Journal of Classification 30 (2013)
DOI: 10.1007/s00357-013-9130-y

Incorporating Student Covariates in Cognitive Diagnosis Models

Elizabeth Ayers

American Institutes for Research, Washington DC

Sophia Rabe-Hesketh

University of California, Berkeley CA and Institute of Education, University of London UK

Rebecca Nugent

Carnegie Mellon University, Pittsburgh PA

Abstract: In educational measurement, cognitive diagnosis models have been developed to allow assessment of specific skills that are needed to perform tasks. Skill knowledge is characterized as present or absent and represented by a vector of binary indicators, or the skill set profile. After determining which skills are needed for each assessment item, a model is specified for the relationship between item responses and skill set profiles. Cognitive diagnosis models are often used for diagnosis, that is, for classifying students into the different skill set profiles. Generally, cognitive diagnosis models do not exploit student covariate information. However, investigating the effects of student covariates, such as gender, SES, or educational interventions, on skill knowledge mastery is important in education research, and covariate information may improve classification of students to skill set profiles. We extend a common cognitive diagnosis model, the DINA model, by modeling the relationship between the latent skill knowledge indicators and covariates. The probability of skill mastery is modeled as a logistic regression model, possibly with a student-level random intercept, giving a higher-order DINA model with a latent regression. Simulations show

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through grant R305A070254 to University of California, Berkeley. The opinions expressed are those of the authors and do not represent the view of the Institute or the U.S. Department of Education.

Authors' Addresses: E. Ayers, American Institutes for Research, 1000 Thomas Jefferson St. NW, Washington DC, e-mail: eayers@air.org; S. Rabe-Hesketh, Graduate School of Education, University of California, 1501 Tolman Hall, University of California, Berkeley, CA 94720-1670 USA, e-mail: sophiarh@berkeley.edu; and R. Nugent, Department of Statistics, Baker Hall, Carnegie Mellon University, Pittsburgh, PA 15213 USA, e-mail: rnugent@stat.cmu.edu.

Published online

that parameter recovery is good for these models and that inclusion of covariates can improve skill diagnosis. When applying our methods to data from an online tutor, we obtain reasonable and interpretable parameter estimates that allow more detailed characterization of groups of students who differ in their predicted skill set profiles.

Keywords: Cognitive diagnosis model; Collateral information; Concomitant variables; Covariates; DIF; DINA; Higher order model; Random effect; Skill diagnosis.

1. Introduction

An important goal of educational assessment is identifying students' knowledge of specific skills that are needed to perform tasks. Such 'cognitive diagnosis' is important for planning instruction and providing feedback to students and teachers (Rupp, Templin, and Henson 2010). It also forms an important component in intelligent tutoring systems (e.g. Self 1993) where diagnosing student misconceptions and knowledge gaps is critical for responding with appropriate corrective action such as feedback or remedial instruction.

In recent years, a number of cognitive diagnosis models, often referred to as CDMs, have been developed (Rupp, Templin, and Henson 2010). These models use student responses to assessment items and information about which skills are required by each item to identify students' latent skill knowledge. Student covariate information tends to be ignored in the CDM framework. However, investigating the effects of student covariates, such as gender, socioeconomic status, or educational interventions, on skill knowledge mastery is important in education research. In addition, knowing the demographics of students who will likely have many unmastered skills may allow teachers or other educators to focus early on these at-risk students. Importantly, covariate information may improve diagnosis of individual skill knowledge and the subsequent classification of students into unique skill set profiles. Finally, when assessing differential item functioning (DIF) between focal and reference groups in a cognitive diagnosis model, it may be important to allow skill knowledge mastery probabilities to differ between groups, analogously to allowing for group differences in the latent trait mean in traditional item response models.

Like CDMs, classical latent variable models such as common factor models, latent class models, and item response theory (IRT) models were initially developed without covariates. For each of these model types, initial work on incorporating covariate information focused on multiple group analysis, where some of the model parameters can differ between groups (Jöreskog 1971; Clogg and Goodman 1984; Mislevy 1985). However, inclusion of several covariates, some of which may be continuous, necessitates some form of regression model. Factor models were extended to include linear regressions for common factors by Jöreskog and Goldberger (1975) who called the models ‘multiple indicator multiple cause’ (MIMIC) models. Dayton and Macready (1988), Formann (1992), and Wedel (2002) allowed latent class membership to depend on ‘concomitant’ variables via multinomial logistic regression models, and Mislevy (1987) and Zwindermann (1991) added ‘auxiliary information’, now often referred to as ‘conditioning variables’ to IRT models via a linear regression for the latent trait. Such latent regressions are now routinely used in large scale educational assessments to produce achievement scores for reporting purposes and for secondary data analysis (e.g., Mislevy, Johnson, and Muraki 1992).

In CDMs, both the response variables (item responses) and the latent variables (skill knowledge indicators) are categorical, so these models can be thought of as latent class models. The presence of multiple skills makes CDMs ‘multiple classification’ latent class models (Maris 1999) similar to the ‘latent class factor models’ by Magidson and Vermunt (2001). Several different types of ‘structural’ models have been proposed for the joint probability distribution of skill mastery across the set of skills. Rupp, Templin, and Henson (2010, Chapter 8) classify the structural models into unstructured, log-linear (Maris 1999; Xu and von Davier 2008), unstructured tetrachoric (Hartz 2002), and structured tetrachoric (Templin 2004) models. The structured tetrachoric models include what de la Torre and Douglas (2004) and Templin, Henson, Templin, and Roussos (2008) call a ‘higher order latent trait model’ where the latent skill knowledge indicators are modeled using a common factor model or item response model. Although the possibility of including covariates in the structural model is mentioned by Rupp, Templin, and Henson (2010, Chapter 8), the only reference given is Templin’s (2004) Ph.D. thesis. Von Davier (2010) specifies hierarchical mixtures of CDMs where the skill distributions can differ across latent or observed groups, and where latent group membership can depend on a clustering variable such as schools, as in Vermunt (2003). If groups are observed, the model can be viewed as a multiple group model. The clustering variable can also be viewed as a covariate that has either fixed or random effects on latent group membership. We are not aware of any other work incorporating covariates in CDMs.

Models that are somewhat related to CDMs are mixture IRT models, where the latent trait is continuous, but examinees consist of latent groups who differ in their ability distributions and possibly in their measurement models. Smit, Kelderman, and van der Flier (1999, 2000) investigate the use of covariate information in such models and show that it improves latent class prediction and reduces standard errors. Cho and Cohen (2010) extend these models to multilevel mixture IRT models with both student and school latent classes where the probability of latent class membership can depend on student and school-level covariates respectively.

In this paper, we extend the deterministic inputs, noisy "and" gate (DINA) model (Haertel 1989; Macready and Dayton 1977; Junker and Sijtsma 2001) by incorporating covariates. We begin by presenting a brief overview of the cognitive diagnosis framework and the DINA model in Section 2. In Section 3 we propose a framework to include student-specific covariates in the model followed by estimation methods and skill diagnosis in Section 4. We then explore the performance of these models using a series of simulation studies in Section 5. In Section 6 we use data from the ASSISTment Tutor (Heffernan, Koedinger, and Junker 2001) for a real data application and compare the diagnostic accuracy of DINA models with and without covariates in Section 7. Finally, in Section 8 we offer conclusions and thoughts on future work.

2. The DINA Model

Cognitive diagnosis models (CDMs) can be either compensatory or non-compensatory. Non-compensatory, or conjunctive, models assume that a student must have mastered all of the skills that an item requires in order to answer it correctly. Compensatory models allow for the mastery of some skills to compensate for those that may be lacking. Disjunctive models are the extreme case of compensatory models and only require one or more of the skills to be mastered for a student to answer correctly. In this paper, we consider a popular conjunctive CDM, the deterministic inputs, noisy "and" gate (DINA) model introduced by Haertel (1989) and Macready and Dayton (1977) and further developed by Junker and Sijtsma (2001) who coined the name DINA. More information about CDMs can be found in Rupp, Templin, and Henson (2010).

Student responses are assembled in a $N \times J$ response matrix Y where y_{ij} indicates whether student i answered item j correctly, N is the total number of students, and J is the total number of items. To estimate student skill knowledge, we need to know which skills are required by each item. To do this, we assemble the skill dependencies of each item into a Q-matrix (Tatsuoka 1983; Embretson 1984; Barnes 2003). The Q-matrix, also referred to as a transfer model or skill coding, is a $J \times K$ matrix where $q_{jk} = 1$ if item j requires skill k and 0 if it does not and K is the total number of skills. The Q-matrix is usually an expert-elicited assignment matrix. The work in this paper assumes that the Q-matrix is known and correct. It should be noted that misspecification of the Q-matrix can lead to incorrect estimates of both student and item parameters (Rupp and Templin 2007).

