Cognitive Assessment Models With Few Assumptions, and Connections With Nonparametric Item Response Theory

Brian W. Junker, Carnegie Mellon University

Klaas Sijtsma, Tilburg University

Some usability and interpretability issues for single-strategy cognitive assessment models are considered. These models posit a stochastic conjunctive relationship between a set of cognitive attributes to be assessed and performance on particular items/tasks in the assessment. The models considered make few assumptions about the relationship between latent attributes and task performance beyond a simple conjunctive structure. An example shows that these models can be sensitive to cognitive attributes, even in data designed to well fit the Rasch model. Several stochastic ordering and monotonicity properties are considered that enhance the interpretability of the models. Simple data summaries are identified that inform about the presence or absence of cognitive attributes when the full computational power needed to estimate the models is not available. *Index terms: cognitive diagnosis, conjunctive Bayesian inference networks, multidimensional item response theory, nonparametric item response theory, restricted latent class models, stochastic ordering, transitive reasoning.*

There has been increasing pressure in educational assessment to make assessments sensitive to specific examinee skills, knowledge, and other cognitive features needed to perform tasks. For example, Baxter & Glaser (1998) and Nichols & Sugrue (1999) noted that examinees' cognitive characteristics can and should be the focus of assessment design. Resnick & Resnick (1992) advocated standards- or criterion-referenced assessment closely tied to curriculum as a way to inform instruction and enhance student learning. These issues are considered in fuller detail by Pellegrino, Chudowsky, & Glaser (2001).

Cognitive assessment models generally deal with a more complex goal than linearly ordering examinees, or partially ordering them, in a low-dimensional Euclidean space, which is what item response theory (IRT) has been designed and optimized to do. Instead, cognitive assessment models produce a list of skills or other cognitive attributes that the examinee might or might not possess, based on the evidence of tasks that he/she performs. Nevertheless, these models have much in common with more familiar IRT models.

Interpretability of IRT-like models is enhanced by simple, monotone relationships between model parts. For example, Hemker, Sijtsma, Molenaar, & Junker (1997) considered in detail stochastic ordering of the manifest sum-score by the latent trait (SOM), and stochastic ordering of the latent trait by the manifest sum-score (SOL), in addition to the usual monotonicity assumption (see below). All three properties are considered here for two conjunctive cognitive assessment models. Additionally, a new monotonicity condition is considered, which asserts that the more task-relevant skills an examinee possesses, the easier the task should be.

Some Extensions of IRT Models for Cognitive Assessment

Consider J dichotomous item response variables for each of N examinees. Let $X_{ij} = 1$ if examinee i performs task j well, and 0 otherwise, where $i = 1, 2, \ldots, N$, and $j = 1, 2, \ldots, J$. Let θ_i be the person parameter (possibly multidimensional) and β_j be the item (difficulty) parameter (possibly multidimensional). The item response function (IRF) in IRT is $P_j(\theta_i) = P[X_{ij} = 1|\theta_i, \beta_j]$.

Most parametric IRT and nonparametric IRT (NIRT) models satisfy three fundamental assumptions:

1. Local independence (LI),

$$P(X_{i1}=x_{i1}, X_{i2}=x_{i2}, \ldots, X_{iJ}=x_{iJ}, |\theta_i, \beta_1, \beta_2, \ldots, \beta_J) = \prod_{i=1}^{N} \prod_{j=1}^{J} P_j(\theta_i)^{x_{ij}} \left[1 - P_j(\theta_i)\right]^{1-x_{ij}} , \tag{1}$$

   for each i.

2. Monotonicity, in which the IRFs $P_j(\theta_i)$ are nondecreasing as a function of θ_i or, if θ_i is multidimensional, nondecreasing coordinate-wise (i.e., nondecreasing in each coordinate of θ_i, with all other coordinates held fixed).
3. Low dimensionality, in which the dimension K of θ_i is small relative to the number of items J. In the Rasch model, for example, θ_i and β_j are unidimensional real-valued parameters, and logit $P_j(\theta_i) = \theta_i - \beta_j$.

Many attempts (see, e.g., Mislevy, 1996) to blend IRT and cognitive measurement are based on a linear decomposition of β_j or θ_i. In the linear logistic test model (LLTM; e.g., Draney, Pirolli, & Wilson, 1995; Fischer, 1995; Huguenard, Lerch, Junker, Patz, & Kass, 1997), β_j is rewritten as a linear combination of K basic parameters η_k with weights q_{jk} and

$$\text{logit } P_j(\theta_i) = \theta_i - \sum_{k=1}^{K} q_{jk}\eta_k , \tag{2}$$

where $\mathbf{Q} = [q_{jk}]$ is a matrix usually obtained a priori based on an analysis of the items into the requisite cognitive attributes needed to complete them, and η_k is the contribution of attribute k to the difficulty of the items involving that attribute.

Multidimensional compensatory IRT models (e.g., Adams, Wilson, & Wang, 1997; Reckase, 1997) follow the factor-analytic tradition; they decompose the unidimensional θ_i parameter into an item-dependent linear combination of underlying traits,

$$\text{logit } P_j(\theta_i) = \sum_{k=1}^{K} B_{jk}\theta_{ik} - \beta_j . \tag{3}$$

Compensatory IRT models, like factor analysis models, can be sensitive to relatively large components of variation in θ. However, they are generally not designed to distinguish finer components of variation among examinees that are often of interest in cognitive assessment. Models like the LLTM can be sensitive to these finer components of variation among items, but they also are not designed to be sensitive to components of variation among examinees—person parameters are often of little direct interest in an LLTM analysis.

Noncompensatory approaches, such as Embretson's (1997) multicomponent latent trait model (MLTM), are intended to be sensitive to finer variations among examinees in situations in which several cognitive components are required simultaneously for successful task performance. For the MLTM, successful performance on an item/task involves the conjunction of successful performances on several subtasks, each of which follows a separate unidimensional IRT model (e.g., the Rasch model),

$$P\left(X_j = 1|\theta_i\right) = \prod_{k=1}^{K} P\left(X_{jk} = 1|\theta_{ik}\right) = \prod_{k=1}^{K} \frac{\exp(\theta_{ik} - \beta_{jk})}{1 + \exp(\theta_{ik} - \beta_{jk})} \ . \tag{4}$$

Generally, conjunctive approaches have been preferred in cognitive assessment models that focus on a single strategy for performing tasks (Corbett, Anderson, & O'Brien, 1995; Tatsuoka, 1995; VanLehn & Niu, in press; VanLehn, Niu, Siler, & Gertner, 1998). Multiple strategies are often accommodated with a hierarchical latent-class structure that divides examinees into latent classes according to strategy. A different model is used within each class to describe the influence of attributes on task performance (e.g., Mislevy, 1996; Rijkes, 1996). Within a single strategy, models involving more-complicated combinations of attributes driving task performance are possible (e.g., Heckerman, 1998), but they can be more challenging to estimate and interpret. The present paper focuses on two discrete latent space analogues of the MLTM that make few assumptions about the relationship between latent attributes and task performance beyond a stochastic conjunctive structure.

