Cognitive Assessment Models with Few Assumptions, and Connections with Nonparametric IRT

Brian Junker[1]
Department of Statistics
Carnegie Mellon University
Pittsburgh PA 15213, USA
`brian@stat.cmu.edu`

Klaas Sijtsma
Department of Methodology and Statistics, FSW
Tilburg University
5000 LE Tilburg
The Netherlands
`k.sijtsma@kub.nl`

December 27, 2000

[1] Some of the work reported here was performed in the course of preparing a commissioned paper for the National Research Council Committee on Educational and Psychological Foundations of Assessment, United States National Academy of Sciences. Preparation of this paper was supported in part by NSF grants #SES-99.07447 and #DMS-97.05032, and the hospitality of the Learning Research and Development Center, University of Pittsburgh. We thank Kerry Kravec for assistance with computations for examples in Section 4, and we thank Mark Schervish for a discussion greatly simplifying one of the monotonicity conditions discussed in Section 4.

Abstract

In recent years, as cognitive theories of learning and instruction have become richer, and computational methods to support assessment have become more powerful, there has been increasing pressure to make assessments truly criterion referenced, that is, to “report” on student achievement relative to theory-driven lists of examinee skills, beliefs and other cognitive features needed to perform tasks in a particular assessment domain. Cognitive assessment models must 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. In this paper we consider some usability and interpretability issues for single-strategy cognitive assessment models that posit a stochastic conjunctive relationship between a set of cognitive attributes to be assessed, and performance on particular items or tasks in the assessment. The attributes are coded as present or absent in each examinee, and the tasks are coded as performed correctly or incorrectly. The models we consider make few assumptions about the relationship between latent attributes and task performance beyond a simple conjunctive structure: all attributes relevant to task performance must be present to maximize probability of correct performance of the task. We show by example that these models can be sensitive to cognitive attributes even in data that was designed to be well-fit by the Rasch model, and we consider several stochastic ordering and monotonicity properties that enhance the interpretability of the models. We also identify some simple data summaries that are informative about the presence or absence of cognitive attributes, when the full computational power needed to estimate the models is not available.

1 Introduction

In recent years, as cognitive theories of learning and instruction have become richer, and computational methods to support educational assessment have become more powerful, there has been increasing pressure to make assessments sensitive to specific examinee skills, knowledge and other cognitive features needed to perform tasks in a particular assessment domain. For example Baxter and Glaser (1998) and Nichols and Sugrue (1999) present compelling cases that examinees' cognitive characteristics can and should be the focus of assessment design. Similarly, Resnick and Resnick (1992) advocate standards-referenced or criterion-referenced assessment closely tied to curriculum, as a way to inform instruction and enhance student learning. Cognitive assessment models generally deal with a more complex goal than ranking or locating examinees in a low-dimensional Euclidean space, which is what item response theory (IRT) has been designed and optimized to do. Rather, cognitive assessment models produce, for each examinee, a list of skills or other cognitive attributes that the examinee may or may not possess, based on the evidence of tasks performed by the examinee.

In this paper we consider some usability and interpretability issues for single-strategy cognitive assessment models that posit a conjunctive relationship between a set of cognitive attributes to be assessed, and performance on particular items or tasks in the assessment. The models we consider make few assumptions about the relationship between latent attributes and task performance beyond a simple conjunctive structure: all attributes relevant to task performance must be present to maximize probability of correct performance of the task.

Interpretability of IRT-like models is enhanced by simple, monotone relationships between parts of the model. For example, in addition to the usual monotonicity (M) assumption of IRT modeling (defined below), 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) were considered in detail by Hemker, Sijtsma, Molenaar and Junker (1997). We consider all three properties for two conjunctive cognitive assessment models, and we also consider a natural new monotonicity condition, which asserts that the more task-relevant skills an examinee possesses, the easier the task should be.

To fix notation, let us consider J dichotomous item response variables for each of N examinees, and let $X_{ij} = 1$ if subject i performs task j well, and 0 otherwise, with values x_{ij}, $i = 1, \ldots, N$, $j = 1, \ldots, J$. Let

θ_i denote person parameters (possibly multidimensional) and β_j denote item parameters (possibly multidimensional). We denote the usual item response function (IRF) in IRT as $P_j(\theta_i) = P[X_{ij} = 1|\theta_i, \beta_j]$.

Most parametric and nonparametric IRT models satisfy the three fundamental assumptions: *local independence (LI)*

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

for each i, *monotonicity (M)*, i.e., the item response functions (IRF's) $P_j(\theta_i)$ are nondecreasing as a function of θ_i, or coordinatewise nondecreasing, if θ_i is multidimensional; and *low dimensionality (LD)*, i.e. 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 each unidimensional real-valued parameters, and logit $P_j(\theta_i) = \theta_i - \beta_j$.

Many attempts (e.g., as surveyed Mislevy, 1996) to blend IRT and cognitive measurement are based on a linear decomposition of item parameters β_j or of person parameters θ_i. For example in the linear logistic test model (LLTM; e.g. Fischer, 1995; Huguenard, Lerch, Junker, Patz & Kass, 1997; Draney, Pirolli, & Wilson, 1995), the difficulty parameters β_j in the Rasch model are rewritten as linear combinations of K basic parameters η_k with weights q_{jk}, and logit $P_j(\theta_i) = \theta_i - \sum_{k=1}^{K} q_{jk}\eta_k$. The matrix $Q = [q_{jk}]$ is usually obtained a priori on the basis of an analysis of the items into 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. On the other hand, multidimensional compensatory IRT models (e.g. Adams, Wilson, & Wang, 1997; Reckase, 1997) follow the factor analytic tradition by decomposing the unidimensional θ_i parameter into an item-dependent linear combination of underlying traits, e.g. logit $P_j(\theta_i) = \sum_{k=1}^{K} B_{jk}\theta_{ik} - \beta_j$.

Compensatory IRT models, like factor analysis models, can be sensitive to relatively large components of variation in examinee ability or propensity to answer items correctly, but they are generally not designed to distinguish finer components of variation among examinees that are often of interest in cognitive assessment. LLTM-style models can be sensitive to these finer components of variation *among items*, but they are also not designed to be sensitive to components of variation *among examinees*—indeed the person parameters are often of little direct interest in an LLTM analysis.