We use α_{ik} to denote student i's underlying knowledge of skill k, where $\alpha_{ik} = 1$ if student i has mastered skill k and 0 if they have not. Then,

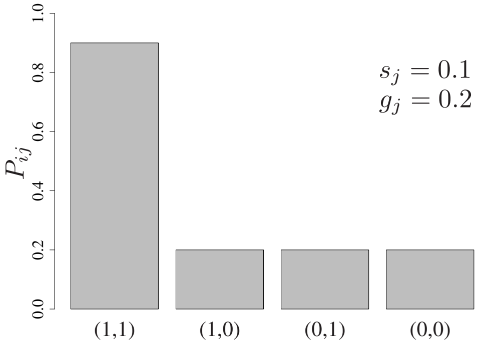

Figure 1. Probability of a correct response under the DINA model for fixed slip and guess parameters for different skill set profiles $(\alpha_{i1}, \alpha_{i2})$.

a student's true underlying skill set profile $\boldsymbol{\alpha}_i$ is a binary vector of length K that indicates whether or not the student has mastered each of the K skills. The probability of a correct response depends on the individual skills via

$$\xi_{ij} = \prod_{k=1}^{K} \alpha_{ik}^{q_{jk}}. \tag{1}$$

If student i has mastered all skills required by item j, $\xi_{ij} = 1$; if the student has not mastered all of the skills, $\xi_{ij} = 0$. Note that ξ_{ij} has also been represented as η_{ij} in the literature. Each item j is characterized by two parameters - the slip and the guess. The slip parameter, $s_j = P(Y_{ij} = 0 \mid \xi_{ij} = 1)$, is the probability of a student answering item j incorrectly even if (s)he has mastered all required skills. The guess parameter, $g_j = P(Y_{ij} = 1 \mid \xi_{ij} = 0)$, is the probability of a student answering item j correctly even if (s)he has not mastered all the required skills. The slip and guess parameters may vary across the items, including those requiring the same skill(s), but for a given item they are the same for all students.

Student responses y_{ij} are modeled as

$$P(Y_{ij} = 1 \mid \boldsymbol{\alpha}_i, s_j, g_j) = (1 - s_j)^{\xi_{ij}} g_j^{1-\xi_{ij}}. \tag{2}$$

The probability of answering correctly depends on whether or not the student has mastered all of the required skills. If student 1 is missing only one skill needed for item j and student 2 is missing two skills needed for item j, they both have the same probability g_j of answering correctly. This idea is represented in the bar-plot in Figure 1, adapted from de la Torre and Chiu (2009). Suppose we have an item that requires two skills, the slip parameter is $s_j = 0.1$, and the guess parameter is $g_j = 0.2$. The y-axis is P_{ij}, the probability that student i will answer item j correctly, and the x-axis shows the student skill set profiles (α_{i1}, α_{i2}). For example, the profile $(0, 1)$ indicates that the student has not mastered Skill 1 but has mastered Skill 2. We can see that only the students in the $(1, 1)$ skill set profile have a high probability of answering the item correctly ($1 - s_j = 0.9$). Students in the other three skill set profiles have a low probability of answering correctly ($g_j = 0.2$) as they have not mastered all of the skills required by the item. Note that students in the $(1, 0)$, $(0, 1)$, and $(0, 0)$ skill set profiles have the same (low) probability of answering correctly since none of them have mastered all of the skills required by the item.

3. Extending the DINA Model to Include Covariates

3.1 Including Covariates in the Structural Part of the Model

In Section 2, we described the measurement part of the model, the relationship between item responses and latent skill knowledge indicators. We now turn to the structural part of the model for the latent skill knowledge indicators. The simplest model assumes independence,

$$\alpha_{ik} \sim \text{Bernoulli}(p_{ik}), \tag{3}$$

where p_{ik} is the probability that student i has mastered skill k. Usually, p_{ik} is allowed to vary across skills but is forced to be the same for all students (i.e., $p_{ik} = p_k$).

In this paper, we extend the structural model by adding student-specific covariates to the model. Explicitly, we model the p_{ik} using logistic regression,

$$\text{logit}(p_{ik}) = \mathbf{x}_i' \boldsymbol{\beta} - \delta_k, \tag{4}$$

where $\mathbf{x}_i$ is a vector of covariates for student i, $\boldsymbol{\beta}$ is a vector of coefficients, and δ_k is the baseline difficulty of skill k. The use of a negative sign allows δ_k to be interpreted as the skill difficulty. This model is similar to the multinomial logit structural model for latent class membership typically used in exploratory latent class models (Dayton and Macready 1988) which becomes a binary logistic regression when there are two latent classes.

The model in (4) assumes that the student covariates have the same effect on all skills (i.e., $\boldsymbol{\beta}_k = \boldsymbol{\beta}$). Such a model may be meaningful if the mastery of all skills is assumed to depend on some overall ability, and the covariates affect the mastery of the individual skills only through the overall ability. To relax this assumption, we could include interactions between skills and covariates

$$\text{logit}(p_{ik}) = \mathbf{x}_i' \boldsymbol{\beta} + \mathbf{z}_i' \gamma_k - \delta_k, \tag{5}$$

where $\mathbf{z}_i$ are covariates with skill-specific coefficients $\boldsymbol{\gamma}_k$, whereas $\mathbf{x}_i$ are covariates with identical coefficients across skills. Thinking of the skills as analogous to items in IRT, models with skill-specific coefficients $\boldsymbol{\gamma}_k$ are analogous to models with differential item functioning (e.g., Muthén and Lehman 1985; Thissen, Steinberg, and Wainer 1988). Since it is the skill indicators, not item responses, whose probabilities can differ between groups, Li and Cohen (2006) refer to this idea as differential skill functioning.

We can also think of the skills as analogous to different traits or dimensions in multidimensional IRT. In multidimensional IRT with latent regressions, the usual approach used for National Assessment of Educational Progress (NAEP) and Program for International Student Assessment (PISA) is to allow regression coefficients to be different for each dimension (e.g., Mislevy, Johnson, and Muraki 1992), corresponding to having no $\mathbf{x}_i'$ in Equation 5.

A variety of structural models have been proposed to relax the independence assumption in (3) (see Rupp, Templin, and Henson 2010, Chapter 8, for a review). In the unstructured model, the probabilities for all possible skill set profiles are free parameters. In log-linear models, the log-frequencies of the skill set profiles are modeled in terms of main effects, two-way interactions, and possibly higher order interactions (Maris 1999; Xu and von Davier 2008). Hartz (2002) specifies a multivariate probit model in which a continuous latent response underlies each skill such that the skill is mastered if the latent response exceeds zero and not mastered if the latent response is less than zero. A multivariate normal distribution is specified for the latent responses whose unstructured correlation matrix, the matrix of tetrachoric correlations, expresses the dependence among the skills. Templin (2004) structures the tetrachoric correlation matrix using a common factor model. Similarly, de la Torre and Douglas (2004) specify an item response model for the skill knowledge indicators. While this paper mostly considers the simplest case of conditionally independent skill knowledge indicators given the covariates, it would be relatively straightforward to include covariates in any of these alternative structural models. Templin (2004) considers this inclusion for structured tetrachoric correlations. Von Davier (2010) discusses a multiple group version of the log-linear structural model with main effects and two-way interactions, but treats the groups as latent in his application.

We fit a model with covariates and a random intercept to the ASSISTment data in Section 6. The model is a higher-order DINA model as considered by de la Torre and Douglas (2004), but with a latent regression for the higher order trait and can be written as

$$\text{logit}(p_{ik}) = \mathbf{x}_i'\boldsymbol{\beta} + \zeta_i + \mathbf{z}_i'\boldsymbol{\gamma}_k - \delta_k, \quad \zeta_i \sim N(0, \psi), \tag{6}$$

where ζ_i is a random intercept. The model can alternatively be written by replacing the first two terms by a continuous latent trait θ_i and specifying the latent regression $\theta_i = \mathbf{x}_i'\boldsymbol{\beta} + \zeta_i$.

3.2 Including Covariates in the Measurement Part of the Model

The models described in Section 3.1 assume that covariates affect the item responses only indirectly through the probabilities of skill mastery. Conditional on skill mastery, the covariates have no direct effects on the item responses. This assumption of *measurement invariance* can be relaxed by allowing the slip and guess parameters to depend on covariates.

In a conference paper, Li and Cohen (2006) define DIF as group differences in slip and guess parameters in a higher order DINA model while allowing for group differences in the mean of the higher order trait. Their model-based approach to DIF detection is analogous to the model-based approach in traditional item response theory used by Muthén and Lehman (1985) and Thissen, Steinberg, and Wainer (1988). In a masters thesis, Bozard (2010) uses a model-based approach to investigate DIF using a linear model for the item responses with main effects and two-way interactions between skill indicators. In this approach, subsets of parameters are allowed to differ between groups.

In a Ph.D. thesis, Zhang (2006) uses predicted skill set profiles as matching criterion in a Mantel-Haenszel test for DIF. However, the DINA model used to predict the profiles does not allow the mastery probabilities to depend on group membership. Generally, it is important to allow for group differences in the latent proficiency distribution when investigating DIF (e.g., Millsap and Everson 1993; Meredith 1993). Otherwise, what is interpreted as DIF may actually be due to group differences in proficiency. We therefore suggest modifying Zhang's procedure by including a dummy variable for group (and possibly other covariates) in the structural model as discussed in Section 3.1. We apply this method to the ASSISTment data in Section 6.

