Important Cognitive Considerations for CPR

Importantly, desiderata about rich cognitive information for interpreting assessment results always come at a price and the level of detail required for CPR always depends on the purpose for which an assessment is conducted. In CPR that seeks to validate psychological constructs, for example, response process models that have been developed from laboratory studies in experimental cognitive science might be required. The development of accurate processing models from first principles in cognitive science is laborious and time-consuming, however, and mapping functions between psychological constructs and mathematical model parameters may be hard to establish. Moreover, since components of cognitive processes need to be explicitly operationalized through statistical variables in CPMs, this may not only require that an explicit response process model for a skill such as "spatial rotation" or "basic number addition" be available along with hypotheses about what its constituent components are, but, also, that information be available concerning whether the cognitive components are linked in a compensatory or noncompensatory fashion along with how they are weighed in their conjunction (e.g., Embretson, 1998; Samejima, 1997).

At other times, an exact process model may not be needed as it would be too fine-grained; instead, the specification of a more proximate model of the psychological response process may suffice. For example, this level of application of a CPM may draw on a general theory about textual comprehension and postulate the existence of certain skills that are required by learners to perform these tasks, which are then integrated into a CPM (e.g., Embretson & Wetzel, 1987; Buck, Tatsuoka, & Kostin, 1997).

The different levels of specificity of cognitive information for the approaches discussed above is, thus, driven by available statistical methods and their requisite data structures, the available substantive theories with respect to the tasks of interest, and, most importantly, the specific purposes of the cognitive modeling exercise. They are specifically linked to the *grainsize of the feedback* that is desired about learners when those models are used for cognitive diagnosis of learner characteristics. Such feedback can range from cognitively coarse—as in state-wide accountability reports of educational performance—to cognitively fine-grained—as in reports within individualized tutoring systems (Rupp & Mislevy, in press) and requires process models, CPMs, and reporting mechanisms at different levels of detail for each case.

As the previous discussions have shown, the process of modeling data for CPR faces numerous challenges that are comparable to those known in most modeling enterprises in statistics but also go beyond those. Therefore, the same criteria for responsible and informed modeling that hold for these general statistical methodologies developed outside the realm of CPR should be applicable to modeling processes in CPR. Specifically, investigations should aim to collect empirical evidence that is gathered through the analysis of Monte-Carlo studies and real-life data-sets that show the degree of stability and replicability of results across domains. This is, of course, not a novel insight per se, but a glance at the current literature points to a lack of an official coherent reminder of this important issue. Further, there are no structured support systems that can facilitate informed discussions about systematic investigations with these specific purposes, which is why the guide in this article is presented.[1]

Conceptual Frameworks Relevant for CPR and the Development of this Guide

The literature does, however, provide frameworks for CPR that guide researchers in the design of assessments, scoring mechanisms, and reporting platforms. For example, Nichols (1994) describes an overarching framework for *cognitively diagnostic assessment*, Embretson (1983, 1994, 1998) describes a similar framework in multiple steps, coined the *cognitive design system*, and Mislevy, Steinberg, and Almond (2003) describe another overarching framework for structured assessment system development, coined *evidence-centered design*. All of these frameworks are indispensable for structuring researchers' thinking about what kinds of inferences they desire, how to collect evidence for these inferences in a principled manner, and how to translate this evidence into statements about learner and assessment system characteristics. Despite their commonalities, however, they differ in the types of research they are predominantly concerned with. Specifically, the work by Nichols (1994) was strongly motivated by *cognitive diagnosis*, the work by Embretson (1983, 1994, 1998) was strongly motivated by *construct validation* and *automatic item generation*, and the work by Mislevy et al. (2003) was strongly motivated by *intelligent tutoring systems* and *authentic assessments of complex skills*.

Therefore, the guide presented herein finds its place amidst these existing framework by focusing specifically on the statistical and cognitive kernels of CPMs; thus, it can be seen as complementary to them. As stated earlier, the guide is presented as a series of questions that facilitate a deeper understanding of the strengths and limitations of the statistical models and cognitive considerations that

[1] From a statistical viewpoint, the elegant development of a modeling guide from first principles is exemplified by the work in the *William Stout Institute for Measurement* (e.g., Stout, 2002), which is concerned predominantly with the investigation of the dimensionality structures of assessments and its differential functioning across multiple populations and measurement conditions. However, it is not a guide for cognitive psychometric research in the sense discussed herein.

are at the core of CPR and show areas of research where unanswered questions await empirical study.

DESCRIPTION OF THE GUIDE

The decision to present the guide as a set of questions goes back to a similar guide for reviews of software programs by Gierl and Ackerman (1996). As is evident from Figure 1, the guiding questions are concerned with three distinct aspects of CPMs: (1) their cognitive foundations, (2) their statistical foundations, and (3) their application foundations.

It is important to note that the categories used for some questions were not designed to be exhaustive, which is why an "other" category is always included. Consequently, the practical value in the guide lies not in its completeness, but in its function as a *consciousness-raising device* and as a *formal basis for discussions among experts* from different disciplines charged with developing a CPR program.

Consequently, answers to the questions listed in this guide for different CPMs empower different agents involved in interdisciplinary CPR in different ways. For example, more applied measurement specialists are aided in determining which CPM is most suitable for their specific purpose, more theoretical measurement specialists are aided in designing studies that seek to investigate the sensitivity of such process across experimental conditions, and cognitive scientists are aided in synthesizing relevant research in such a manner that results can be operationalized for different CPMs. It is hoped that once a body of research on different CPMs has been collected over the years, the inferential limits of each CPM for a given purpose will become more transparent and researchers will be better able to make judicious choices about which models to select or to develop further for their purposes.

Further below, rationales for the inclusion of individual questions are presented and they are applied to the rule-space methodology. Before that can be done, however, a few words on the rule-space methodology itself are necessary.

Overview of the Rule-Space Methodology

The rule-space methodology was developed in a program of research by Kukimi Tatsuoka (1983, 1985, 1986, 1993, 1995), and combines general latent variable modeling within an IRT framework with Bayesian cluster analysis. Its primary aim is to classify learners into one of several possible distinct "knowledge" or "attribute" states. Work on the methodology was motivated by the analysis of "buggy rules," the different types of errors that learners make as a result of their misconceptions in certain domains, specifically basic arithmetic (e.g., Tatsuoka & Eddins, 1985). The idea behind it is as simple as it is appealing, which has led to its wide-

I. Cognitive Foundations

1. What are the purposes of the cognitive psychometric research with the CPM?
   (a) Construct validation
   (b) Learner diagnosis and profiling
   (c) Item analysis and generation
   (d) Psychometric analysis of model properties
   (e) Other
2. What is the cognitive grainsize of the feedback about learners that is desired with the CPM?
   (a) Very coarse (e.g., norm-referenced state-wide accountability reports)
   (b) Coarse (e.g., norm-referenced summative school-level and classroom assessment)
   (c) Fine (e.g., criterion-referenced formative classroom assessment)
   (d) Very fine (e.g., criterion-referenced individualized diagnosis)
   (e) Other
3. What specific inferences are desired with the CPM?
   (a) About learners
   (b) About items
   (c) About assessment conditions
   (d) Other
4. What is the desired mapping between constructs in the nomological network and the statistical structure of the CPM?
   (a) Mathematical model as data reduction instrument or filtering device
   (b) Mathematical model as approximate representation of response process
   (c) Mathematical model as exact representation of response process
   (d) Other
5. Which variables are used to operationalize different characteristics for learners and items in the CPM?

