Article

A Cognitive Diagnosis Model for Identifying Coexisting Skills and Misconceptions

Applied Psychological Measurement
1–13
© The Author(s) 2017
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0146621617722791
journals.sagepub.com/home/apm
SAGE

Bor-Chen Kuo[1], Chun-Hua Chen[1], and Jimmy de la Torre[2]

Abstract

At present, most existing cognitive diagnosis models (CDMs) are designed to either identify the presence and absence of skills or misconceptions, but not both. This article proposes a CDM that can be used to simultaneously identify what skills and misconceptions students possess. In addition, it proposes the use of the expectation-maximization algorithm to estimate the model parameters. A simulation study is conducted to evaluate the viability of the proposed model and algorithm. Real data are analyzed to demonstrate the applicability of the proposed model, and compare it with existing CDMs. Furthermore, a real data–based simulation study is conducted to determine how the correct classification rates in the context of the proposed model can be improved. Issues related to the proposed model and future research are discussed.

Keywords

cognitive diagnosis model, DINA, Bug-DINO, expectation-maximization, agreement rate

Introduction

In mathematics and science domains, misconceptions are viewed as students' conceptions of various phenomena or prior knowledge of these domains that differ from expert knowledge (Andersson, 1986; Smith, diSessa, & Roschelle, 1993). Misconceptions can originate from students' prior learning in their interaction with the world or in the classroom (Smith et al., 1993; Smolleck & Hershberger, 2011; Thompson & Logue, 2006). For example, in elementary mathematics, most misconceptions are incorrect generalizations of prior knowledge that students use to solve new problems (Nesher, 1987); in Newtonian mechanics, students' misconceptions regarding force and motion are formed by their everyday experiences in the physical world (Clement, 1982b, 1987). Students' misconceptions in formal mathematics or science instruction are persistent and resistant to remediation by conventional teaching (Clement, 1987; Potvin, Masson, Lafortune, & Cyr, 2015; Shaughnessy, 1977; Smith et al., 1993). Because misconceptions may produce systematic errors and interfere with learning (Bradshaw & Templin, 2014;

[1]National Taichung University of Education, Taiwan
[2]The University of Hong Kong

Corresponding Author:
Chun-Hua Chen, Graduate Institute of Educational Information and Measurement, National Taichung University of Education, No. 140, Minsheng Rd., West Dist., Taichung City, Taiwan 40306, R.O.C.
Email: cch419@gmail.com

Nesher, 1987; Smith et al., 1993), they must be identified to aid teachers in overcoming them and to improve students' learning.

In mathematics and science education, cognitively diagnostic assessments (CDAs; de la Torre & Minchen, 2014) can be designed to specifically measure students' skills and misconceptions (de la Torre, 2008, 2009; Masters, 2014; Minstrell, 2001). Analyzing data collected using CDAs requires statistical models that are referred to as cognitive diagnosis models (CDMs; de la Torre & Minchen, 2014). In contrast to traditional (i.e., unidimensional) item response theory models, which are used for ordering students along a latent trait continuum, CDMs are latent class models used for classifying students based on the set of skills that they have and have not mastered in a particular domain. CDMs can provide finer grained information more relevant to instruction and learning (de la Torre, 2009). In comparison to many CDMs used for diagnosing skills, only a few CDMs exist for diagnosing misconceptions. Examples of CDMs for diagnosing skills include the deterministic inputs, noisy ''and'' gate model (DINA; Haertel, 1989; Junker & Sijtsma, 2001) and the deterministic input, noisy ''or'' gate model (DINO; Templin & Henson, 2006); examples of CDMs for diagnosing misconceptions include the bug deterministic inputs, noisy ''or'' gate model (Bug-DINO; Kuo, Chen, Yang, & Mok, 2016), and the scaling individuals and classifying misconceptions model (Bradshaw & Templin, 2014).

At present, most of the existing CDMs are designed to either identify skills or misconceptions, but not both. However, an incorrect response to an item may be due to the student lacking the required skills, possessing some misconceptions, or both. Therefore, for feedback from CDMs to be complete and practically useful, it should inform not only on the skills students possess but also misconceptions they have. On such model is the generalized diagnostic classification models for multiple-choice option-based scoring (GDCM-MC), which accounts for both the students' desired (i.e., skills) and problematic (i.e., misconceptions) thinking processes (DiBello, Henson, & Stout, 2015). As it currently stands, the GDCM-MC has a complex formulation, and its parameters cannot be estimated by expectation-maximization (EM) algorithm, as in, it employs Markov chain Monte Carlo. In most cases, the former is more efficient than the latter. More importantly, the possible coexistence of skills and misconceptions is not adequately accounted for when this generalized model reduces to some dichotomous models such as the extended version of the DINA model (called the EDINA-MC model) for modeling dichotomous data. For example, the probabilities of choosing the correct option and the incorrect option that, respectively, measure one skill and one misconception by two students who have both or neither the skill and the misconception are indistinguishable in the EDINA-MC model. However, in practice, misconceptions may sometimes remain for some students who can generate correct solutions (Clement, 1982a). In addition, recent studies have shown that students' misconceptions, such as nonscientific conceptions, may coexist with scientific concepts (Stavy et al., 2006; Stavy & Tirosh, 1996), even after a conceptual change has occurred and correct answers are produced (Brault Foisy, Potvin, Riopel, & Masson, 2015; Potvin et al., 2015; Vosniadou & Verschaffel, 2004).

