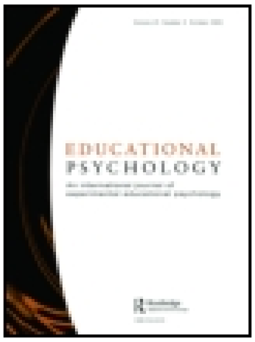

Routledge
Taylor & Francis Group

Educational Psychology

An International Journal of Experimental Educational Psychology

ISSN: 0144-3410 (Print) 1469-5820 (Online) Journal homepage: http://www.tandfonline.com/loi/cedp20

Cognitive diagnostic models for tests with multiple-choice and constructed-response items

Bor-Chen Kuo, Chun-Hua Chen, Chih-Wei Yang & Magdalena Mo Ching Mok

To cite this article: Bor-Chen Kuo, Chun-Hua Chen, Chih-Wei Yang & Magdalena Mo Ching Mok (2016): Cognitive diagnostic models for tests with multiple-choice and constructed-response items, Educational Psychology, DOI: 10.1080/01443410.2016.1166176

To link to this article: http://dx.doi.org/10.1080/01443410.2016.1166176

Published online: 18 Apr 2016.

Educational Psychology, 2016
http://dx.doi.org/10.1080/01443410.2016.1166176

Routledge
Taylor & Francis Group

Cognitive diagnostic models for tests with multiple-choice and constructed-response items

Bor-Chen Kuo[a]*, Chun-Hua Chen[a], Chih-Wei Yang[a] and
Magdalena Mo Ching Mok[b]

[a]Graduate Institute of Educational Information and Measurement, National Taichung University of Education, Taichung City, Taiwan, ROC; [b]Assessment Research Centre, Department of Psychological Studies, Christian Fellowship & Development Centre, The Hong Kong Institute of Education, Hong Kong, Hong Kong, China

(*Received 17 July 2014; final version received 9 March 2016*)

Traditionally, teachers evaluate students' abilities via their total test scores. Recently, cognitive diagnostic models (CDMs) have begun to provide information about the presence or absence of students' skills or misconceptions. Nevertheless, CDMs are typically applied to tests with multiple-choice (MC) items, which provide less diagnostic information than constructed-response (CR) items. This paper introduces new CDMs for tests with both MC and CR items, and illustrates how to use them to analyse MC and CR data, and thus, identify students' skills and misconceptions in a mathematics domain. Analyses of real data, the responses of 497 sixth-grade students randomly selected from four Taiwanese primary schools to eight direct proportion items, were conducted to demonstrate the application of the new models. The results show that the new models can better determine students' skills and misconceptions, in that they have higher inter-rater agreement rates than traditional CDMs.

Keywords: cognitive diagnosis; multiple-choice item; constructed-response item

Feedback is defined as 'information provided by an agent (e.g. teacher, peer, book, parent, self, experience) regarding aspects of one's performance or understanding' (Hattie & Timperley, 2007, p. 81). According to recent research (Golke, Dörfler, & Artelt, 2015; Harks, Rakoczy, Hattie, Besser, & Klieme, 2014; Lee, 2016; Pekrun, Cusack, Murayama, Elliot, & Thomas, 2014; Timmers, Walraven, & Veldkamp, 2015; Van der Kleij, Feskens, & Eggen, 2015), quality feedback provided to the student is one of the most important predictors of subsequent learning. Good feedback has been found to affect later learning significantly (Rust, 2007). Nevertheless, in traditional assessment, teachers typically provide grades using total scores. This type of feedback is too gross, and may not be specific enough to provide diagnostic information to support later learning.

Cognitively diagnostic assessments (CDAs) are an integration of cognitive psychology and educational measurement that can provide feedback with informative diagnostic information to help teachers understand the learning status of students (Ketterlin-Geller & Yovanoff, 2009; Roberts et al., 2014). In mathematics education,

*Corresponding author. Email: kbc@mail.ntcu.edu.tw

CDAs are designed to measure students' domain-specific knowledge states, which are formed by a combination of skills and misconceptions. Skills are required for successfully completing cognitive tasks, whereas misconceptions are persistent incorrect ideas that produce error patterns (Ashlock, 1994; Huang & Wu, 2013).Because students who lack skills or possess misconceptions may struggle in learning and systematically provide incorrect answers to items on a test rather than accidental errors due to slips or guesses (Arieli-Attali & Liu, 2015; Bradshaw & Templin, 2014), identifying skills and misconceptions is essential and can help teachers adjust their classroom instructions to improve students' learning.