Noncompensatory approaches, such as Embretson's (e.g. Embretson, 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. The MLTM asserts that successful performance on an item or task involves the conjunction of successful performances on several subtasks, each of which follow a separate unidimensional IRT model, such as the Rasch model,

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

Conjunctive approaches have generally been preferred in cognitive assessment models that focus on a single student strategy for performing tasks, in assessments (e.g. Tatsuoka, 1995; VanLehn, Niu, Siler & Gertner, 1998), and in intelligent tutoring systems (e.g. Corbett, Anderson & O'Brien, 1995; VanLehn & Niu, 1999). Multiple strategies are often accommodated with a hierarchical latent class structure that divides the examinee population into latent classes according to strategy, and uses a different model within each latent strategy class to describe the influence of attributes on task performance within that strategy class (e.g. Mislevy, 1996; Rijkes, 1996). Within a single strategy, models involving more complicated combinations of attributes driving task performance are certainly possible (e.g. Heckerman, 1996), but these become more challenging to estimate and interpret.

In this paper we focus 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. In Section 2 we review the assessment of transitive reasoning in young children considered by Sijtsma and Verweij (1999); and we discuss a simple task analysis leading to six cognitive attributes that we will use to illustrate several points in the paper. In Section 3 we introduce the models, and explore whether they satisfy a monotonicity condition in the latent attributes, by fitting them to the transitive reasoning data. In Section 4 we begin to build a nonparametric IRT perspective on the models: what properties of the models facilitate interpretation of parameter estimates? How does this help us find simple data summaries when the full computational power needed to estimate the models is not available? A brief discussion follows in Section 5.

2 An Assessment of Transitive Reasoning in School Children

Sijtsma and Verweij (1999) analyzed data from a set of transitive reasoning tasks, in which children are shown objects A, B, C, ..., with physical attributes Y_A, Y_B, Y_C, ... For example the objects might be sticks, and the attributes might be lengths. Relationships between attributes of all pairs of adjacent objects in an ordered series, such as $Y_A < Y_B$ and $Y_B < Y_C$, are shown to each child. The child is then asked to reason about the relationship between some pair not shown, e.g. Y_A vs. Y_C in this example. Reasoning directly from the premises $Y_A < Y_B$ and $Y_B < Y_C$ to the conclusion $Y_A < Y_C$, without guessing and without using any other information, is an example of *transitive reasoning* (see Sijtsma & Verweij, 1999; and Verweij, Sijtsma & Koops, 1999, for summaries of the relevant developmental psychology).

The data consist of the responses of 417 second-, third- and fourth-grade students, on nine transitive reasoning tasks of the sort just described. The tasks were generated by considering three types of objects featuring different physical attributes: wooden sticks differing in length by 0.2cm per pair of sticks, wooden disks differing in diameter by 0.2cm per pair of disks, and clay balls differing in weight by 30g per pair of balls. 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), and one item, AC (asking 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. Examinees' explanations for each item response were also recorded, to evaluate strategy use. A summary of the tasks is given in Table 1.

Sijtsma and Verweij (1999) showed that if each item within a task was scored as correct when *both* the correct answer was given *and* a correct deductive strategy based on transitive reasoning was evidenced in the examinee's explanation—referred to by them as the DEDSTRAT data—and if these dichotomous item scores were summed within tasks to give task scores, then the resulting task response data were well-fit by a polytomous monotone homogeneity model (that is, by a model assuming only local independence, unidimensionality and monotonicity; cf. e.g. van der Ark, 2001).

In order to facilitate analyses with binary-response models, we have recoded the Sijtsma and Verweij DEDSTRAT data somewhat further. In our recoding, a task is scored as correct if *all* the items within that task were answered correctly using a correct deductive strategy; otherwise the task was scored as incorrect. This leads to a 417 × 9 array of 1's (correct) and 0's (incorrect). In this array, the scores for all examinees on tasks 5 and 6, both involving size of disks, were 0's. Relatively large visual differences between the sizes

	Number & Type	Featured	Number of	Number of	Rasch Difficulties	
Task	of Objects	Attribute	Premises	Items	EAP	PSD
Task 1	3 Sticks	Length	2	1	–0.38	0.16
Task 2	4 Sticks	Length	3	2	1.88	0.17
Task 3	5 Sticks	Length	4	3	6.06	0.50
Task 4	3 Disks	Size	2	1	–1.78	0.17
Task 5	4 Disks	Size	3	2	12.60	5.12
Task 6	5 Disks	Size	4	3	12.40	4.86
Task 7	3 Balls	Weight	2	1	–3.40	0.22
Task 8	4 Balls	Weight	3	2	3.95	0.25
Task 9	5 Balls	Weight	4	3	8.07	1.23

Table 1: The nine transitive reasoning tasks of Sijtsma and Verweij (1999). Expected a posteriori (EAP) estimates and posterior standard deviations (PSD) for the difficulty parameters in a Bayesian fit of the Rasch model to the data are recorded on the right.

of the disks (disk diameters varied linearly, so disk areas varied quadratically) seem to have encouraged examinees to arrive at a correct answer for some items by direct visual comparison of the disks in the item, rather than by a deductive strategy; such responses would be coded as 0's (deductive strategy not used) in the DEDSTRAT data.

These nine binary DEDSTRAT tasks were fit well by a dichotomous monotone homogeneity model: After deleting tasks 5 and 6, whose responses were all zeros, the program MSP (Molenaar & Sijtsma, 2000) reported a very high scaling coefficient ($H = 0.82$) for the seven remaining tasks and 417 respondents of $H = 0.82$ and taskwise scaling coefficients H_j between 0.78 and 1.00 (cf. e.g. Sijtsma, 1998), and found no sample violations of manifest monotonicity (Junker & Sijtsma, 2000). A Rasch model was also fitted, using RSP (Glas & Ellis, 1994); again tasks 5 and 6 were deleted, and examinees with all-zero responses were deleted causing item 9 to have all zero responses in the reduced data set, so it was deleted as well. For the remaining six items and 382 respondents, standard Rasch fit statistics (e.g. as discussed by Glas & Verhelst, 1995) indicated very good fit as well. The Rasch model was refitted using BUGS 0.6 (Spiegelhalter, Thomas, Best & Gilks, 1997) which uses a Bayesian formulation of the model that does not necessitate deleting any items or persons; similarly good fit was found. The item difficulty parameters β_j estimated by BUGS, based on a fixed normal θ distribution and a common $N(\mu_\beta, \sigma^2_\beta)$ prior for the β's with weak hyperpriors

	Context			Premise		
	Length	Size	Weight	1^{st} & 2^{nd}	3^{rd}	4^{th}
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

Table 2: Decomposition of the Sijtsma and Verweij (1999) tasks ($j = 1, \ldots, 9$) into hypothetical cognitive attributes ($k = 1, \ldots, 6$). $Q_{jk} = 1$ if and only iff task j requires attribute k.