To address the aforementioned issues, this article aims (a) to propose a new model that can simultaneously identify skills and misconceptions that sometimes may coexist in students, (b) to present an efficient (i.e., EM) algorithm for estimating the parameters of the proposed model, and (c) to compare the performance of the proposed model against existing CDMs that identify skills and misconceptions separately. The remaining sections of this article are structured as follows. The second section discusses the DINA and Bug-DINO models, whereas the third section introduces a CDM that can simultaneously account for the coexistence of skills and misconceptions. The fourth section evaluates the performance of the proposed model using a simulation study, whereas the fifth section illustrates the application of the proposed model using fraction

multiplication data. The sixth section demonstrates how the correct classification rates in the context of the proposed model can be improved using a real data-based simulation study. This article concludes with a discussion of the implications of the findings and directions for future research.

Background

This section gives a background on two existing CDMs: the DINA model for identifying skills and the Bug-DINO model for identifying misconceptions. Let J be the number of items, K the number of skills or misconceptions of a test, and $\mathbf{X}_i$ the vector of J dichotomous item responses of student i. Both the DINA and Bug-DINO models relate $\mathbf{X}_i$ to $\boldsymbol{\alpha}_i = \{\alpha_{i1}, \alpha_{i2}, \ldots, \alpha_{iK}\}'$, the attribute vector of examinee i, by assuming that the elements of $\mathbf{X}_i$ are statistically independent given $\boldsymbol{\alpha}_i$. For implementing the DINA and Bug-DINO models, a Q-matrix (Tatsuoka, 1983) is required. The Q-matrix is a $J \times K$ matrix, and the element $q_{jk} = 1$, if that skill or misconception k is measured by item j; otherwise $q_{jk} = 0$.

The DINA Model

In the CDM literature, the DINA model is a conjunctive model that is commonly used for identifying the presence or absence of skills. The conjunctive nature of the DINA model assumes that students can be expected to correctly answer an item only when they have mastered all the skills required by the item; otherwise, they cannot, even if they only lack one of the required skills. Nevertheless, students who have mastered all the required skills may slip and answer the item incorrectly; in the same manner, students who lack one or more of the required skills may guess and answer the item correctly. In the DINA model, the probability of a correct response to item j by student i is defined as follows:

$$P\left(X_{ij} = 1 | \boldsymbol{\alpha}_i\right) = \left(1 - s_j\right)^{\eta_{ij}} g_j^{(1-\eta_{ij})}, \tag{1}$$

where η_{ij} is the deterministic latent response variable that divides students into two groups, and is mathematically expressed as $\eta_{ij} = \prod_{k=1}^{K} \alpha_{ik}^{q_{jk}}$, with $\eta_{ij} = 1$ as an indicator that student i masters all the required skills of item j, and $\eta_{ij} = 0$ otherwise. The item parameters s_j and g_j introduce stochasticity in the response process, and are defined as $s_j = P(X_{ij} = 0 | \eta_{ij} = 1)$ and $g_j = P(X_{ij} = 1 | \eta_{ij} = 0)$, respectively. The item parameter s_j represents the probability of incorrectly answering item j due to slips, and the item parameter g_j the probability of correctly answering item j by guessing.

The Bug-DINO Model

In contrast to the DINA model, the Bug-DINO model is a disjunctive model and has been used for identifying misconceptions. The disjunctive nature of the Bug-DINO model assumes that students are expected to correctly answer an item only if they do not possess any of the misconceptions measured by the item. In other words, students cannot correctly answer the item if they possess one or more of the measured misconceptions. However, students may slip and incorrectly answer the item even they do not possess any measured misconceptions, or they may guess the correct answer to the item even they possess one or more misconceptions. In the Bug-DINO model, the probability of correctly answering item j by student i is defined as follows:

$$P(X_{ij}=1|\boldsymbol{\alpha}_i)=\left(1-s_j^*\right)^{1-\gamma_{ij}}\left(g_j^*\right)^{\gamma_{ij}}, \tag{2}$$

