JEM JOURNAL of EDUCATIONAL MEASUREMENT NCME national council on measurement in education

Journal of Educational Measurement
Spring 2014, Vol. 51, No. 1, pp. 75–97

The Random-Effect DINA Model

Hung-Yu Huang
University of Taipei
Wen-Chung Wang
The Hong Kong Institute of Education

The DINA (deterministic input, noisy, and gate) model has been widely used in cognitive diagnosis tests and in the process of test development. The outcomes known as slip *and* guess *are included in the DINA model function representing the responses to the items. This study aimed to extend the DINA model by using the random-effect approach to allow examinees to have different probabilities of slipping and guessing. Two extensions of the DINA model were developed and tested to represent the random components of slipping and guessing. The first model assumed that a random variable can be incorporated in the slipping parameters to allow examinees to have different levels of caution. The second model assumed that the examinees' ability may increase the probability of a correct response if they have not mastered all of the required attributes of an item. The results of a series of simulations based on Markov chain Monte Carlo methods showed that the model parameters and attribute-mastery profiles can be recovered relatively accurately from the generating models and that neglect of the random effects produces biases in parameter estimation. Finally, a fraction subtraction test was used as an empirical example to demonstrate the application of the new models.*

Cognitive diagnosis models (CDMs) are of growing interest in test development and in the measurement of human performance. In contrast with item response theory (IRT) models where examinees are calibrated along a latent trait continuum (Lord, 1980), CDMs provide profile information of finer-grained skills or attributes of examinees (de la Torre, Hong, & Deng, 2010). Two features of CDMs (but possibly not the only ones) related to item response functions are termed *slip* and *guess* (Rupp & Templin, 2008; Rupp, Templin, & Henson, 2010). Slipping indicates that a person with mastery of all (or at least one, depending on the model) of the attributes that an item requires fails to answer the item correctly. In contrast, guessing indicates that a correct response to an item is given by a person who lacks all (or at least one, depending on the model) of the attributes that are required by an item. Slipping and guessing in CDMs often are treated as item characteristics that are fixed across persons. However, the level of slipping may depend on the person's degree of caution, and the probability of a correct guess may be determined by the person's ability to eliminate distractors among the options included in an item. The variations in slipping and guessing across examinees should be considered, and a much more general model can be developed by treating these item parameters as random effects rather than fixed effects, as in random-effect IRT models (De Boeck & Wilson, 2004).

Among CDMs, the deterministic input, noisy and gate (DINA) model (Haertel, 1989; Junker & Sijtsma, 2001) is probably the most popular because of its simple structure and easy interpretation (de la Torre, 2009, 2011; de la Torre & Douglas, 2004, 2008; de la Torre et al., 2010; de la Torre & Lee, 2010; Henson, Templin, & Willse, 2009). For this reason, we extended the DINA model with a random-effect approach in this study. However, the extensions are not limited to the DINA model and other CDMs are possible. Let Y_{ij} be the response of person i ($i = 1, \ldots, I$) to item j ($j = 1, \ldots, J$). Let α_{ik} be the binary variable for person i on attribute k ($k = 1, \ldots, K$), where a value of 1 indicates that person i shows mastery of attribute k and a value of 0 indicates nonmastery, and let $\boldsymbol{\alpha}_i$ be the vector of attributes of person i. Let $\mathbf{Q}$ be a J by K matrix, with element q_{jk} indicating whether attribute k is required to answer item j correctly. If the attribute is required, $q_{jk} = 1$. If not, $q_{jk} = 0$. In the DINA model, a person's attribute vector and the Q-matrix constitute a latent response variable for person i to item j:

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

It can be shown that ξ_{ij} is equal to 1 if person i possesses all of the attributes that are required for item j and is equal to 0 otherwise. The probability of a correct response to item j for person i is defined as

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

where the slipping parameter s_j is the probability of an incorrect response to item j if all of the required attributes of item j have been mastered, that is, $s_j = P(Y_{ij} = 0|\xi_{ij} = 1)$, and the guessing parameter g_j is the probability of a correct response to item j if a person lacks at least one of the required attributes for item j, that is, $g_j = P(Y_{ij} = 1|\xi_{ij} = 0)$. Because examinees who master all of the attributes required by an item have a greater probability of answering the item correctly, the DINA model is regarded as a type of noncompensatory model.

In the context of cognitive diagnosis, each α_k can be viewed as a knowledge state (Tatsuoka, 1995). The acquisition of these attributes can be assumed to be a function of one or more general abilities. This assumption means that the model consists of two levels (de la Torre & Douglas, 2004). At Level 1, the item responses are determined by the attributes according to Equation 2. At Level 2, the attributes are determined by the examinee's general ability according to

$$P(\alpha_{ik} = 1|\theta_i) = \frac{\exp[\lambda_{1k}(\theta_i - \lambda_{0k})]}{1 + \exp[\lambda_{1k}(\theta_i - \lambda_{0k})]}, \tag{3}$$

where λ_{0k} and λ_{1k} are the location and discrimination parameters, respectively, for attribute k, and θ_i is the level of general ability for person i. The level of general ability follows the standard normal distribution to identify the model. In addition to a higher-order structure, the use of tetrachoric formulations represents an alternative for constructing the relationship between attributes (de la Torre & Douglas, 2004; Hartz, 2002). In this study, we used and extended the higher-order DINA model; however, the nonhigher-order DINA model can be applied to our new models in a straightforward way. In addition to the (higher-order) DINA model, many CDMs have been developed. Readers who are interested in a variety of CDMs can refer to Rupp et al. (2010).

Both the slipping and guessing parameters are treated as fixed effects in these CDMs. In the DINA model, for example, s_j and g_j do not have a subscript of *i*. This feature of the model specifies that all persons have the same slipping parameter and the same guessing parameter. Specifically, all examinees have the same probability of slipping and the same probability of making a lucky guess in response to item *j*. In practice, different persons may have different probabilities of slipping as well as different amounts of luck at guessing.

There are two reasons to develop the random-effect DINA model. First, the phenomenon of slipping is substantially affected by the motivation of examinees. For example, the 2005 National Assessment of Educational Progress (NAEP) showed that 12th-grade students failed to make progress but gained better grades in more advanced courses according to the recent Nation's Report Card (Grigg, Donahue, & Dion, 2007). Because the NAEP is a low-stakes test, most of the students taking it were not strongly motivated (Cao & Stokes, 2008). Performance deficits can be explained by motivational rather than cognitive issues (Graham & Hudley, 2005; Pintrich, 2003; Wentzel & Wigfield, 2009). Low motivation results in a decline in the level of caution when examinees respond to test items, and caution usually affects the probability that a slip will occur. If an examinee's level of caution is lower, the slipping probability of that examinee will be higher. Because examinees have different degrees of motivation, caution can vary among examinees.