$\mu_\beta \sim N(0, 100)$ and $\sigma_\beta^{-2} \sim \text{Gamma}(0.01, 0.01)$, are recorded in the last two columns of Table 1.

If, however, we desire to use the transitive reasoning scale 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, because they only tell us the ranks or locations of examinees on a unidimensional latent scale. Instead we must explicitly model the task performance in terms of presence or absence of particular cognitive attributes related to transitive reasoning.

To illustrate, let us consider the rough task analysis for these nine tasks in Table 2, corresponding to the task summary in Table 1. The first three attributes are simply the ability to recognize or reason about transitivity in the context of length, size and weight. In addition, the tasks place differential load on subjects' working memory capacity (e.g. Carpenter, Just & Shell, 1990; Kyllonen & Christal, 1990). Thus the next three cognitive attributes correspond to three levels of working memory capacity: manipulating the first two premises given in a task in working memory; manipulating a third task premise, if it is given; and manipulating a fourth task premise, if it is given.

The issue here is not strictly model-data fit. If we wish to know whether particular students can focus on a transitive reasoning strategy in the context of weight problems, then the total score on the nine items—the central examinee statistic in Rasch and monotone homogeneity models generally—will not help. Similarly, an LLTM can tell us whether additional working memory load makes tasks more difficult on average, but it cannot tell us whether a particular student has difficulty with maintaining a third premise in solving transitive reasoning problems. In order to answer these questions we need to turn to models that partition the data into signal and noise differently than do unidimensional IRT models. Even if the data was designed for a unidimensional IRT model, as was the DEDSTRAT data, this repartitioning can be informative (e.g. Tatsuoka, 1995); and of course if we have in mind to use a cognitive assessment model we may wish to design assessment tasks, data collection, and data coding for analysis with such a model.

3 Two IRT-like Cognitive Assessment Models

We focus here on two discrete latent attribute models, that allow both for modeling the cognitive loads of items and inferences about the cognitive attributes of examinees. In both models the latent variable is a vector of 0's and 1's for each examinee, indicating the absence or presence of particular cognitive attributes for that examinee, and we use Table 2 to determine which attributes the examinee needs to perform each task correctly. As in Section 1, we will assume there are N examinees and J binary task performance variables, and in addition we suppose there is a fixed set of K cognitive attributes involved in performing these tasks; different subsets of attributes may be involved in different tasks. For both models we define

$$\begin{array}{rcl} X_{ij} & = & \text{1 or 0, indicating whether or not student } i \text{ performed task } j \text{ correctly;} \\ Q_{jk} & = & \text{1 or 0, indicating whether or not attribute } k \text{ is relevant to task } j\text{; and} \\ \alpha_{ik} & = & \text{1 or 0, indicating whether or not student } i \text{ possesses attribute } k. \end{array} \tag{1}$$

The values Q_{ij} are fixed in advance, like the design matrix in an LLTM model. The Q_{ij} can in fact be assembled into a Q-matrix of the type discussed by Tatsuoka (1995). Figure 1 illustrates the structure defined by X_{ij}, Q_{jk} and α_{ik} graphically as a Bayes network.

We wish to make inferences about the latent variables α_{ik}, or to make inferences about the relationship between these attributes and observed task performance. Both models are most easily specified using the latent response framework elaborated by Maris (1995), which is closely related to the notion of data augmentation in statistical estimation (e.g. Tanner, 1996).

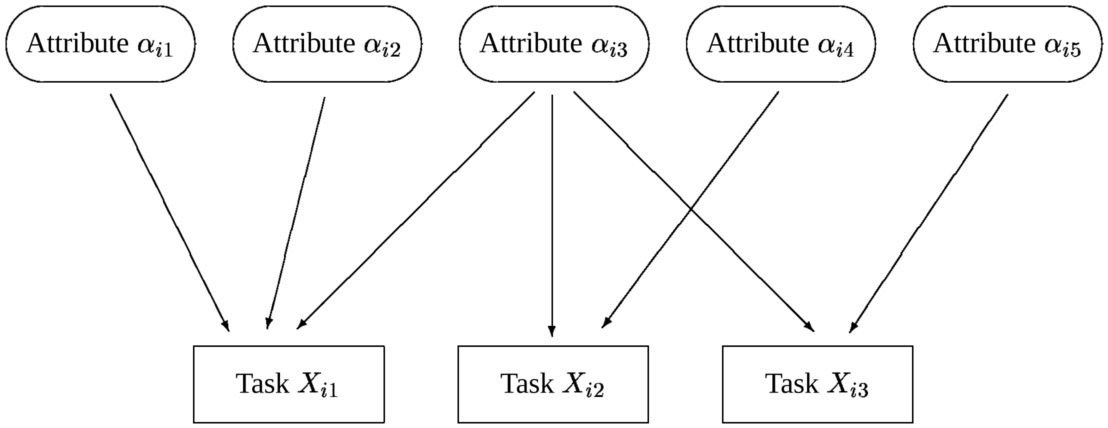

Figure 1: One-layer Bayes network for discrete cognitive attributes models. For examinee i, $\alpha_{ik} = 1$ or 0 describes presence or absence of latent attribute k; $X_{ij} = 1$ or 0 describes success or failure performing task j; and $Q_{jk} = 1$ or 0 describes the presence or absence of edges in the graph. Tasks are conditionally independent given attributes (local independence); attributes relevant to a task may combine conjunctively, as in models (3) and (5), or nonconjunctively, to influence task performance.