Assessing Transitive Reasoning in Children

Method

Sijtsma & Verweij (1999) analyzed data from a set of transitive reasoning tasks. The data consisted of the responses to nine transitive reasoning tasks from 417 students in second, third, and fourth grade. Examinees were shown objects A, B, C, ..., with physical attributes Y_{A}, Y_{B}, Y_{C}, ... Relationships between attributes of all pairs of adjacent objects in an ordered series, such as $Y_{\mathrm{A}} < Y_{\mathrm{B}}$ and $Y_{\mathrm{B}} < Y_{\mathrm{C}}$, were shown to each examinee. The examinee was asked to reason about the relationship between some pair not shown, for example, Y_{A} and Y_{C}. Reasoning that $Y_{\mathrm{A}} < Y_{\mathrm{C}}$ from the premises $Y_{\mathrm{A}} < Y_{\mathrm{B}}$ and $Y_{\mathrm{B}} < Y_{\mathrm{C}}$, without guessing or using other information, is an example of transitive reasoning (for relevant developmental psychology, see Sijtsma & Verweij, 1999; Verweij, Sijtsma, & Koops, 1999).

The tasks were generated by considering three types of objects (wooden sticks, wooden disks, and clay balls) with different physical attributes (sticks differed in length by .2 cm per pair, disks differed in diameter by .2 cm per pair, and balls differed in weight by 30 g per pair). Each task involved three, four, or five of the same type of object.

For a three-object task, there were two premises, AB (specifying the relationship between Y_{A} and Y_{B}) and BC (similarly for Y_{B} and Y_{C}). There was one item, AC, which asked for the relationship between Y_{A} and Y_{C}. For a four-object task, there were three premises (AB, BC, and CD) and two items (AC and BD). For a five-object task, there were four premises (AB, BC, CD, DE) and three items (AC, BD, and CE). Tasks, premises, and items within tasks were presented to each examinee in random order. Explanations for each answer were recorded to evaluate the use of strategy. Table 1 summarizes the nine tasks.

Results

Sijtsma & Verweij (1999) showed that the task response data fit a polytomous monotone homogeneity model (a model assuming only LI, unidimensionality, and monotonicity; see Van der Ark, 2001) well when (1) each item within a task was scored as correct—when a correct response and a

Table 1
Nine Transitive Reasoning Tasks and Expected A-Posteriori (EAP) Rasch Difficulties and Corresponding Posterior Standard Deviations (PSD)

					Rasch Difficulties	
Task	Objects	Attribute	Premises	Items	EAP	PSD
1	3 Sticks	Length	2	1	−.38	.16
2	4 Sticks	Length	3	2	1.88	.17
3	5 Sticks	Length	4	3	6.06	.50
4	3 Disks	Size	2	1	−1.78	.17
5	4 Disks	Size	3	2	12.60	5.12
6	5 Disks	Size	4	3	12.40	4.86
7	3 Balls	Weight	2	1	−3.40	.22
8	4 Balls	Weight	3	2	3.95	.25
9	5 Balls	Weight	4	3	8.07	1.23

correct deductive strategy based on transitive reasoning were given (referred to as DEDSTRAT data); and (2) the dichotomous item scores were summed within tasks to give task scores.

The data were recoded by the present authors for analysis with binary models. A task was considered correct (scored 1) if all the items within that task were answered correctly using a correct deductive strategy; otherwise, the task was considered incorrect (scored 0). This led to 417 × 9 scores. The scores for all examinees on Tasks 5 and 6, involving disk sizes, were 0. Relatively large visual differences between disk sizes (diameters varied linearly, so disk areas varied quadratically) seemed to encourage examinees to arrive at a correct answer for some items by direct visual comparison, rather than by a deductive strategy. These responses were coded 0 because a deductive strategy was not used.

After deleting Tasks 5 and 6, which had all 0 responses, the computer program MSP5 (Molenaar & Sijtsma, 2000) reported a very high scaling coefficient ($H = .82$) for the remaining seven tasks. The scaling coefficients (Sijtsma, 1998) for the tasks, H_j, were between .78 and 1.00. No sample violations of manifest monotonicity (Junker & Sijtsma, 2000) were found. The program RSP (Glas & Ellis, 1994) was used to fit a Rasch model to the data. Again Tasks 5 and 6 were deleted along with examinees who had all zero responses. This caused Item 9 to have all zero responses in the reduced dataset, so it was deleted as well. For the remaining six items and 382 examinees, standard Rasch fit statistics (Glas & Verhelst, 1995) indicated good fit. The Rasch model was refitted using BUGS (Spiegelhalter, Thomas, Best, & Gilks, 1997). BUGS uses a Bayesian formulation of the model that does not require items or persons to be deleted. Good fit again was found. The item difficulty parameters (β_j) estimated by BUGS are shown in Table 1. β_j was based on a fixed normal θ distribution and a common $\mathrm{N}(\mu_\beta, \sigma_\beta^2)$ prior for those with weak hyperpriors $\mu_\beta \sim \mathrm{N}(0, 100)$ and $\sigma_\beta^{-2} \sim \Gamma(.01, .01)$.

If the transitive reasoning scale is to be used as evidence in designing or improving an instructional program for children or to provide feedback on particular aspects of transitive reasoning to teachers and students, then analyses with the monotone homogeneity model and the Rasch model will not help. They only provide the ranks or locations of examinees on a unidimensional latent scale. Instead, task performance must be explicitly modeled in terms of the presence or absence of particular cognitive attributes related to transitive reasoning.

To illustrate, consider the preliminary analysis of the nine tasks in Table 2. The first three attributes are the ability to recognize or reason about transitivity in the context of length, size, and weight. The tasks also place differential load on an examinee's working memory capacity (Carpenter, Just, & Shell, 1990; Kyllonen & Christal, 1990). Thus, the next three cognitive attributes correspond to three levels of working memory capacity: (1) manipulating the first two premises given in a task in working memory; (2) manipulating a third task premise, if it is given; and (3) manipulating a fourth task premise, if it is given.