Second, it is plausible that guessing is person dependent rather than item dependent because examinees with high ability are expected to have a larger probability of making a correct guess than their counterparts with low ability (San Martín, del Pino, & De Boeck, 2006). In addition to the inclusion of random effects in the slipping parameters, random effects can be incorporated in the guessing parameters to depict the variation among persons in making a guess on an item. In contrast to the random-slip approach described above, the random-guess approach is not justifiable or theoretically appealing because a correct guess on a test item usually depends on the examinee's ability rather than on the use of a random guess. Empirical analyses of educational tests demonstrated that the examinees tended to use their ability to guess and that an IRT model with ability-based guessing fit the data better than traditional IRT models (San Martín et al., 2006). Although these findings resulted from the applications of IRT models, the different degrees of caution and different probabilities of guessing observed in the IRT analyses also can be observed in CDMs. In fact, IRT-based large-scale assessments such as the Trends in Mathematics and Science Study (TIMSS) have been fitted to a CDM (e.g., the DINA model) in conjunction with a means of defining the Q-matrix (Lee, Park, & Taylan, 2011). This result implies that in some cases a CDM (e.g., the DINA model) may be applicable to IRT-based data and that different types of responding behavior occurring with test items in IRT models can apply to a CDM. It is known that remedial instruction is an important feature of CDMs or the DINA model. If randomness in the item parameters is not considered, the parameter estimates would be biased. As a result, the instructional resources might be allocated in a misleading way.

The main purposes of this study were to extend the DINA model to a general class of cognitive diagnosis models by using the random-effect approach, to assess the efficiency of the new models through Bayesian estimation, and to compare the new models with the convenient DINA model with respect to parameter recovery upon application of the extended models. The remainder of this article consists of five sections. First, the rationale and development of the new models are introduced. Second, Bayesian estimation with Markov chain Monte Carlo (MCMC) methods is described briefly. Third, we discuss simulations that were conducted to assess model parameter recovery, the effects of model misspecification, and model-data fit indices with MCMC estimates. Fourth, we review a 15-item fraction subtraction test that was used as an empirical example to demonstrate the applications of these extensions. Finally, conclusions about these new models are drawn and other possible extensions are discussed.

The Random-Slip DINA Model

The slipping parameter s_j in the DINA model does not involve persons. Under this parameterization, all persons share the same slipping parameter for an item. This assumption may be too stringent to describe individual differences because different persons may have different degrees of caution (and slipping). If so, a new model is needed. We propose the random-slip DINA model (denoted as RS-DINA) to incorporate the individual differences in the slipping parameter:

$$P(Y_{ij} = 1|\boldsymbol{\alpha}_i) = (1 - s_{ij})^{\xi_{ij}} g_j^{(1-\xi_{ij})}, \tag{4}$$

$$s_{ij} = \frac{\exp(\delta_j - \gamma_i)}{1 + \exp(\delta_j - \gamma_i)}, \tag{5}$$

$$\gamma_i \sim N(0, \sigma_\gamma^2), \tag{6}$$

where δ_j is the threshold parameter of item *j* for slipping and γ_i is the level of caution of person *i*, on a logit scale; the other parameters are defined in the same way as their counterparts in the DINA model. If all persons have the same degree of caution (i.e., all zeros), the RS-DINA becomes the DINA, with

$$s_j = \frac{\exp(\delta_j)}{1 + \exp(\delta_j)}. \tag{7}$$

As is the case for the higher-order DINA model, a combination of Equation 3 and the RS-DINA model leads to the higher-order RS-DINA model. Similar approaches are found in IRT models in which an additional random variable is added to describe the variation in responses resulting from different responding behaviors. This approach serves to maintain local item independence (De Boeck & Wilson, 2004; Wang & Jin, 2010a, 2010b; Wang & Liu, 2007; Wang, Wilson, & Shih, 2006).

The level of caution depends on the aptitude that examinees possess. If randomness in slipping parameters is ignored, the probability of a correct answer to an item by a careful examinee who gives the item thoughtful consideration may be underestimated. In contrast, another examinee with a low level of caution may show an overestimate of the probability of a correct response to the same item. These estimation errors may be found for all examinees who have mastered the attributes required by the item.

The Ability-Guessing DINA Model

Guessing on multiple-choice (MC) items is highly complex. As in the three-parameter logistical model (3PLM; Birnbaum, 1968), the guessing parameter g_j in the DINA model does not involve persons. This parameterization suggests that guessing is a characteristic of an item, not of a person. In reality, guessing is an interaction between the propensity of an item to elicit guessing and the person's level of ability. For this reason, guessing on MC items may involve ability. Examinees often are instructed to make a wise guess rather than a random guess, and so more capable examinees may have a greater probability of making a correct guess than less capable ones. In contrast, less capable examinees are more prone to be affected by distractors than more capable examinees.

Two main procedures for responding to MC items can be defined (San Martín et al., 2006). In the process (P) procedure, an examinee answers an MC item according to his/her ability (knowledge). In the guessing (G) procedure, an examinee guesses an answer to an MC item. Two arrangements of P and G are possible. In the first arrangement, P occurs first and G follows if the correct response is not found. In this arrangement, an examinee first works on an MC item according to his/her knowledge. If the correct response is not found, the examinee makes a guess. In the second arrangement, G occurs first and depending on the result, P follows. The first arrangement (P, then G) appears to be more appealing than the second (G, then P) because if an examinee makes a guess initially, it does not appear that he/she would apply his/her knowledge afterward. If the second arrangement and item-dependent guessing parameters apply, the 3PLM results. In contrast, San Martín et al. (2006) combined the first arrangement with guessing based on ability and developed the one-parameter logistic model with ability-based guessing (1PL-AG). This model is given by

$$P_{ij} = \frac{\exp(\theta_i - b_j)}{1+\exp(\theta_i - b_j)} + (1 - \frac{\exp(\theta_i - b_j)}{1+\exp(\theta_i - b_j)}) \times \frac{\exp(\omega\theta_i + d_j)}{1+\exp(\omega\theta_i + d_j)}, \quad (8)$$

where P_{ij} is the probability of being correct on item j for person i, θ_i is the latent trait of person i, b_j is the difficulty parameter of item j, ω is the weight of ability in making a guess, and d_j describes the propensity of item j to provoke guessing. In such a formulation of Equation 8, the first term, $\frac{\exp(\theta_i - b_j)}{1+\exp(\theta_i - b_j)}$, is the same as the Rasch model (Rasch, 1960) and can be identified with the P procedure. The next term, $1 - \frac{\exp(\theta_i - b_j)}{1+\exp(\theta_i - b_j)}$, is 1 minus the value of the first term (i.e., $1 - P$). The last term, $\frac{\exp(\omega\theta_i + d_j)}{1+\exp(\omega\theta_i + d_j)}$, can be identified with the G procedure in which ability is involved in making a guess. Depending on the ability-weighted parameter, the item response function can be used to formulate different probabilities of a correct guess on an item by different persons. This property of the item response function means that the random-effect approach serves to make the guessing parameters variable across examinees in the 1PL-AG model.

In the DINA model, a third arrangement of P and G occurs. In this arrangement, the P procedure is determined by the slipping parameters, whereas the G procedure is determined by the guessing parameters. The P and G procedures are not consecutive. Although there are different response functions, guessing may involve ability, as the 1PL-AG assumes. To incorporate the effect of general ability on guessing behavior, the guessing parameter g_j in the DINA model can be replaced with $\frac{\exp(\omega\theta_i+d_j)}{1+\exp(\omega\theta_i+d_j)}$ on a logit scale, so that the resulting ability-guessing DINA (AG-DINA) model is

$$P\left(Y_{ij} = 1|\boldsymbol{\alpha}_i\right) = \left(1 - s_j\right)^{\xi_{ij}} \left(\frac{\exp(\omega\theta_i + d_j)}{1 + \exp(\omega\theta_i + d_j)}\right)^{(1-\xi_{ij})}, \tag{9}$$

where θ_i is the general ability of person i, defined as in Equation 3; d_j is the guessing parameter of item j on a logit scale; and the other parameters are defined as above. If $\omega = 0$, meaning that general ability has no impact on the guessing parameters, the AG-DINA model eventually becomes the DINA model.