where γ_{ij} is the deterministic latent response variable that divides students into two groups, and can be mathematically computed as $\gamma_{ij}=1-\prod_{k=1}^{K}(1-\alpha_{ik})^{q_{jk}}$. In the Bug-DINO model, $\alpha_{ik}=1$ (or 0) represents that student i possesses (or does not possess) misconception k, and $\gamma_{ij}=0$ (or 1) indicates that student i possesses none (or at least one) of the misconceptions measured by item j. The parameters s_j^* and g_j^*, also referred to as the slip and guessing probabilities, are defined as $s_j^*=P(X_{ij}=0|\gamma_{ij}=0)$ and $g_j^*=P(X_{ij}=1|\gamma_{ij}=1)$, respectively. The Bug-DINO model is a modification of the DINO model (Templin & Henson, 2006), which has been used for identifying disorders in the medical, clinical, and psychological fields (de la Torre, van der Ark, & Rossi, 2015; Templin & Henson, 2006). The Bug-DINO model differs from the DINO model in that, although the formulas for γ_{ij} in the Bug-DINO and DINO models are the same, the association of γ_{ij} with s_j^* and g_j^* in the Bug-DINO model differs from that in the DINO model. Specifically, $\gamma_{ij}=0$ and $\gamma_{ij}=1$ are associated with the s_j^* and g_j^* in the Bug-DINO model, whereas the reverse is true for the DINO model. It can be noted that the Bug-DINO model can also be viewed as a reformulation of the DINA model. Specifically, if $\eta_{ij}=\prod_{k=1}^{k}\alpha_{ik}{}^{qjk}$ for the DINA model is modified as $\eta_{ij}=\prod_{k=1}^{k}(1-\alpha'_{ik})^{qjk}$, the modified DINA model is equal to the Bug-DINO model. However, to also acknowledge the existence of the model in the literature, the name ''Bug-DINO'' has been retained in this study.

A New Model for Simultaneously Identifying Skills and Misconceptions (SISM)

As discussed above, the DINA model and the Bug-DINO model can be separately used to identify skills and misconceptions, respectively. However, solely identifying skills may ignore whether or not students have misconceptions, and vice versa. Therefore, to provide more complete diagnostic information, simultaneously identifying skills and misconceptions, which can possibly coexist, is necessary. This section introduces a new model for simultaneously identifying skills and misconceptions (SISM), to accomplish this goal.

As discussed earlier, some misconceptions (e.g., those resulting from instruction) cannot coexist with some skills. That is, some misconceptions are related to (or not independent of) certain skills so students cannot have these skills and misconceptions at the same time. Other misconceptions (e.g., nonscientific conceptions originating from intuitive thinking) can coexist with some skills. This is to say that some misconceptions are not related to (or independent of) some skills so students may acquire skills while still hold on to some related misconceptions. Thus, an item that simultaneously measures skills and misconceptions have four associated success probabilities: (a) success probability of students who have mastered all the measured skills and possess none of the measured misconceptions, (b) success probability of students who have mastered all the measured skills but possess some of the measured misconceptions, (c) success probability of students who have not mastered all the measured skills and possess none of the measured misconceptions, and (d) success probability of students who have not mastered all the measured skills and possess at least one of the measured misconceptions. To account for the different success probabilities, the item response function of the SISM model requires four parameters.

For notational convenience, it is assumed without loss of generality that the first K_S elements of $\boldsymbol{\alpha}_i$ correspond to the skills, and the last $K-K_S=K_M$ attributes correspond to the misconceptions. The item response function of the SISM model is defined as follows:

$$P\left(X_{ij}=1|\boldsymbol{\alpha}_i\right)=h_j^{\eta_{ij}\left(1-\gamma_{ij}\right)}\omega_j^{\eta_{ij}\gamma_{ij}}g_j^{\left(1-\eta_{ij}\right)\left(1-\gamma_{ij}\right)}\varepsilon_j^{\left(1-\eta_{ij}\right)\gamma_{ij}}, \tag{3}$$

where the parameters η_{ij} and γ_{ij} are deterministic latent response variables defined as $\eta_{ij}=\prod_{k=1}^{K_S}\alpha_{ik}^{q_{jk}}$ and $\gamma_{ij}=1-\prod_{k=Ks+1}^{K}(1-\alpha_{ik})^{q_{jk}}$, respectively, and can be interpreted as before. The SISM model divides the examinees into four latent groups based on the different combinations of η_{ij} and γ_{ij}, and its four parameters, h_j, ω_j, g_j, and ε_j, are defined as follows:

$$h_j=P\left(X_{ij}=1\left|\eta_{ij}=1,\gamma_{ij}=0\right.\right), \tag{4}$$

$$\omega_j=P\left(X_{ij}=1\left|\eta_{ij}=1,\gamma_{ij}=1\right.\right), \tag{5}$$

$$g_j=P\left(X_{ij}=1\left|\eta_{ij}=0,\gamma_{ij}=0\right.\right), \text{ and} \tag{6}$$

$$\varepsilon_j=P\left(X_{ij}=1\left|\eta_{ij}=0,\gamma_{ij}=1\right.\right). \tag{7}$$

An EM algorithm was developed to estimate the SISM model parameters. The computational details of the algorithm and the associated standard errors (*SE*s) are given in Appendix A in the supplementary material.

Special Cases

If an item simultaneously measures skills and misconceptions, and some of these misconceptions are not related to all the measured skills, the four item parameters are needed in the SISM model. However, if all these measured misconceptions are related to the measured skills, the item parameter ω_j can be dropped. In such a case, the SISM model becomes a reduced model, which can be referred to as the reduced SISM (rSISM) model.