The issue is not strictly model-data fit. If the objective is to know whether particular students can focus on a transitive reasoning strategy in the context of weight problems, the total score on the nine items—the central examinee statistic in Rasch and monotone homogeneity models—will not help. Similarly, an LLTM can determine whether additional working memory load makes tasks more difficult on average, but it cannot indicate whether a particular examinee has difficulty maintaining a third premise in solving transitive reasoning problems. Models that partition the data into signal and noise differently than unidimensional IRT models are clearly needed.

Two IRT-Like Cognitive Assessment Models

Two discrete latent attribute models are described. These allow both for modeling the cognitive loads of items and for inferences about the cognitive attributes of examinees. In both models, the latent variable is a vector of 0s and 1s for each examinee, indicating the absence or presence of particular cognitive attributes. Table 2 shows which attributes the examinee needed to perform each task correctly.

Table 2
Decomposition of Tasks Into
Hypothetical Cognitive Attributes

	Context			Premise		
	Length	Size	Weight	1st/2nd	3rd	4th
Q_{jk}	1	2	3	4	5	6
1	1	0	0	1	0	0
2	1	0	0	1	1	0
3	1	0	0	1	1	1
4	0	1	0	1	0	0
5	0	1	0	1	1	0
6	0	1	0	1	1	1
7	0	0	1	1	0	0
8	0	0	1	1	1	0
9	0	0	1	1	1	1

To describe these models, consider N examinees and J binary task performance variables. A fixed set of K cognitive attributes are involved in performing these tasks (different subsets of attributes might be involved in different tasks). For both models,

$$X_{ij} = \text{ 1 or 0, indicating whether examinee } i \text{ performed task } j \text{ correctly;}$$

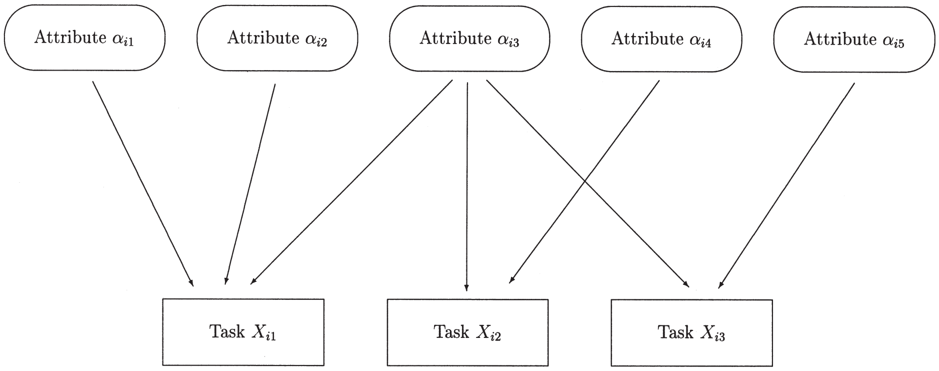
$$Q_{jk} = \text{ 1 or 0, indicating whether attribute } k \text{ is relevant to task } j\text{; and}$$

$$\alpha_{ik} = \text{ 1 or 0, indicating whether examinee } i \text{ possesses attribute } k. \qquad (5)$$

Q_{jk} are fixed in advance, similar to the design matrix in an LLTM. The Q_{jk} can be assembled into a **Q** matrix (Tatsuoka, 1995). Figure 1 illustrates the structure defined by X_{ij}, Q_{jk} and α_{ik} as a Bayesian network.

The objective is to make inferences about the latent variables α_{ik}, indicating cognitive attributes that examinees do or do not possess, or inferences about the relationship between these attributes and observed task performance. Both models are easily specified using the latent response framework

Figure 1
A One-Layer Bayesian Network for Conjunctive Discrete Cognitive Attributes Models



(Maris, 1995), which is closely related to the notion of data augmentation in statistical estimation (Tanner, 1996).

The DINA Model

The deterministic inputs, noisy "and" gate model (called the DINA model) has been the foundation of several approaches to cognitive diagnosis and assessment (Doignon & Falmagne, 1999; Tatsuoka, 1995). It was considered in detail by Haertel (1989; also Macready & Dayton, 1977), who identified it as a restricted latent class model. In the DINA model, latent response variables are defined as

$$\xi_{ij} = \prod_{k:Q_{jk}=1} \alpha_{ik} = \prod_{k=1}^{K} \alpha_{ik}^{Q_{jk}} , \tag{6}$$

indicating whether examinee i has all the attributes required for task j. In Tatsuoka's (1995) terminology, the latent vectors $\alpha_{i\cdot} = (\alpha_{i1}, \alpha_{i2}, \ldots , \alpha_{iK})$ are called *knowledge states*, and the vectors $\xi_{i\cdot} = (\xi_{i1}, \xi_{i2}, \ldots , \xi_{iJ})$ are called *ideal response patterns*—they represent a deterministic prediction of task performance from each examinee's knowledge state.

The latent response variables ξ_{ij} are related to observed task performances X_{ij} according to the probabilities

$$s_j = P\left(X_{ij} = 0|\xi_{ij} = 1\right) \tag{7}$$

and

$$g_j = P\left(X_{ij} = 1|\xi_{ij} = 0\right) , \tag{8}$$

where s_j and g_j are error probabilities—false negative and false positive rates—in a simple signal detection model for detecting ξ_{ij} from noisy observations X_{ij}. s_j and g_j were selected to be mnemonic, thinking of examinees' slips and guesses, but genuine slipping and guessing behavior might be the least important reason for observing $X_{ij} \neq \xi_{ij}$. Other reasons include poor wording of the task description, inadequate specification of the **Q** matrix, use of an alternative solution strategy by the examinee, and general lack of model fit. DiBello, Stout, & Roussos (1995) addressed this issue in their discussion of the positivity of a task with respect to a cognitive attribute (see below).