4. Model Estimation and Skill Diagnosis

In likelihood-based inference, the skill knowledge indicators can be conceptualized as latent variables, and it is natural to integrate them out yielding a marginal likelihood. This likelihood can be maximized using an EM algorithm, treating the skill knowledge indicators as missing data. For a complete description of the EM algorithm in the CDM framework, see de la Torre (2009). Templin, Henson, and Douglas (2007) describe macros to estimate the models in Mplus (Muthén and Muthén 2010). Alternatively, priors can be specified for all model parameters yielding a hierarchical Bayesian model in which the regression parameters ($\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$) and the skill difficulties ($\boldsymbol{\delta}$) can be viewed as hyperparameters. The model can then be estimated by Markov chain Monte Carlo, the approach taken here.

4.1 Markov Chain Monte Carlo

We use Markov chain Monte Carlo (MCMC) estimation as implemented in WinBUGS (Spiegelhalter, Thomas, and Best 2003; see Appendix B for sample code). WinBUGS is run from R (R Development Core Team, 2004) and most of the subsequent analyses are completed in R. For all regression coefficients $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ and difficulty parameters $\boldsymbol{\delta}$, vague Normal(0, σ^2=10) priors are used. The conditional distribution of α_{ik}, given $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, and $\boldsymbol{\delta}$, is then Bernoulli(p_{ik}) where the probability that student i has mastered skill k is a logistic regression model as described in Section 3.1. The priors on the slip parameters s_j are Uniform(0,1). To ensure that students who have mastered all the skills required by an item have a higher probability of a correct response than those who have not mastered all of the skills, we constrain $1 - s_j > g_j$ (Junker and Sijtsma 2001). To satisfy this constraint, the prior on g_j is Uniform$(0, 1 - s_j)$.

For each model considered in this paper, a three chain MCMC is run. Each chain has a burn-in of 1000 iterations and is run for an additional 5000 iterations. After thinning, 334 iterations from each chain are used giving a total of 1002 draws from the posterior for estimating the posterior means and standard deviations. The chains are compared to assure that the chains converged to the same values. In the simulation studies (Section 5), the first estimation for each model is analyzed for convergence; no problems are found.

4.2 Skill Diagnosis

A fundamental goal of cognitive diagnosis modeling is the prediction of the underlying skill set profile. Many prediction methods, such as the posterior mean (or empirical Bayes if the model is estimated by maximum likelihood) yield a continuous prediction $\widehat{\alpha}_{ik} \in [0, 1]$. However, it is common to round the $\widehat{\alpha}_{ik}$ to 0/1 and make inferences based on these hard codings (Chiu 2008; Chiu, Douglas, and Li 2009). If $\widehat{\alpha}_{ik}$ is the posterior mean, the rounded prediction corresponds to the posterior mode, the most common prediction method in latent class modeling. Alternatively, the $\widehat{\boldsymbol{\alpha}}_i$ may be clustered (Chiu 2008; Ayers, Nugent, and Dean 2009) and clusters of students assigned to the closest profile using a decision rule (such as rounding).

Because the classification of students into skill set profiles relies on the individual skill knowledge parameter estimates, any method that improves the prediction of the α_{ik}, is also likely to improve the classification of students into skill set profiles.

5. Simulation Studies

5.1 Simulation Conditions and Models Estimated

To explore the use of covariates and evaluate parameter recovery, three simulation studies were run. We used one binary covariate, X_1, sampled from Bernoulli(0.5), and one continuous covariate, X_2, sampled from Uniform(0,1). For concreteness of the example, let X_1 represent a dummy variable for gender (Male; 1 = male, 0 = female) and X_2 be the score on a pre-test (Pre). To generate the latent skill set profiles, α_{ik} is drawn from a Bernoulli(p_{ik}) where p_{ik} is calculated from models 1-4 presented below. The true latent skill set profiles used for simulation studies 1 and 2 are the same; new skill set profiles are generated for simulation study 3. In the first simulation study, we generate response data for $N = 1000$ students, $J = 30$ items, and $K = 2$ skills. All items are single skill items, and each skill is required by 15 items. In the second simulation study, we only changed the design of the Q-matrix. In this case, we have 10 single skill items for each skill and 10 items that require both skills. In the third simulation study, we generated data for $N = 1000$ students, $J = 68$ items and $K = 5$ skills. The Q-matrix had three single skill items for each skill, two items for each unique skill pair, triple, and quadruple, and three items that required all five skills. Each skill was therefore required by 34 items.

We considered four different logistic regressions:

Model 1:

$$\text{logit}(p_{ik}) = \beta_{\text{Male}} \cdot \text{Male}[i] - \delta_k \tag{7}$$

Model 2:

$$\text{logit}(p_{ik}) = \beta_{\text{Male}} \cdot \text{Male}[i] + \beta_{\text{Pre}} \cdot \text{Pre}[i] - \delta_k \tag{8}$$

Model 3:

$$\begin{aligned} \text{logit}(p_{ik}) &= \beta_{\text{Male}} \cdot \text{Male}[i] + \beta_{\text{Pre}} \cdot \text{Pre}[i] \\ &+ \beta_{\text{Male}\cdot\text{Pre}} \cdot \text{Male}[i] \cdot \text{Pre}[i] - \delta_k \end{aligned} \tag{9}$$

Model 4:

$$\text{logit}(p_{ik}) = \gamma_{\text{Male}\cdot\text{Skill}k} \cdot \text{Male}[i] + \beta_{\text{Pre}} \cdot \text{Pre}[i] - \delta_k \tag{10}$$

In the first three models, there is no interaction between the skill and the covariates. For these models, we used $\beta_{\text{Male}} = -0.5$, $\beta_{\text{Pre}} = 0.25$. Thus, the probability of mastering each skill is lower for males and higher for those doing better on the pre-test. In the fourth model, a covariate-skill interaction was added. In this model $\gamma_{Male \cdot Skill1} = -0.5$ and $\gamma_{Male \cdot Skill2} = 0.5$, indicating that females have a higher probability of mastering Skill 1 than males, whereas males have a higher probability of mastering Skill 2 than females.

For the $K = 2$ skill examples, the skill difficulties were set to $\delta_1 = -1$ and $\delta_2 = 1.5$. Students have a higher probability of having Skill 1 compared to Skill 2. For the $K = 5$ skill example, the skill difficulties were set to $\boldsymbol{\delta}' = (-1.5, -0.75, 0, 0.75, 1.5)$. Skill 1 has the highest probability of mastery, and Skill 5 has the lowest. The true parameter values for the simulations were given in Tables 1, 4, and 5. The slip and guess parameters were simulated independently from Uniform(0,0.20) distributions. To assess the effect of larger slip and guess parameters on parameter recovery, these parameters were also generated from a Uniform(0.2, 0.4) distribution for Model 2 in the first simulation study.

For each model, 25 response datasets were generated. The slip (s_j) and guess (g_j) parameters, the skill difficulties ($\boldsymbol{\delta}$), and the coefficients of the covariates ($\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$) were held constant across all replications. In each replication, only the true skill set profiles and the responses are randomly drawn. Both the correct logistic regression DINA model and the standard DINA model (i.e., without covariates) are estimated for each generated dataset. For more information about data generation, refer to Appendix A.

5.2 Parameter Recovery

In this section we discuss the estimation of the coefficients for the covariates and the skill difficulties. Accuracy of the predicted skill knowledge indicators is discussed in Section 7.

Table 1 summarizes the results for each model for simulation study 1. For each parameter, the first line shows the mean of the 25 estimated posterior means for β and δ, with the standard deviation over the replications in parentheses. The second line shows the mean of the 25 standard error estimates, with standard deviation over the replications in parentheses. To assess parameter recovery for δ and β, the 95% posterior credible intervals for all replications for each parameter are estimated. As 25 replications are not enough to obtain a clear picture for any individual parameter, we calculate the coverage of the credible intervals for all replications for all parameters for each model. For example, for Model 1 there will be a total of 75 intervals (25 replications times 3 parameters). The proportion of credible intervals that contain their respective true parameter value is given in the

Table 1. Simulation Study 1: Mean (standard deviation) of the posterior mean estimates (first row) and posterior standard deviation estimates (second row) over the replications for models 1-4. The last row gives the proportion of 95% posterior credible intervals (CI) that contained the true parameter value across all replications and all parameters for that model.