Traditionally, CDAs use multiple-choice (MC) items, which can be scored easily and quickly to measure students' skills or misconceptions. Recently, constructed-response (CR) items, which have been found to be useful in assessing complex problem-solving skills, have received more attention (Sykes & Hou, 2003; Williamson, Bejar, & Hone, 1999; Williamson, Bejar, & Sax, 2004; Yang, Kuo, & Liao, 2011), especially in large-scale assessments (Attali & Burstein, 2006; Neidorf, Binkley, Gattis, & Nohara, 2006; Zenisky & Sireci, 2002). Although CR items can provide more valuable diagnostic information than MC items (Attali, Powers, Freedman, Harrison, & Obetz, 2008; Yang et al., 2011), writing out the problem-solving process for CR items may require a great deal of time on the part of the candidates taking the test. Therefore, a test with all CR items is not feasible because of limited testing time. In practice, a test should include both MC and CR items (Ercikan et al., 1998). In addition, using expert graders to score CR items is time consuming and costly. In recent years, the growth of computer technologies has made scoring CR items by computer more economically efficient, alleviating the graders' load (Yang et al., 2011). However, until such system is firmly set in place, the use of CR items may not always be practicable.

To analyse MC and CR data, CDAs require psychometric approaches, such as item response theory (IRT) and cognitive diagnostic models (CDMs). As opposed to ranking students on a latent trait continuum in IRT (Lord, 1980), CDMs can identify the presence or absence of students' skills or misconceptions (Bradshaw & Templin, 2014; de la Torre, 2009b) and provide sufficient diagnostic information to aid teachers in designing remedial instruction (de la Torre, 2009b). Although CDMs can provide more diagnostic information in designing better instruction, most of them are developed for tests with MC items (de la Torre, 2009a, 2011; Junker & Sijtsma, 2001), and cannot handle the richness of the diagnostic information provided by CR items.

This article focuses on the application of CDMs to tests with both MC and CR items for determining students' skills and misconceptions. This is illustrated by contextualising the diagnosis process in an example: primary students solving mathematics problems in the direct proportion unit. Therefore, the two aims of this paper are (a) to introduce new models for tests with both MC and CR items, and (b) to illustrate how these new models can be used to analyse MC and CR data and thus determine students' skills and misconceptions. The new models are modifications of two CDMs. One is the deterministic inputs, noisy 'and' gate (DINA) model (Haertel, 1989; Junker & Sijtsma, 2001), and the other is the deterministic input, noisy 'or' gate (DINO) model (Templin & Henson, 2006). The reason for using the direct proportion unit is that proportional reasoning is an important mathematical thinking ability for comprehending phenomena in the real world (Arieli-Attali & Liu, 2015; Lesh, Post, & Behr, 1988; Modestou & Gagatsis, 2007).

The remainder of the paper is composed of six sections. The DINA and DINO models are introduced in 'Cognitive Diagnosis Models' section. The proposed new models are presented in the 'The Bug-DINO Model' and 'New Models for tests with both MC and CR Items' sections. Analyses of real data from the direct proportion unit are described in the 'Methods' section. The results of the real data analyses are given in the 'Results' section to demonstrate the application of the new models. Conclusions and future work are provided in the final section.

Cognitive diagnosis models

This section introduces the DINA and DINO models. Both are well-known CDMs. The objective of CDMs is to identify whether students possess certain attributes needed in a learning unit. The term 'attribute' is a generic term that refers to something within a specific domain, such as a skill or a misconception in mathematics education or a disorder in psychology. For instance, in response to an item 'In the proportion 8:6 = 4:A, what is the value of A?' in the direct proportion unit, if students possess the skill 'to solve proportion problems using cross products', they may give the correct answer, '3'; however, if students possess the misconception 'believing that ratios of outer and inner terms in the proportion are equivalent', they may give the wrong answer of '12'. In this example, both the skill and misconception are referred to as attributes.