II. Statistical Foundations

Design Components

6. What are mathematically admissible (i.e., theoretically tractable) and empirically admissible (i.e., practically estimable) measurement scales in the CPM?
   (a) For each score variable associated with items
   (b) For each latent variable associated with learner characteristics

FIGURE 1 A guide for describing and investigating the conceptual foundations and statistical properties of cognitive psychometric models.

7. What data-collection designs can the CPM accommodate?
   (a) Cross-sectional
   (b) Longitudinal
   (c) Other
8. What sampling structures can the CPM accommodate?
   (a) Single-level
   (b) Multi-level
   (c) None
   (d) Other
9. What dependencies can the CPM accommodate?
   (a) Testlet dependencies
   (b) Strategy dependencies
   (c) Stage development dependencies
   (d) None
   (e) Other

Estimation Components

10. Which software packages are available for estimating the CPM?
11. What sample size requirements exist for stable parameter estimation with the CPM?
12. What specifications about the data structure are available for the CPM?
    (a) What distributional assumptions are made about item score variables?
    (b) What distributional assumptions are made about latent variables?
    (c) What other assumptions are required to estimate the model?
    (d) How sensitive are parameter estimates to violations of these assumptions?
    (e) What practical implications do these sensitivities have?
    (f) Other
13. What fit statistics are available to estimate model fit?
    (a) Global fit
    (b) Person fit
    (c) Item fit
    (d) None
    (e) Other

FIGURE 1 *(Continued)*

14. What types of missing data are tractable with the software for the CPM?
    (a) MCAR
    (b) MAR
    (c) MNAR
    (d) None
    (e) Other

III. Application Foundations

15. What domains and cognitive processes is the CPM most suitable for?
16. What extensions of the CPM would be most beneficial?
17. What guidelines should be given for reporting the results of an analysis with the CPM?

FIGURE 1 *(Continued)*

spread applicability for CPR; it can be best described as a series of steps as follows (see Gierl et al. 2000).

First, test developers specify an overall list of so-called cognitive "attributes," which are characteristics of learners that are required to answer a set of items, along with all of their direct and indirect relationships. In addition, they specify exactly which subset of attributes is required, conjunctively, to answer each item in order to obtain a specific maximum score (e.g., a score of "1" on a simple binary scale or a score of "3" on a graded scale). These pieces of information are encoded in different matrices: the *direct* attribute relationships (i.e., which attributes are dependent on other attributes when pairs are considered) are encoded in an *attribute matrix* **A**, the *direct and indirect* relationships (i.e., the implied dependencies of attributes arising from the pairwise dependencies) are encoded in a *reachability matrix* **R**.[2] Finally, as stated earlier, the *item-by-attribute pattern* is encoded in the incidence matrix **Q**.

The number of possible attribute combinations and, thus, *attribute mastery states* is $2^k - 1$ if all attributes are independent (i.e., if **R** is an identity matrix) but is much less if attribute dependencies exist (i.e., if **R** has also off-diagonal entries of 1s). The full **Q**-matrix can, thus, often be *reduced* by using Boolean algebra that capitalizes on the dependencies between the attributes as specified in **R**. It is important to note at this point, however, that the process of specifying the different matrices does *not* require any actual data but, instead, rather sophisticated cognitive theories about response processes.

Once a reduced **Q**-matrix is available and learners have answered the assessment items, the data are analyzed *independently* with an appropriate IRT model,

[2] Note that the **R** matrix can also be numerically obtained from the **A** matrix.

typically a unidimensional Rasch model (i.e., a model with one continuous latent variable typically denoted θ that represents an overall proficiency, aptitude, or disposition). Using the attribute specifications for the items on the assessment in the **Q**-matrix, one can then compute *ideal total scores* and *ideal latent variable scores* for each attribute configuration, which are the scores that a learner with a given attribute pattern would obtain if he or she responded without error.

Then, using the latent variable estimates from the IRT model a *standardized residual function*, denoted ζ, is estimated (Tatsuoka & Tatsuoka, 1987). It estimates how "atypical" the actual response pattern of a given learner is compared to the expected response pattern predicted jointly by the IRT model and the Q-matrix specification for the items.[3] Subsequently one can plot each learner using his or her latent variable estimate, $\hat{\theta}$, and his or her standardized residual function estimate, $\hat{\zeta}(\hat{\theta})$, in a two-dimensional plot called the *rule-space*. Sets of points tend to cluster elliptically around the *ideal score centroids* that correspond to the specific attribute-mastery states. In the rule-space, each learner is, thus, represented by a point with two coordinates $(\hat{\theta}, \hat{\zeta}(\hat{\theta}))$.

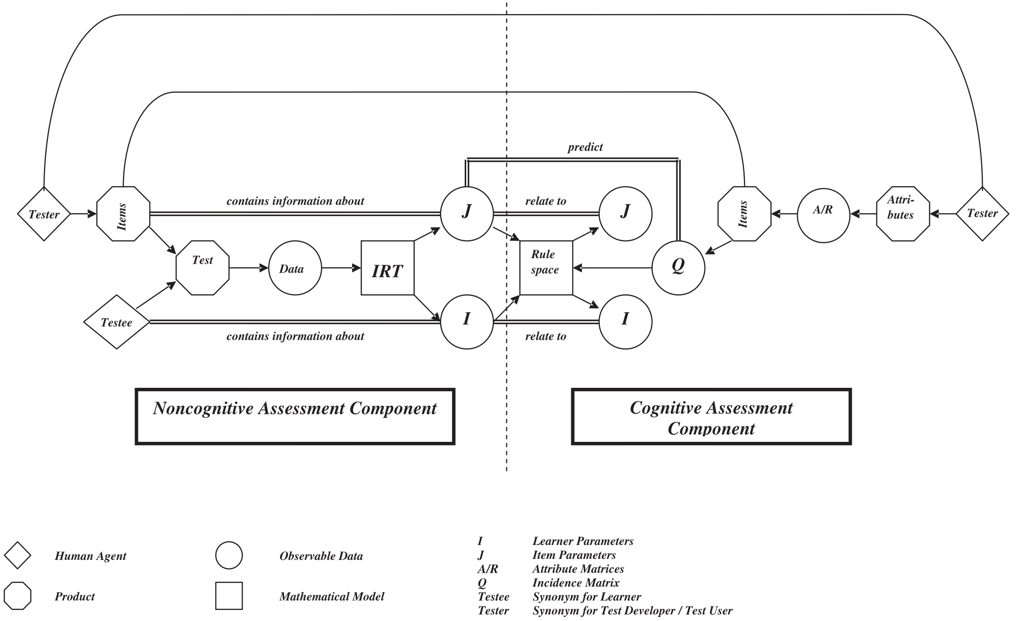
Under the reasonable assumption that these points follow a bivariate normal distribution and through the utilization of a bivariate distance measure, most learners can then be classified into one attribute-mastery state with the highest probability. This effectively assigns each learner a *profile* of which attributes he or she has most likely mastered and which ones he or she has most likely yet to master. Note, however, that learners with highly unusual response patterns may remain unclassifiable.

Based on these classifications, *attribute-mastery plots* can be developed for the sample, which show the relationship between the latent variable estimate of learners and the probability of mastering each individual attribute. That is, the plots show how "easy" or "difficult" it is to attain a certain attribute for learners with different levels of proficiency, aptitude, or disposition. In addition, so-called *proficiency scales for attributes* can be developed, which show how the mastery of different attribute combinations is related to different estimated proficiency scores from IRT revealing optimal pathways for attribute mastery (Tatsuoka, 1995; see also Templin & Henson, 2004).

To summarize how different agents, entities, and data components interact and influence in the flow of information in an application of the rule-space methodology, see Figure 2, which contains an explanation of the different elements at the bottom.