The IRF for a single task is

$$P\left(X_{ij} = 1|\boldsymbol{\alpha}, s, g\right) = (1 - s_j)^{\xi_{ij}} g_j^{1-\xi_{ij}} \equiv P_j(\alpha_{i\cdot}) \ . \tag{9}$$

Each ξ_{ij} acts as an "and" gate (i.e., it is a binary function of binary inputs with value 1 if and only if all the inputs are 1s), combining the deterministic inputs $\alpha_{ik}^{Q_{jk}}$. Each X_{ij} is modeled as a noisy observation of each ξ_{ij} (cf. VanLehn et al., 1998). Equation 9 makes it clear that $P_j(\alpha_{i\cdot})$ is coordinate-wise monotone in $\alpha_{i\cdot}$ if and only if $1 - s_j > g_j$. Assuming LI among examinees, the joint likelihood for all responses under the DINA model is

$$\begin{aligned} P(X_{ij} = x_{ij}, \forall\, i, j|\boldsymbol{\alpha}, s, g) &= \prod_{i=1}^{N} \prod_{j=1}^{J} P_j(\alpha_{i\cdot})^{x_{ij}} [1 - P_j(\alpha_{i\cdot})]^{1-x_{ij}} \\ &= \prod_{i=1}^{N} \prod_{j=1}^{J} \left[(1 - s_j)^{x_{ij}} s_j^{1-x_{ij}}\right]^{\xi_{ij}} \left[g_j^{x_{ij}} (1 - g_j)^{1-x_{ij}}\right]^{1-\xi_{ij}} \ . \end{aligned} \tag{10}$$

The NIDA Model

The noisy inputs, deterministic "and" gate model (called the NIDA model) was recently discussed by Maris (1999) and has been used as a building block in more elaborate cognitive diagnosis models (DiBello et al., 1995). In the NIDA model, X_{ij}, Q_{jk}, and α_{ik} are taken from Equation 5 and the latent variable η_{ijk} = 1 or 0 is defined, indicating whether examinee i's performance in the context of task j is consistent with possessing attribute k.

The η_{ijk} are related to the examinee's $\alpha_{i\cdot}$ according to the probabilities

$$s_k = P\left(\eta_{ijk} = 0|\alpha_{ik} = 1, Q_{jk} = 1\right) \ , \tag{11}$$

$$g_k = P\left(\eta_{ijk} = 1|\alpha_{ik} = 0, Q_{jk} = 1\right) \ , \tag{12}$$

and

$$P\left(\eta_{ijk} = 1|\alpha_{ik} = a, Q_{jk} = 0\right) \equiv 1 \ , \tag{13}$$

regardless of the value a of α_{ik}. The definition in Equation 13 simplifies writing several expressions below, and does not restrict the model in any way. s_k and g_k are mnemonically named false negative and false positive error probabilities in a signal detection model for detecting α_{ik} from noisy η_{ijk}. Observed task performance is related to the latent response variables through

$$X_{ij} = \prod_{k:Q_{jk}=1} \eta_{ijk} = \prod_{k=1}^{K} \eta_{ijk} \ . \tag{14}$$

Thus, the IRF is

$$\begin{aligned} P(X_{ij} = 1|\boldsymbol{\alpha}, s, g) &= \prod_{k=1}^{K} P\left(\eta_{ijk} = 1|\alpha_{ik}, Q_{jk}\right) = \prod_{k=1}^{K} \left[(1 - s_k)^{\alpha_{ik}} g_k^{1-\alpha_{ik}}\right]^{Q_{jk}} \\ &= \prod_{k=1}^{K} \left(\frac{1 - s_k}{g_k}\right)^{\alpha_{ik} Q_{jk}} \prod_{k=1}^{K} g_k^{Q_{jk}} \equiv P_j(\alpha_{i\cdot}) \ . \end{aligned} \tag{15}$$

For the NIDA model, noisy inputs η_{ijk}, reflecting attributes α_{ik} in examinees, are combined in a deterministic "and" gate X_{ij}. Again, the IRF is monotone in the coordinates of $\alpha_{i\cdot}$ as long as $(1 - s_k) > g_k$. The joint model for all responses in the NIDA model is

$$P(X_{ij} = x_{ij}, \forall\, i, j | \boldsymbol{\alpha}, s, g) = \prod_{i=1}^{N} \prod_{j=1}^{J} P_j(\alpha_{i\cdot})^{x_{ij}} [1 - P_j(\alpha_{i\cdot})]^{1-x_{ij}}$$

$$= \prod_{i=1}^{N} \prod_{j=1}^{J} \left\{ \prod_{k=1}^{K} \left[(1 - s_k)^{\alpha_{ik}} g_k^{1-\alpha_{ik}} \right]^{Q_{jk}} \right\}^{x_{ij}} \left\{ 1 - \prod_{k=1}^{K} \left[(1 - s_k)^{\alpha_{ik}} g_k^{1-\alpha_{ik}} \right]^{Q_{jk}} \right\}^{1-x_{ij}} . \qquad (16)$$

Exploring Monotonicity

The DINA and NIDA models are stochastic conjunctive models for task performance. Under monotonicity $(1 - s > g)$, examinees must possess all attributes listed for each task to maximize the probability of successful performance. The DINA and NIDA models also are restricted latent class models (Haertel, 1989), and therefore closely related to IRT models, as suggested by Equations 10 and 16. [If $P_j(\alpha_{i\cdot})$ were replaced with $P_j(\theta_i)$, the setting would be IRT: $\alpha_{i\cdot}$ plays the role of the latent variable θ_i, and s_k and g_k play the role of β_j.] These models also can be seen as one-layer Bayesian inference networks for discrete variables (Mislevy, 1996; VanLehn et al., 1998) for task performance (see Figure 1). In general, Bayesian network models do not need to be conjunctive (e.g., Heckerman, 1998), but when examinees are presumed to be using a single strategy, conjunctive models seem natural (e.g., DiBello et al., 1995).

Method. To explore whether monotonicity actually holds in real data, BUGS (Version 0.6; Spiegelhalter et al., 1996) was used to fit the DINA and NIDA models to the dichotomous DEDSTRAT data using the **Q** matrix in Table 2. Bayesian formulations of the models were used. Population probabilities $\pi_k = P[\alpha_{ik} = 1]$ were assumed to have independent, uniform priors Unif[0, 1] on the unit interval. Independent, flat priors Unif[0, g_{max}] and Unif[0, s_{max}] also were used on the false positive error probabilities $g_1, g_2, \ldots$, and false negative error probabilities $s_1, s_2, \ldots$, in each model. When g_{max} and s_{max} are small, these priors tend to favor error probabilities satisfying $1-s > g$. g_{max} and s_{max} also were estimated in the model, using Unif[0, 1] hyperprior distributions.