The AG-DINA model can be further generalized to include random slipping by combining Equations 4 and 9, whereas the combination of two different random-effect models is more complicated. For the sake of simplicity and ease of interpretation, we did not investigate such a complex combination. Instead, we explored the two models separately.

MCMC Methods

Bayesian estimation has been widely applied to IRT models for decades (Fox, 2010) and has attracted considerable attention in recent years in the use of CDMs for parameter estimation (Rupp et al., 2010). If high dimensionality is involved, Bayesian estimation is much more appealing and efficient than the traditional estimation methods (e.g., marginal maximum likelihood estimation). Several studies relating to CDMs or the DINA model have applied Bayesian estimation to data analysis (de la Torre & Douglas, 2004, 2008; Henson et al., 2009; Rupp et al., 2010). In our extensions, the use of Bayesian estimation to calibrate the data set is necessary because many random effects must be incorporated.

In Bayesian estimation, specifications of a statistical model, prior distributions of model parameters, and observed data produce a joint posterior distribution for the model parameters. MCMC methods are applied to simulate the joint posterior distribution of the unknown quantities and obtain simulation-based estimates of the posterior parameters of interest. There are two main procedures for constructing a Markov chain through the transition density to the target density: the Metropolis-Hastings (Chib & Greenberg, 1995) and the Gibbs sampling (Gelfand & Smith, 1990; Geman & Geman, 1984; Tanner & Wong, 1987) algorithms. MCMC methods are both computationally intensive and cumbersome to program. The freeware WinBUGS (Spiegelhalter, Thomas, & Best, 2003) furnishes simple user-friendly tools for constructing a Markov chain for model parameters. Thus, deriving the posterior distribution of the model parameters becomes less important because users do not need to implement complex algorithms. In this study, the WinBUGS computer program was used to calibrate model parameters and was assessed for its efficiency in parameter estimation.

As noted above, the prior distributions for model parameters should be specified before employing MCMC estimation. Normal priors were adopted for fixed-effect parameters on the logit scale, lognormal priors for parameters with positive values, gamma priors for the variance of random variables, and beta priors for probabilities. The setting of prior distributions for model parameters was consistent with the methods of previous studies involving MCMC methods (Bolt & Lall, 2003; Cao & Stokes, 2008; de la Torre & Douglas, 2004, 2008; Fox, 2010; Li, Bolt, & Fu, 2006). In view of these considerations, we specified the prior distributions for all model parameters as follows: $\lambda_{0k} \sim N\,(0, 4)$, $\lambda_{1k} \sim \text{Log}N\,(0, 1)$, $\theta_i \sim N\,(0, 1)$, $\delta_j \sim N\,(-3, 4)$, $\gamma_i \sim N\left(0, \sigma_\gamma^2\right)$, $\sigma_\gamma^{-2} \sim G\,(1, 1)$, $d_j \sim N\,(-2, 4)$, $\omega \sim \text{Log}N\,(0, 1)$, $s_j \sim B\,(1, 1)$, $g_j \sim B\,(1, 1)$.

Method

Simulation Design

We conducted a series of simulations to examine the performance and relative effectiveness of the proposed models. The item responses were generated according to the RS-DINA model and the AG-DINA model, with five attributes in a higher-order structure in which the location parameters were set at −2, −1.5, −1, −.5, and 0, respectively. Four major independent variables were manipulated: (a) analysis model: the data-generating model and the DINA model; (b) test length: 20 and 40 items; (c) sample size: 1,000 and 2,000 persons; and (d) the second-order discrimination parameters: all were set at 1 and 2. The Q-matrix was generated according to a probability of .5 for each attribute. This approach was similar to the procedure used by Cheng (2009). The general ability of persons was randomly sampled from a standard normal distribution.

To express the item parameters on a logit scale, the slipping and guessing parameters were sampled from a uniform distribution between −5 and −.85 (i.e., a probability between 0 and .3), and −3 and −1.4 (i.e., a probability between .05 and .2), respectively. The random variables used to model the caution of the examinees were generated with a standard normal distribution in the RS-DINA model. The weight of ability in the guessing process for the AG-DINA model was set at 2 to show the stronger effect of general ability on the occurrence of a correct guess. The specifications of item parameters in these models were consistent with those commonly found in practice (de la Torre & Douglas, 2008; San Martín et al., 2006; Wang & Jin, 2010a, 2010b; Wang & Liu, 2007; Wang et al., 2006).

Twenty replications were conducted for each data set. A total of 640 analyses were required to obtain the Bayesian estimates (20 data sets × 2 generating models × 2 analysis models × 2 test lengths × 2 sample sizes × 2 types of discrimination parameters). We found that the number of replications was sufficient to obtain reliable inferences because the sampling variation across replications appeared to be relatively small and the mean square error of the estimator was close to the variance plus the square of the bias across replications. Actually, many IRT studies using Bayesian estimation conducted only 10 replications (Bolt & Lall, 2003; Cao & Stokes, 2008;

Li et al., 2006), and 25 replications were conducted for the DINA model and its extensions (de la Torre & Douglas, 2004, 2008). Moreover, we increased the number of replications to a value greater than 20 for several conditions and found that the differences among them were trivial and could be neglected.

We wrote a Matlab computer program to generate item responses. In addition, a fraction subtraction test (Tatsuoka, 2002; Tatsuoka, 1990) was used as an empirical example of the fitting of the proposed models to real data.

Analysis

The computer program WinBUGS 1.4 (Spiegelhalter et al., 2003), based on Bayesian estimation with MCMC methods, was used to calibrate the data sets with all parameter values specified. We used 15,000 iterations for calculating the posterior mean from the parameter estimates. The first 5,000 iterations were used as burn-in. This approach was based on the multivariate potential scale reduction factor (Brooks & Gelman, 1998). Three parallel chains were used for the first simulated data set across all analysis models. The WinBUGS commands are available from the first author.

For each estimator, the bias, root mean square error (RMSE), and root mean square deviation (RMSD) were computed as

$$Bias(E(\zeta)) = \sum_{r=1}^{R} (E(\zeta_r) - \zeta)/R, \tag{10}$$

$$\text{RMSE}(E(\zeta)) = \sqrt{\sum_{r=1}^{R} (E(\zeta_r) - \zeta)^2/R}, \tag{11}$$

$$\text{RMSD}(E(\theta)) = \sqrt{\sum_{i=1}^{N_p} (E(\theta_i) - \theta_i)^2/N_p}, \tag{12}$$

where ζ and $E(\zeta)$ were the generating value and the estimated parameter from the posterior distribution, respectively; θ and $E(\theta)$ were the true general ability and its corresponding posterior mean, respectively; and R and N_p were the number of simulation replications and the sample size in each data set, respectively. The percentage of correct attribute classification was calculated to assess the recovery of the mastery profile. To examine the fit of the models to the data, the Akaike Information Criterion (AIC; Akaike, 1974) and the Bayesian Information Criterion (BIC; Schwarz, 1978) were computed to compare different models. In addition to assessing the recovery of the parameters under each condition, we also were interested in the effects of using the DINA model to fit the data generated from the complex models. It was expected that the parameters could be recovered well with each proposed model, that all the parameter estimates would be severely biased if the DINA model was fit to the data generated by the extended models, and that the AIC and BIC indices would select the true model as the best-fitting model.