One can argue that the cognitive and noncognitive components provide an overarching frame that contains the rule-space methodology and its constituent IRT model as the core, which have a central "conductor" role for the data in both direc-

[3] This function measures the total residual discrepancy between observed scores and predicted scores for individual items as well as the total residual discrepancy between the predicted scores for individual items and the *average* predicted scores across all items.



In the ***non-cognitive assessment component***, test-developers ('testers') develop items ('items') that are assembled in an assessment ('test') that is administered to learners ('testee'). Once the responses ('data') from the assessment have been calibrated with an IRT model ('IRT'), estimates of learners' proficiencies ('I') and operating characteristics of items ('J') are available.

In the ***cognitive assessment component***, test-developers ('testers') specify characteristics of learners ('attributes') and encode their relationship to one another in matrices ('A/R'). These matrices are used as a blueprint for item development ('items'), which is encoded in a Q-matrix ('Q'). The rule-space methodology then uses the estimates from learns and items generated by the IRT model as well as the Q-matrix provided by the experts to produce item characteristics ('J') and learner attribute classifications ('I'). These can be related to the IRT estimates via separate plots and regression models.

The arches at the top of the figure highlight that the test-developers are involved in both assessment components—perhaps in different teams—and that the item blueprint from the cognitive assessment component is linked to the items themselves.

FIGURE 2 Flow of information within the rule-space methodology.

tions. What the diagram does not show, however, is the important process of interpreting the results from such an analysis, which is key for the validity of the inferences drawn from the assessment; this will be addressed further below.

Explanations of Questions and Application to the Rule-Space Methodology

In this section of the article, the intention of each question in the guide will be discussed either individually or in sets before they will be applied to the rule-space methodology for illustration.

Questions 1–3: Purpose, Feedback, and Specific Inferences

Modeling of data for CPR is statistically located at the intersection of two measurement spaces—those of learners and items (e.g., Holland, 1990; Zimmerman & Zumbo, 2001; Zumbo & Rupp, 2004). These questions, therefore, encourage a structured thinking about the kinds of statements that are of interest for learners and items as well as the cognitive dimension that is overlaid onto them. In other words, the questions are included to make explicit the breadth and the relative weighting of different desired inferential statements. More importantly, since it is the purpose of the CPR that guides the grainsize of the cognitive information and their operationalizations through model parameters, answering these questions carefully is one of the most important steps in designing a cognitively grounded assessment and selecting an appropriate CPM for the purpose at hand.

The first question lists multiple purposes for the CPR, because in practice different stakeholders typically have competing objectives. It concerns specifically the reasons why cognitive considerations are relevant to the assessment in the first place, which is directly related to the grainsize of the cognitive information that is required to produce the desired types of feedback. The semantic labeling of "very coarse" to "very fine" is not as important here as is the fact that it should serve as a basis for discussion among the experts of whether multiple grainsizes of cognitive detail are needed and which CPMs and data structures can help to provide them. Most of the current modeling processes in CPR are designed for simpler skills (e.g., basic arithmetic, logical reasoning, spatial rotation) that can be decomposed into their componential subskills with relative ease. Much confusion about the utility and inferential capacities of CPMs can arise if the shades of "cognition" are not clearly and, most importantly, *explicitly* differentiated within the context that the CPM is applied. Again, this question is included to force researchers to unearth their hidden assumptions about what "cognition" means to them to underscore that most assessment contexts would probably not require an analysis of which brain areas serve which mental processes but, rather, an analysis of processes in the form of production rules.

The third question then serves to force the specialists to specify which types of statements are exactly required. It is motivated by the understanding that inferences can be categorized according to criteria such as time (e.g., whether they concern description of present abilities or prediction of future performances), referencing (e.g., whether they concern individual learners as in criterion-referencing or groups of learners as in norm-referencing), or purpose (e.g., whether diagnostic or admission decisions are based on them).

Rule-Space Analysis

Most rule-space analyses have been used to develop diagnostic profiles of learners with any information about the construct seen as a byproduct of the enterprise. Moreover, the term "attribute" has been used with various meanings in applications with the rule-space methodology and its semantic scope has included (a) a componential subskill possessed by learners (e.g., adding a term to both sides of an equation for a basic arithmetic task; see Birenbaum, Kelly, Tatsuoka, & Gutvirtz, 1994), (b) an interaction of subskills that are, arguably, representative of higher-order skills (e.g., making inferences across sentences; see Buck, Tatsuoka, & Kostin, 1997), and (c) an attribute that resides in an item (e.g., the type of input such as picture or diagram for a complex architecture task; see Katz, Martinez, Sheehan, & Tatsuoka, 1998).

As discussed earlier, the output of the rule-space methodology is primarily *probabilities*, which allow one to classify learners into attribute-mastery states and, secondarily, a representation of the entire attribute-mastery space that shows how learners can progress through the space toward mastery of all attributes (see, e.g., Birenbaum, Kelly, & Tatsuoka, 1993; Tatsuoka & Tatsuoka, 1997). What gives the rule-space methodology its widespread appeal is its conceptual flexibility, because there is nothing in its statistical structure that prevents a researcher from applying it to any assessment where attribute relationships and item-by-attribute patterns can be defined.

It should not be forgotten, however, that response data are the result of an *interaction* between items and learners. Therefore, the usage of a term like "knowledge state" or "attribute state," which implies that the assessment instrument had a neutral function of simply encouraging the learners to bring to bear their true competencies, can be misleading. At a minimum, discussions about applications of the rule-space methodology need to include careful considerations about confounding factors that affect response processes over and above the attributes that the assessment is supposed to tap. This could be partially investigated by empirically assessing the stability of the attribute matrix specifications across learner populations and assessment contexts through carefully manipulated experimental designs.

Questions 4–5: Mapping and Operationalization of Constructs

These questions are two of the most fundamental questions that modelers will have to answer. For example, for CPMs used for construct validation, primary concern may be with an exact mapping of the procedural cognitive realm to the static statistical realm so that, by implication, the statistical modeling structure should reflect or capture the cognitive process structure as in componential models or, perhaps, neural networks.

Some argue that the choice of a CPM is mostly a matter of the measurement scale of constituent latent and observed variables and, as such, is of minor importance, because the mathematical model is a mere "filtering device" that transforms input from one measurement scale into output on an alternative measurement scale (e.g., Junker, 1999, p. 10). It is true that a latent variable is merely a *statistical construction* and does not, in and of itself, represent any *psychological construct*. The relationship between classification probabilities or numerical scores and conceptual entities needs to be made primarily rationally and can only be strengthened with empirical evidence (Borsboom et al., 2004; Messick, 1989, 1995); believing anything else amounts to wishful thinking and is not responsible modeling practice. Therefore, though, a choice of measurement model is actually more than a convenience choice between models with variables on different measurement scales; it entails a commitment to a set of subtle assumptions about causality, representations, as well as inter- and intra-individual differences (Borsboom et al., 2003). The inclusion of these questions encourages researchers to commit to their theoretical stance toward the substantive nature of the modeling enterprise.

Rule-Space Analysis

In the rule-space methodology, the constituent IRT model is a *data-reduction instrument* that transforms observed scores onto a continuous latent scale that allows for quantitative norm-referenced comparisons between learners with desirable mathematical properties. While one could argue that the specification of a **Q**-matrix is a representation of a process model, it is certainly only a very coarse one. This coarseness stems from the fact that the **Q-**matrix is a *static* entity that is often superimposed post-hoc on sets of items that do *not* carefully map out all possible corners of a solution space.

Specifically, the attribute specifications through the **A**, **R**, and **Q**-matrices are viewed as *exact representations* of the psychological *components* that the response process draws on. However, the cognitive component of the model is viewed only as an *approximate representation* of the psychological response *process*, because the manner in which attributes combine in the process remains unspecified. Overall, then, the rule-space methodology is viewed as an *approximate mapping* of certain constructs in a nomological network onto a mathematical structure.