For each model, the Markov chain monte carlo (MCMC) algorithm compiled by BUGS ran five times, for 3,000 steps each, from various randomly selected starting points. The first 2,000 steps of each chain were discarded as burn-in, and the remaining 1,000 steps were thinned by retaining every fifth observation. Thus, there were 200 observations per chain. Both models showed evidence of under-identification (slow convergence and multiple maxima), as was expected (Maris, 1999; Tatsuoka, 1995).

Results. Tables 3 and 4 list tentative expected a posteriori (EAP) and posterior standard deviations (PSDs) for each set of error probabilities in the two models, using 1,000 MCMC steps obtained by pooling the five thinned chains for each model. Most of the point estimates satisfied monotonicity $[1 - s > g$ (or equivalently, $g + s < 1$)]. The exceptions were the error probabilities for Tasks 4 and 8 under the DINA model. The posterior probabilities in each model that $1 - s > g$ for each task (DINA model) or latent attribute (NIDA model) were near .50. Although this did not contradict the hypothesis that monotonicity held, it was not strongly confirmed.

In the DINA model, Tasks 5 and 6 (all examinees scored 0) yielded the estimates $\hat{g}_j = \hat{P}[X_{ij} = 1|\xi_{ij} = 0] = .002$ (PSD = .002). Except for these two tasks, all error probabilities in the DINA model were near their prior means with fairly large PSDs, suggesting that the attributes outlined in Table 2 were not very predictive of successful task performance.

Table 3
Tentative EAP Estimates and PSDs
for $\hat{g}_j$ and $\hat{s}_j$ in the DINA Model

	$\hat{g}_j$		$\hat{s}_j$			$[(1-\hat{g}_j)/\hat{g}_j]\times$
j	EAP	PSD	EAP	PSD	$1-\hat{s}_j > \hat{g}_j$	$[(1-\hat{s}_j)/\hat{s}_j]$
1	.478	.167	.486	.277	yes	1.15
2	.363	.162	.487	.281	yes	1.85
3	.419	.255	.479	.292	yes	1.51
4	.657	.199	.488	.279	no	.55
5	.002	.002	.462	.270	yes	581.09
6	.002	.002	.464	.270	yes	576.43
7	.391	.420	.486	.274	yes	1.65
8	.539	.242	.489	.275	no	.89
9	.411	.162	.480	.283	yes	1.55
Maximum	.910	.081	.910	.079		

However, the error probabilities in the NIDA model seemed to move farther from their prior means, in some cases with relatively small PSDs. Attributes 4, 5, and 6, indicating increasing cognitive load, had decreasing g_ks and generally increasing s_ks, reflecting the successively increasing difficulty of tasks involving these attributes. The EAP estimates of g_{max} and s_{max} in both models were above .870 with small PSDs. This reflects the large PSDs (and, therefore, large estimation uncertainty) associated with at least some of the error probabilities in each model. It also suggests that the prior preference for monotonicity $(1 - s > g)$ was not very strong—the mild evidence for monotonicity seen in the model fit might reflect the data and not the prior distribution choices.

Table 4
Tentative EAP Estimates and PSDs for
$\hat{g}_k$ and $\hat{s}_k$ in the NIDA Model

	$\hat{g}_k$		$\hat{s}_k$				
k	EAP	PSD	EAP	PSD	$1-\hat{s}_k > \hat{g}_k$	$(1-\hat{s}_k)/\hat{g}_k$	$\log(1-\hat{s}_k)/\hat{g}_k$
1	.467	.364	.369	.392	yes	1.351	.301
2	.749	.207	.161	.125	yes	1.120	.113
3	.764	.246	.005	.009	yes	1.302	.264
4	.364	.319	.163	.318	yes	2.299	.833
5	.176	.168	.785	.129	yes	1.222	.200
6	.061	.115	.597	.294	yes	6.607	1.888
Maximum	.877	.109	.877	.108			

A NIRT Perspective on Cognitive Assessment Models

One strength of the NIRT approach is that it encourages researchers to consider fundamental model properties that are important for inference about latent variables from observed data.

Data Summaries Relevant to Parameter Estimation

The DINA model. Junker (2001) considered the DINA model as a possible starting place for formulating a NIRT for cognitive assessment models. Using calculations for the complete conditional distributions often employed in MCMC estimation algorithms, he showed that:

1. Estimation of the "slip" probabilities s_j were sensitive only to an examinee's X_{ij} on tasks for which he/she was hypothesized to have all the requisite cognitive attributes $(\xi_{ij} = 1)$.

2. Estimation of the "guessing" probabilities g_j depended only on an examinee's X_{ij} on tasks for which one or more attributes was hypothesized to be missing ($\xi_{ij} = 0$).
3. Estimation of α_{ik}, indicating possession of attribute k by examinee i, was sensitive only to performance on those tasks for which examinee i was already hypothesized to possess all other requisite cognitive attributes.

The posterior odds of $\alpha_{ik} = 1$, conditional on the data and all other parameters are (Junker, 2001)

$$\prod_{j=1}^{J}\left(\frac{s_j}{1-g_j}\right)^{\xi_{ij}^{(-k)}Q_{jk}}\cdot\prod_{j=1}^{J}\left[\left(\frac{1-g_j}{g_j}\cdot\frac{1-s_j}{s_j}\right)^{\xi_{ij}^{(-k)}Q_{jk}}\right]^{x_{ij}}\cdot\frac{\pi_{ik}^{\alpha}(1)}{\pi_{ik}^{\alpha}(0)}\ , \tag{17}$$

where

$$\xi_{ij}^{(-k)} = \prod_{\ell\neq k:Q_{j\ell}=1}\alpha_{i\ell}\ , \tag{18}$$

which indicates the presence of all attributes needed for task j except attribute k. $\pi_{ik}^{\alpha}(1)/\pi_{ik}^{\alpha}(0)$ are the prior odds. The first product in Equation 17 is constant in the data. The second product shows that the odds of $\alpha_{ik} = 1$ are multiplied by $[(1-g_j)/g_j] \times [(1-s_j)/s_j]$ for each additional correct task j, assuming that task j involves attribute k, and that all other attributes needed for task j have been mastered. Otherwise, there is no change in the odds. If monotonicity holds, this multiplier is greater than 1. Table 3 shows that these multipliers ranged from .55 to 1.85, except for Tasks 5 and 6. (Tasks 5 and 6 had very high multipliers because the model was able to estimate g_js near zero, because no one correctly answered those tasks.) Combining the influence of these multipliers with the effect of $\xi_{ij}^{(-k)}$ (Equation 18), it can be seen that correctly answering additional tasks in this model might not appreciably change the odds that an examinee possesses any one of the latent attributes (cf. VanLehn et al., 1998).