3.1 Deterministic Inputs, Noisy "And" Gate (DINA)

The first model we consider has been the foundation of several approaches to cognitive diagnosis and assessment (see the references in Tatsuoka, 1995; and Doignon & Falmagne, 1999). It was considered in detail by Haertel (1989; see also Macready & Dayton, 1977) who identified it as a restricted latent class model. In this model, we also define latent response variables

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

indicating whether or not student i has all the attributes required for task j. In Tatsuoka's terminology, the latent vectors $\alpha_{i\cdot} = (\alpha_{i1}, \ldots, \alpha_{iK})$ are called "knowledge states", and the vectors $\xi_{i\cdot} = (\xi_{i1}, \ldots, \xi_{iJ})$ are called "ideal response patterns", since 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 $s_j = P[X_{ij} = 0|\xi_{ij} = 1]$ and $g_j = P[X_{ij} = 1|\xi_{ij} = 0]$. Note that s_j and g_j are merely error probabilities—the false negative and false positive rates—in a simple signal detection model for detecting

ξ_{ij} from noisy observations X_{ij}. We have chosen the symbols s_j and g_j to be mnemonic, thinking of examinees' slips and guesses, but genuine slipping and guessing behavior may be the least important reason for observing $X_{ij} \neq \alpha_{ik}^{Q_{jk}}$. Other reasons include poor wording of the task description for examinees, inadequate specification of the Q matrix, use of an alternative solution strategy by the examinee, and lack of model fit generally. DiBello, Stout, & Roussos (1995) address this issue in detail in their discussion of what they call the *positivity* of a task with respect to a cognitive attribute, to which we shall return below in Section 4.

The IRF for a single task is

$$P[X_{ij} = 1|\boldsymbol{\alpha}, \boldsymbol{s}, \boldsymbol{g}] \;=\; (1 - s_j)^{\xi_{ij}} g_j^{1-\xi_{ij}} \;\equiv\; P_j(\alpha_{i\cdot}), \tag{2}$$

Note that each ξ_{ij} functions as an "and" gate (i.e., it is a binary function of binary inputs whose value is 1 if and only if all the inputs are 1's), combining the deterministic inputs $\alpha_{ik}^{Q_{jk}}$; and each task performance X_{ij} is modeled as a noisy observation of each ξ_{ij} (cf. VanLehn, Niu, Siler & Gertner, 1998). We refer to the model hereafter as the DINA (deterministic inputs, noisy "and") model. From (2) it is also clear that $P_j(\alpha_{i\cdot})$ will be coordinatewise monotone (M) in $\alpha_{i\cdot}$ if and only if $1 - s_j > g_j$. Assuming local independence and independence among examinees, the joint likelihood for all responses under the DINA model is

$$\begin{aligned} P[X_{ij} = x_{ij},\ \forall\, i, j \mid \boldsymbol{\alpha}, \boldsymbol{s}, \boldsymbol{g}] \;&=\; \prod_{i=1}^{N}\prod_{j=1}^{J} P_j(\alpha_{i\cdot})^{x_{ij}}[1 - P_j(\alpha_{i\cdot})]^{1-x_{ij}} \\ &=\; \prod_{i=1}^{N}\prod_{j=1}^{J} \left[(1 - s_j)^{x_{ij}} s_j^{1-x_{ij}}\right]^{\xi_{ij}} \left[g_j^{x_{ij}}(1 - g_j)^{1-x_{ij}}\right]^{1-\xi_{ij}} \end{aligned} \tag{3}$$

3.2 Noisy Inputs, Deterministic "And" Gate (NIDA)

The second model we consider was recently discussed by Maris (1999) for example, and has been used as a building block in more elaborate cognitive diagonsis models (e.g. DiBello et al., 1995). In this model we take X_{ij}, Q_{jk} and α_{ik} as in (1), and define latent response variables

η_{ijk} = 1 or 0 indicating whether or not student i's performance in the context of task j is consistent with possessing attribute k

The latent response variables η_{ijk} are related to the student's knowledge state $\alpha_{i\cdot}$ according to the probabilities $s_k = P[\eta_{ijk} = 0|\alpha_{ik} = 1, Q_{jk} = 1]$ and $g_k = P[\eta_{ijk} = 1|\alpha_{ik} = 0, Q_{jk} = 1]$; and for completeness we define $P[\eta_{ijk} = 1|\alpha_{ik} = a, Q_{jk} = 0] \equiv 1$, regardless of the value a of α_{ik}. Again, s_k and g_k are merely mnemonically-named false negative and false positive error probabilities in a signal detection model for detecting α_{ik} from noisy η_{ijk}; as discussed above for the DINA model, the real causes for observing $\eta_{ijk} \neq \alpha_{ik}^{Q_{jk}}$ are likely much more varied than simple slipping and guessing behavior. Observed task performance is related to the latent response variables via $X_{ij} \equiv \prod_{k:\ Q_{jk}=1} \eta_{ijk} = \prod_{k=1}^{K} \eta_{ijk}$, so that the IRF for this model is

$$\begin{aligned} P[X_{ij} = 1|\boldsymbol{\alpha}, \boldsymbol{s}, \boldsymbol{g}] &= \prod_{k=1}^{K} P[\eta_{ijk} = 1|\alpha_{ik}, Q_{jk}] \\ = \prod_{k=1}^{K} \left[(1 - s_k)^{\alpha_{ik}} g_k^{1-\alpha_{ik}}\right]^{Q_{jk}} &= \prod_{k=1}^{K} \left(\frac{1 - s_k}{g_k}\right)^{\alpha_{ik} Q_{jk}} \prod_{k=1}^{K} g_k^{Q_{jk}} \equiv P_j(\alpha_{i\cdot}). \end{aligned} \quad (4)$$

In this model, noisy inputs η_{ijk} reflecting possession of attributes α_{ik} by examinees are combined in a deterministic “and” gate X_{ij}; we refer to the model hereafter as the NIDA model (noisy inputs, deterministic “and”). Again it is clear that the vector $\alpha_{i\cdot}$ plays the role of the latent variable θ_i, that s_k and g_k play the role of β_j, and that this 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

$$\begin{aligned} P[X_{ij} = x_{ij},\ \forall\, i, j \mid \boldsymbol{\alpha}, \boldsymbol{s}, \boldsymbol{g}] &= \prod_{i=1}^{N} \prod_{j=1}^{J} P_j(\alpha_{i\cdot})^{x_{ij}} [1 - P_j(\alpha_{i\cdot})]^{1-x_{ij}} \\ = \prod_{i=1}^{N} \prod_{j=1}^{J} \left\{ \prod_{k=1}^{K} \left[(1 - s_k)^{\alpha_{ik}} g_k^{1-\alpha_{ik}}\right]^{Q_{jk}} \right\}^{x_{ij}} & \left\{ 1 - \prod_{k=1}^{K} \left[(1 - s_k)^{\alpha_{ik}} g_k^{1-\alpha_{ik}}\right]^{Q_{jk}} \right\}^{1-x_{ij}} . \end{aligned} \quad (5)$$