Numerically, it can be observed that the global proficiencies or characteristics that the assessment is supposed to measure are captured, via a coarse metric, through the creation of the *continuous latent variable* in the IRT model and, via a more fine-grained metric, through the creation of the *discrete attribute vectors* that contain the attribute specifications for items and the mastered attributes for learners in different attribute-mastery states. Furthermore, the derived (θ, ζ) coordinates for individual items capture their characteristics such as difficulty and discrimination beyond the IRT item parameters. For example, unusually easy or difficult items will be located in isolated spots in the rule-space and sets of items that discriminate sharply between learners will be located further away from each other in the rule-space.

Question 6: Admissible Measurement Scale

While the previous questions encourage researchers to identify relevant variables via linguistic labels, it is important to explicitly consider the measurement scales that they can be on. It is important to distinguish between those measurement scales that are theoretically possible, which show the potential of a modeling process for accommodating different inferential and operational needs, and those that are currently practically possible, which show the practical limitations of the modeling process due to estimation constraints.

In a **Q**-matrix the typical coding scheme in many applications is binary (e.g., attributes in an incidence matrix are coded with “0” for “absent” or “not required” and with “1” for “present” or “required”). Nevertheless, one could also encode them either on a graded ordinal scale or a continuous interval or ratio scale (see, e.g., Templin et al., 2003; and von Davier, 2005, for different models with similar principles), even though many researchers in CPR would probably only advocate this for characteristics that can be measured with little amounts of error (e.g., processing speed, working-memory capacity). If factors that contribute to the cognitive complexity of items can, theoretically, be coded on any scale that possesses at least ordinal qualities, it would be interesting to investigate how the loss of information in binary scales is offset by the loss of information that results from coding an attribute on an interval or ratio scale perhaps less reliably.

Rule-Space Analysis

For the rule-space methodology, the response variables for items can be on *any* scale that the latent IRT model accommodates, which are nominal (e.g., for IRT models that accommodate multiple choices), ordinal (e.g., for IRT models that accommodate dichotomous or polytomous scores from items scored for correctness), counts (e.g., for IRT models that accommodate unbounded counts such as the mistakes learners make in segments of written text), or continuous (e.g., for IRT models that accommodate interval or ratio measures such as response time). The latent variable in the constituent IRT model is measured on an interval scale, which implies that the resulting ζ index, as a function of it, is as well. Hence, the resulting rule-space is a *2-dimensional space of interval-measured variables*, which is why the assumption of bivariate normality for classification purposes is justifiable. The indicator variables for attributes are binary variables measured on a nominal/ordinal scale. They could *not* be on an interval or ratio scale from either a conceptual-theoretical or practical estimation perspective, because the mathematical classification algorithm relies on the application of Boolean algebra.

Question 7: Data-Collection Designs

This question relates to the question of how association statements or causal statements about variable relationships and, by implication, the constructs that they are measuring, can be justified based on data. Elegant discussions about the use of causal versus association statements in latent variable modeling have recently appeared in the literature (e.g., Borsboom et al., 2003, 2004; Edwards & Bagozzi, 2000) and are of fundamental concern not only for any given empirical study but also, for the design of data collection mechanisms for future studies. Specifically, it needs to be remembered that the application of a latent variable model does not allow one to claim causal relationships between variables unless data come from an experimental design structure, very strong assumptions are made, or mediating statistical mechanisms such as potential-outcome analysis for causal inference are employed.

Rule-Space Analysis

The rule-space methodology can be applied to data from *cross-sectional observational designs* only. While *items* can be designed systematically through its logical approach of generating **A**, **R**, and **Q**-matrices, no statistical mechanism exists to compare attribute classifications, the structure of the attribute-space, or the psychological response-process structure as captured through the above matrices across multiple populations, time points, or conditions. Therefore, only *association statements* are possible for any given application of the model.

Question 8: Sampling Structures

The manner in which observational units, which comprise items and learners, are sampled determines the types of generalizations that are possible from the data structure and the types of statistical models that properly allow for them.

On a general level, Holland (1990) elaborated on the differences between a *stochastic subject* interpretation of probability and a *random sampling* interpretation of probability and showed that the former is consistent with joint and conditional maximum likelihood estimation, whereas the latter is consistent with marginal maximum likelihood estimation. In large-scale assessments, lower-level observational units such as students may be nested within higher-order observational units such as classrooms and this hierarchical data structure with its dependencies might require cognitive hierarchical linear models that adequately represent unit dependencies and lead to unbiased standard error estimates (see Raudenbush & Bryk, 2002, for a general introduction to hierarchical linear models). Furthermore, few discussions of modeling processes discuss the degree to which a violation of the random sampling assumptions for items, learners, occasions, or contexts impacts the accuracy of model parameter estimates specifically in the context of CPMs.

Rule-Space Analysis

To estimate the parameters of the constituent IRT model, either *random sampling without replacement* of learners from populations or the existence of a probabilistic mechanism that is captured through the chosen IRT model are assumed; weakened assumptions such as mere exchangeability of items or learners are not addressed. For the cognitive component of the rule-space methodology, no sampling assumptions are made about the attributes themselves; yet, from a nonstochastic *domain-sampling* perspective in an assessment design phase, it is probably assumed that the set of items for an assessment maps out all corners of the attribute space and that an individual item is a representative member of all items with identical attribute requirements.

To classify learners into attribute mastery states, it is assumed that their $(\hat{\theta}, \hat{\zeta}(\hat{\theta}))$ coordinates are *randomly sampled* from a bivariate normal distribution that has a bivariate mean at a (θ, ζ) coordinate corresponding to an ideal-score response pattern as well as a 2 × 2 covariance matrix that is based on the joint statistical information about the estimators. Interestingly, one may wonder what the backward implications of this random sampling assumption are for the data-generation mechanism, because this random sampling process is overlaid on the random sampling process of the constituent IRT model. The methodology does *not* explicitly take into account complex sampling structures such as those arising from a nested structure of lower-level sampling units within higher-level sampling units (e.g., students within schools), although this could be partially accommodated through using a hierarchical IRT model instead of a basic IRT model as its constituent component (e.g., Fox & Glas, 2001; see also Janssen, Tuerlinckx, Meulders, & de Boeck, 2000). The methodology also does *not* explicitly accommodate domain-stratified assessments or matrix-sampled data through specific statistical mechanisms. The model further assumes that the assessment administration condition is representative of a typical administration condition and that changes in the administration condition do not introduce any additional measurement error.

Question 9: Response Dependencies

Exemplary applications of statistical models are often done with independent responses that are dichotomously scored and answered by a homogeneous population of learners unless the statistical model was decidedly developed for more complex structures. Such structures include item bundles or testlets, which are sets of questions whose responses are dependent, to some degree, due to their reference to a common stimulus (e.g., multiple items that measure reading comprehension referring to a common reading passage). Moreover, they can include the use of different strategies by groups of learners, which would be a dependence of the model parameters for each group on the chosen strategy.

Rule-Space Analysis

Since the constituent measurement model for the rule-space methodology is an IRT model that produces estimated ability values, models that include parameters for testlet dependencies (e.g., Wainer & Wang, 2001; Wang, Bradlow, & Wainer, 2002) could be used. In order to model multiple strategies or stages of development, models are also available (e.g., Wilson, 1989) even though the ability estimates from the different classes would have to be comparable in meaning for the rule-space analysis to be interpretable also. However, applications for these situations have, so far, not appeared in the literature and the software to estimate the rule-space methodology does not formally allow the user to incorporate these complex models.