The NIDA model. A Bayesian version of the NIDA model is considered. Equation 16 is multiplied by unspecified, independent priors

$$\pi(s) = \prod_k \pi_k^s(s_k),\, \pi(g) = \prod_k \pi_k^g(g_k) \tag{19}$$

and

$$\pi(\boldsymbol{\alpha}) = \prod_{ik}\pi_{ik}^{\alpha}(\boldsymbol{\alpha}_{ik})\ . \tag{20}$$

The complete conditional distribution for any parameter, such as the "guessing" probability g_k, is proportional to the product of those factors in Equation 16 containing g_k and the prior density $\pi_k^g(g_k)$. For g_k, this is

$$\prod_{i:\alpha_{ik}=0}\ \prod_{j:Q_{jk}=1}\left(c_k^i g_k\right)^{x_{ij}}\left(1-c_k^i g_k\right)^{1-x_{ij}}\pi_k^g(g_k)\ , \tag{21}$$

where

$$c_k^i = \prod_{\ell\neq k}\left[(1-s_\ell)^{\alpha_{i\ell}}g_\ell^{1-\alpha_{i\ell}}\right]^{Q_{j\ell}}\ . \tag{22}$$

Similarly, the complete conditional distribution for each s_k is proportional to

$$\prod_{i:\alpha_{ik}=1}\ \prod_{j:Q_{jk}=1}\left[c_k^i(1-s_k)\right]^{x_{ij}}\left[1-c_k^i(1-s_k)\right]^{1-x_{ij}}\pi_k^s(s_k)\ . \tag{23}$$

Estimates of g_k depend precisely on those task responses for which attribute k was required of the examinee but he/she did not possess. s_k depends on those task responses for which attribute k was required of and possessed by the examinee.

The complete conditional distribution for each latent attribute indicator α_{ik} is proportional to

$$\left[c_k^i(1-s_k)^{\alpha_{ik}}g_k^{1-\alpha_{ik}}\right]^{m_k^i}\left[1-c_k^i(1-s_k)^{\alpha_{ik}}g_k^{1-\alpha_{ik}}\right]^{n_k-m_k^i}\pi_{ik}^{\alpha}(\alpha_{ik})\ , \tag{24}$$

where

$$m_k^i=\sum_{j:Q_{jk}=1}x_{ij}=\sum_{j=1}^{J}x_{ij}Q_{jk}=\text{ number of tasks correct involving attribute } k, \tag{25}$$

and

$$\eta_k=\sum_{j=1}^{J}Q_{jk}=\text{ total number of tasks involving attribute } k. \tag{26}$$

The posterior odds of $\alpha_{ik}=1$, conditional on the data and all other parameters, is equal to

$$\left(\frac{1-s_k)}{g_k}\right)^{m_k^i}\left[\frac{1-c_k^i(1-s_k)}{1-c_k^ig_k}\right]^{n_k-m_k^i}\cdot\frac{\pi_{ik}^{\alpha}(1)}{\pi_{ik}^{\alpha}(0)}\ , \tag{27}$$

for the NIDA model.

When monotonicity $(1-s_k>g_k)$ holds, the first term (in parentheses) in Equation 27 is greater than 1 and the second term (in brackets) is less than 1. Thus, the odds of $\alpha_{ik}=1$ increase as m_k^i increases. Essentially, the conditional odds of $\alpha_{ik}=1$ are multiplied by $(1-s_k)/g_k$ for each additional correct task involving attribute k. This is done regardless of the examinee's status on the other attributes. (c_k^i in Equation 22 is typically less than 10^{-5}, so the second term in Equation 27 is negligible.)

Table 4 shows that these multipliers ranged approximately from 1.1–1.4, except for the higher multipliers for Attribute 4 (cognitive capacity to handle the first two premises in a task) and Attribute 6 (cognitive capacity to handle the fourth premise in a task). Attribute 4 had moderately low estimated guessing and slip probabilities; Attribute 6 had a very low estimated guessing probability. This increased the model's certainty that each of these two attributes was possessed when an examinee correctly accomplished a task depending on that attribute.

Hartz, DiBello, & Stout (2000) noted that $(1-s_k)/g_k$ measures what DiBello et al. (1995) call positivity, which is approximately the extent to which task performance is a deterministic function of the knowledge state $\alpha_{i\cdot}$. Analysis of Equation 27 shows that attributes with high positivity are strongly credited in the NIDA model when the corresponding tasks are performed well. Comparing Equation 17 with Equation 27 shows that the posterior odds of $\alpha_{ik}=1$ tend to be more sensitive to the data under the NIDA model than under the DINA model.

Three NIRT Monotonicity Properties

For models satisfying LI, monotonicity, and low dimensionality, it follows immediately from Lemma 2 of Holland & Rosenbaum (1986) that for any nondecreasing summary $g(\mathbf{X})$ of $\mathbf{X} = (X_1, \ldots, X_J)$, $E[g(\mathbf{X})|\alpha_{i\cdot}]$ is nondecreasing in each coordinate α_{ik} of $\alpha_{i\cdot}$. This implies SOM (Hemker et al., 1997)—$P[X_+ > c|\alpha_{i\cdot}]$ is nondecreasing in each coordinate α_{ik} of $\alpha_{i\cdot}$. Little is known about SOL (Hemker et al., 1997)—$P[\alpha_{i1} > c_1, \ldots, \alpha_{ik} > c_k|X_+ = s]$—when the latent trait is multidimensional. A weaker property related to SOL is that

$$P\left[\alpha_{ik} = 1 \middle| \alpha_{i1}, \ldots, \alpha_{i(k-1)}, \alpha_{i(k+1)}, \ldots \alpha_{iK} \quad \text{and} \quad \sum_{j:Q_{jk}=1} X_{ij} = s\right] \tag{28}$$

is nondecreasing in s, with all other parameters fixed.

For the NIDA model, Equation 28 is immediate from Equation 27, because by Equation 25, $m_k^i = \sum_{j:Q_{jk}=1} X_{ij}$ in Equation 28. However, Equation 28 does not need to hold for the DINA model, as Equation 17 shows. If the products of odds $[(1 - g_j)/g_j] \times [(1 - s_j)/s_j]$ vary greatly, Equation 17 does not need to be monotone in $m_k^i = \sum_{j:Q_{jk}=1} X_{ij}$.