	Truth	Model 1	Model 2	Model 3	Model 4
δ_1	-1	-0.951 (0.084)	-0.985 (0.084)	-0.990 (0.143)	-1.006 (0.118)
		0.089 (0.003)	0.128 (0.004)	0.156 (0.005)	0.154 (0.004)
δ_2	1.5	1.530 (0.116)	1.503 (0.107)	1.477 (0.144)	1.448 (0.108)
		0.100 (0.004)	0.135 (0.004)	0.162 (0.004)	0.168 (0.005)
β_{Male}	-0.5	-0.473 (0.132)	-0.504 (0.110)	-0.557 (0.153)	
		0.109 (0.003)	0.108 (0.003)	0.209 (0.005)	
β_{Pre}	0.25		0.241 (0.153)	0.228 (0.286)	0.227 (0.164)
			0.184 (0.005)	0.256 (0.007)	0.238 (0.006)
$\beta_{\text{Male}\cdot\text{Pre}}$	-1.15			-1.002 (0.276)	
				0.369 (0.008)	
$\gamma_{\text{Male}\cdot\text{Skill1}}$	-0.5				-0.528 (0.100)
					0.129 (0.004)
$\gamma_{\text{Male}\cdot\text{Skill2}}$	0.5				0.457 (0.129)
					0.149 (0.004)
95% CI Coverage		0.91	0.97	0.96	0.98

last line of the table. These proportions range from 0.91 to 0.98 across the models and do not differ significantly from 0.95 at the 5% level. Overall, we are satisfied with the recovery of the true parameter values across the various models.

Turning to the slip and guess parameters, recall that these parameters vary across the items j, but were held constant across the $R = 25$ replications. For example, the slip parameter for item 1 is always $s_1 = 0.14$. For a particular item j, the estimated bias and root mean squared error (RMSE) of the slip parameter are obtained as

$$\widehat{\text{bias}} = \frac{1}{R}\sum_{r=1}^{R}(\hat{s}_j^r - s_j) \qquad \text{RMSE} = \sqrt{\frac{1}{R}\sum_{r=1}^{R}(\hat{s}_j^r - s_j)^2}, \tag{11}$$

where $\hat{s}_j^r$ is the estimate of s_j for the r^{th} simulated dataset. Analogous expressions are used for the guess parameter. Table 2 shows the minimum, mean, and maximum of the estimated bias for both slip and guess parameters by model for simulation study 1. Summaries of the RMSE for each model are shown in Table 3. For both the bias and the RMSE, the results are similar across the models. Further analysis did not reveal any discernible relationship between either the estimated bias or RMSE and the true value of the parameter.

Table 2. Summary of estimated slip and guess parameter bias for Simulation Study 1

		Model 1	Model 2	Model 3	Model 4
s_j	min	-0.003	-0.004	-0.008	-0.002
	mean	0.003	0.002	0.003	0.003
	max	0.013	0.012	0.010	0.013
g_j	min	-0.005	-0.003	-0.003	-0.005
	mean	0.001	0.001	0.002	0.002
	max	0.009	0.011	0.011	0.009

Table 3. Summary of slip and guess parameter RMSE for Simulation Study 1

		Model 1	Model 2	Model 3	Model 4
s_j	min	0.006	0.005	0.005	0.004
	mean	0.017	0.016	0.017	0.014
	max	0.035	0.033	0.039	0.029
g_j	min	0.004	0.004	0.005	0.005
	mean	0.010	0.011	0.010	0.011
	max	0.022	0.022	0.017	0.018

Table 4 reports the results for simulation study 2. The proportions of 95% credible intervals that contained the true value ranges from 0.91 to 0.98 across the models. In addition, we can note that both the parameter and standard deviation estimates from the single skill items only (simulation study 1) and both single and multiple skill items (simulation study 2) replications are similar. We might have expected that having fewer single skill items would lead to larger posterior standard deviations. In a conjunctive model, when a student incorrectly answers an item that requires two (or more) skills, the model does not distinguish which skill the student did not know. Thus it would be reasonable to expect estimation of model parameters and skill knowledge indicators to worsen as the number of multiple skill items increases. The results for the bias and RMSE of the slip and guess parameters are also similar to those from simulation study 1 and are not shown.

Table 5 shows the results for simulation study 3. Only 10 replications were completed because the computational time for each estimation was significantly longer than in either simulation study 1 or 2. The proportions of 95% credible intervals that contained the true value ranges from 0.85 to 0.99 across the models. The coverage for model 4 is significantly lower than 0.95 at the 5% level, While the estimated bias and RMSE for the slip and guess parameters is slightly larger than those for simulation studies 1 and 2, there is still no systematic relationship between either the bias or RMSE and the true value of the parameter.

Table 4. Simulation Study 2: Mean (standard deviation) of the posterior mean estimates (first row) and posterior standard deviation estimates (second row) over the replications for models 1-4. The last row gives the proportion of 95% posterior credible intervals (CI) that contained the true parameter value across all replications and all parameters for that model.

	Truth	Model 1	Model 2	Model 3	Model 4
δ_1	-1	-0.952 (0.085)	-0.985 (0.084)	-0.989 (0.143)	-1.002 (0.119)
		0.089 (0.002)	0.128 (0.003)	0.159 (0.004)	0.135 (0.003)
δ_2	1.5	1.529 (0.115)	1.501 (0.107)	1.478 (0.143)	1.448 (0.107)
		0.099 (0.004)	0.134 (0.004)	0.164 (0.004)	0.142 (0.004)
β_{Male}	-0.5	-0.472 (0.132)	-0.505 (0.109)	-0.556 (0.153)	
		0.109 (0.003)	0.109 (0.002)	0.212 (0.006)	
β_{Pre}	0.25		0.240 (0.154)	0.229 (0.285)	0.231 (0.163)
			0.184 (0.004)	0.257(0.006)	0.175 (0.003)
$\beta_{\text{Male}\cdot\text{Pre}}$	-1.15			-1.002 (0.273)	
				0.372 (0.011)	
$\gamma_{\text{Male}\cdot\text{Skill1}}$	-0.5				-0.525 (0.101)
					0.139 (0.004)
$\gamma_{\text{Male}\cdot\text{Skill2}}$	0.5				0.456 (0.128)
					0.147 (0.004)
95% CI Coverage		0.91	0.97	0.96	0.98

The results for Model 2 with the Q-matrix from simulation study 1 but with large slip and guess parameters, simulated from Uniform(0.2, 0.4), are shown in Table 6. Comparing the results with those for Model 2 in Table 1, we see that the sampling standard deviations are considerably larger than would be expected, especially for the skill difficulty parameters, but that the mean estimates are still close to the generating parameters. The slip and guess bias and RMSE are slightly larger than those shown for Model 2 in Tables 2 and 3. Similar to the other simulation studies, there is no systematic relationship between either the bias or RMSE and the true value of the parameter. The effect on recovery of the skill set profiles will be discussed along with the diagnostic accuracy results of the other simulations in Section 7.

A standard DINA model (without covariates) was also estimated for each of the replications in the simulation studies above. The slip and guess bias and RMSE are similar to the logistic regression DINA model results presented in Tables 2 and 3. In the logistic regression DINA model, the skill difficulties δ_k represent minus the conditional log-odds of having the skills, given that the covariates are zero, and are hence not comparable with the marginal probabilities of skill mastery, p_k, from the standard DINA model. However, we can note that the models place the skills in the same order of difficulty. The recovery of the skill knowledge indicators is summarized in

Table 5. Simulation Study 3: Mean (standard deviation) of the posterior mean estimates (first row) and posterior standard deviation estimates (second row) over the replications for models 1-4. The last row gives the proportion of 95% posterior credible intervals (CI) that contained the true parameter value across all replications and all parameters for that model.

	Truth	Model 1	Model 2	Model 3	Model 4
δ_1	-1.5	-1.509 (0.076)	-1.469 (0.079)	-1.490 (0.123)	-1.428 (0.191)
		0.090 (0.002)	0.106 (0.004)	0.120 (0.002)	0.133 (0.009)
δ_2	-0.75	-0.763 (0.077)	-0.742 (0.069)	-0.689 (0.143)	-0.732 (0.114)
		0.074 (0.002)	0.094 (0.002)	0.109 (0.002)	0.112 (0.006)
δ_3	0	-0.005 (0.081)	-0.005 (0.106)	0.057 (0.123)	-0.010 (0.134)
		0.073 (0.002)	0.091 (0.002)	0.109 (0.002)	0.107 (0.002)
δ_4	0.75	0.755 (0.090)	0.726 (0.086)	0.821 (0.161)	0.735 (0.111)
		0.078 (0.003)	0.096 (0.002)	0.112 (0.002)	0.111 (0.002)
δ_5	1.5	1.511 (0.089)	1.496 (0.137)	1.607 (0.130)	1.526 (0.117)
		0.094 (0.004)	0.107 (0.003)	0.128 (0.004)	0.127 (0.006)
β_{Male}	-0.5	-0.512 (0.084)	-0.507 (0.050)	-0.442 (0.123)	
		0.066 (0.001)	0.065 (0.002)	0.130 (0.002)	
β_{Pre}	0.25		0.268 (0.105)	0.387 (0.198)	0.213 (0.170)
			0.113 (0.002)	0.158 (0.002)	0.168 (0.003)
$\beta_{\text{Male}\cdot\text{Pre}}$	-1.15			-1.335 (0.235)	
				0.234 (0.005)	
$\gamma_{\text{Male}\cdot\text{Skill1}}$	-0.5				-0.402 (0.146)
					0.162 (0.006)
$\gamma_{\text{Male}\cdot\text{Skill2}}$	0.5				0.544 (0.164)
					0.148 (0.003)
$\gamma_{\text{Male}\cdot\text{Skill3}}$	0.5				0.487 (0.133)
					0.133 (0.002)
$\gamma_{\text{Male}\cdot\text{Skill4}}$	0.25				0.247 (0.118)
					0.132 (0.003)
$\gamma_{\text{Male}\cdot\text{Skill5}}$	-0.75				-0.742 (0.168)
					0.185 (0.009)
95% CI Coverage		0.97	0.99	0.90	0.85

Section 7 along with the classification rates from the logistic regression model.