3.3 Exploring Monotonicity

Both the DINA and NIDA models are stochastic conjunctive models for task performance: under the monotonicity conditions $1 - s > g$, examinees must possess all attributes listed for each task, in order to maximize the probability of successful performance of that task. Both the DINA and NIDA models are restricted latent class models (cf. Haertel, 1989), and therefore closely related to IRT models, as we have tried to suggest in the joint likelihood expressions (3) and (5). Indeed, if one replaces the IRF’s $P_j(\alpha_{i\cdot})$ in these likelihoods with standard logistic IRF’s $P_j(\theta_i)$ one immediately returns to a familiar IRT setting. They can also both be seen as simple one-layer Bayes inference networks for discrete variables (e.g. Mislevy, 1996; VanLehn, Niu, Siler & Gertner, 1998) for task performance, as illustrated in Figure 1. In general Bayes net models need not be conjunctive (e.g. Heckerman, 1996) but when examinees are presumed to be using a single strategy, conjunctive models seem natural; DiBello et al. (1995) provide a detailed analysis of conjunctive cognitive modeling in the development of their "Unified Model" for cognitive assessment.

To explore whether monotonicity actually holds in real data, we used BUGS 0.6 (Spiegelhalter et al., 1996) to fit the DINA and NIDA models to the dichotomous DEDSTRAT data described in Section 2, using the Q-matrix in Table 2. We used Bayesian formulations of the models, in which 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}] were also 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 the upper bounds g_{max} and s_{max} are small, these prior choices tend to favor error probabilities satisfying the monotonicity conditions $1-s > g$; these upper bounds were also estimated in the model, using Unif[0,1] hyperprior distributions. For each model we ran five Markov Chain Monte Carlo (MCMC) chains of length 3000 from various randomly-chosen start points; the first 2000 steps of each chain were discarded as burn-in, and the remaining 1000 steps were thinned by retaining every fifth observation, for a total of 200 observations per chain. Both models showed some evidence of under-identification (slow convergence and multiple maxima), as would be expected from Tatsuoka (1995) and Maris (1999). Tables 3 and 4 list tentative estimated posterior means (expected a posteriori values, or EAP's) and posterior standard deviations (PSD's) for each set of error probabilities in the two models, using 1000 MCMC steps obtained by pooling the five thinned chains for each model.

Several observations can be made regarding Tables 3 and 4. First, most of the point estimates satisfy the monotonicty condition $1 - s > g$, or equivalently, $g + s < 1$ (the exceptions are the error probabilities for tasks 4 and 8 under the DINA model). Examination of the posterior distributions in each model shows that the posterior probability that $1 - s > g$ for each task (DINA model) or latent attribute (NIDA model) is near 0.50 in each case: this certainly does not contradict the hypothesis that M holds, but neither is it a strong confirmation of M. Second, in the DINA model, tasks 5 and 6 on which all examinees scored zeros yield the very plausible estimates $\hat{g}_j = \widehat{P}[X_{ij} = 1|\xi_{ij} = 0] = 0.002$ $(PSD = 0.002)$; clearly there is no evidence of guessing or alternate solution strategies in the performance data for these tasks! Third, except for these two tasks, all the error probabilities in the DINA model are near their prior means, with fairly large

	$\hat{g}_j$		$\hat{s}_j$			
Task (j)	EAP	PSD	EAP	PSD	$1 - \hat{s}_j > \hat{g}_j$?	$[(1 - \hat{g}_j)/\hat{g}_j] \cdot [(1 - \hat{s}_j)/\hat{s}_j]$
1	0.478	0.167	0.486	0.277	√	1.15
2	0.363	0.162	0.487	0.281	√	1.85
3	0.419	0.255	0.479	0.292	√	1.51
4	0.657	0.199	0.488	0.279	×	0.55
5	0.002	0.002	0.462	0.270	√	581.09
6	0.002	0.002	0.464	0.270	√	576.43
7	0.391	0.420	0.486	0.274	√	1.65
8	0.539	0.242	0.489	0.275	×	0.89
9	0.411	0.162	0.480	0.283	√	1.55
	$\hat{g}_{max}$		$\hat{s}_{max}$			
	0.910	0.081	0.910	0.079		

Table 3: Tentative expected a posteriori (EAP) estimates and posterior standard deviations (PSD) for the task-wise error probabilities g_j and s_j in the DINA model, for the transitive reasoning data, with Q-matrix as in Table 2. The last column is discussed in Section 4.2.

	$\hat{g}_k$		$\hat{s}_k$				
Attrib (k)	EAP	PSD	EAP	PSD	$1 - \hat{s}_k > \hat{g}_k$?	$\frac{1-\hat{s}_k}{\hat{g}_k}$	$\log \frac{1-\hat{s}_k}{\hat{g}_k}$
1	0.467	0.364	0.369	0.392	√	1.351	0.301
2	0.749	0.207	0.161	0.125	√	1.120	0.113
3	0.764	0.246	0.005	0.009	√	1.302	0.264
4	0.364	0.319	0.163	0.318	√	2.299	0.833
5	0.176	0.168	0.785	0.129	√	1.222	0.200
6	0.061	0.115	0.597	0.294	√	6.607	1.888
	$\hat{g}_{max}$		$\hat{s}_{max}$				
	0.877	0.109	0.877	0.108			

Table 4: Tentative expected a posteriori (EAP) estimates and posterior standard deviations (PSD) for the attribute-wise error probabilities g_k and s_k in the NIDA model, for the transitive reasoning data, with Q-matrix as in Table 2. The last two columns are discussed in Section 4.2.

posterior SD's, suggesting that in the DINA model the attributes outlined in Table 2 are not very predictive of successful task performance. On the other hand, the error probabilities in the NIDA model seem to have moved farther from their prior means, and in some cases with relatively small PSD's. For example, attributes 4, 5 and 6, indicating increasing cognitive load, have decreasing g_k's and roughly increasing s_k's, reflecting the increasing difficulty of tasks involving these attributes. Finally we note that the EAP estimates of g_{max} and s_{max} in both models are above 0.870, with small PSD's. This reflects the large PSD's (and hence large estimation uncertainty) associated with at least some of the error probabilities in each model; in addition it suggests that the prior preference for the monotonicity condition $1 - s > g$ was not very strong, so that the evidence for monotonicity we see in the model fits may reflect the data and not the prior distribution choices.