In the SISM model, if an item measures skills only and does not measure any misconceptions, the value of γ_{ij} is always equal to 0, and the formula of the SISM model can be written as $P(X_{ij}=1|\boldsymbol{\alpha}_i)=h_j^{\eta_{ij}}g_j^{(1-\eta_{ij})}$, which is equivalent to the DINA model. In contrast, if an item only measures misconceptions, the value of η_{ij} is always equal to 1, and the formula of the SISM model can be written as $P(X_{ij}=1|\boldsymbol{\alpha}_i)=h_j^{(1-\gamma_{ij})}\omega_j^{\gamma_{ij}}$, which is equivalent to the Bug-DINO model. Therefore, the DINA and Bug-DINO models can be obtained from the SISM model when γ_{ij} and η_{ij} are set to 0 and 1, respectively.

Simulation Study

In this section, a simulation study was designed to evaluate the performance of the SISM model. Specifically, it sought to examine how well the item parameters of the proposed model can be estimated, and student classification can be recovered.

Design

In the simulation study, 12 conditions were examined by combining three factors: (a) test length (J = 20 and 40 items), (b) sample size (I = 500, 1,000, and 2,000 students), and (c) item quality (high and low). For each of simulation conditions, 500 data sets were generated and analyzed using the SISM model.

Two Q-matrices were created to mimic realistic short and long tests with 20 and 40 items, respectively. The Q-matrix for the short test is given in Table B1 (see Appendix B in the supplementary material), which measured four skills and three misconceptions. The skills and misconceptions were related as follows: (a) Misconception 1 is related to Skills 1 and 2, but not to Skills 3 and 4; (b) Misconception 2 is related to Skill 4, but not Skills 1, 2, and 3; and (c) Misconception 3 is not related to any of the skills.

In the Q-matrix given in Table B1, Items 1 through 4 measured one skill only, Items 5 through 7 measured one misconception only, Items 8 through 11 measured one skill and one misconception, Items 12 through 15 measured two skills and one misconception, Items 16 through 18 measured two skills and two misconceptions, and Items 19 through 20 measured three skills and two misconceptions. Each of the skills and misconceptions was measured by seven items. The Q-matrix for $J = 20$ was doubled to create the Q-matrix for $J = 40$.

Because $K = 7$, the maximum number of possible attribute patterns is $2^7 = 128$. However, due to the relationships between the skills and misconceptions, some attribute patterns are deemed unreasonable. Specifically, attribute patterns that include the presence of both Skill 1 and Misconception 1, or Skill 2 and Misconception 1, or both Skill 4 and Misconception 2 were dropped. As a result, only 60 possible attribute patterns were deemed reasonable, and an equal number of students were generated for each of these attribute patterns. Note that when using the EM algorithm to estimate the SISM parameters, the prior probabilities of these reasonable attribute patterns can be set as equal and the other unreasonable attribute patterns are set to zeros.

Across all the items, high-quality items were assigned the parameters $h = 0.95$, $\omega = 0.15$, $g = 0.35$, and $\varepsilon = 0.05$, whereas low-quality items the parameters $h = 0.85$, $\omega = 0.2$, $g = 0.4$, and $\varepsilon = 0.1$. These settings were similar to those in de la Torre and Lee (2010) and Huo and de la Torre (2014). It should be noted that not all items in both two tests have four item parameters. For example, items that did not measure misconceptions (i.e., Items 1-4) were fitted with the SISM model with the two parameters h and g only (i.e., the DINA model), items that did not measure skills (i.e., Items 5-7) were fitted with the SISM model with the two parameters h and ω only (i.e., the Bug-DINO model), and items that measured dependent skills and misconceptions (i.e., Items 8, 9, 11, 12, 19, and 20) were fitted with the SISM model with the three parameters h, g, and ε only (i.e., the rSISM model). Only the items (i.e., Items 10, 13, 14, 15, 16, and 17) in that the measured misconception B3 is independent of the measured skills and Item 18 in which the measured misconception B2 is also independent of the measured skills (i.e., Skills 1 and 3), were fitted with the full SISM model.

Evaluation Indices

To examine how well the model parameters were estimated, item parameter recovery was assessed using the absolute deviation (AD), *SE*, and *SD*, whereas for skill and misconception, classification accuracy was assessed using correct attribute classification rate (CACR) and correct pattern classification rate (CPCR). Note that the CACR and CPCR indices can be, respectively, referred to as correct attribute agreement rate (CAAR) and correct pattern agreement rate (CPAR; Kuo et al., 2016) when comparing attributes identified via fitted models with classification results produced by human raters.

Results

Table 1 shows the ADs of the item parameter estimates for the condition where the test length was 20, the sample sizes were 500, 1,000, and 2,000, and the item quality was low. The ADs between the true and estimated item parameters for the three sample sizes were quite small—most ADs

Table 1. ADs for I = 500, 1,000, 2,000; J = 20; and Low-Quality Items (500 Simulation Trials).