Results

Parameter Recovery of the RS-DINA Model and Consequence of Ignoring Random Slip

Because of space constraints, the bias and RMSE for the individual item parameters are not reported; instead, their mean, standard deviation, and range are shown. Table 1 shows the parameter recovery of the RS-DINA model and the DINA model for test items following the RS-DINA model with the discrimination parameters equal to 1. If the sample size was 1,000 and the test length was 20, the RMSE of the discrimination parameters was .089 for the RS-DINA model and 1.149 for the DINA model, the RMSE of the location parameters was .079–.136 (M = .105) for the RS-DINA model and .539–3.762 (M = 2.338) for the DINA model, the RMSE of the slipping parameters was .004–.024 (M = .013) for the RS-DINA model and .280–.822 (M = .625) for the DINA model, the RMSE of the guessing parameters was .005–.029 (M = .013) for the RS-DINA model and .244–.778 (M = .553) for the DINA model, and the RMSE of the variance of the caution variables was .160 for the RS-DINA model. Apparently, the RS-DINA model yielded better parameter recovery than the convenient DINA model. Similar results can be found for different sample sizes and test lengths. The biases were close to zero for the RS-DINA model. In contrast, the DINA model had much larger biases and the slipping and guessing parameters were substantially overestimated. The principal reason for this outcome was that random slipping (i.e., the caution variables) was present but was not considered. If the test length was increased to 40 items, the estimates of the DINA model were also poor. However, the problems with these estimates appeared to be less severe than those observed on the short test. The RS-DINA model estimated the parameter values more accurately as the sample size increased from 1,000 to 2,000 persons.

Table 2 summarizes the parameter recovery for the same conditions as above but with the discrimination parameters equal to 2. Apparently, the parameter recovery was satisfactory in the RS-DINA model across all conditions but was poor in the DINA model. For example, if the sample size was 1,000 and the test length was 20, the RMSE of the discrimination parameters was .146 for the RS-DINA model and 1.556 for the DINA model, the RMSE of the location parameters was .039–.225 (M = .107) for the RS-DINA model and .771–4.765 (M = 2.572) for the DINA model, the RMSE of the slipping parameters was .002–.028 (M = .012) for the RS-DINA model and .142–.904 (M = .588) for the DINA model, the RMSE of the guessing parameters was .006–.045 (M = .016) for the RS-DINA model and .360–.914 (M = .620) for the DINA model, and the RMSE of the variance of the caution variables was .187 for the RS-DINA model. The analysis of the bias showed that the slipping and guessing parameters again were substantially overestimated by the DINA model. The same conclusions resulted if the sample size was 2,000 and the test length was 40. In addition, the shorter test lengths produced poorer calibration of the model parameters by the misleading DINA model, and the larger sample sizes produced more accurate estimates from the RS-DINA model. The difference in parameter recovery between the high and low discrimination parameters was not substantial, and the degree of the discrimination parameters had little effect on parameter estimation.

Table 1

Bias and RMSE (Multiplied by 1,000) when the RS-DINA and DINA Models were Fit to RS-DINA Data with Discrimination Parameters Equal to 1

Size	1,000								2,000							
Length	20				40				20				40			
Model	RS-DINA		DINA		RS-DINA		DINA		RS-DINA		DINA		RS-DINA		DINA	
Parameter	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE
Discrim	47	89	511	1,149	−26	79	−73	222	3	65	159	940	−30	60	288	1,416
Location																
Mean	18	105	−779	2,338	−48	119	−241	759	−8	88	−1,020	2,101	−38	108	−539	1,712
SD	22	21	1,138	1,231	47	59	220	613	18	27	1,137	1,220	41	37	844	1,013
Max	44	136	890	3,762	6	229	−27	1,763	24	140	533	3,618	24	154	897	3,375
Min	−16	79	−1,962	539	−134	57	−613	103	−25	61	−2,388	892	−89	60	−1,364	234
Slipping																
Mean	0	13	549	625	−1	12	70	164	−2	10	348	477	−2	10	295	433
SD	3	6	103	110	5	7	39	114	3	5	143	162	3	6	172	179
Max	6	24	749	822	15	30	118	304	3	20	561	719	5	26	644	765
Min	−8	4	252	280	−14	2	6	7	−11	2	87	123	−10	2	50	83
Guessing																
Mean	0	13	481	553	1	14	29	92	0	9	249	390	1	10	218	365
SD	5	6	148	169	4	8	28	79	3	4	118	117	3	5	98	104
Max	9	29	677	778	10	47	90	257	5	20	541	694	7	27	488	627
Min	−10	5	210	244	−14	4	−26	4	−6	3	125	206	−7	3	112	185
Variance	61	160	–	–	36	92	–	–	27	111	–	–	46	80	–	–

Note. Length = test length; Size = sample size; Discrim = discrimination parameter; Variance = variance of caution.

Table 2

Bias and RMSE (Multiplied by 1,000) when the RS-DINA and DINA Models were Fit to RS-DINA Data with Discrimination Parameters Equal to 2

Size	1,000								2,000							
Length	20				40				20				40			
Model	RS-DINA		DINA		RS-DINA		DINA		RS-DINA		DINA		RS-DINA		DINA	
Parameter	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE
Discrim	−48	146	265	1,556	−40	130	−105	1,038	−10	57	1,293	1,727	−11	73	−148	535
Location																
Mean	−26	107	−880	2,572	−7	67	−538	1,609	−25	54	71	1,529	−7	55	−288	1,310
SD	29	65	640	1,540	17	27	552	1,006	10	20	480	1,020	19	19	390	894
Max	5	225	−12	4,765	13	106	219	2,718	−14	77	687	3,063	16	92	270	2,670
Min	−76	39	−1,598	771	−28	23	−1,071	114	−40	25	−694	179	−34	39	−818	52
Slipping																
Mean	−2	12	529	588	−2	12	161	309	−1	9	560	599	−1	8	50	109
SD	4	7	231	215	4	8	65	128	3	5	202	213	3	5	38	114
Max	3	28	903	904	6	33	291	522	8	20	836	879	4	20	115	309
Min	−12	2	129	142	−13	3	28	50	−8	2	143	154	−12	1	5	6
Guessing																
Mean	2	16	549	620	3	16	131	297	2	12	621	674	2	11	23	69
SD	8	12	138	135	5	12	36	65	5	8	122	134	5	7	29	80
Max	22	45	914	914	15	72	183	403	20	36	818	859	10	43	95	258
Min	−14	6	336	360	−10	4	54	179	−6	3	374	397	−24	4	−26	5
Variance	18	187	–	–	−40	90	–	–	23	49	–	–	8	56	–	–

Note. Length = test length; Size = sample size; Discrim = discrimination parameter; Variance = variance of caution.