6. ASSISTments: An Online Tutor

6.1 Data and Models

In this section we analyze a subset of the ASSISTment tutoring data (Heffernan, Koedinger, and Junker 2001). The ASSISTment tutor is used by students in Worcester, Massachusetts to prepare for the mathematics portion of their end-of-year assessment exams. A similar subset of the ASSISTment data was previously analyzed using a Rasch model in Ayers and Junker (2008). In this subset of data from the spring 2004 semester, there are

Table 6. Simulation Study - Large Slip and Guess: Mean (standard deviation) of the posterior mean estimates (first row) and posterior standard deviation estimates (second row) over the replications for Model 2 using a Uniform(0.2, 0.4) for both the slip and guess parameters. The last row gives the proportion of 95% posterior credible intervals (CI) that contained the true parameter value across all replications and all parameters for that model.

	Truth	Model 2
δ_1	-1	-1.033 (0.165)
		0.154 (0.003)
δ_2	1.5	1.503 (0.166)
		0.157 (0.007)
β_{Male}	-0.5	-0.541 (0.119)
		0.123 (0.003)
β_{Pre}	0.25	0.299 (0.179)
		0.209 (0.005)
95% CI Coverage		0.95

$N = 435$ 8^{th} grade students, $J = 26$ items, and $K = 3$ skills. The skills used are Evaluating Functions, Multiplication, and Unit Conversion and occur in 8, 20, and 2 items respectively. In the dataset, 55% of the student responses are missing and here are treated as missing at random (Rubin, 1987).

At the beginning of the school year, all students were given a pre-test designed by the researchers to assess pre-tutor knowledge. In addition, several student demographics were collected. For covariates, we use a dummy variable for being male, a dummy variable for free lunch, and a scaled pre-test score. 49.9% of the sample are male, and 56.6% of the sample are entitled to a free lunch. The raw pre-test sum-score is scaled to range from 0 to 1 resulting in a mean of 0.376 and a standard deviation of 0.194.

Three models are estimated: a logistic regression DINA model, a higher-order DINA model with a latent regression (see equation 6), referred to here as ‘latent regression DINA model’, and a standard DINA model.

6.2 Estimates of Structural Model

Table 7 shows the results of the logistic and latent regression DINA models. For the logistic regression DINA model, the estimated coefficient for gender is -0.269 with a posterior standard deviation of 0.351, indicating no evidence for a gender difference, controlling for free lunch and pre-test. The estimated coefficient for free lunch is -0.823 with a posterior standard deviation of 0.378. Controlling for gender and pre-test score, students who receive free lunch have a lower probability of skill mastery. Controlling for gender and free lunch, a standard deviation (0.194) increase in the pre-test is

Table 7. Bayesian estimates for ASSISTment data using logistic and latent regression DINA models.

		Logistic Regression		Latent Regression	
Parameter	Variable	Estimates	SD	Estimates	SD
δ_1	Evaluating Functions	1.076	0.672	1.308	0.822
δ_2	Multiplication	0.754	0.815	0.649	0.799
δ_3	Unit Conversion	-2.142	2.251	-2.059	2.320
β_1	Male	-0.269	0.351	-0.481	0.469
β_2	Free Lunch	-0.823	0.378	-0.929	0.478
β_3	Pre-test	8.292	1.494	9.712	1.871
$\sqrt{\psi_\zeta}$	SD of Random Intercept	NA	NA	1.318	0.409

associated with a 5.0-fold increase in the odds of mastering the skills, with a 95% credible interval (2.8, 8.8).

For the latent regression DINA model, the estimated skill difficulties (the δ's) and the coefficients of the covariates (the β's) are similar to those for the logistic regression DINA model. The random intercept standard deviation is estimated as 1.38. Following Larsen, Petersen, Budtz-Jørgensen, and Endahl (2000), we can interpret the estimate by considering two randomly chosen students with identical covariate values. The median of the odds ratios, comparing the student with the larger random intercept to the other student is estimated as 3.52.

For the standard DINA model, the α_{ik} are modeled using a Bernoulli(p_k) distribution, where p_k is the probability of having mastered skill k. The estimated values of p_k are 0.67, 0.99, and 0.55 for the three skills respectively. We note that the standard DINA model puts the skills in a different order of difficulty from the logistic and latent regression DINA models. In the logistic and latent regression models, Evaluating Functions has the highest δ_k value and would be considered the hardest skill, and in the standard DINA model, Unit Conversion has the lowest probability of mastery and would be considered the hardest skill.

6.3 Estimates of Slip and Guess Parameters

For the logistic regression DINA model, the estimated slip parameters range from 0.003 to 0.71 and the estimated guess parameters range from 0.06 to 0.98. Item 7 involves only multiplication and was answered correctly by all students. This item has an estimated slip parameter of 0.003 and an estimated guess parameter of 0.98 and could be removed from the analysis. In fact, there are nine items, all requiring only multiplication, with an estimated guess parameter over 0.90. Omitting these nine items, the estimated guess parameters range from 0.06 to 0.71.

For both the slip and guess, the correlation of the estimates between the logistic and latent regression DINA models is 0.99. Comparing the standard DINA and logistic regression DINA models, the correlation of the estimates is 0.97 for the slip parameters and 0.91 for the guess parameters.

6.4 Skill Diagnosis

When looking at the skill set profile predictions, each of the models assigns students to four of the eight possible skill set profiles. The logistic and latent regression DINA models yield the same four skill set profiles, but do not agree perfectly in their assignment of students to profiles. However, the standard DINA model yields a different subset of skill set profiles. The frequency distributions of the estimated skill set profiles for the three models are given in Table 8.

All three models assign students to the $(0, 1, 1)$ and $(1, 1, 1)$ skill set profiles. However, the frequencies differ between the models. To further explore the differences, we look at the covariate distributions for groups of students assigned to different skill set profiles. For simplicity, Table 9 distinguishes only between students assigned to $(1, 1, 1)$ and students not assigned to this skill set profile by the three models. For the logistic and latent regression DINA models, the proportion of students assigned to the $(1, 1, 1)$ skill set profile who are male is lower than among students not assigned to the $(1, 1, 1)$ skill set profile ($\chi^2 = 4.26$, df=1, p=0.04; $\chi^2 = 5.54$, df=1, p=0.02). In addition, among students in the $(1, 1, 1)$ skill set profile, the proportion who qualified for free lunch is lower than among those not assigned to the $(1, 1, 1)$ skill set profile ($\chi^2 = 61.47$, df=1, $p < 0.001$; $\chi^2 = 66.37$, df=1, $p < 0.001$). Finally, in the logistic and latent regression DINA models, the mean scaled pre-test score is significantly higher for students placed in the $(1, 1, 1)$ skill set profile, compared to those who are not (t=18.57, df=407, $p < 0.001$; t=18.01, df=407, $p < 0.001$). No significant differences were found at the 5% level for any of the covariates under the standard DINA model.

For further comparison across the models, we can use the Adjusted Rand Index to compare the skill set profile assignments between the logistic regression, latent regression, and standard DINA models. The Adjusted Rand Index (ARI; Hubert and Arabie 1985) is a common measure of agreement between two partitions that adjusts for chance agreements. Under random partitioning, the expected value of the ARI is zero. The maximum value is one, with larger values indicating better agreement. The ARI between the logistic and latent regression DINA models is 0.90, indicating a

Table 8. Skill set profile assignments for logistic regression, latent regression, and standard DINA models using the ASSISTment dataset

Skill set profile	Logistic	Latent	standard DINA
(0,0,0)	0	0	0
(0,0,1)	100	107	0
(0,1,0)	0	0	6
(0,1,1)	162	69	37
(1,0,0)	0	0	0
(1,0,1)	38	23	0
(1,1,0)	0	0	33
(1,1,1)	235	236	359

Table 9. Distribution of covariates across the predicted skill set profiles

Predicted	Standard DINA		Logistic DINA		Latent DINA	
Skill Set Profile	(1,1,1)	not (1,1,1)	(1,1,1)	not (1,1,1)	(1,1,1)	not (1,1,1)
Number of students	359	76	235	200	236	199
Percent Male	50.6	47.3	45.1	55.5	44.3	56.6
Percent Free Lunch	54.7	64.2	39.1	77.0	38.0	78.8
Mean (sd) of Scaled	.374	.386	.494	.239	.491	.241
Pre-test Score	(.193)	(.203)	(.172)	(.112)	(.175)	(.113)

high level of agreement between the methods. The ARI between the logistic regression and the standard DINA models is 0.05 and the ARI between the latent regression and the standard DINA model is 0.06. These are quite low and indicate poor agreement between either the logistic or latent regression DINA model and the standard DINA model.