	AD (h)			AD (ω)			AD (g)			AD (ε)		
	I			I			I			I		
Item	500	1,000	2,000	500	1,000	2,000	500	1,000	2,000	500	1,000	2,000
1	.00	.00	.00	—	—	—	.00	.00	.00	—	—	—
2	.00	.00	.00	—	—	—	.00	.00	.00	—	—	—
3	.02	.01	.00	—	—	—	.04	.02	.02	—	—	—
4	.01	.00	.00	—	—	—	.00	.00	.00	—	—	—
5	.00	.00	.00	.02	.01	.01	—	—	—	—	—	—
6	.00	.00	.00	.02	.01	.01	—	—	—	—	—	—
7	.00	.00	.00	.00	.00	.00	—	—	—	—	—	—
8	.01	.00	.00	—	—	—	.02	.01	.01	.00	.00	.00
9	.01	.01	.00	—	—	—	.02	.01	.01	.00	.00	.00
10	.00	.00	.00	.01	.00	.00	.01	.00	.00	.00	.00	.00
11	.01	.01	.00	—	—	—	.03	.02	.01	.00	.00	.00
12	.00	.00	.00	—	—	—	.00	.00	.00	.00	.00	.00
13	.01	.00	.00	.01	.00	.00	.00	.00	.00	.00	.00	.00
14	.00	.00	.00	.02	.01	.00	.00	.00	.00	.00	.00	.00
15	.00	.00	.01	.01	.00	.00	.00	.00	.00	.00	.00	.00
16	.01	.00	.00	.00	.01	.00	.00	.00	.00	.00	.00	.00
17	.00	.01	.00	.02	.01	.00	.00	.00	.00	.00	.00	.00
18	.01	.01	.01	.00	.00	.00	.00	.00	.00	.00	.00	.00
19	.02	.00	.00	—	—	—	.00	.00	.00	.00	.00	.00
20	.02	.00	.00	—	—	—	.01	.00	.00	.00	.00	.00
M	.01	.00	.00	.01	.01	.00	.01	.00	.00	.00	.00	.00

Note. AD = absolute deviation; I = sample size; J = test length.

were .00 and the maximum AD was .04. When the AD was averaged across the 20 items, most of the values were .00 and the maximum was .01.

For an overall view, Table 2 summarizes the results of average ADs across all items (also called MADs) for 12 simulation conditions. As shown in Table 2, all the average AD values were small—most values were .00 and the maximum was .01. Although the improvement was rather small, the average AD decreased with longer test length, larger sample sizes, and higher item quality. As a whole, these results showed that the item parameters of the SISM model can be accurately estimated using the proposed EM algorithm.

The *SE* and *SD* were used to examine the theoretical and empirical *SE*s, respectively, of the item parameter estimates. Table 3 gives the average *SE* and *SD* of h, ω, g, and ε across all items for the different simulation conditions. The average *SE*s were from .01 to .06 (when item quality was high) and from .02 to .09 (when item quality was low) under the three sample sizes and two test lengths; in contrast, the average *SD*s were from .01 to .07 and from .01 to .13 under the same conditions. An increase in the test length, sample size, or item quality resulted in smaller *SE*s and *SD*s. Except for one condition (i.e., I = 500, J = 40, high-quality items and for h), the *SE* was consistently smaller than or equal to the *SD*. This indicates that the theoretical *SE* tended to underestimate the empirical *SE*. The discrepancy was most obvious when small sample size, short test, and low item quality were involved. With larger sample size, longer test, and better item quality, the reported *SE* approximated the variability that can be expected from the item parameter estimates across different samples.

Table 4 gives the average CACR ($\overline{\text{CACR}}$) and CPCR and their *SD* for the different simulation conditions. The $\overline{\text{CACR}}$ is the average of the CACRs of all attributes across all students,

Table 2. Average ADs of Item Parameters Across All Items for the 12 Simulation Conditions.

		MAD (*h*)			MAD (ω)			AD (*g*)			AD (ε)		
		I			*I*			*I*			*I*		
J	Item quality	500	1,000	2,000	500	1,000	2,000	500	1,000	2,000	500	1,000	2,000
20	HQ	.01	.00	.00	.00	.00	.00	.01	.00	.00	.00	.00	.00
	LQ	.01	.00	.00	.01	.01	.00	.01	.00	.00	.00	.00	.00
40	HQ	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
	LQ	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00

Note. AD = absolute deviation; *J* = test length; MAD = mean absolute deviation; *I* = sample size; HQ = high quality; LQ = low quality.

Table 3. Average *SE*s and *SD*s of Item Parameter Estimates Across All Items for the 12 Simulation Conditions.

			h		ω		*g*		ε	
I	*J*	Item quality	*SE*	*SD*	*SE*	*SD*	*SE*	*SD*	*SE*	*SD*
500	20	HQ	.04	.04	.06	.07	.04	.04	.02	.02
		LQ	.06	.09	.09	.13	.04	.07	.03	.04
	40	HQ	.04	.03	.05	.05	.03	.03	.02	.02
		LQ	.05	.05	.06	.07	.04	.04	.03	.03
1,000	20	HQ	.03	.03	.04	.04	.03	.03	.01	.02
		LQ	.05	.06	.06	.08	.03	.04	.02	.03
	40	HQ	.02	.02	.03	.04	.02	.02	.01	.01
		LQ	.03	.04	.04	.05	.03	.03	.02	.02
2,000	20	HQ	.02	.02	.03	.03	.02	.02	.01	.01
		LQ	.03	.04	.04	.05	.02	.03	.02	.02
	40	HQ	.01	.01	.02	.02	.02	.02	.01	.01
		LQ	.02	.03	.03	.03	.02	.02	.01	.01

Note. *I* = sample size; *J* = test length; HQ = high quality; LQ = low quality.