Table 3
Mean Accuracy Rates for Attributes and Profiles and Mean Root Mean Square Deviation for Person Measures (Multiplied by 1,000) when the RS-DINA and DINA Models were Fit to RS-DINA Data

Model	RS-DINA							
Sample Size	1,000				2,000			
Test Length	20		40		20		40	
Discrimination	1	2	1	2	1	2	1	2
Attribute 1	930	953	970	975	934	958	970	976
Attribute 2	952	969	985	992	957	971	988	991
Attribute 3	916	946	928	956	922	949	928	953
Attribute 4	981	987	980	992	982	988	982	991
Attribute 5	968	990	995	998	966	990	995	999
Profile	830	884	889	929	835	889	891	925
General Ability (Mean RMSD)	766	622	757	613	761	624	753	618
Model	**DINA**							
Sample Size	1,000				2,000			
Test Length	20		40		20		40	
Discrimination	1	2	1	2	1	2	1	2
Attribute 1	805	696	963	932	872	682	812	928
Attribute 2	469	664	980	926	815	609	902	898
Attribute 3	724	704	919	852	771	600	823	816
Attribute 4	391	453	978	761	455	196	458	711
Attribute 5	405	99	896	755	692	252	826	796
Profile	85	5	795	574	332	89	278	533
General Ability (Mean RMSD)	1,292	1,380	781	857	1,060	1,504	1,005	917

Table 3 summarizes the mean percentage of the attributes correctly classified and the mean RMSD for general ability under the RS-DINA model compared with the DINA model. These results indicate that the RS-DINA model consistently showed a higher correct classification rate for the five attributes and for the entire profile and yielded a lower RMSD for general ability across all conditions. For example, if the sample size was 2,000 and the test length was 20, the recovery rate was .922–.982 for the attributes and .835 for the entire profile in the RS-DINA model and .455–.872 for the attributes and .332 for the entire profile in the DINA model. The RMSD of general ability was .761 in the RS-DINA model and 1.060 in the DINA model. If the test length was increased to 40 items, the percentage of correct attribute classification increased and the precision of the estimate of general ability improved in the RS-DINA model, and the recovery of the person measures in the DINA model improved but was inferior to the corresponding RS-DINA model. In addition, higher values of the discrimination parameters were associated with improved recovery of attributes and general abilities in the RS-DINA model. The sample size had little effect on the estimation of attributes and general abilities.

Parameter Recovery of the AG-DINA Model and Consequence of Ignoring Ability Guessing

Table 4 presents the parameter recovery of the AG-DINA model and the DINA model for test items following the AG-DINA model with discrimination parameters equal to 1. The parameters in the AG-DINA model can be recovered better than those in the DINA model across all conditions. For example, if the sample size was 1,000 and the test length was 20, the RMSE of the discrimination parameters was .071 for the AG-DINA model and .706 for the DINA model, the RMSE of the location parameters was .080–.177 ($M = .121$) for the AG-DINA model and .478–3.410 ($M = 1.840$) for the DINA model, the RMSE of the slipping parameters was .004–.022 ($M = .015$) for the AG-DINA model and .144–.601 ($M = .363$) for the DINA model, the RMSE of the guessing parameters was .017–.039 ($M = .027$) for the AG-DINA model and .151–.483 ($M = .268$) for the DINA model, and the RMSE of the ability-weighted parameter was .211 for the AG-DINA model. The AG-DINA model had a smaller bias than the DINA model, and the slipping and guessing parameters were overestimated under most conditions if the DINA model was fit to the AG-DINA data. The principal reason for this outcome was that the guessing parameters of each item were treated as item characteristics rather than person characteristics in the DINA model; the effect of ability on the occurrence of a correct guess was not considered. Similar results were found for parameter recovery if the test length was 40 and the sample size was 2,000. The estimation by the DINA model was much poorer in the short test than in the long test. The reason for this outcome is that the bias was less serious in the long test if the DINA model was fit to the AG-DINA data and the ability-relative guess was not considered. In addition, as in the RS-DINA model, the larger sample size was associated with better parameter recovery by the AG-DINA model.

Table 5 shows the parameter recovery under the same conditions as above except that the discrimination parameters all were set to 2. The AG-DINA model again yielded better parameter recovery than the DINA model. If the sample size was 1,000 and the test length was 20, for example, the RMSE of the discrimination parameters was .077 for the AG-DINA model and 1.537 for the DINA model, the RMSE of the location parameters was .038–.476 ($M = .157$) for the AG-DINA model and .128–5.579 ($M = 2.146$) for the DINA model, the RMSE of the slipping parameters was .002–.306 ($M = .126$) for the AG-DINA model and .123–.845 ($M = .535$) for the DINA model, the RMSE of the guessing parameters was .036–.138 ($M = .086$) for the AG-DINA model and .270–.726 ($M = .490$) for the DINA model, and the RMSE of the ability-weighted parameter was .834 for the AG-DINA model. Similar results can be found for different test lengths and sample sizes. The analysis of the bias showed that the slipping and guessing parameters were overestimated under most conditions by mistakenly using the DINA model. In summary, the shorter test length resulted in poorer calibration of the model parameters by the DINA model. The

Table 4

Bias and RMSE (Multiplied by 1,000) when the AG-DINA and DINA Models were Fit to AG-DINA Data with Discrimination Parameters Equal to 1

Size	1,000								2,000							
Length	20				40				20				40			
Model	AG-DINA		DINA		AG-DINA		DINA		AG-DINA		DINA		AG-DINA		DINA	
Parameter	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE
Discrim	24	71	63	706	8	71	699	2,032	9	52	2,045	2,141	−37	45	452	1,548
Location																
Mean	−33	121	−624	1,840	−21	100	53	205	−12	80	453	619	−44	83	90	598
SD	24	38	382	1,164	32	23	74	129	23	20	507	413	28	28	319	673
Max	5	177	−202	3,410	27	136	152	424	7	105	1,175	1,227	−13	130	719	1,916
Min	−59	80	−1,084	478	−69	76	−31	99	−54	52	−282	234	−91	53	−159	82
Slipping																
Mean	1	15	189	363	2	14	71	210	1	12	628	659	1	11	107	269
SD	4	5	100	128	5	7	25	66	3	5	144	150	4	5	45	81
Max	9	22	381	601	12	30	117	311	6	19	835	876	14	23	200	428
Min	−6	4	58	144	−9	2	17	61	−6	3	332	349	−5	2	29	80
Guessing																
Mean	−17	27	61	268	−16	22	−7	226	−2	15	578	613	−13	18	8	234
SD	5	6	80	95	6	5	31	41	4	5	121	128	3	4	41	56
Max	−6	39	241	483	0	34	57	302	8	29	830	877	−8	28	141	401
Min	−29	17	−31	151	−31	13	−64	148	−12	6	348	372	−20	10	−72	137
Weight	−170	211	–	–	−178	190	–	–	0	89	–	–	−140	150	–	–

Note. Length = test length; Size = sample size; Discrim = discrimination parameter; Weight = ability-weighted parameter.