To identify which skills and misconceptions that students possess, test items are constructed that measure these skills and misconceptions. Some items are constructed to measure only one skill or misconception, while others are constructed to measure multiple skills and misconceptions. When implementing CDMs, a Q-matrix (Tatsuoka, 1983) is used to describe the relationships between items and measured skills or misconceptions. To illustrate what a Q-matrix for skills is, an example is given in Table 1. This Q-matrix contains three direct proportion items to detect three skills required in the direct proportion unit. The three skills are C1 (writing a ratio in simplest form), C2 (writing a ratio as a fraction) and C3 (solving proportion problems using cross products). The Q-matrix, as shown in Table 1, consists of 0s and 1s. If the solution of a certain item requires a skill, then the corresponding cell in the Q-matrix contains a '1'. Otherwise, the cell contains a '0'. In the Q-matrix, Item 1 requires Skill C1, Item 2 requires Skill C2 and Item 3 requires Skills C1 and C3.

The DINA model

One of the commonly used CDMs is the DINA model (Haertel, 1989; Junker & Sijtsma, 2001), which is used for diagnosing skills needed in education. The DINA

Table 1. The Q-matrix for three direct proportion items.

		Skills		
No.	Item	C1	C2	C3
1	A is 5, B is 10, what is the ratio of A to B?	1	0	0
2	A to B is equal to 2:3, what is the fraction of A to B?	0	1	0
3	A to B is equal to 3:5, A is 6, what is the value of B?	1	0	1

Note: C1 = Writing a ratio in simplest form; C2 = Writing a ratio as a fraction; C3 = Solving proportion problems using cross products.

model assumes that a student can answer a particular item correctly only if s/he possesses all the skills required for the item. This assumption is called the conjunctive condition. Furthermore, the absence of one required skill cannot be made up for by the presence of other required skills. This second condition is called the non-compensatory condition (de la Torre & Douglas, 2004; Maris, 1999; Rupp & Templin, 2008). For example, as shown in Table 1, a student who only knows how to write a ratio in simplest form (Skill S1), but does not know how to write a ratio as a fraction (S2) and solve proportion problems using cross products (S3) will correctly answer Item 1, but incorrectly answer Items 2 and 3.

However, even if the student possesses all the skills required for an item, s/he may slip and get the wrong answer to the item because of carelessness. Furthermore, even if s/he only has a subset of the skills required to solve the item, s/he may get the correct answer because of guessing if the item is an MC item (de la Torre, 2009a). Therefore, the DINA model has two parameters (i.e. slip and guessing), for each item (see Appendix 1).

The DINO model

In contrast to the DINA model, the DINO model (Templin & Henson, 2006) is more popular when diagnosing major symptoms in the medical, clinical and psychological fields (de la Torre, van der Ark, & Rossi, 2015; Templin & Henson, 2006). In these fields, the term ‘attribute’ is referred to as disorder in CDMs. For example, de la Torre et al. (2015) used the DINO model as one of the CDMs to analyse the data obtained with the Dutch version of the Millon Clinical Multiaxial Inventory-III (Rossi, Sloore, & Derksen, 2008) to diagnose respondents’ disorders. In addition, because psychological tests are not educational tests, the response of 1 or 0 to an item is said to be ‘positive’ or ‘negative’ rather than ‘correct’ or ‘incorrect’.

In a psychological test, the DINO model assumes that a respondent will endorse positively a particular item if s/he possesses at least one of the underlying disorders measured by the item. This assumption is called the disjunctive condition. Furthermore, the absence of one required disorder can be made up for by the presence of other required disorders. This second condition is called the compensatory condition (Templin & Henson, 2006). For example, as shown in Table 2, four disorders, defined as D1 (‘The respondent lies to family or friends to conceal the extent of his/

Table 2. The Q-matrix for four pathological gambling items.

		Disorder			
No.	Item	D1	D2	D3	D4
1	I am private about my gambling experiences	1	0	0	0
2	I have gotten into trouble over things I have done to finance my gambling	0	1	1	0
3	I have gambled with money that I intended to spend on something else	0	0	0	1
4	I have gone to great lengths to obtain money for gambling	0	1	1	1

Note: D1 = The respondent lies to family or friends to conceal the extent of his/her gambling; D2 = The respondent has broken the law to finance his/her gambling; D3 = The respondent has lost relationships because of his/her gambling; D4 = The respondent has been aided by others to overcome financial hardships caused by his/her gambling.