Table 4. Mean and *SD* of CCRs Across the 12 Simulation Conditions (500 Simulation Trials).

		$\overline{\text{CACR}}$			*SD* ($\overline{\text{CACR}}$)			CPCR			*SD* (CPCR)		
		I			*I*			*I*			*I*		
J	Item quality	500	1,000	2,000	500	1,000	2,000	500	1,000	2,000	500	1,000	2,000
20	HQ	.91	.92	.92	.01	.00	.00	.55	.58	.59	.02	.02	.01
	LQ	.83	.85	.85	.01	.01	.00	.29	.32	.34	.03	.02	.01
40	HQ	.97	.97	.97	.00	.00	.00	.80	.82	.82	.02	.01	.01
	LQ	.91	.92	.92	.01	.00	.00	.53	.56	.57	.02	.02	.01

Note. CCR = correct classification rate; CACR = correct attribute classification rate; CPCR = correct pattern classification rate; *I* = sample size; *J* = test length; HQ = high quality; LQ = low quality.

and CPCR means the correct pattern classification rates of whole patterns across all students. Regardless of the sample size, high item quality resulted in $\overline{\text{CACR}}$ and CPCR that were at least .91 and .55, respectively, when the test was short. The corresponding *SD*s were .01 and .02,

respectively, indicating that results were consistent across replications. When longer test was involved under the same conditions, huge improvements in classification accuracy were observed (i.e., the $\overline{\text{CACR}}$ and CPCR that were at least .97 and .80); only slight improvements, if at all, in the corresponding *SD*s can be observed. As expected, lower classification accuracies were obtained when lower quality items were used. However, the $\overline{\text{CACR}}$ can be considered acceptable because it was never lower than .83. It appears that the item quality had relatively small impact on the *SD* of the $\overline{\text{CACR}}$ and CPCR. The table also shows that increasing the sample size can improve the attribute classification accuracy and decrease the associated *SD* particularly when CPCR was involved. Overall, it can be concluded that the SISM model can provide high attribute classification accuracy if high-quality items are involved, notwithstanding a small sample size. Moreover, even if the item quality was low, the attribute classification accuracy can be high provided when a long test was used.

Fraction Multiplication Data

Description of Data

To demonstrate the applicability of the model to real data, a subset of the original data described and used by Lin (2012) was analyzed. The data comprised of responses of 286 students that were selected from three Taiwanese elementary schools to seven fraction multiplication items. Each item was an open-ended problem. Item 3 is given as an example in Figure C1 (given in Appendix C in the supplementary material). When students answered the item, they were required to write down their problem-solving process, as well as select an answer choice. One of the original purposes of the fraction multiplication test was to identify students' skills and misconceptions through their problem-solving process. Because writing the problem-solving process for item required a lot of time, only seven items were used in the test. Although the different steps involved in the problem-solving process can provide more diagnostic information, only the correct/incorrect responses were used for the purpose of comparing three CDMs: DINA model for identifying skills only, Bug-DINO for identifying misconceptions only, and SISM model for identifying both skills and misconceptions. The model-based classifications were compared with the skill and misconception classifications provided by human raters who were experts in this area.

The fraction multiplication test measured four skills and three misconceptions. The four skills were S1 (ability to multiply a whole number by a fraction), S2 (ability to multiply a fraction by a fraction), S3 (ability to reduce the answer to its lowest terms), and S4 (ability to solve a two-step problem); and the three misconceptions were B1 (turning the second fraction upside down when multiplying a fraction by a fraction), B2 (solving only the first step of a two-step problem), and B3 (performing incorrect arithmetic operations when confused about the relational terms). The relationships between the four skills and misconceptions are as defined in the simulation study. The Q-matrix for the fraction multiplication data is shown in Table B2.

Analysis

In this study, EM algorithms were used to obtain the parameter estimates of the DINA, Bug-DINO, and SISM models. The Q-matrix shown in Table B2 was used in conjunction with the SISM model. For the DINA and Bug-DINO model, two submatrices were extracted from the original Q-matrix. The first submatrix that involved only the four skills was used in conjunction with the DINA model; and the second submatrix that only involved the three misconceptions

Table 5. CAARs of Skills and Misconceptions for Fraction Multiplication Data Fitted by the DINA, Bug-DINO, and SISM Models.

	Attribute						
	Skill				Misconception		
Fitted model	α_1	α_2	α_3	α_4	α_5	α_6	α_7
DINA	.703	.818	.801	.731	—	—	—
Bug-DINO	—	—	—	—	.825	.591	.437
SISM	.717	.839	.818	.769	.895	.951	.549

Note. CAAR = correct attribute agreement rate; DINA = deterministic inputs, noisy "and" gate model; DINO = deterministic input, noisy "or" gate model; Bug-DINO = bug deterministic inputs, noisy "or" gate model; SISM = simultaneously identifying skills and misconceptions.

was used in conjunction with the Bug-DINO model. The CAAR index was used to evaluate the agreement between model-based and human rater classifications for each attribute.