Table 5

Bias and RMSE (Multiplied by 1,000) when the AG-DINA and DINA Models were Fit to AG-DINA Data with Discrimination Parameters Equal to 2

Size	1,000								2,000							
Length	20				40				20				40			
Model	AG-DINA		DINA		AG-DINA		DINA		AG-DINA		DINA		AG-DINA		DINA	
Parameter	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE
Discrim	−8	77	104	1,537	−23	139	40	155	−4	69	304	1,424	−66	76	−670	1,041
Location																
Mean	15	157	−1,178	2,146	−9	74	20	77	−11	55	−968	2,662	−36	57	−1,524	2,618
SD	68	161	1,222	1,897	12	19	22	24	14	24	1,415	2,000	14	16	1,304	1,862
Max	146	476	18	5,579	9	99	42	107	13	100	966	5,265	−15	77	35	4,875
Min	−38	38	−3,295	128	−27	56	−21	52	−23	33	−2,681	87	−52	33	−3,058	94
Slipping																
Mean	39	126	429	535	2	13	2	13	0	9	479	571	1	9	189	344
SD	44	129	241	242	4	6	4	6	3	4	227	240	3	4	125	199
Max	99	306	759	845	11	28	12	29	6	15	780	872	9	18	389	612
Min	−9	2	79	123	−8	3	−7	3	−5	1	90	127	−4	2	0	2
Guessing																
Mean	−5	86	321	490	−26	35	−95	95	−12	26	385	529	−20	30	34	242
SD	22	33	124	102	10	11	37	36	9	13	113	116	10	12	102	108
Max	29	138	621	726	−5	66	−24	168	0	57	582	688	−6	86	183	458
Min	−53	36	182	270	−53	23	−168	26	−35	13	197	269	−64	15	−168	36
Weight	99	834	–	–	−159	186	–	–	−63	118	–	–	−166	211	–	–

Note. Length = test length; Size = sample size; Discrim = discrimination parameter; Weight = ability-weighted parameter.

Table 6

Mean Accuracy Rates for Attributes and Profiles and Mean Root Mean Square Deviation for Person Measures (Multiplied by 1,000) when the AG-DINA and DINA Models were Fit to AG-DINA Data

Model	AG-DINA							
Sample Size	1,000				2,000			
Test Length	20		40		20		40	
Discrimination	1	2	1	2	1	2	1	2
Attribute 1	953	966	994	997	935	969	996	997
Attribute 2	996	998	1,000	1,000	983	999	1,000	1,000
Attribute 3	935	963	941	963	907	966	944	965
Attribute 4	985	995	991	998	950	995	991	997
Attribute 5	974	931	999	1,000	934	995	999	1,000
Profile	879	864	928	959	787	932	934	960
General Ability (Mean RMSD)	649	601	614	567	567	585	610	562
Model	**DINA**							
Sample Size	1,000				2,000			
Test Length	20		40		20		40	
Discrimination	1	2	1	2	1	2	1	2
Attribute 1	874	898	953	996	632	896	859	985
Attribute 2	882	806	951	1,000	399	782	948	999
Attribute 3	881	835	889	961	473	786	905	912
Attribute 4	891	633	925	997	277	252	924	896
Attribute 5	649	218	908	999	252	510	907	700
Profile	525	188	825	955	76	186	735	575
General Ability (Mean RMSD)	894	1,185	825	604	1,584	1,264	832	745

AG-DINA model yielded better parameter recovery for the large sample size than for the small sample size. The discrimination parameters had little effect on parameter estimation for most of the estimators.

The mean percentage of the attributes correctly classified and the mean RMSD for general ability under the AG-DINA model and the DINA model are shown in Table 6. Similar results were found for the RS-DINA model and the AG-DINA model. For example, if 2,000 persons and a 20-item test were used, the recovery rate was .907–.983 for the attributes and .787 for the entire profile in the AG-DINA model and .252–.632 for the attributes and .076 for the entire profile in the DINA model, and the RMSD of general ability was .567 in the AG-DINA model and 1.584 in the DINA model. These results indicated that the AG-DINA model outperformed the DINA model in the estimation of the attributes and general ability. The same conclusions apply to different sample sizes and test lengths. In addition, the bias in the person measures became less serious in the DINA model as the test length increased. In summary, the longer test length was associated with better recovery of the attributes and general abilities in the AG-DINA model. Moreover, higher values of the discrimination parameters corresponded to more accurate estimates of the person measures.

Fitting the RS-DINA and AG-DINA to DINA Data

Was it harmful to fit unnecessarily complicated models to simple data sets to estimate the model parameters and the attribute profiles of the examinees? What were the consequences of fitting the RS-DINA and AG-DINA to the simple DINA data? To answer these questions, a simulation was performed with 2,000 simulees and a 20-item test. For 20 replications in which the generating model was the DINA model, the estimate of variance for the caution variables was .005–.027 ($M = .016$), a value very close to zero. This result indicated that the unnecessary random variables did little harm. The fitting of the AG-DINA model yielded an estimate of the ability-weighted parameter of .061–.130 ($M = .081$), suggesting that the unnecessary ability-weighted parameter did little harm. In short, fitting the RS-DINA model or the AG-DINA model to DINA data was harmless, whereas fitting the DINA model to RS-DINA or AG-DINA data yielded very poor parameter estimates. Note that because the RS-DINA model and the AG-DINA model were conceptually very different, we did not fit the AG-DINA model to the RS-DINA model, or vice versa. Had we done so, poor parameter estimates would have been yielded.

These results indicate that identifying the most suitable model is an important issue. To address this, the AIC and BIC indices were computed to compare the models. The results of this comparison are not shown in detail because of space limitations. When the data followed the RS-DINA model, the RS-DINA model always yielded the smallest AIC and BIC values (mean AIC = 23024.5 and mean BIC = 23293.5) followed by the AG-DINA model (mean AIC = 24,278 and mean BIC = 24,538) and the DINA model (mean AIC = 31,862 and mean BIC = 32,118). When the data followed the AG-DINA model, the AG-DINA model always yielded the smallest AIC and BIC values (mean AIC = 24,610 and mean BIC = 24,876) followed by the RS-DINA model (mean AIC = 26,684.5 and mean BIC = 26,953.5) and the DINA model (mean AIC = 35,748 and mean BIC = 36,005). In other words, the AIC and BIC indices were very powerful in selecting the true model.

An Empirical Example

To measure the skills involved in subtracting fractions, Tatsuoka (1990) and Tatsuoka (2002) developed the fraction subtraction test. A total of 536 middle school students were recruited to complete the test. Recently, de la Torre and Douglas (2008), de la Torre (2009), and de la Torre and Lee (2010) analyzed a 15-item fraction subtraction test with five attributes. We used the same framework to demonstrate the application of the extended models. The RS-DINA, AG-DINA, and DINA models were fit. The AIC and BIC values were 20,750 and 20,790, respectively, for the RS-DINA models; 20,870 and 21,080, respectively, for the AG-DINA model; and

22,450 and 22,660, respectively, for the DINA model. The RS-DINA appeared to have the best fit.