4 A Nonparametric IRT Perspective on Cognitive Assessment Models

One of the strengths of the nonparametric IRT (NIRT) approach is that it encourages researchers to consider fundamendal model properties that are important for inference about latent variables from observed data. In this section we consider the DINA and NIDA models, described in (3) and (5) respectively, in this context. We begin by reviewing how parameter estimates in the two models depend on the data, and then continue with a broader discussion of NIRT-like properties, propelled by these two models. We examine standard NIRT that facilitate interpretation of parameter estimates in these cognitive assessment models, and explore other general properties that suggest themselves as we compare the DINA and NIDA models.

4.1 Data Summaries Relevant to Parameter Estimation

Junker (2001) considered the DINA model, (3), as a possible starting place for formulating a nonparametric IRT for cognitive assessment models. Using calculations for the complete conditional distributions often employed in Markov Chain Monte Carlo (MCMC) estimation algorithms, he showed that (a) estimation of the "slip" probabilities s_j were sensitive only to examinees' performances X_{ij} on tasks for which they were hypothesized to have all the requisite cognitive attributes ($\xi_{ij} = 1$); similarly (b) estimation of the "guessing" probabilities g_j depended only on examinees' task performances X_{ij} corresponding to tasks for which one or more attributes were missing ($\xi_{ij} = 0$); and finally (c) estimation of α_{ik}, indicating posession 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.

Indeed, we may calculate the posterior odds of $\alpha_{ik} = 1$, conditional on the data and all other parameters, from Junker's (2001) complete conditional distribution for α_{ik}. These odds turn out to be

$$\prod_{j=1}^{J} \left(\frac{s_j}{1-g_j} \right)^{\xi_{ik}^{(-k)} Q_{jk}} \cdot \prod_{j=1}^{J} \left[\left(\frac{1-g_j}{g_j} \cdot \frac{1-s_j}{s_j} \right)^{\xi_{ik}^{(-k)} Q_{jk}} \right]^{x_{ij}} \cdot \frac{\pi_{ik}^{\alpha}(1)}{\pi_{ik}^{\alpha}(0)} \qquad (6)$$

where $\xi_{ij}^{(-k)} = \prod_{\ell \neq k : Q_{j\ell}=1} \alpha_{i\ell}$, which indicates presence of all attributes needed for task j, *except* for attribute k, and $\pi_{ik}^{\alpha}(1)/\pi_{ik}^{\alpha}(0)$ are the prior odds. The first product in (6) is constant in the data. From the second product, we see that the odds of $\alpha_{ik} = 1$ are multiplied by $[(1-g_j)/g_j] \cdot [(1-s_j)/s_j]$ 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 ($1 - s_j > g_j$) holds, this multiplier is greater than one. Table 3 shows that these multipliers range from 0.55 to 1.85, except for tasks 5 and 6. (Tasks 5 and 6 have very high multipliers because the model was able to estimate g_j's near zero, since no one got those tasks correct.) Combining the influence of these multipliers with the effect of the $\xi_{ij}^{(-k)}$ in (6), we see that getting additional tasks right in this model may not appreciably move the odds that an examinee possesses any one of the latent attributes, a phenomenon previously noted by VanLehn et al. (1998).

We now turn to a similar analysis of the NIDA model. We consider a Bayesian version of the model, in which the likelihood (5) is multiplied by unspecified, independent priors $\pi(\boldsymbol{\alpha}) = \prod_{ik} \pi_{ik}^{\alpha}(\alpha_{ik})$, $\pi(\boldsymbol{g}) = \prod_k \pi_k^g(g_k)$, and $\pi(\boldsymbol{s}) = \prod_k \pi_k^s(s_k)$. We again consider the complete conditional distributions that figure in the development of MCMC algorithms; these are the distributions of each parameter in turn, conditional on the values of all other parameters as well as the data.

The complete conditional distribution for any parameter, such as the "guessing" probability g_k, is proportional to the product of those factors in (5) containing that parameter, g_k, and the prior density $\pi_k^g(g_k)$ for the parameter. For g_k, this turns out to be

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

where

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

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

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

We see at once that estimates of g_k depend precisely on those task responses for which attribute k was called for but not present in the responding examinees; similarly s_k depends on those task responses for which attribute k was called for and was present in the responding examinee.

The complete conditional distribution for each latent attribute indicator α_{ik_0} is proportional to

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

where

$$\begin{aligned} m_k^i &= \textstyle\sum_{j:\, Q_{jk}=1} x_{ij} \;=\; \sum_{j=1}^J x_{ij} Q_{jk} \\ &= \text{number of tasks correct involving attribute } k \\ n_k &= \textstyle\sum_{j=1}^J Q_{jk} \\ &= \text{total number of tasks involving attribute } k \end{aligned} \tag{7}$$

and with c_k^i defined as before. After a little algebra, we conclude that the posterior odds of $\alpha_{ik} = 1$, conditional on the data and all other parameters, is equal to

$$\left(\frac{1 - s_k}{g_k} \right)^{m_k^i} \left(\frac{1 - c_k^i (1 - s_k)}{1 - c_k^i g_k} \right)^{n_k - m_k^i} \cdot \frac{\pi_{ik}^{\alpha}(1)}{\pi_{ik}^{\alpha}(0)} \tag{8}$$

in the NIDA model. When monotonicity $(1 - s_k > g_k)$ holds, the first term in parentheses is greater than one and the second term in parentheses is less than one, so that the odds of $\alpha_{ik} = 1$ increase as m_k^i, the number of tasks involving attribute k that the examinee has gotten correct, increases. Essentially, the conditional odds (8) of $\alpha_{ik} = 1$ are multiplied by $(1 - s_k)/g_k$ for each additional correct task involving attribute k—regardless of the examinee's status on the other attributes. (The quantity c_k^i that appears in (8) is typically less than 10^{-5}, so the second term in parentheses in (8) is negligible.) From Table 4, we see that these multipliers range from approximately 1.1 to approximately 1.4, except for the higher multipliers for attributes 4 (cognitive capacity to handle the first two premises in a task) and 6 (cognitive capacity to handle the fourth premise in a task). Attribute 4 has moderately low estimated guessing and slip probabilities, and attribute 6 has a very low estimated guessing probability, which increase the model's certainty that each of these two attributes is possessed, when an examinee gets a task correct that depends on that attribute.