Results

Table 5 gives the CAAR values for the three fitted CDMs. The results show that the SISM model had higher agreement rates with the human raters than the DINA model when classifying skills. Similarly, the SISM model had higher agreement rates with the human raters than the Bug-DINO model when classifying misconceptions. It can be noted the larger discrepancies can be observed in classifying misconceptions.

Real Data–Based Simulation Study

Design

As can be seen in Table 5, in the SISM model, only the CAAR of Misconception 6 was higher than .9 and the minimum CAAR was lower than .6, which may be considered relatively low for practical purposes. These results may be due to the short test length and small sample size. To determine how the correct classification rates in the context of the SISM model can be improved, a targeted, real data-based simulation study with four conditions that examined three factors (i.e., sample size, test length, and Q-matrix design) was carried out.

A multinomial distribution based on the posterior probabilities of all reasonable attribute patterns for the fraction multiplication data estimated by the SISM model was used to generate the attribute patterns for the simulation study. In the first condition, each data set involved 286 simulated students, seven items, and seven attributes (including four skills and three misconceptions). This condition also used the original Q-matrix (denoted as Q1) and item parameters estimated from the real data. The second condition was identical to the first condition except that, instead of 286, 1,000 simulated students were used. The third condition was identical to the second condition except that, instead of seven, 14 items were used. The Q-matrix for this condition (denoted by Q3) was created by stacking one Q1 on top of another Q1. In addition, identical parameters for Item 1 through 7 were used for Items 8 through 14. Finally, the fourth condition was identical to the third condition except for the Q-matrix. To obtain Q4, the Q-matrix for the fourth condition, Q3 was a modified to include more items with simpler (i.e., fewer) attribute specifications. In particular, the specifications for Items 1 through 7 and 13 were modified (see

Table 6. CACR for Real Data–Based Simulation Studies.

			Attribute						
			Skill				Misconception		
Condition	*I*	*J*	S1	S2	S3	S4	B1	B2	B3
1	286	7	.823	.915	.877	.832	.956	.962	.807
2	1,000	7	.829	.919	.878	.840	.968	.984	.813
3	1,000	14	.851	.939	.902	.870	.991	.983	.870
4	1,000	14	.933	.930	.926	.915	.999	.985	.939

Note. CACR = correct attribute classification rate; *I* = sample size; *J* = test length.

Table B3). Note that the parameters for items with modified attribute specifications were the relevant parameters from the corresponding items in the third condition. For example, the parameters of Item 1 in Condition 4, h_1 and g_1, were obtained from (h_1, ω_1, g_1, and ε_1), which were the parameters of Item 1 in Condition 3.

Results

Table 6 shows the CACR results for the four conditions of the real data-based simulation study. Results from Conditions 1 and 2 show that increasing the sample size by almost fourfold can only slightly improve the CACRs across all attributes. Results from Conditions 2 and 3 show that doubling the test length can result in obvious improvements in CACRs. Finally, results from Conditions 3 and 4 show that a carefully constructed Q-matrix can result in a dramatic improvement in the overall CACR—although the CACR of S2 in Condition 4 was lower than that in Condition 3 (.939 vs. .930), the mean CACR in Condition 4 was much higher than mean CACR in Condition 3 (.947 vs. .915).

Discussion and Conclusion

In many disciplines, such as mathematics and science education, both skills and misconceptions crucially affect students' learning. In the CDM literature, most models are used to exclusively identify either skills or misconceptions. For simultaneously identifying skills and misconceptions, which may coexist, the SISM model was proposed in this article. An EM algorithm was also proposed for estimating the SISM model parameters. The simulation study showed that accurate parameter estimates of the SISM model can be obtained, and the corresponding *SE*s can be reliable, particularly when the sample size was relatively large, the test was long, or the item quality was high. Furthermore, in the first simulation study, an equal number of students were generated for each pattern. Because the number of students for estimating parameter ω is smaller than those for other parameters, the *SE* and *SD* of estimates of parameter ω were the largest compared with other parameters. The classification accuracy of skills and misconceptions provided by the SISM model ranged from moderately high to high.

The fraction multiplication data analysis demonstrated that the classification agreements for both skills and misconceptions provided by the SISM model were better than those provided separately by the DINA and Bug-DINO models. In addition, the results indicated that the SISM model had acceptable classification agreement even when short tests were involved. Finally, the results of the real data-based simulation study showed that the classification accuracy can be improved mostly by increasing the test length, and using a better-designed Q-matrix.

The SISM model is a simple and flexible dichotomous model that can be applied with diagnostic assessment for the purpose of diagnosing either or both skills and misconceptions. Although promising, additional work is needed to better understand the model properties before it can be used with sufficient confidence in practice. For example, in many applied settings, the total number of skills and misconceptions can be large. It remains to be seen to what extent the computational challenge associated with such a situation can affect the practical viability of the SISM model. In addition, as is true in many domains, students use various strategies in solving different problems. Such strategies may involve different subsets of skills and misconceptions. Diagnostic feedback that students and teachers are provided with would be more useful if, on top of skills and misconceptions, multiple ways of solving a problem can be incorporated in the CDM.