To verify the fit of the RS-DINA model to the data, posterior predictive model checking (PPMC; Gelman, Meng, & Stern, 1996; Rubin, 1984) was used to assess the plausibility of posterior predictive replicated data in the context of the observed data within a Bayesian framework for several test statistics. The first test statistic was a Bayesian chi-square test. In this test, the discrepancy between the observed and expected responses of examinees for all items was evaluated by comparing the observed data with the replicated data from a large number of iterations (Sinharay, 2005; Sinharay, Johnson, & Stern, 2006). The results indicated that no test items showed a poor fit in terms of the model. Another Bayesian chi-square test that has a test statistic that asymptotically follows a chi-square distribution was used to allow the application of a hypothesis test to evaluate the fit of the model to the data (Johnson, 2004). In this example,

$$R(\psi) = \sum_{j=1}^{15} \frac{\left[\sum_i Y_{ij} - \sum_i P(Y_{ij} = 1|\psi)\right]^2}{\sum_i P(Y_{ij}|\psi)} \tag{13}$$

and the statistic $R(\psi)$ has an asymptotic chi-square distribution with 14 degrees of freedom. The proportion of the values of this statistic that exceeded the 95th percentile of the chi-square distribution based on MCMC estimation was zero. This result showed that the RS-DINA model produced an excellent fit to the data. Based on these findings, the RS-DINA model was selected to fit the fraction subtraction test data. Apparently, the examinees had different slipping parameters and these different levels of caution should be considered.

Under the RS-DINA model, the estimates were 8.45 for the discrimination parameters, -1.01 to $-.13$ ($M = -.64$) for the location parameters, .03–.44 ($M = .19$) for the slipping parameters, and .00–.29 ($M = .07$) for the guessing parameters. In addition, the estimate for the variance of the caution variables was 1.94. This value clearly was greater than the variance of general ability, and the variation in the caution variables implied that the randomness of the slipping parameters should not be neglected. The analysis of the person measures showed that the percentage of examinees mastering each attribute was .84, .77, .79, .53, and .54 for attributes 1 to 5, respectively. This result suggested that attributes 4 and 5 (borrowing one from a whole number to a fraction and converting a whole number to a fraction, respectively) were insufficient for nearly half of the examinees. The significance of this finding was that it furnished important information that can help teachers and consultants provide timely remedial instruction.

To show the practical differences among the three models, those attribute estimates derived from the RS-DINA model were treated as the gold standard with which those derived from the AG-DINA and DINA models were compared. The AG-DINA model had an agreement rate with the RS-DINA model between .68 and .96 on the five attributes and .49 on the entire profile. The DINA model had an agreement rate with the RS-DINA model between .85 and .96 on the five attributes and .80 on the entire profile. These agreement rates demonstrated the impact of selection of an appropriate model. Although the DINA model had the worst fit among the three models, it yielded attribute estimates more similar to those of the RS-DINA model than did the AG-DINA model. This was because the conceptual similarity between the DINA and RS-DINA models was higher than that between the AG-DINA and RS-DINA models.

Discussion and Conclusions

In this study, the DINA model was extended to yield a more general and flexible response function with fewer constraints. Examinees may respond to test items in a way that differs substantially from the assumptions of the DINA model. First, the level of the examinees' caution can be variable rather than constant, and the probability of slipping can be included in the model by incorporating a random-effect variable to represent the slipping parameters. This approach yields the RS-DINA model. Similarly, the guessing parameters can differ among the examinees. The guessing parameters no longer are viewed as item characteristics. Instead, an ability-weighted parameter is introduced to allow the guessing process to be affected by the level of the examinee's ability. The resulting AG-DINA model is more convincing and justifiable than the DINA model because the guessing parameters are regarded as a characteristic of the person. An examinee of relatively high general ability may have a higher probability of making a correct guess without mastering the required attributes of an item. In this respect, the AG-DINA model shares the relevant properties of compensatory CDMs. These properties imply that general ability can compensate for the examinee's insufficiency on the attributes that are required by an item. As a result, general ability can improve the examinee's probability of responding correctly to that item. The assumptions of the extended DINA models do not share the stringent characteristics of the DINA assumptions and represent a practical approach to the real testing situation.

In a series of simulations with MCMC methods, the RS-DINA and AG-DINA models were used to generate item responses. The generating model and the DINA model both were fit to the simulated data given different values of test length, sample size, and discrimination power. Ignoring the randomness of slipping and using the DINA model to fit the RS-DINA data produced a bias in item parameter estimation, a lower correct attribute classification rate, and a higher RMSD for general ability, regardless of test length, sample size, or discrimination power. When fitting the RS-DINA model to the RS-DINA data, we can conclude the following. As the sample size increased, the item parameter estimation became more precise. As the test length increased, the attribute classification and the estimation of general ability improved. The use of high values of the discrimination parameters slightly improved the precision of the estimation of the person parameters; however, the value of the discrimination parameter had little effect on the estimation of the item parameters. In addition, the results showed that the slipping and guessing parameters were overestimated substantially by the DINA model. The principal reason for this outcome was that the different levels of caution of the examinees were not considered.

The major conclusions drawn from the RS-DINA model also were appropriate in the case of the AG-DINA model. If the DINA model was fit to the AG-DINA data, the estimates of the item and person parameters all were biased. In contrast, the parameter recovery was satisfactory independent of test length, sample size, and discrimination power if the generating model and the analysis model were identical (i.e., the AG-DINA model was fit to the AG-DINA data). The effects of different discrimination parameters, test lengths, and sample sizes on parameter estimation in the AG-DINA model were the same as the corresponding effects in the RS-DINA model. Overestimation of the slipping and guessing parameters also occurred if the DINA model was fit to the AG-DINA data. Although the estimation problems resulting from the use of the misleading DINA model to fit the RS-DINA or AG-DINA data became less serious if the test was relatively long, the efficiency of the two proposed models was better than that of the regular DINA model. In addition, the AIC and BIC indices represent an efficient method of comparing the proposed models with the DINA model if different types of responses to test items are considered.

The DINA model has been applied to the fraction subtraction test in several studies (e.g., de la Torre, 2009, 2011; de la Torre & Douglas, 2004, 2008; de la Torre & Lee, 2010; Henson et al., 2009). However, these studies did not consider the possible random effects of slipping or guessing. We demonstrated that the RS-DINA model fitted the data better than the AG-DINA model and the DINA model according to the lower AIC and BIC values and the PPMC statistics. These findings indicated that the examinees had different levels of caution and that the slipping parameters should not be considered constant across persons. The difference in attribute estimation between the RS-DINA model and the DINA model or the AG-DINA model should not be neglected because a large number of the examinees would be diagnosed differently and the mastery profile of examinees obtained from the DINA model or the AG-DINA model may be misleading. Note that the random-effect approach produces a model that furnishes a plausible fit to the data analyzed from the fraction subtraction test in the context of the DINA model. Other CDMs may fit the data more adequately than the DINA model if different response functions are specified. For example, the NIDA model (noisy inputs, deterministic, and gate; Junker & Sijtsma, 2001; Maris, 1999) represents a plausible alternative.

Generalizations of the DINA model are the focus in this study. Future studies can aim at developing more general DINA models that accommodate both random-slip and ability-weighted parameters simultaneously. In addition, mixture random-effect DINA models can be developed to account for situations where some examinees follow the RS-DINA model, some follow the AG-DINA model, and some follow both. The formulation of other CDMs with a random-effect approach is straightforward. For this purpose, one simply must specify an item response function of interest and then impose the random effects on item parameters if necessary. For example, the log-linear cognitive diagnosis model (Henson et al., 2009) and the generalized DINA model (de la Torre, 2011) can be extended readily by incorporating the random-slip and ability-weighted parameters. Finally, other estimation methods that have been applied to the DINA model—such as marginal likelihood estimation (de la Torre, 2009, 2011; de la Torre et al., 2010; de la Torre & Lee, 2010)—deserve assessment in terms of the efficiency of parameter estimation relative to that furnished by MCMC methods.