Hartz, DiBello and Stout (2000) point out that the quantities $(1 - s_k)/g_k$ measure what DiBello et al. (1995) call *positivity*, which is roughly the extent to which the task performance is a determinstic function of the knowledge state $\alpha_{i\cdot}$. Our analysis of (8) shows that attributes with high positivity are strongly credited in the NIDA model when the corresponding tasks are performed well. More generally, comparing the posterior odds (6) for the DINA model with the posterior odds (8) for the NIDA model, we see 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.

4.2 Three NIRT Monotonicity Properties

We have seen already that LI holds by construction in these models, and LD holds as long as we keep the number of cognitive attributes K small, relative to the number of tasks J. As discussed in Section 3, M holds as long and $1 - s > g$ for each task or latent attribute, and there some evidence in favor of this property in the example data set. We briefly consider here the SOM and SOL properties (defined below) discussed by Hemker et al. (1997), and introduce a new monotonicity property that has some of the intuitive appeal of monotonicity in a conventional unidimensional IRT model.

For models satisfying LI, M and LD, it follows immediately from Lemma 2 of Holland and Rosenbaum (1986) that for any non-decreasing summary $g(\boldsymbol{X})$ of $\boldsymbol{X} = (X_1, \ldots, X_J)$, $E[g(\boldsymbol{X}) \mid \alpha_{i\cdot}]$ is non-decreasing in each coordinate α_{ik} of $\alpha_{i\cdot}$; this implies the SOM (Stochastic Ordering of the Manifest score X_+ by the latent trait) property of Hemker et al. (1997), that $P[X_+ > c \mid \alpha_{i\cdot}]$ is non-decreasing in each coordinate α_{ik} of $\alpha_{i\cdot}$. Not much is known about the inverse and more useful property SOL (Stochastic Ordering of the Latent variables by the manifest score X_+; Hemker et al., 1997), that $P[\alpha_{i1} > c_1, \ldots, \alpha_{iK} > c_K \mid X_+ = s]$ is nondecreasing in s, when the latent "trait" is multidimensional. We consider here a weaker, related property that can be studied using our work in Section 4.1, namely that

$$P\left[\, \alpha_{ik} = 1 \;\middle|\; \alpha_{i1}, \ldots, \alpha_{i(k-1)}, \alpha_{i(k+1)}, \ldots, \alpha_{iK}, \text{ and } \sum_{j:Q_{jk}=1} X_{ij} = s \,\right] \tag{9}$$

is non-decreasing in s, with all other model parameters fixed.

For the NIDA model, property (9) is immediate from inspecting the posterior odds (8) of $\alpha_{ik} = 1$, since by (7), m_k^i equals the sum $\sum_{j:Q_{jk}=1} X_{ij}$ on the right in (9). However, (9) need not hold for the DINA model, as inspection of the corresponding odds (6) of $\alpha_{ik} = 1$ in the DINA model shows. Indeed, if the products of odds $[(1 - g_j)/g_j] \cdot [(1 - s_j)/s_j]$ vary greatly, it need not be the case that (6) is monotone in $m_k^i = \sum_{j:Q_{jk}=1} X_{ij}$, the number of correctly performed tasks involving attribute k.

Finally we turn to a new kind of monotonicity that seems plausible for some cognitive assessment models. In a standard monotone unidimensional IRT model of the kind reviewed in Section 1, the more of θ there is, the higher the probability of getting a task right. What corresponds to this in the NIDA and DINA models? One answer might be, the more task-relevant latent attributes the examinee has, the higher the probability of correct task performance should be. In other words, we are asking whether each of the IRF's in equations (2) and (4) is nondecreasing in

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

This monotonicity property is easy to see for the DINA model, since in that model $P_j(\alpha_{i\cdot}) = (1 - s_j)^{\xi_{ij}} g_j^{1-\xi_{ij}}$ equals g_j as long as $m_{ij} < \sum_{k=1}^K Q_{jk}$, and switches to $1 - s_j$ as soon as $m_{ij} = \sum_{k=1}^K Q_{jk}$. For the NIDA model it is not so obvious, and in fact not true in general.

In the NIDA model, $P_j(\alpha_{i\cdot}) = \prod_{k=1}^K [(1 - s_k)/g_k]^{\alpha_{ik} Q_{jk}} \prod_{k=1}^K g_k^{Q_{jk}}$ varies with m_{ij} through the first term, since 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} in the NIDA model is equivalent to

$$\min_{\boldsymbol{e}:e_+=s+1} \sum_{k=1}^{K} e_k p_k \geq \max_{\boldsymbol{e}:\ e_+=s} \sum_{k=1}^{K} e_k p_k, \tag{10}$$

for each s, where $e_+ = \sum_k e_k$. This constrains the variability of the $p_k = \log(1 - s_k)/g_k$. When the e_k are unrestricted, it is easy to see that the condition (10) is equivalent to the single inequality

$$\sum_{k=1}^{s_0+1} p'_k \geq \sum_{k=K-s_0+1}^{K} p'_k \tag{11}$$

where the 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 $(K - 1)/2$. In other words, (10) holds for all a and all $\boldsymbol{e}$, if and only if it holds for s_0 (the largest possible s), and the $\boldsymbol{e}$'s that allocate the smallest $s_0 + 1$ p's to one sum and the largest s_0 p's to the other. If either (10) or (11) holds, then all the IRF's P_j in the NIDA model will be monotone in m_{ij}.