Question 10: Software Packages

Obviously, the estimability of CPMs is one of the important practical desiderata for CPR. It is the unavailability of freely accessible and, more importantly, user-friendly software programs within, for example, a Windows™ environment that has hindered their widespread use both in practice and in research.

Rule-Space Analysis

To perform a complete rule-space analysis one currently requires the program BUGSHELL (Tatsuoka, Varandi, & Tatsuoka, 1992) which is available as a license through TANAR software for research purposes only,[4] which, unfortunately, makes its practical value for applied specialists limited. Furthermore, the software is written to run in DOS on restricted systems only, which also limits its user-friendliness. Freely available, however, are a tutorial for understanding the conceptual foundations of the rule-space methodology by the CRAME Research Center at the University of Alberta.[5]

Question 11: Sample Size Requirements

Unstable parameter estimates certainly point to estimation error and the amount of estimation error is inversely proportional to the amount of statistical information in the data. Yet, it is important to ask how a lack of statistical information translates to a lack of conceptual information about learner, items, and their interaction. This impacts the classes of inferences that are drawn about them, because the CPM is applied to data that are a result of the interaction of items and learners; any uncertainties in model parameters are thus related to all classes of inferences in a complex manner.

The question of required sample size and model sensitivity is an issue that has been extensively investigated for basic IRT models (e.g., Stone, 1992; Kirisci, Hsu, & Yu, 2001; see also Rupp, 2003) but, again, needs to be investigated more extensively for CPMs as well. While one can read comments about data requirements in selected articles, systematic investigations of the sensitivity of sample size requirements for CPMs seem to be largely lacking at this point with only a few notable exceptions (Hartz, 2002; Douglas & de la Torre, 2005).

Rule-Space Analysis

For the constituent IRT model in the rule-space methodology, sample size requirements for learners depend on the number of items, the complexity of the model, and the information about item and learner parameters that is available in the data. Simplistic estimates may range anywhere from 100 learners for a 1-parameter IRT model of small or moderate length (e.g., 40–60 items) to around 1000 for a three-parameter IRT model of similar length (see Rupp, 2003, and studies cited therein). For the cognitive component of the rule-space methodology, classification accuracy appears to depend on the number of (θ, ζ) coordinates that cluster around the ideal-score centroids, because they are the data that allow for a reliable estimation of the bivariate covariance matrix. The cluster patterns, in turn, depend on the number of possible attribute-mastery states, which depends on the

[4] Interested researchers should contact Curtis Tatsuoka (tatsuoka@prodigy.net) to discuss licensing options.

[5] The tutorial can be downloaded from www.education.ualberta.ca/educ/psych/crame/research.htm

number of attributes and the number of items that require learners to tap each attribute or combination of attributes.

Question 12: Model Assumptions

This question has traditionally been one of the driving forces for the development of robust and nonparametric statistical methods (e.g., Conover, 1999) and can be investigated through simulation studies (see, e.g., Muthén & Muthén, 2002). For example, while Likert scales are ordinal in nature and are thus, theoretically not appropriate for estimation processes that assume an interval or ratio scale, Monte-Carlo studies on various models have shown repeatedly that for about five score points and relatively symmetrical score distributions, inferences from procedures for interval or ratio scale data applied to Likert scales are basically identical to the ideal continuous conditions for all practical purposes (e.g., Muthén & Kaplan, 1985, 1992).

Item response theory models are nowadays typically estimated with marginal maximum likelihood in a fully Bayesian framework and are full-information models that accommodate discrete and continuous response variables. If IRT models are estimated through an equivalent partial-information factor analytic model that decomposes covariance structures, either assumptions about normality of responses have to be made, tetrachoric correlation matrices have to be analyzed, or factor-analytic models for ordinal scales have to be employed (e.g., du Toit, 2003; McDonald, 1999).

Rule-Space Analysis

While the sensitivity of many standard statistical models to violations of response variable distributions has been investigated, the impact of such violations on classification accuracy has not been investigated extensively for the rule-space methodology. It is worth remembering that additional possibilities of model misspecification exist. For example, item response functions in the IRT model may not be monotonic or the specified dimensionality of the model may not match the dimensionality of the data-generation mechanism. For the cognitive component, the attribute and attribute relationship specifications may be incorrect and/or incomplete and may not map onto the response process that learners actually engage in. Therefore, studies such as those by Baker (1993), Hartig (2004), or Hartz (2002) are welcome and generally needed for CPMs and for the rule-space methodology specifically.

Question 13: Fit Statistics

To choose a certain IRT model that "best" fits the data, fit statistics may be employed that are sensitive to violations of the assumptions discussed in the previous questions. A variety of such statistics exists for latent variable models generally (e.g., Hu & Bentler, 1998, 1999) and IRT models more specifically (e.g., Meijer & Sijtsma, 2001; Orlando & Thissen, 2000; Sinharay & Johnson, 2003). However, they are not always easy to compute and typically only a restricted set is implemented into software programs. The fit of CPMs is, of course, also determined by the specification of the Q-matrix, but it is still a relatively open question which fit indices are best suited to differentiate model misfits in terms of model structure and Q-matrix specification.

Rule-Space Analysis

For the constituent IRT models, fit statistics should be reported, but seldom are in applications of the methodology. It appears as if the cognitive focus of the studies supersedes the need for a comprehensive reporting of statistical information. While most studies report the classification rates that the methodology is able to attain for a given data set, no cut-off values exist that would allow readers or analysts to judge what an "acceptable" classification rate is that reflects adequate model fit. Similarly, regression analyses are often used to regress the Q-matrix variables onto the estimated IRT difficulty values. Apart from underestimating the estimation error in the difficulty values in this case, there is also no available cut-off that would help determine whether a misspecification of the Q-matrix exists and what the nature of this misspecification is.

Question 14: Missing Data

This question concerns the types of missing data that are accommodated by the software that is currently available for a given CPM. Research on missing data has formalized the types of mechanisms that have given rise to missing data (e.g., missing completely at random, missing at random, nonignorable) and associated estimation algorithms that can accommodate these structures quite generally have been developed (Little & Rubin, 2002; Schafer, 1997). As a consequence, algorithms for the imputation of missing response data or the indirect imputation of latent variable values are typically already included in software programs such as Mplus (Muthén & Muthén, 2004), BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 1996), or ConQuest (Wu, Adams, & Wilson, 1998) within the area of general latent variable modeling and their relevance for CPR needs to be documented in more detail as well.

Rule-Space Analysis

Since missing data in the rule-space methodology will stem from the response matrix, the same considerations as for general IRT models apply. It would be worthwhile to investigate how the additional uncertainty in model estimates arising from imputed data impacts classification accuracy for different learner and item sample sizes and different amounts and patterns of missing response data.

Questions 15–16: Practical Applicability and Extensions

In a straightforward manner, answers to these questions can reveal fundamental beliefs about the generalizability of inferences from CPMs across domains. Therefore, considering domains for which a CPM is potentially applicable encourages researchers to further unearth their often subtle hidden beliefs about the generality of postulated psychological response processes and their beliefs about how studies could be framed to investigate the sensitivity of the CPM that is selected to capture these psychological processes.

Rule-Space Analysis

The rule-space methodology can be used in two primary ways, (1) *a priori* as a tool to develop assessments that cover attributes and their relationships exhaustively through appropriately designed assessment items and, (2) *a posteriori* as a tool to specify likely attributes and their relationships for an already existing assessment. As shown in an application of the rule-space methodology to a diagnostic form of computer adaptive testing (Tatsuoka & Tatsuoka, 1997), the former application path is preferable. This reasoning stems from the fact that there is even less of a guarantee that one can ensure that attributes derived from already existing assessments are sound or that one can empirically infer their constituent attribute *relationships* post-hoc. Therefore, it can be argued that the rule-space methodology "forces researchers and practitioners to go beyond identifying general, typically imprecise, cognitive skills [...] to carefully identifying and ordering the cognitive skills required to solve problems in a specific content area" (Gierl et al., 2000, pp. 36–37). Of course, this is a strength that is shared by other CPMs, but the rule-space methodology possesses a particular elegance and practical appeal.