Acknowledgment

This study was partly supported by the National Science Committee (101–2410-H-133–001).

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*, 716–723.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinees' ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–479). Reading, MA: Addison-Wesley.

Bolt, D. M., & Lall, V. F. (2003). Estimation of compensatory and noncompensatory multidimensional item response models using Markov chain Monte Carlo. *Applied Psychological Measurement*, *27*, 395–414.

Brooks, S. P., & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, *7*, 434–455.

Cao, J., & Stokes, S. L. (2008). Bayesian IRT guessing models for partial guessing behaviors. *Psychometrika*, *73*, 209–230.

Cheng, Y. (2009). When cognitive diagnosis meets computerized adaptive testing: CD-CAT. *Psychometrika*, *74*, 619–632.

Chib, S., & Greenberg, E. (1995). Understanding the Metropolis-Hastings algorithm. *American Statistician*, *49*, 327–335.

De Boeck, P., & Wilson, M. (Eds.). (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. New York, NY: Springer.

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. (2011). The generalized DINA model framework. *Psychometrika*, *76*, 179–199.

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

de la Torre, J., & Douglas, J. (2008). Model evaluation and multiple strategies in cognitive diagnosis: An analysis of fraction subtraction data. *Psychometrika*, *73*, 595–624.

de la Torre, J., Hong, Y., & Deng, W. (2010). Factors affecting the item parameter estimation and classification accuracy of the DINA model. *Journal of Educational Measurement*, *47*, 227–249.

de la Torre, J., & Lee, Y,-S. (2010). A note on the invariance of the DINA model parameters. *Journal of Educational Measurement*, *47*, 115–127.

Fox, J. P. (2010). *Bayesian item response modeling*. New York, NY: Springer.

Gelfand, A. E., & Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, *85*, 398–409.

Gelman, A., Meng, X.-L., & Stern, H. S. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, *6*, 733–807.

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *6*, 721–741.

Graham, S., & Hudley, C. (2005). Race and ethnicity in the study of motivation and competence. In A. Elliot & C. Dweck (Eds.), *Handbook of competence and motivation* (pp. 392–413). New York, NY: Guilford Press.

Grigg, W., Donahue, P., & Dion, G. (2007). *The nation's report card: 12th-grade reading and mathematics 2005*. Washington, DC: U.S. Department of Education, National Center for Education Statistics.

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* (Unpublished doctoral dissertation). University of Illinois, Urbana-Champaign.

Henson, R., Templin, J., & Willse, J. (2009). Defining a family of cognitive diagnosis models using log linear models with latent variables. *Psychometrika*, *74*, 191–210.

Johnson, V. (2004). A Bayesian χ^2 test for goodness of fit. *Annals of Statistics*, *32*, 2361–2384.

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

Lee, Y.-S., Park, Y. S., & Taylan, D. (2011). A cognitive diagnostic modeling of attribute mastery in Massachusetts, Minnesota, and the U.S. national sample using the TIMSS 2007. *International Journal of Testing*, *11*, 144–177.

Li, Y., Bolt, D. M., & Fu, J. (2006). A comparison of alternative models for testlets. *Applied Psychological Measurement*, *30*, 3–21.

Lord, F. M. (1980). *Application of item response theory to practical testing problems*. Hillside, NJ: Lawrence Erlbaum.

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

Pintrich, P. R. (2003). Motivation and classroom learning. In W. Reynolds & G. Miller (Eds.), *Handbook of psychology, Vol. 7: Educational psychology* (pp. 103–122). New York, NY: Wiley.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Institute of Educational Research (Expanded edition, 1980). Chicago, IL: University of Chicago Press.

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Annals of Statistics*, *12*, 1151–1172.

Rupp, A. A., & Templin, J. L. (2008). Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art. *Measurement*, *6*, 219–262.

Rupp, A. A., Templin, J. L., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications*. New York, NY: Guilford Press.

San Martín, E., del Pino, G., & De Boeck, P. (2006). IRT models for ability-based guessing. *Applied Psychological Measurement*, *30*, 193–203.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*, 461–464.

Sinharay, S. (2005). Assessing fit of unidimensional item response theory models using a Bayesian approach. *Journal of Educational Measurement*, *42*, 375–394.

Sinharay, S., Johnson, M. S., & Stern, H. S. (2006). Posterior predictive assessment of item response theory models. *Applied Psychological Measurement*, *30*, 298–321.

Spiegelhalter, D. J., Thomas, A., & Best, N. (2003). *WinBUGS version 1.4* [Computer Program.]. Cambridge, UK: MRC Biostatistics Unit, Institute of Public Health.

Tanner, M. A., & Wong, W. H. (1987). The calculation of posterior distributions by data augmentation (with discussion). *Journal of the American Statistical Association*, *82*, 528–550.

Tatsuoka, C. (2002). Data-analytic methods for latent partially ordered classification models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *51*, 337–350.

Tatsuoka, K. K. (1990). Toward an integration of item-response theory and cognitive error diagnosis. In N. Frederiksen, R. Glaser, A. Lesgold, & M. Safto (Eds.), *Monitoring skills and knowledge acquisition* (pp. 453–488). Hillsdale, NJ: Lawrence Erlbaum.

Tatsuoka, K. K. (1995). Architecture of knowledge structures and cognitive diagnosis: A statistical pattern recognition and classification approach. In P. D. Nichol, S. F. Chipman, & R. L. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 327–359). Hillsdale, NJ: Lawrence Erlbaum.

Wang, W.-C., & Jin, K.-Y. (2010a). Multilevel, two-parameter, and random-weights generalizations of the model with internal restrictions on item difficulty. *Applied Psychological Measurement*, *34*, 46–65.

Wang, W.-C., & Jin, K.-Y. (2010b). A generalized model with internal restrictions on item difficulty for polytomous items. *Educational and Psychological Measurement*, *70*, 181–198.

Wang, W.-C., & Liu, C.-Y. (2007). Formulation and application of the generalized multilevel facets model. *Educational and Psychological Measurement*, *67*, 583–605.

Wang, W.-C., Wilson, M. R., & Shih, C.-L. (2006). Modeling randomness in judging rating scales with a random-effects rating scale model. *Journal of Educational Measurement*, *43*, 335–353.

Wentzel, K. R., & Wigfield, A. (Eds.) (2009). *Handbook of motivation at school*. New York, NY: Taylor & Francis.

Authors

HUNG-YU HUANG is Assistant Professor, Department of Psychology and Counseling, University of Taipei, No.1, Ai-Guo West Road, Taipei, 10048, Taiwan; hyhuang@go.utaipei.edu.tw. His primary research interests include cognitive diagnosis modeling, item response theory, and computerized adaptive testing.

WEN-CHUNG WANG is Director of Assessment Research Centre, the Hong Kong Institute of Education, 10 Lo Ping Rd., Tai Po, New Territories, Hong Kong; wcwang@ied.edu.hk. His primary research interests include cognitive diagnosis modeling, item response theory, and computerized adaptive testing.