If the true skill set profiles are dependent on the student covariates, the standard DINA model lacks the ability to capture this information and the associations are not found for the predicted profiles. The reason for the latter could be because of the relatively small sample size, the missing data, and small number of items requiring some of the skills (8, 20, and 2 items per skill) leading to imprecise prediction. In Section 7, we explore the diagnostic accuracy of the logistic regression DINA model for data simulated from the logistic regression DINA model with parameters equal to the estimates in Table 7 and having the same Q-matrix, number of students, and missing data.

6.5 Assessment of Differential Item Functioning

Following the work of Zhang (2006), we perform Mantel-Haenszel tests for differential item functioning (DIF) with predicted skill set profiles as matching criterion (using Stata 11, StataCorp 2009). We use predicted skill set profiles from the logistic and latent regression and standard DINA model. The null hypothesis for the test is that the item does not exhibit DIF.

For free lunch, five items show significant DIF (at the 5% level) when skill set profile predictions from the standard DINA model are used. For four of these items, students receiving free lunch have a significantly lower odds of responding correctly within strata defined by the skill set profiles. For these items, the logistic and latent regression DINA models found no DIF. It may be that the skill set profiles for these models already reflect the lower odds of having the skills for free lunch students. For the item where free lunch students have a greater odds of responding correctly within strata defined by the standard DINA skill set profiles, the pooled odds ratio based on logistic and latent regression DINA skill set profiles is larger than that based on the standard DINA model as would be expected. These results suggest that it is important to allow for covariate effects in skill prediction when investigating DIF. For gender, which has a much smaller regression coefficient, the results are less clear with two items exhibiting DIF in opposite directions according to the logistic and latent regression DINA models and no item exhibiting DIF according to the standard DINA model.

7. Diagnostic Accuracy

In this section, we explore the prediction of the student skill knowledge indicators $\boldsymbol{\alpha}_i$ for the datasets simulated in Section 5. In addition, we simulate skill set profiles and responses from the logistic regression DINA model estimated in Section 6 to allow for assessment of diagnostic accuracy for unbalanced Q-matrix designs with few items requiring some of the skills, smaller samples of students, and some large slip and guess parameters. Since we generate the students' true skill set profiles, we can compare the predictions based on both the logistic DINA model and the standard DINA model to the truth.

Tables 10 and 11 show the mean ARI values with their standard deviations over the 25 replications for the logistic regression and standard DINA models from simulation studies 1 to 3 described in Section 5. The first line of the header row indicates the number of skills K, and the second line indicates the design of the Q-matrix. In the *Single* design, only single skill items occur in the Q-matrix and in the *Single, Multiple* design both single and multiple skill items occur in the Q-matrix. The ARI is 1 for all 25 replications under the $K = 2$ *Single* design (simulation study 1). This is not surprising since each skill has 15 occurrences in single skill items; we expect this high number of single skill observations to provide good information. Under the $K = 2$ *Single, Multiple* design (simulation study 2), the

Table 10. Mean ARI values with standard deviation over the 25 replications for logistic regression DINA model for simulations 1 to 3 described in Section 5.

Number of Skills	$K = 2$	$K = 2$	$K = 5$
Q-matrix design	*Single*	*Single, Multiple*	*Single, Multiple*
Model 1	1 (0)	0.9998 (0.0007)	0.9230 (0.0137)
Model 2	1 (0)	0.9998 (0.0007)	0.9278 (0.0099)
Model 3	1 (0)	0.9998 (0.0004)	0.8709 (0.0108)
Model 4	1 (0)	0.9999 (0.0003)	0.9646 (0.0087)

Table 11. Mean ARI values with standard error over the 25 replications for standard DINA model for simulations 1 to 3 described in Section 5.

Number of Skills	$K = 2$	$K = 2$	$K = 5$
Q-matrix design	*Single*	*Single, Multiple*	*Single, Multiple*
Model 1	1 (0)	0.9994 (0.0013)	0.9229 (0.0121)
Model 2	1 (0)	0.9991 (0.0018)	0.9294 (0.0124)
Model 3	1 (0)	0.9993 (0.0017)	0.8686 (0.0105)
Model 4	1 (0)	0.9995 (0.0012)	0.9663 (0.0076)

mean ARI is 0.9998 or 0.9999 for all four models under the logistic regression DINA model and only slightly lower (between 0.9991 and 0.9995) for the standard DINA model. In the $K = 5$ example (simulation study 3), the mean ARI ranges from 0.8686 to 0.9663 for the four models. We note the lower ARI values and larger standard deviations in this example. These findings are similar to previous results of Ayers, Nugent, and Dean (2009), who found that the ARIs using the (standard) DINA model drop as the ratio of the number of single skill items decreases with respect to the number of multiple skill items in the Q-matrix.

Recall that an additional 25 replications with large slip and guess values were run using Model 2 in simulation study 1. The mean ARI values are 0.759 and 0.756 for the logistic and standard DINA models, respectively, both with standard deviations of 0.023. The ARI values are therefore lower due to the larger slip and guess values, but somewhat surprisingly, the covariates do not improve diagnosis appreciably.

Table 12 shows the results of the parameter estimation for the simulation study based on the ASSISTment data that uses the values in Table 7 as the true generating values for the skill difficulties and covariate coefficients. The Q-matrix described in Section 6 was used with 8, 20, and 2 items requiring Skills 1, 2, and 3, respectively. Recall that the slip parameters range from 0.003 to 0.70 and the guess parameters range from 0.06 to 0.98. We generated the true skill set profiles and a student response matrix with no missing values. The logistic regression and standard DINA models were estimated using these datasets with no missing data. The results appear in

Table 12. ASSISTment simulation study: Mean (standard deviation) of the posterior mean estimates (first row) and posterior standard deviation estimates (second row) over the replications. The last row gives the proportion of 95% postertior credible intervals (CI) that contained the true parameter value across all replications and all parameters for that model.

	Truth	NO missing data	Missing data
δ_1	1.076	1.053 (0.296)	0.849 (0.665)
		0.345 (0.019)	0.917 (0.194)
δ_2	0.754	0.620 (0.288)	-0.750 (1.475)
		0.337 (0.017)	1.049 (0.684)
δ_3	-2.142	-2.002 (1.229)	-0.445 (1.022)
		1.780 (0.335)	2.534 (0.316)
β_{Male}	-0.269	-0.286 (0.214)	-0.164 (0.338)
		0.208 (0.010)	0.483 (0.216)
$\beta_{\text{FreeLunch}}$	-0.823	-0.794 (0.214)	-0.928 (0.356)
		0.228 (0.011)	0.542 (0.200)
β_{PreTest}	8.292	8.007 (0.817)	7.461 (1.323)
		0.861 (0.063)	1.513 (0.226)
95% CI Coverage		0.953	0.933

Table 13. Mean ARI values and percent agreement for the individual skills with standard deviations over the 25 replications

		NO missing data		Missing data	
		Logistic DINA	standard DINA	Logistic DINA	standard DINA
ARI		0.752 (0.047)	0.680 (0.048)	0.381 (0.077)	0.042 (0.061)
Percent Agreement	Skill 1	0.942 (0.010)	0.927 (0.010)	0.783 (0.025)	0.714 (0.021)
	Skill 2	0.947 (0.010)	0.930 (0.010)	0.831 (0.036)	0.774 (0.025)
	Skill 3	0.954 (0.054)	0.947 (0.025)	0.847 (0.146)	0.773 (0.304)

Column 3 of Table 12. To match the missing data pattern of the ASSISTment data, we removed any response that was missing in the real dataset (approximately 45% of the responses). The logistic regression DINA model and standard DINA model were again estimated, and these results appear in Column 4 of Table 12. We note that the standard deviations of the parameter estimates across the replications are higher than in the previous simulation studies and these larger values are approximately captured by the estimated posterior standard deviations. Note that the dataset with the missing student responses was created by removing responses from the full response matrix; thus the results for the two conditions (no missing data, missing data) are not independent.

Table 13 shows the mean ARI values and percent agreement for each skill with the standard deviation over the replications for both the logistic regression DINA model and the standard DINA model for both the full and reduced datasets. We acknowledge that percent agreement does not correct for chance agreement between the partitions. However, for many replications, the models placed all students in either the mastery or non-mastery category for one or more of the skills, resulting in Cohen's Kappa (Cohen 1960) values of zero even when percent agreement was good. For both the ARI and percent agreement, significantly higher values are achieved using the logistic regression DINA model when compared to the standard DINA model. While the logistic regression DINA model and the standard DINA model perform comparably in the first three simulation studies, we see that the logistic DINA model outperforms the standard DINA model for the simulation based on the ASSISTment data. One possible reason is that, in the latter case, the item responses alone provide insufficient information about the skills due to the sparse Q-matrix and some large slip and guess parameters. In such cases, it would be expected that covariate information can improve predictions considerably.