Finally, a new type of monotonicity condition seems plausible for some cognitive assessment models. In a standard monotone unidimensional IRT model, higher θ is associated with higher probability of correctly performing a task. A corresponding property in NIDA and DINA models might focus on the relationship between the number of task-relevant latent attributes the examinee has and the probability of correct task performance. It might be required that the IRFs in Equations 9 and 15 be nondecreasing in

$$m_{ij} = \sum_{k=1}^{K} \alpha_{ik} Q_{jk} = \text{ number of task-relevant attributes possessed.} \tag{29}$$

This monotonicity property is immediate for the DINA model when $1 - s_j > g_j$, because

$$P_j(\alpha_{i\cdot}) = (1 - s_j)^{\xi_{ij}} g_j^{1-\xi_{ij}} \tag{30}$$

equals g_j as long as $m_{ij} < \sum_{k=1}^{K} Q_{jk}$, and changes to $1 - s_j$ when $m_{ij} = \sum_{k=1}^{K} Q_{jk}$.

For the NIDA model, this monotonicity condition is generally not true. In the NIDA model,

$$P_j(\alpha_{i\cdot}) = \prod_{k=1}^{K} [(1 - s_k)/g_k]^{\alpha_{ik} Q_{jk}} \prod_{k=1}^{K} g_k^{Q_{jk}} \tag{31}$$

varies with m_{ij} through the first term, because j is held fixed. The logarithm of this term is $\sum_{k=1}^{K} \alpha_{ik} Q_{jk} \log(1 - s_k)/g_k$. Fixing i and j, setting $e_k = \alpha_{ik} Q_{jk}$ and $p_k = \log(1 - s_k)/g_k$, monotonicity of $P_j(\alpha_{i\cdot})$ in m_{ij} is equivalent to

$$\min_{e:e_+=s+1} \sum_{k=1}^{K} e_k p_k \geq \max_{e:e_+=s} \sum_{k=1}^{K} e_k p_k \ , \tag{32}$$

for each s, where $e = (e_1, \ldots, e_K)$ and $e_+ = \sum_k e_k$. This constrains the variability of $p_k = \log(1 - s_k)/g_k$. When the e_ks are unrestricted, Equation 32 is equivalent to

$$\sum_{k=1}^{s_0+1} p'_k \geq \sum_{k=K-s_0+1}^{K} p'_k \ , \tag{33}$$

where p'_k are the p_k renumbered so that $p'_1 \leq p'_2 \leq \ldots \leq p'_K$, and s_0 is the largest integer not exceeding $(K-1)/2$. Equation 32 holds for all s and all e if and only if it holds for s_0 and those es that allocate the smallest s_0+1 ps to one sum and the largest s_0 ps to the other. When Equation 32 or Equation 33 holds, all IRFs in the NIDA model are monotone in m_{ij}.

For the NIDA parameter estimates in Table 4, $p'_1 + p'_2 + p'_3 = .577 < 2.721 = p'_5 + p'_6$. Thus, there is no guarantee of monotonicity for all $P_j(\alpha_{i\cdot})$ in m_{ij}. However, the e_ks are restricted by the Q_{jk}s. In the transitive reasoning data, Q_{jk} limited the number of attributes that could affect each task to two, three, or four. The two-attribute tasks (Tasks 1, 4, and 7) had IRFs that were monotone in m_{ij}. On the other hand, none of the other tasks had monotone IRFs. In Table 4, the problem is the vast disparity between Attribute 4 (maintaining the first two premises of a task), with $p_4 = .833$, and Attribute 5 (maintaining the third premise), with $p_5 = .200$. Task 2 involved Attributes 1, 4, and 5, for example, and $p_1 + p_5 < p_4$, violating the condition in Equation 32. Hence, $P_2(\alpha_{i\cdot})$ cannot be monotone in m_{i2}.

Conclusions

Even when the fit is good, standard unidimensional IRT modeling might not be as relevant as some discrete attributes models, if the goal of testing is cognitive assessment or diagnosis. Two conjunctive cognitive attributes models, the DINA and NIDA models, have been shown to satisfy familiar multidimensional generalizations of standard IRT assumptions. Thus, intuitions about the behavior and interpretation of multidimensional IRT models carry over, at least in part, to these newer models.

In a transitive reasoning example, interesting structure was found at the cognitive attributes level, despite the data having been designed to fit the Rasch model. It is probable that data designed to be informative about a handful of cognitive attributes through the DINA or NIDA models would fare quite well in terms of model fit and ability to infer the presence or absence of particular attributes.

Relating model parameters to simple and useful data summaries is important when computational machinery is not available (e.g., in embedded assessments; cf. Wilson & Sloane, 2000). For example, a natural new monotonicity condition was considered, which asserts that the more task-relevant skills an examinee possesses, the easier the task should be. This property comes "almost for free" in one of the two models considered here, and it places interesting constraints on the parameters of the other model. Some model parameters also were related here to simple and useful data summaries, such as the number of tasks correctly performed involving a particular attribute. This is a beginning toward a clearer theory of which data summaries are relevant to the cognitive inferences desired over a wide variety of cognitive assessment models (cf. Junker, 2001). Such a theory would be an important contribution from the interface between NIRT and PIRT methodology.

References

Adams, R. J., Wilson, M., & Wang, W.-C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21,* 1–23.

Baxter, G. P., & Glaser, R. (1998). Investigating the cognitive complexity of science assessments. *Educational Measurement: Issues and Practice, 17,* 37–45.

Carpenter, P. A., Just, M. A., & Shell, P. (1990). What one intelligence test measures: A theoretical account of processing in the Raven's Progressive Matrices Test. *Psychological Review, 7,* 404–431.

Corbett, A. T., Anderson, J. R., & O'Brien, A. T. (1995). Student modeling in the ACT programming tutor. In P. D. Nichols, S. F. Chipman, & R. L. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 19–41). Hillsdale NJ: Erlbaum.

DiBello, L. V., Stout, W. F., & Roussos, L. A. (1995). Unified cognitive/psychometric diagnostic assessment likelihood-based classification techniques. In P. D. Nichols, S. F. Chipman, & R. L. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 361–389). Hillsdale NJ: Erlbaum.