Historically, the rule-space methodology has been applied primarily to mathematics, specifically arithmetic (e.g., Birenbaum & Tatsuoka, 1987; Birenbaum et al., 1993; Birenbaum, Nasser, & Tatsuoka, 2005; Birenbaum, et al., 1994; Tatsuoka, 1986; Tatsuoka, Birenbaum, & Arnold, 1989; Tatsuoka, Corter, & Tatsuoka, 2005), but it has recently also been applied to more complex skill domains such as reading comprehension (e.g., Buck et al., 1997; Kasai, 1997; Scott, 1999), listening comprehension (Buck & Tatsuoka, 1998), and architecture (Katz et al., 1998). Theoretically, the model can be applied to any domain for which response process models are available and for which response process components can be operationalized via constituent attributes and attribute relationships. Not surprisingly, however, as with many cognitively analyzed tasks, the rule-space methodology has been predominantly applied to simple tasks within broader task domains, because it is easiest to define subcomponents for those. Once the rule-space model gets applied to complex tasks, interpretations of its output become more difficult to formulate because its mathematical structure struggles to accommodate the psychological task complexity.

On the one hand, one may argue that, descriptively, the classification percentages and the predictive values of the attribute codes for item difficulty values are relatively high in such studies seemingly showing that this may not be as large a problem as perhaps expected. On the other hand, one may argue that it is difficult to make sense of the complexity of the attribute *combinations* in light of the complexity of the attributes themselves, which makes it challenging, at best, to reconstruct a process model from them. In other words, it is difficult to precisely specify the types of simple components that learners draw on when solving a complex task as well as to argue how they interact to form a response process for already existing assessments; yet, the specification of the attribute matrices and its relationship to cognitive processes forms the key to a successful "attribute-driven" assessment task analysis.

Akin to the reparametrized unified/fusion model, it would be beneficial to connect separate aspects of the model into a unifying probabilistic guide. At the moment, the fitting of the IRT model is independent of the attribute specifications resulting, typically, in a separate regression of IRT item difficulty parameters onto the **Q**-matrix indicator variables. However, this approach has typically not been applied to other IRT item parameters and the **Q**-matrix is not integrated into the IRT model as in, for example, the LLTM (see Fisher, 1973).

Question 17: Reporting Guidelines

Undeniably, the manner in which results from assessment studies are reported has an enormous effect on the way these studies are critically disseminated by different stakeholders—especially the larger public (see, e.g., the linear logistic test method special issue on reporting mechanisms by Kohler & Schrader, 2004). Feedback mechanisms for diagnostic studies nowadays exist in different countries (e.g., the *Klassencockpit* system in Switzerland, accessible at www.klassencockpit.ch, or the *Pearson Progress Assessment* system, accessible at http://pearsonpaseries.com). Such reporting mechanisms reveal that the term "diagnostic" is used quite broadly in practice and that it neither guarantees that a fine-tuned cognitive process analysis has been used to develop the assessments nor that a CPM was used to analyze the results. Furthermore, the reporting mechanisms show the importance of translating the rather technical information from cognitive analyses of learner performance so that it becomes practically relevant to the learners, their parents, or their teachers. One of the core objectives of individualized diagnostic assessments is to foster learning and to support both weaker and stronger students in developing optimal learning pathways. Care has to be taken to consider the consequences for these stakeholders in constructing summary statements of performances through attribute profiles, especially when those are used to develop remedial support.

Rule-Space Analysis

The output of an analysis with the rule-space methodology holds great potential for practical use due to the fact that attribute-mastery plots, proficiency scales for attributes, and individual attribute profiles are relatively easy to interpret. However, the sensitivity of the results to many factors discussed earlier requires that analysts provide more comprehensive evidence that the classifications of the model are reliable and trustworthy for individuals, which is not always the case in the applications that have been published.

CONCLUSION

The area of CPR is currently on a brink toward moving from infancy to early childhood and continues to have many advocates that highlight its obvious potential. This means that the area is likely to be more open to numerous novel developments in cognitive psychology and statistical science than areas with a longer and more distinct history. However, researchers may also feel overwhelmed by the breadth and depth of interdisciplinary knowledge that may be required of them to fully understand the tools for CPR so that they can select them judiciously and develop them further.

To support researchers in the complex process of working with CPMs, this article has introduced a guide with a comprehensive set of guiding questions for describing and investigating modeling processes for CPR that brings together conceptual and statistical considerations from research concerned with responsible modeling practice. It was developed to facilitate communication between different agents engaged in CPR so that properties of CPMs become better understood for a larger number of people. The rule-space methodology, a practically and theoretically appealing tool for CPR, was then used to illustrate the kinds of answers the guiding questions in this guide can elicit and the directions for future research it helps to uncover.

The set of questions that are presented in this guide are, of course, not individually original. Quite to the contrary, it attempted to show a few intersections of related statistical and psychometric modeling endeavors to demonstrate how issues in those realms are also relevant for CPR. While some of the questions are addressed selectively in different articles concerned with specific models, authors seem to be reluctant to address them as a set. This probably stems from the fact that they are rather difficult to answer comprehensively and require extensive simulation work and empirical replication. In closing, it should be underscored that the list of questions in the guide is open to discussion and modification. Indeed, it is desirable that it become refined over time so that it can serve as an acceptable procedural guide for comparative investigations of CPMs that lead to detailed descriptions of their conceptual and statistical potential and limitations.

REFERENCES

Baker, F.B. (1993). Sensitivity of the linear logistic test model to misspecification of the weight matrix. *Applied Psychological Measurement, 17,* 201–210.

Birenbaum, M., & Tatsuoka, K.K. (1987). Open-ended versus multiple-choice response formats—it does make a difference for diagnostic purposes. *Applied Psychological Measurement, 4*, 385–395.

Birenbaum, M., Kelly, A.E., & Tatsuoka, K.K. (1993). Diagnosing knowledge states in algebra using the rule-space model. *Journal of Research in Mathematics Education, 24*, 442–459.

Birenbaum, M., Kelly, A. E., Tatsuoka, K.K., & Gutvirtz, Y. (1994). Attribute-mastery patterns from rule space as the basis for student models in algebra. *International Journal of Human-Computer Studies, 40*, 497–508.

Birenbaum, M., Nasser, F., & Tatsuoka, C. (2005). Large-scale diagnostic assessment: Mathematics performance in two educational systems. *Educational Research and Evaluation, 11*, 487–507.

Bolt, D., & Fu, J. (2004, April). *A polytomous extension of the fusion model and its Bayesian parameter estimation.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Borsboom, D., Mellenbergh, G.J., & van Heerden, J. (2003). The theoretical status of latent variables. *Psychological Review, 110*, 203–219.

Borsboom, D., Mellenbergh, G.J., & van Heerden, J. (2004). The concept of validity. *Psychological Review, 111,* 1061–1071.

Buck, G., & Tatsuoka, K. (1998). Application of the rule-space procedure to language testing: Examining attributes of a free response listening test. *Language Testing, 15*, 119–157.

Buck, G., Tatsuoka, K., & Kostin, I. (1997). The subskills of reading: Rule-space analysis of a multiple-choice test of second language reading comprehension. *Language Learning, 47*, 423–466.