Acknowledgments

The authors are very grateful to the editors and anonymous reviewers for providing insightful comments and valuable suggestions.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research is partially supported by the Ministry of Science and Technology, Taiwan, under Grants MOST 102-2511-S-142-008-MY3 and MOST 105-2511-S-142-009-MY3.

Supplemental Material

Supplementary material is available for this article online.

References

Andersson, B. (1986). The experiential gestalt of causation: A common core to pupils' preconceptions in science. *European Journal of Science Education*, *8*, 155-171.

Bradshaw, L., & Templin, L. (2014). Combining item response theory and diagnostic classification models: A psychometric model for scaling ability and diagnosing misconceptions. *Psychometrika*, *79*, 403-425.

Brault Foisy, L.-M., Potvin, P., Riopel, M., & Masson, S. (2015). Is inhibition involved in overcoming a common physics misconception in mechanics? *Trends in Neuroscience & Education*, *4*, 26-36.

Clement, J. (1982a). Algebra word-problems solutions: Thought processes underlying a common misconception. *Journal for Research in Mathematics Education*, *13*, 16-30.

Clement, J. (1982b). Students' preconceptions in introductory mechanics. *American Journal of Physics*, *50*, 66-71.

Clement, J. (1987, April). *The use of analogies and anchoring intuitions to remediate misconceptions in mechanics*. Paper presented at the Annual Meeting of American Educational Research Association, Washington, DC.

de la Torre, J. (2008). An empirically based method of Q-matrix validation for the DINA model: Development and applications. *Journal of Educational Measurement*, *45*, 343-362.

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., & Huo, Y. (2014). Estimating a cognitive diagnostic model for multiple strategies via the EM algorithm. *Applied Psychological Measurement*, *38*, 464-485.

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.

de la Torre, J., & Minchen, N. (2014). Cognitively diagnostic assessments and the cognitive diagnosis model Framework. *Psicología Educativa*, *20*, 89-97.

de la Torre, J., van der Ark, L. A., & Rossi, G. (2015). Analysis of clinical data from cognitive diagnosis modeling framework. *Measurement and Evaluation in Counseling and Development*. Advance online publication. doi:10.1177/0748175615569110

DiBello, L. V., Henson, R. A., & Stout, W. F. (2015). A family of generalized diagnostic classification models for multiple choice option-based scoring. *Applied Psychological Measurement*, *39*, 62-79.

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

Huo, Y., & de la Torre, J. (2014). Estimating a cognitive diagnostic model for multiple strategies via the EM algorithm. *Applied Psychological Measurement*, *38*, 464-485.

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.

Kuo, B.-C., Chen, C.-H., Yang, C.-W., & Mok, M. M. C. (2016). Cognitive diagnostic models for tests with multiple choice and constructed response items. *Educational Psychology*, *36*, 1115-1133.

Lin, H.-S. (2012). *An analysis on the effect of different on-line diagnostic test items of multiplication and division of fraction* (Unpublished master's thesis). National Taichung University of Education, Taiwan.

Masters, J. (2014, April). *The diagnostic geometry assessment system: Results from a randomized controlled trial*. Paper presented at the annual meeting of the American Educational Research Association, Philadelphia, PA.

Minstrell, J. (2001). Facets of students' thinking: Designing to cross the gap from research to standards-based practice. In K. Crowley, C. D. Schunn, & T. Okada (Eds.), *Designing for science: Implications for professional, instructional, and everyday science* (pp. 415-443). Mahwah, NJ: Lawrence Erlbaum.

Nesher, P. (1987). Towards an instructional theory: The role of students' misconceptions. *For the Learning of Mathematics*, *7*, 33-40.

Potvin, P., Masson, S., Lafortune, S., & Cyr, G. (2015). Persistence of the intuitive conception that heavier objects sink more: A reaction time study with different levels of interference. *International Journal of Science and Mathematics Education*, *13*, 21-43.

Shaughnessy, J. M. (1977). Misconceptions of probability: An experiment with a small-group, activity-based model building approach to introductory probability. *Educational Studies in Mathematics*, *8*, 295-316.

Smith, J. P., diSessa, A. A., & Roschelle, J. (1993). Misconception reconceived: A constructivist analysis of knowledge in transition. *Journal of the Learning Science*, *3*, 115-163.

Smolleck, L., & Hershberger, V. (2011). Playing with science: An investigation of young children's science conceptions and misconceptions. *Current Issues in Education*, *14*. Retrieved from http://cie.asu.edu/ojs/index.php/cieatasu/article/view

Stavy, R., Babai, R., Tsamir, P., Tirosh, D., Lin, F.-L., & McRobbie, C. (2006). Are intuitive rules universal? *International Journal of Science and Mathematics Education*, *4*, 417-436.

Stavy, R., & Tirosh, D. (1996). Intuitive rules in science and mathematics: The case of ''more of A—More of B''. *International Journal of Science Education*, *18*, 653-667.

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. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods*, *11*, 287-305.

Thompson, F., & Logue, S. (2006). An exploration of common student misconceptions in science. *International Education Journal*, *7*, 553-559.

Vosniadou, S., & Verschaffel, L. (2004). Extending the conceptual change approach to mathematics learning and teaching. *Learning and Instructions*, *14*, 445-451.