8. Conclusions and Future Work

We present an extension of the standard DINA model that includes student-specific covariates in the estimation of the student skill knowledge parameters. This approach was motivated by a desire for possible improved prediction of latent skill knowledge as well as better characterization of the students assigned to each skill set profile. Simulation studies show that both the recovery of model parameters and the prediction of student skill knowledge is good. In addition, the estimated regression coefficients can help us understand how skill knowledge is associated with covariates. This is particularly important for investigating the effects of interventions by including treatment dummy variables in the logistic regression.

In the ASSISTment data example, we show that the logistic and latent regression DINA models, unlike the standard DINA model, tend to place students with higher pre-test scores in the skill set profile where all skills are mastered. Because the pre-test was designed to be predictive of skill knowledge, we find this result consistent with expectation.

With respect to comparison of skill set profile recovery between our approach and the standard DINA model, performance is comparable when there are a large number of items per skill and many students with no missing data (the first three simulation studies). However, when data is generated using a Q-matrix design where some skills are measured by few items, and there are missing data, we see that the logistic regression DINA model outperforms the standard DINA model in terms of classification of students to skill set profiles (Table 13). As we expect this final simulation study to mirror real classroom data (both in Q-matrix design and missing data), we believe that the inclusion of covariates will often be valuable for improving classification accuracy.

For the ASSISTment data, we also incorporated covariates within the higher-order DINA model introduced by de la Torre and Douglas (2004). In addition, we assessed differential item functioning using a Mantel-Haenszel test using the predicted skill set profiles from the logistic regression, latent regression, and standard DINA models as matching criterion. We feel that recent developments for detecting DIF in cognitive diagnosis models are a good starting point, but that further work is needed to compare different methods.

The work in this paper focused only on the conjunctive DINA cognitive diagnosis model. Exploring the effect of incorporating student-specific covariates in other CDMs would be of interest. For example, does assuming a compensatory relationship lead to similar results? Also, the DINA model has a slip and guess parameter per item. How does the use of student-specific covariates affect CDMs that use a slip and guess parameter for each skill? Given the success of this framework within the DINA model, we are cautiously optimistic that work with other CDMs will be fruitful.

Appendix A

Below is the R code used to generate the true underlying skill set profiles and the student responses. As inputs, the function requires the number of students `N`, the number of items `J`, the number of skills `K`, the Q-matrix denoted `Q`, the slip and guess parameters denoted `sj` and `gj`, the skill difficulties denoted `delta`, a matrix of student covariates denoted `X`, and the effects of the covariates denoted `beta`. The function returns the response matrix `Y`, the true latent classes `A` used to generate responses, and the matrix `pi.ik` indicating the probability that student i has mastered skill k. While the slip and guess and the Q-matrix are used as inputs, they are also returned and stored with the data generation so that the true values are known for future analysis.

```
Generate = function(N,J,K,Q,sj,gj,delta,X,beta){

  # find the probabilities of student i having skill k
  pi.ik = matrix(-7,nrow=N,ncol=K)
  for(i in 1:N){
    for(k in 1:K){
      pi.ik[i,k] = exp(X[i,] %*% beta - delta[k]) /
        (1 + exp(X[i,] %*% beta - delta[k]) )
    }
  }

  # next generate a student’s true latent class
  A = matrix(0,nrow=N,ncol=K)
  for(i in 1:N){
    for(k in 1:K){
     A[i,k] = rbinom(1,1,pi.ik[i,k])
    }
  }
  #calc. if students have all skills needed for each item

  xi = matrix(0,nrow=N,ncol=J)
  for(i in 1:N){
    for(j in 1:J){
      xi[i,j] = prod(A[i,] ^ Q[j,])
    }
  }

  #generate probability correct and sample responses

  prob.correct = matrix(0,N,J)
  Y = matrix(0,nrow=N,ncol=J)

  for(i in 1:N){
    for(j in 1:J){
      prob.correct[i,j] = ((1-sj[j])^xi[i,j] )
                                  * (gj[j] ^ (1-xi[i,j]) )
      Y[i,j] = rbinom(1,1,prob.correct[i,j])
    }
  }

  return(list(resp=Y,trueLatent=A,pi.ik=pi.ik,sj=sj,
                                              gj=gj,Qmat=Q))
}
```

Appendix B

WinBUGS code for simulation study 1 Model 1 is shown below. Code for the remaining models is similar. `response` is a vector containing all student responses and the vectors `students` and `question` contain the identifiers to link the `response` vector entries to students and questions. An entry of the vector `p` indicates the probability of a correct response on the corresponding entry of `response` vector. The other variables used correspond to those from the generating code in Appendix A. Note that in WinBUGS, the Normal distribution uses $\tau = \frac{1}{\sigma^2}$. Thus the variance in the priors is 10.

```
model {

#response
for (i in 1:nResp){
  response[i] ~ dbern(p[i])
  p[i] <- pow((1-s[question[i]]), xi[students[i],
          question[i]] ) * pow(g[question[i]],
          1-xi[students[i],question[i]])
}

#alphas
for(k in 1:K){
  for(i in 1:N){
   logit(pi.ik[i,k]) <- gender[i] * betaGender - delta[k]
   }
}

#priors for beta and delta
betaGender ~ dnorm(0,.1)

for(k in 1:K){
  delta[k] ~ dnorm(0,.1)
}

for(k in 1:K){
  for(i in 1:N){
    alpha[i,k] ~ dbern(pi.ik[i,k])
  }
}

# eta, one line for each item
for(i in 1:N){
  eta[i, 1 ] <-alpha[i, 1 ]
  eta[i, 2 ] <-alpha[i, 1 ]
  eta[i, 3 ] <-alpha[i, 1 ]
  eta[i, 4 ] <-alpha[i, 1 ]
  eta[i, 5 ] <-alpha[i, 1 ]
  eta[i, 6 ] <-alpha[i, 1 ]
  eta[i, 7 ] <-alpha[i, 1 ]
  eta[i, 8 ] <-alpha[i, 1 ]
  eta[i, 9 ] <-alpha[i, 1 ]
  eta[i, 10 ] <-alpha[i, 1 ]
  eta[i, 11 ] <-alpha[i, 1 ]
  eta[i, 12 ] <-alpha[i, 1 ]
  eta[i, 13 ] <-alpha[i, 1 ]
  eta[i, 14 ] <-alpha[i, 1 ]
  eta[i, 15 ] <-alpha[i, 1 ]
  eta[i, 16 ] <-alpha[i, 2 ]
  eta[i, 17 ] <-alpha[i, 2 ]
  eta[i, 18 ] <-alpha[i, 2 ]
  eta[i, 19 ] <-alpha[i, 2 ]
  eta[i, 20 ] <-alpha[i, 2 ]
  eta[i, 21 ] <-alpha[i, 2 ]
  eta[i, 22 ] <-alpha[i, 2 ]
```

```
    eta[i, 23 ] <-alpha[i, 2 ]
    eta[i, 24 ] <-alpha[i, 2 ]
    eta[i, 25 ] <-alpha[i, 2 ]
    eta[i, 26 ] <-alpha[i, 2 ]
    eta[i, 27 ] <-alpha[i, 2 ]
    eta[i, 28 ] <-alpha[i, 2 ]
    eta[i, 29 ] <-alpha[i, 2 ]
    eta[i, 30 ] <-alpha[i, 2 ]
  }

  # item slip and guess parameters
  for(j in 1:J){
    s[j] ~ dunif(0,1)
    max.gj[j] <- 1 - s[j]
    g[j] ~ dunif(0,max.gj[j])
  }

  }
```

References

AYERS, E., and JUNKER, B.W. (2008), “IRT Modeling of Tutor Performance to Predict End-of-year Exam Scores”, *Educational and Psychological Measurement, 68*, 972–987.

AYERS, E., NUGENT, R., and DEAN, N. (2009), “A Comparison of Student Skill Knowledge Estimates”, in *Educational Data Mining 2009: 2nd International Conference on Educational Data Mining, Proceedings*, eds. T. Barnes, M. Desmarais, C. Romero, and S. Ventura, Cordoba, Spain, pp.1-10, http://www.educationaldatamining.org/EDM2009/uploads/proceedings/ayers.pdf.

BARNES, T.M. (2003), “The Q-matrix Method of Fault-Tolerant Teaching in Knowledge Assessment and Data Mining”, Ph.D. Thesis, Department of Computer Science, North Carolina State University, NC.

BOZARD, J. (2010), “Invariance Testing in Diagnostic Classification Models”, Masters Thesis, The University of Georgia, Athens, GA.

CHIU, C. (2008), “Cluster Analysis for Cognitive Diagnosis: Theory and Applications”, Ph.D. Thesis, Educational Psychology, University of Illinois, Urbana Champaign, IL.