Conover, W. J. (1999). *Practical nonparametric statistics.* New York: Wiley.

de Boeck, P., & Wilson, M. (2004). *Explanatory measurement: Generalized linear and non-linear mixed models for item response data.* New York: Springer-Verlag.

de la Torre, J., & Douglas, J. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika, 69,* 333–353.

diBello, L.V., Stout, W.F., & Roussos, L. (1995). Unified cognitive psychometric assessment likelihood-based classification techniques. In P.D. Nichols, S.F. Chipman, & R.L. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 361–390). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Diehl, K.A. (2002). *Algorithmic item generation and problem solving strategies in matrix completion problems.* Dissertation Abstracts International: Section B: The Sciences & Engineering, 64(8-B), Lawrence, Kansas: The University of Kansas.

Douglas, J., & de la Torre, J. (2005, April). *Modeling multiple strategies in cognitive diagnosis.* Paper presented at the annual meeting of the National Council on Measurement in Education, Montréal, Canada.

du Toit, M. (Ed.). (2003). *IRT from SSI: BILOG, MULTILOG, PARSCALE, TESTFACT.* Lincolnwood, IL: Scientific Software International.

Edwards, J.R., & Bagozzi, R.P. (2000). On the nature and direction of relationships between constructs and measures. *Psychological Methods, 5,* 155–174.

Embretson, S.E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin, 93,* 179–197.

Embretson, S.E. (1994). Applications of cognitive design systems to test development. In C.R. Reynolds (Ed.), *Cognitive assessment: A multidisciplinary perspective* (pp. 107–135). New York: Plenum Press.

Embretson, S.E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods, 3,* 380–396.

Embretson, S.E. (1999a). Cognitive psychology applied to testing. In F.T. Durso, R.S. Nickerson, R.W. Schvaneveldt, S.T. Dumais, D.S. Lindsay, & M.T.H. Chi (Eds.), *Handbook of applied cognition* (pp. 629–658). New York: Wiley.

Embretson, S.E. (1999b). Generating items during testing: Psychometric issues and models. *Psychometrika, 64,* 407–433.

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

Embretson, S.E., & Wetzel, D. (1987). Component latent trait models for paragraph comprehension tests. *Applied Psychological Measurement, 11,* 175–193.

Fischer, G.H. (1973). Linear logistic test model as an instrument in educational research. *Acta Psychologica, 37,* 359–374.

Fischer, G.H. (1993). A reply to M. Wilson's paper "The 'Saltus model' misunderstood." *Methodika, 7,* 5–7.

Fischer, G.H. (1997). Unidimensional linear logistic Rasch models. In W.J. van der Linden & R.K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 225–244). New York: Springer-Verlag.

Fox, J., & Glas, C.A.W. (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika, 66,* 271–288.

Gelman, A., Carlin, J.B., Stern, H.S., & Rubin, D.B. (1995). *Bayesian data analysis.* New York: Chapman & Hall.

Gierl, M. J., & Ackerman, T. (1996). XCALIBRE: Marginal maximum-likelihood estimation program Windows version 1.10 [software review]. *Applied Psychological Measurement, 20,* 303–307.

Gierl, M.J., Leighton, J.P., & Hunka, S.M. (2000). Exploring the logic of Tatsuoka's rule-space model for test development and analysis. *Educational Measurement: Issues and Practice, 19*, 34–44.

Goldstein, H., & Wood, R. (1989). Five decades of item response modeling. *British Journal of Mathematical and Statistical Psychology, 42,* 139–167.

Gorin, J.S. (2005). Manipulation of processing difficulty on reading comprehension questions: The feasibility of verbal item generation. *Journal of Educational Measurement, 42*(4)*,* 351–373.

Haertel, E.H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement, 26,* 333–352.

Hartig, J. (2004, July). *Assessing the appropriateness of specifications in LLTM weight matrices.* Paper presented at the biannual meeting of the Society for Multivariate Data Analysis in the Behavioral Sciences (SMABS), Jena, Germany.

Hartz, S.M. (2002). *A Bayesian guide for the unified model for assessing cognitive abilities: Blending theory with practicality.* Unpublished doctoral dissertation, University of Illinois, Department of Statistics, Urbana-Champaign, IL.

Hartz, S., Roussos, L., & Stout, W. (2002). *Skill diagnosis: Theory and practice* (user manual for Arpeggio software). Princeton, NJ: Educational Testing Service.

Holland, P.W. (1990). On the sampling theory foundations of item response theory models. *Psychometrika, 55,* 577–602.

Hoskens, M., & de Boeck, P. (1995). Componential IRT models for polytomous items. *Journal of Educational Measurement, 32,* 364–384.

Hoskens, M., & de Boeck, P. (2001). Multidimensional componential item response theory models for polytomous items. *Applied Psychological Measurement, 25,* 19–37.

Hu, L., & Bentler, P.M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparametrized model specification. *Psychological Methods, 3,* 424–453.

Hu, L., & Bentler, P.M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6,* 1–55.

Irvine, S.H., & Kyllonen, P. (2002). *Item generation for test development.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Janssen, R., & de Boeck, P. (1996a). Psychometric modeling of componentially designed synonym tasks. *Applied Psychological Measurement, 21,* 37–50.

Janssen, R., & de Boeck, P. (1996b). The contribution of a response-production component to a free-response synonym task. *Journal of Educational Measurement, 33,* 417–432.

Janssen, R., Tuerlinckx, F., Meulders, M., & de Boeck, P. (2000). A hierarchical IRT model for criterion-referenced measurement. *Journal of Educational and Behavioral Statistics, 25,* 230–285.

Junker, B.W. (1999). *Some statistical models and computational methods that may be useful for cognitively-relevant assessment.* Unpublished manuscript. Retrieved March 6, 2007 from http://www.stat.cmu.edu/~brian/nrc/cfa

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement, 25,* 258–272.

Kasai, M. (1997). Application of the Rule Space Model to the reading comprehension section of the Test of English as a Foreign Language (TOEFL). *Dissertation Abstracts International Section A: Humanities & Social Sciences, 58,* 6-A. Urbana-Champaign, IL: University of Illinois at Urbana-Champaign: University Microfilms International.

Katz, I.R., Martinez, M.E., Sheehan, K.M., & Tatsuoka, K.K. (1998). Extending the rule space methodology to a semantically-rich domain: Diagnostic assessment in architecture. *Journal of Educational and Behavioral Statistics, 24,* 254–278.

Kirisci, L., Hsu, T.-C., & Yu, L. (2001). Robustness of item parameter estimation programs to assumptions of unidimensionality and normality. *Applied Psychological Measurement, 25,* 146–162.

Kohler, B., & Schrader, F.-W. (2004). Ergebnisrückmeldung und Rezeption [Reporting mechanisms and their reception]. *Empirische Pädagogik, 18,* 3–17.

Little, R.J.A., & Rubin, D.B. (2002). *Statistical analysis with missing data.* Hoboken, NJ: John Wiley & Sons.

Maris, E. (1995). Psychometric latent response models. *Psychometrika, 60,* 523–547.

McDonald, R.P. (1999). *Test theory: A unified treatment.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Meijer, R.R., & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement, 25,* 107–135.

Mellenbergh, G.J. (1994). Generalized linear item response theory. *Psychological Bulletin, 115,* 300–307.

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement* (pp. 13–103). New York: Macmillan.

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

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

Mislevy, R.J., Steinberg, L.S., & Almond, R.G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1,* 3–62.

Muthén, B.O. (2002). Beyond SEM: General latent variable modeling. *Behaviormetrika, 29,* 81–117.