Doignon, J.-P., & Falmagne, J.-C. (1999). *Knowledge spaces.* New York: Springer-Verlag.

Draney, K. L., Pirolli, P., & Wilson, M. (1995). A measurement model for a complex cognitive skill. In P. D. Nichols, S. F. Chipman, R. L. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 103–125). Hillsdale NJ: Erlbaum.

Embretson, S. E. (1997). Multicomponent response models. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 305–321). New York: Springer-Verlag.

Fischer, G. H. (1995). The linear logistic test model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications* (pp. 131–155). New York: Springer-Verlag.

Glas, C. A. W., & Ellis, J. (1994). *RSP: Rasch scaling program.* Groningen, The Netherlands: ProGAMMA.

Glas, C. A. W., & Verhelst, N. D. (1995). Testing the Rasch model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications* (pp. 69–95). New York: Springer-Verlag.

Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement, 26,* 301–321.

Hartz, S., DiBello, L. V., & Stout, W. F. (2000, July). *Hierarchical Bayesian approach to cognitive assessment: Markov chain monte carlo application to the Unified Model.* Paper presented at the Annual North American Meeting of the Psychometric Society, Vancouver, Canada.

Heckerman, D. (1998). A tutorial on learning with Bayesian networks. In M. Jordan (Ed.), *Learning in graphical models* (pp. 301–354). Dordrecht, The Netherlands: Kluwer.

Hemker, B. T., Sijtsma K., Molenaar, I. W., & Junker, B. W. (1997). Stochastic ordering using the latent trait and the sum score in polytomous IRT models. *Psychometrika, 62,* 331–347.

Holland, P. W., & Rosenbaum, P. R. (1986). Conditional association and unidimensionality in monotone latent trait models. *Annals of Statistics, 14,* 1523–1543.

Huguenard, B. R., Lerch, F. J., Junker, B. W., Patz, R. J., & Kass, R. E. (1997). Working memory failure in phone-based interaction. *ACM Transactions on Computer-Human Interaction, 4,* 67–102.

Junker, B. W. (2001). On the interplay between nonparametric and parametric IRT, with some thoughts about the future. In A. Boomsma, M. A. J. Van Duijn, & T. A. B. Snijders (Eds.), *Essays on item response theory* (pp. 274–276). New York: Springer-Verlag.

Junker, B. W., & Sijtsma, K. (2000). Latent and manifest monotonicity in item response models. *Applied Psychological Measurement, 24,* 65–81.

Kyllonen, P., & Christal, R. (1990). Reasoning ability is (little more than) working memory capacity? *Intelligence, 14,* 389–394.

Macready, G. B., & Dayton, C. M. (1977). The use of probabilistic models in the assessment of mastery. *Journal of Educational Statistics, 2,* 99–120.

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

Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika, 64,* 187–212.

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

Molenaar, I. W., & Sijtsma, K. (2000). *MSP5 for Windows* [Computer program]. Groningen, The Netherlands: ProGAMMA.

Nichols, P., & Sugrue, B. (1999). The lack of fidelity between cognitively complex constructs and conventional test development practice. *Educational Measurement: Issues and Practice, 18,* 18–29.

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

Reckase, M. D. (1997). A linear logistic multidimensional model for dichotomous item response data. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 271–286). New York: Springer-Verlag.

Resnick, L. B., & Resnick, D. P. (1992). Assessing the thinking curriculum: New tools for educational reform. In B. R. Gifford & M. C. O'Connor (Eds.), *Changing assessments: Alternative views of aptitude, achievement, and instruction* (pp. 37–75). Norwell MA: Kluwer.

Rijkes, C. P. M. (1996). *Testing hypotheses on cognitive processes using IRT models.* Unpublished doctoral dissertation, University of Twente, The Netherlands.

Sijtsma, K. (1998). Methodology review: Nonparametric IRT approaches to the analysis of dichotomous item scores. *Applied Psychological Measurement, 22,* 3–31.

Sijtsma, K., & Verweij, A. (1999). Knowledge of solution strategies and IRT modeling of items for transitive reasoning. *Applied Psychological Measurement, 23,* 55–68.

Spiegelhalter, D. J., Thomas, A., Best, N. G., & Gilks, W. R. (1997). *BUGS: Bayesian inference using Gibbs sampling, Version 0.6* [Computer program]. Cambridge, UK: MRC Biostatistics Unit.

Tanner, M. A. (1996). *Tools for statistical inference: Methods for the exploration of posterior distri-*

butions and likelihood functions (3rd ed.). New York: Springer-Verlag.

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. 327–359). Hillsdale NJ: Erlbaum.

Van der Ark, L. A. (2001). An overview of relationships in polytomous item response theory and some applications. *Applied Psychological Measurement, 25,* 273–282.

VanLehn, K., & Niu, Z. (in press). Bayesian student modeling, user interfaces and feedback: A sensitivity analysis. *International Journal of Artificial Intelligence in Education.*

VanLehn, K., Niu, Z., Siler, S., & Gertner, A. (1998). Student modeling from conventional test data: A Bayesian approach without priors. In B. P. Goettle, H. M. Halff, C. L. Redfield, & V. J. Shute (Eds.), *Proceedings of the Intelligent Tutoring Systems Fourth International Conference, ITS 98* (pp. 434–443). Berlin: Springer-Verlag.

Verweij, A., Sijtsma, K., & Koops, W. (1999). An ordinal scale for transitive reasoning by means of a deductive strategy. *International Journal of Behavioral Development, 23,* 241–264.

Wilson, M., & Sloane, K. (2000). From principles to practice: An embedded assessment system. *Applied Measurement in Education, 13,* 181–208.

Acknowledgments

This work was initiated during preparation of a commissioned paper for the National Research Council Committee on the Foundations of Assessment, United States National Academy of Sciences, while the first author was on leave at the Learning Research and Development Center, University of Pittsburgh, and was completed with partial support by National Science Foundation grants SES-99.07447 and DMS-97.05032. The authors thank Kerry Kravec for her computational help and Mark Schervish for helpful discussion of a monotonicity condition.

Authors' Addresses

Send requests for reprints or further information to Brian Junker, Department of Statistics, Carnegie Mellon University, 232 Baker Hall, Pittsburgh PA 15213, U.S.A.; or Klaas Sijtsma, Department of Methodology and Statistics, FSW, Tilburg University, P.O. Box 90153, 5000 LE Tilburg, The Netherlands. Email: brian@stat.cmu.edu; k.sijtsma@kub.nl.