For the NIDA parameter estimates in Table 4, we find that $p'_1 + p'_2 + p'_3 = 0.577 < 2.721 = p'_5 + p'_6$. Thus there is no guarantee of monotonicity for all the fitted IRF's $P_j(\alpha_{i\cdot})$ in m_{ij}. However, the e_k are restricted in the NIDA model by the constants Q_{jk}; in the case of the transitive reasoning data the Q_{jk}'s limit the number of attributes that can affect each task to two, three or four. It is easy to see that the two-attribute tasks, tasks 1, 4, and 7, have IRF's that are monotone in m_{ij} (indeed any two-attribute task satisfies the condition (11) vacuously). On the other hand, none of the other tasks have monotone IRF's. Looking at Table 4, the problem is the vast disparity between attribute 4 (maintaining the first two premises of a task), with $p_4 = 0.833$ and attribute 5 (maintaining the third premise), with $p_5 = 0.200$. Task two involves attributes 1, 4, and 5, for example, and $p_1 + p_5 < p_4$, violating condition (10); hence the IRF $P_2(\alpha_{i\cdot})$ will *not* be monotone in m_{i2}.

5 Discussion

Standard unidimensional IRT modeling may be less relevant, even if the fit is good, if the goal of testing is cognitive assessment or diagnosis. We have explored two single-strategy, conjunctive latent class models that have appeared in the educational assessment literature (e.g., Haertel, 1989; and Maris, 1999), and showed that they satisfy familiar multidimensional generalizations of standard IRT assumptions. Using data from an investigation of scale construction for transitive reasoning (Sijtsma & Verweij, 1999), we fitted the models and showed that, in particular, IRT monotonicity is plausible for most tasks under both model fits.

The cognitive analysis in our example was not deep psychologically, and the data had been designed originally to fit a standard IRT model, not a cognitive assessment model. Nevertheless both cognitive assesment models found interesting structure at the cognitive attributes level, a phenomenon that has also been observed in applications of Tatsuoka's (1995) rule-space methodology to standardized testing data created with IRT technology. We can imagine that data designed from scratch to be informative about a handful of cognitive attributes through one of the two cognitive assessment models we considered, would fare quite well in terms of model fit and ability to infer about the presence or absences of particular attributes.

We also began considering these models from a nonparametric IRT (NIRT) perspective, seeking to understand both previously-identified, as well as new, nonparametric monotonicity properties, for example. This enterprise could be widened along the lines of Junker (2001), who considers a variety of NIRT properties in the context of cognitive assessment models, and could be applied to natural generalizations of the models we have considered, such as the Unified Model of DiBello et al. (1995) and Hartz et al. (2000). A NIRT perspective can help with the interpretation and usability of cognitive assessment models. 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. The condition holds "almost for free" in one of the models we considered, and it puts interesting constraints on the parameters of the other model.

Finally, we related some model parameters to simple and useful data summaries, such as the number of correctly performed tasks involving a particular attribute. Such summaries are useful when the machinery needed to estimate the models is not available, such as in embedded assessments (e.g. Wilson & Sloane, 2000). We hope that this is a start down the path toward a clear theory of which data summaries are relevant to the cognitive inferences we wish to make, over a wide variety of cognitive assessment models; such a theory would be an important contribution from the interface between nonparametric and parametric IRT 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, 97*.

Corbett, A. T., Anderson, J. R., & O'Brien, A. T. (1995). Student modeling in the ACT programming tutor. Chapter 2 in Nichols, P. D., Chipman, S. F., & Brennan, R. L. (eds.) (1995). *Cognitively diagnostic assessment.* Hillsdale, NJ: Lawrence Erlbaum Associates.

DiBello, L. V., Stout, W. F., & Roussos, L. A. (1995). Unified cognitive/psychometric diagnostic assessment likelihood-based classification techniques. Chapter 15, pp. 361–389, in Nichols, P. D., Chipman, S. F., & Brennan, R. L. (eds). (1995). *Cognitively diagnostic assessment.* 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: Lawrence Erlbaum Associates.

Embretson, S. E. (1997). Multicomponent response models. In W. J. Van der Linden and 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. Chapter 8 (pp. 131–155) in Fischer, G. H., & Molenaar, I. W. (eds.) (1995). *Rasch models: foundations, recent developments, and applications.* New York: Springer-Verlag.

Glas, C. A. W. & Ellis, J. (1994). *RSP. Rasch Scaling Program*. Groningen: iecProGAMMA.

Glas, C. A. W. & Verhelst, N. D. (1995). Testing the Rasch model. Chapter 5 (pp. 69–95) in Fischer, G. H., & Molenaar, I. W. (eds.) (1995). *Rasch models: foundations, recent developments, and applications.* 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). *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, July 8, 2000. Vancouver, BC, Canada.

Heckerman, D. (1996). *A tutorial on learning with Bayesian networks.* Microsoft Research tech. report, MSR-TR-95-06. Obtained November 2000 from www address `ftp://ftp.research.microsoft.com/pub/tr/TR-95-06.PS`

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 Boomsma A., et al. (Eds.) *Essays on Item Response Theory*. 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.* Groningen: iecProGAMMA.

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.

Reckase, M. D. (1997). A linear logistic multidimensional model for dichotomous item response data. Chapter 16, pp. 271–286 in In W. J. Van der Linden, and R. K. Hambleton (Eds.), *Handbook of modern item response theory*. New York: Springer Verlag.

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

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

Sijtsma, K. (1998). Methodology review: Nonparametric IRT approaches to the analysis of dichotomous item scores. *Applied Psychological Measurement, 22,* 3–32.

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. MRC Biostatistics Unit, Cambridge.

Tanner, M. A. (1996). *Tools for statistical inference: methods for the exploration of posterior distributions and likelihood functions. 3rd Edition.* 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, and R. L. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 327–359). Hillsdale, NJ: Lawrence Erlbaum Associates.

van der Ark, L. A. (2001). An overview of relationships in polytomous IRT and some applications. *Applied Psychological Measurement, xx,* xxx–xxx.

VanLehn, K., & Niu, Z. (1999). Bayesian student modeling, user interfaces and feedback: A sensitivity analysis. In press, *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. pp. 434–443 in Goetl, B. et al. (Eds.) (1998). *Proceedings of the Intelligent Tutoring Systems Fourth International Conference, ITS 98.* Berlin, Hiedelberg: 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.