Muthén, B., & Kaplan, D. (1985). A comparison of some methodologies for the factor analysis of non-normal Likert variables. *British Journal of Mathematical & Statistical Psychology, 38,* 171–189.

Muthén, B., & Kaplan, D. (1992). A comparison of some methodologies for the factor analysis of non-normal Likert variables: A note on the size of the model. *British Journal of Mathematical & Statistical Psychology, 45,* 19–30.

Muthén, L.K., & Muthén, B. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling, 9,* 599–620.

Muthén, L. K., & Muthén, B.O. (2004). *Mplus user's guide* (2nd ed.). Los Angeles, CA: Muthén & Muthén.

National Research Council (2001). *Knowing what students know: The science and design of educational assessment.* Washington, DC: National Academy Press.

Nichols, P. (1994). A guide for developing cognitively diagnostic assessments. *Review of Educational Research, 64,* 575–603.

Nichols, P.D., Chipman, S.F., & Brennan, R.L. (Eds.). (1995). *Cognitively diagnostic assessment.* Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement, 24,* 50–64.

Patz, R.J., & Junker, B.W. (1999a). A straightforward approach to Markov Chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics, 24,* 146–178.

Patz, R.J., & Junker, B.W. (1999b). Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses. *Journal of Educational & Behavioral Statistics, 24,* 342–366.

Raudenbush, S.W., & Bryk, A.S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oakes, CA: Sage.

Rupp, A.A. (2002). Feature selection for choosing and assembling measurement models: A building-block based organization. *International Journal of Testing, 2,* 311–360.

Rupp, A.A. (2003). Item response modeling with BILOG-MG and MULTILOG for Windows. *International Journal of Testing, 3,* 365–384.

Rupp, A.A., Dey, D.K., & Zumbo, B.D. (2004) To Bayes or not to Bayes, from whether to when: Applications of Bayesian methodology to modeling. *Structural Equation Modeling, 11,* 424–451.

Rupp, A.A., & Mislevy, R.J. (in press). Cognitive foundations of structured item response theory models. In J. Leighton & M.J. Gierl (Eds.), *Cognitive diagnostic assessment in education: Theory and practice*. Cambridge, UK: Cambridge University Press.

Samejima, F. (1997). Graded response model. In W.J. van der Linden & R.K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 85–100). New York: Springer-Verlag.

Scott, H.S. (1999). Cognitive diagnostic perspectives of a second language reading test. *Dissertation Abstracts International. Section B: The Sciences & Engineering,* 59(11-B). Urbana-Champaign, IL: University of Illinois at Urbana-Champaign: University Microfilms International.

Schafer, J.L. (1997). *Analysis of incomplete multivariate data.* Boca Raton, FL: Chapman & Hall/CRC Press.

Sinharay, S., & Johnson, M.S. (2003). *Simulation studies applying posterior predictive model checking for assessing fit of the common item response theory models.* (Research Report No. RR-03–28). Princeton, NJ: Educational Testing Service.

Snow, R.E., & Lohman, D.F. (1989). Implication of cognitive psychology for education measurement. In R.L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 263–331). New York: Macmillan.

Stone, C.A. (1992). Recovery of marginal maximum likelihood estimates in the two-parameter logistic response model: An evaluation of MULTILOG. *Applied Psychological Measurement, 16,* 1–16.

Stout, W. (2002). Psychometrics: From practice to theory and back. *Psychometrika, 67,* 485–518.

Takane, Y., & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika, 52,* 393–408.

Tatsuoka, C., Varandi, F., & Tatsuoka, K.K. (1992). *BUGSHELL* [computer software]. Ewing, NJ: Tanar Software.

Tatsuoka, K.K. (1983). Rule-space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement, 20,* 345–354.

Tatsuoka, K.K. (1985). A probabilistic model for diagnosing misconceptions by the pattern classification approach. *Journal of Educational Statistics, 10*, *55–73.*

Tatsuoka, K.K. (1986). Diagnosing cognitive errors: Statistical pattern classification based on item response theory. *Behaviormetrika, 19,* 73–86.

Tatsuoka, K.K. (1993). Item construction and psychometric models appropriate for constructed responses. In R.E. Bennet & W.C. Ward (Eds.), *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment* (pp. 107–133). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Tatsuoka, K.K. (1995). Architecture of knowledge structures and cognitive diagnosis: A statistical pattern recognition and classification approach. In P.D. Nichols, S.F. Chipman, & R.L. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 327–360). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Tatsuoka, K.K., Birenbaum, M., & Arnold, J. (1989). On the stability of students' rules of operation for solving arithmetic problems. *Journal of Educational Measurement, 26,* 351–361.

Tatsuoka, K., Corter, J., & Tatsuoka, C. (2005). Patterns of diagnosed mathematical content and process skills in TIMSS-R across a sample of 20 countries. *Educational Research Journal, 41*, 901–926.

Tatsuoka, K.K., & Eddins, J.M. (1985). Computer analysis of students' procedural « bugs » in an arithmetic domain. *Journal of Computer-Based Instruction, 12,* 34–38.

Tatsuoka, K.K., & Tatsuoka, M.M. (1987). Bug distribution and statistical pattern classification. *Psychometrika, 52,* 193–206.

Tatsuoka, K.K., & Tatsuoka, M.M. (1997).Computerized cognitive diagnostic adaptive testing: Effect on remedial instruction as empirical validation. *Journal of Educational Measurement, 34,* 3–20.

Templin, J.L., & Henson, R.A. (2004). *Creating a proficiency scale with models for cognitive diagnosis*. Lawrence, Kansas: The Univeristy of Kansas: Unpublished Technical Report for the ETS External Diagnostic Group.

Templin, J.L., Roussos, L., & Stout, W. (2003). *An extension of the current fusion model to treat polytomous attributes*. Unpublished Technical Report for the ETS External Diagnostic Group.

van der Linden, W.J., & Hambleton, R.K. (Eds.). (1997). *Handbook of modern item response theory.* New York: Springer-Verlag.

von Davier, M. (2005). *A general diagnostic model applied to language testing data* (Research Report No. RR-05–16). Princeton, NJ: Educational Testing Service.

Wainer, H., & Wang, X. (2001). *Using a new statistical model for testlets to score TOEFL* (Technical Report No. TR-16). Princeton, NJ: Educational Testing Service.

Wang, X., Bradlow, E.T., & Wainer, H. (2002). *A general Bayesian model for testlets: Theory and applications* (Research Report No. 02–02). Princeton, NJ: Educational Testing Service.

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

Wilson, M. (1993). The "Saltus model" misunderstood. *Methodika, 7,* 1–4.

Wu, M.L., Adams, R.J., & Wilson, M.R. (1998). *ACER ConQuest: Generalized item response modeling software* [software program]. Melbourne: Australian Council for Educational Research.

Zimowski, M.F., Muraki, E., Mislevy, R.J., & Bock, R.D. (1996). *BILOG-MG: Multiple-group IRT analysis and test maintenance for binary items* [computer software]. Chicago: Scientific Software International.

Zimmerman, D.W., & Zumbo, B.D. (2001). The geometry of probability, statistics, and test theory. *International Journal of Testing, 1,* 283–303.

Zumbo, B.D., & MacMillan, P.D. (1999). An overview and some observations on the psychometric models used in computer-adaptive testing. In M. Chalhoub-Deville (Ed.), *Issues in computer-adaptive testing of reading proficiency* (pp. 216–228). Cambridge, UK: University Press.

Zumbo, B.D., & Rupp, A.A. (2004). Responsible modeling of measurement data for appropriate inferences: Important advances in reliability and validity theory. In D. Kaplan (Ed.), *Handbook of quantitative methodology for the social sciences* (pp. 83–96). Newbury Park, CA: Sage.