CHIU, C., DOUGLAS, J., and LI, X. (2009), “Cluster Analysis for Cognitive Diagnosis: Theory and Applications”, *Psychometrika, 74*, 633–665.

CHO, S-J., and COHEN, A.S. (2010), “A Multilevel Mixture IRT Model with an Application to DIF”, *Journal of Educational and Behavioral Statistics, 35*, 336–370.

CLOGG, C.C., and GOODMAN, L.A. (1984), “Latent Structure Analysis of a Set of Multidimensional Contingency Tables”, *Journal of the American Statistical Association, 79*, 762–771.

COHEN, J. (1960), “A Coefficient of Agreement for Nominal Scales”, *Educational and Psychological Measurement, 20*, 37–46.

DAYTON, C.M., and MACREADY, G.B. (1988), “Concomitant Variable Latent Class Models”, *Journal American Statistical Association, 83*, 173–178.

DE LA TORRE, J., and DOUGLAS, J. (2004), “Higher-order Latent Trait Models for Cognitive Diagnosis”, *Psychometrika, 69*, 333–353.

DE LA TORRE, J. (2009), “DINA Model and Parameter Estimation: A Didactic”, *Journal of Educational and Behavioral Statistics, 34*, 115–130.

DE LA TORRE, J., and CHIU, C.Y. (2009), “A Generalized Index of Item Discrimination for Cognitive Diagnosis Models”, paper presented at the International Meeting of the Psychometric Society, Cambridge, England.

EMBRETSON, S.E. (1984), “A General Latent Trait Model for Response Processes”, *Psychometrika, 49*, 175–186.

FORMANN, A.K. (1992), “Linear Logistic Latent Class Analysis for Polytomous Data”, *Journal of the American Statistical Association, 87*, 476–486.

HAERTEL, E. H. (1989), “Using Restricted Latent Class Models to Map the Skill Structure of Achievement Items”, *Journal of Educational Measurement, 26*, 333–352.

HARTZ, S. (2002), “A Bayesian Framework for the Unified Model for Assessing Cognitive Abilities: Blending Theory with Practicality”, Ph.D Thesis, University of Illinois, Urbana-Champaign, IL.

HEFFERNAN, N.T., KOEDINGER, K.R., and JUNKER, B.W. (2001), “Using Web-Based Cognitive Assessment Systems for Predicting Student Performance on State Exams”, research proposal to the Institute of Educational Statistics, US Department of Education; Department of Computer Science at Worcester Polytechnic Institute, Worcester County, MA.

HUBERT, L., and ARABIE, P. (1985), “Comparing Partitions”, *Journal of Classification, 2*, 193–218.

JÖRESKOG, K.G. (1971), “Simultaneous Factor Analysis in Several Populations”, *Psychometrika, 36*, 409–426.

JÖRESKOG, K.G., and GOLDBERGER, A.S. (1975), “Estimation of a Model with Multiple Indicators and Multiple Causes of a Single Latent Variable”, *Journal of the American Statistical Association, 70*, 631–639.

JUNKER, B.W., and SIJTSMA, K. (2001), “Cognitive Assessment Models with Few Assumptions, and Connections with Nonparametric Item Response Theory”, *Applied Psychological Measurement, 25*, 258–272.

LARSEN, K., PETERSEN, J. H., BUDTZ-JØRGENSEN, E., and ENDAHL, L. (2000), “Interpreting Parameters in the Logistic Regression Model with Random Effects”, *Biometrics, 56*, 909–914.

LI, F., and COHEN, A.S. (2006), “A Higher-Order DINA Rasch Model For Detection of Differential Item Functioning”, paper presented at the annual meeting of the Pacific Rim Objective Measurement Symposium, Hong Kong, People’s Republic of China.

MACREADY, G.B., and DAYTON, C.M. (1977), “The Use of Probabilistic Models in the Assessment of Mastery”, *Journal of Educational Statistics, 2*, 99–120.

MAGIDSON, J., and VERMUNT, J.K. (2001), “Latent Class Factor and Cluster Models, Bi-plots and Related Graphical Displays”, *Sociological Methodology, 31*, 223–264.

MARIS, E. (1999), “Estimating Multiple Classification Latent Class Models”, *Psychometrika, 64*, 187–212.

MEREDITH, W. (1993), “Measurement Invariance, Factor Analysis and Factorial Invariance”, *Psychometrika, 58*, 525-543.

MISLEVY, R.J. (1985), “Estimation of Latent Group Effects”, *Journal of the American Statistical Association, 80*, 993–997.

MISLEVY, R.J. (1987), “Exploiting Auxiliary Information about Examinees in the Estimation of Item Parameters”, *Applied Psychological Measurement, 11*, 81–91.

MISLEVY, R.J., JOHNSON, E.G., and MURAKI, E. (1992), "Scaling Procedures in NAEP", *Journal of Educational Statistics, 17*, 131–154.

MILLSAP, R. E., and EVERSON, H. T. (1993), "Methodology Review: Statistical Approaches for Assessing Measurement Bias", *Applied Psychological Measurement, 17*, 297–334.

MUTHÉN, B., and LEHMAN, J. (1985), "Multiple Group IRT Modeling: Applications to Item Bias Analysis", *Journal of Educational Statistics, 10*, 133–142.

MUTHÉN, L. K., and MUTHÉN, B. O. (2010), *Mplus User's Guide* (Sixth Ed.), Los Angeles, CA: Muthén & Muthén.

R DEVELOPMENT CORE TEAM. (2004), "R: A Language and Environment for Statistical Computing", R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0, http://www.R-project.org.

RUBIN, D.B. (1987), *Multiple Imputation for Nonresponse in Surveys*, New York: Wiley.

RUPP, A., and TEMPLIN, J. (2007), "The Effects of Q-Matrix Misspecification on Parameter Estimates and Classification Accuracy in the DINA Model", *Educational and Psychological Measurement, 68 (1)*, 78–96.

RUPP, A., TEMPLIN, J., and HENSON, R. (2010), *Diagnostic Measurement: Theory, Methods, and Applications*, New York: The Guildford Press.

SELF, J. (1993), "Model-Based Cognitive Diagnosis", *User Modeling and User-Adapted Interaction, 3*, 89–106.

SMIT, A., KELDERMAN, H., and VAN DER FLIER, H. (1999), "Collateral Information and Mixed Rasch Models", *Methods of Psychological Research Online, 4*, 19–32.

SMIT, A., KELDERMAN, H., and VAN DER FLIER, H. (2000), "The Mixed Birnbaum Model: Estimation using Collateral Information", *Methods of Psychological Research Online, 5*, 31–43.

SPIEGELHALTER, D.J., THOMAS, A., and BEST, N.G. (2003), *WinBUGS: Bayesian Inference Using Gibbs Sampling, Manual Version 1.4*, Cambridge: Medical Research Council Biostatistics Unit.

STATACORP. (2009), *Stata Statistical Software: Release 11*, College Station, TX: StataCorp LP.

TATSUOKA, K.K. (1983), "Rule Space: An Approach for Dealing with Misconceptions Based on Item Response Theory", *Journal of Educational Measurement, 20*, 345–354.

TEMPLIN, J. (2004), "Generalized Linear Mixed Proficiency Models", Ph.D. Thesis, University of Illinois, Urbana-Champaign, IL.

TEMPLIN, J., HENSON, R., and DOUGLAS, J. (2007), "General Theory and Estimation of Cognitive Diagnosis Models: Using Mplus to Derive Model Estimates", Manuscript under Review.

TEMPLIN, J.L., HENSON, R.A., TEMPLIN, S.E., and ROUSSOS, L. (2008), "Robustness of Hierarchical Modeling of Skill Association in Cognitive Diagnosis Models, *Applied Psychological Measurement, 32*, 559–574.

THISSEN, D., STEINBERG, L., and WAINER, H. (1988), "Use of Item Response Theory in the Study of Group Differences in Trace Lines", in *Test Validity*, eds. H. Wainer and H. Braun, Hillsdale, NJ: Erlbaum, pp. 147–169.

VERMUNT, J.K. (2003), "Multilevel Latent Class Models", *Sociological Methodology, 33*, 213–239.

VON DAVIER, M. (2010), "Hierarchical Mixtures of Diagnostic Models", *Psychological Test and Assessment Modeling, 52*, 8–28.

XU, X., and VON DAVIER, M. (2008), "Fitting the Structural Diagnostic Model to NAEP Data", Research Report RR-08-27, Princeton, NJ: Educational Testing Service.

WEDEL, M. (2002), "Concomitant Variables in Finite Mixture Models", *Statistica Neerlandica, 56*, 362–375.

ZHANG, W. (2006), "Detecting Differential Item Functioning Using the DINA Model", Ph.D. Thesis, The University of North Carolina at Greensboro, Greensboro, NC.

ZWINDERMAN, A.H. (1991), "A Generalized Rasch Model for Manifest Predictors", *Psychometrika, 56*, 589–600.