her gambling'), D2 ('The respondent has broken the law to finance his/her gambling'), D3 ('The respondent has lost relationships because of his/her gambling') and D4 ('The respondent has been aided by others to overcome financial hardships caused by his/her gambling'), are required to positively endorse four items used in the pathological gambling test (Rupp, Templin, & Henson, 2010) taken from the Gambling Research Instrument (GRI; Feasel, Henson, & Jones, 2004). When a respondent only has broken the law to finance his/her gambling (D2), but does not lie to family or friends to conceal the extent of his/her gambling (D1), has not lost relationships because of his/her gambling (D3), and has not been aided by others to overcome financial hardships caused by his/her gambling (D4), s/he will positively endorse Items 2 and 4, but negatively endorse Items 1 and 3. However, even if a student has at least one of the disorders required for an item and should positively endorse it, s/he may slip and give the negative response because of 'carelessness'. Furthermore, even if s/he does not have any of the disorders required for the item, s/he may give a positive response because of 'guessing', particularly if it is an MC item. Therefore, like the DINA model, the DINO model also has two parameters for each item (see Appendix 2), as in, slip and guessing parameters although their interpretations differ from those of the DINA model.

In learning mathematics, students often engage in 'buggy' thinking, which refers to the fact that students may have some misconceptions or have no strategy at all and simply guess (Roberts et al., 2014). For the purpose of diagnosing misconceptions, this buggy thinking may be treated as a disorder. Consequently, the disjunctive rule of the DINO model can be applied to detect students' bugs in learning. Nevertheless, a positive response of '1' to an item in the DINO model does not adequately to represent the fact that students who have at least one of the measured bugs (misconceptions) regarding a mathematics item can correctly answer the mathematics item. Therefore, to diagnose students' misconception, which is one of the purposes of this paper, the DINO model must be modified. The next section will introduce a modified DINO model that can be used to diagnose students' misconceptions.

The Bug-DINO model

In mathematics, misconceptions can cause systematic errors and make students give incorrect answers to items (Bradshaw & Templin, 2014). Students possessing misconceptions and lacking skills cannot correctly answer items according to the DINA model (unless there is guessing). In remedial instruction, identifying skills can help teachers to determine which skills students do not possess and need to learn, whereas identifying misconceptions can help teachers to determine which misconceptions students possess that need to be corrected. Therefore, identifying misconceptions is as important as identifying skills. This section introduces a new model used for diagnosing students' misconceptions. This new model is called the Bug-DINO model. As mentioned, the term 'bug' means that a student has misconceptions or does not have any of the skills required for solving a problem. That is, if the student has bug(s), s/he cannot successfully solve the problem. Therefore, the Bug-DINO model assumes that a student cannot correctly answer a particular item if s/he possesses at least one of the bugs measured by the item. In contrast, the student is expected correctly answer the item, only if s/he possesses none of the measured bugs, assuming he/she has the required skills for the item. However, even if the student possesses none of the bugs measured by the item, s/he may slip and provide the wrong answer. Furthermore, even if s/he has one or more of the bugs measured by the item, s/he may still get provide correct answer, particularly if the item is of the MC type. The Bug-DINO model is a modification of the DINO model, in which slip and guessing parameters are used for each item. The mathematical formula for the Bug-DINO model can be found in Appendix 3.

As mentioned above, the DINA and Bug-DINO models can be used for diagnosing skills and misconceptions, respectively. Both are used for tests with MC items only, and they utilise MC item responses that are scored as 1s/0s in the analysis process. Because guessing can be used with MC items, we do not know whether a student can or cannot correctly answer a particular MC item, even if his/her answer is right. Thus, the diagnostic information provided by MC items can be ambiguous. In contrast, the diagnostic information provided by CR items is less ambiguous because CR items do not allow for the possibility of a correct response through random guessing. For example, in response to a CR item, if a student provides the right answer, s/he certainly possesses all the measured skills or does not possess all the measured misconceptions of the CR item. Therefore, the diagnostic information provided by CR items is more exact than that provided by MC items.

The first purpose of this paper was to introduce new models for tests with both MC and CR items, which can provide diagnostic information with more certainty than tests with MC items only. Because CR items disallow guessing, the DINA and Bug-DINO models are not appropriate for modelling CR item responses. For analysing tests with both MC and CR items, these models must be modified. Therefore, in this paper, four new models are designed to deal with the guessing issue when handling both MC and CR items. The four new models are modifications of the DINA model or the Bug-DINO model, and are presented and illustrated below.

New models for MC and CR items

In this section, two new models for modelling the responses to MC and CR items that are scored as 1s/0s are introduced. The first model is called the DINA–CR model, and is a modification of the DINA model that can be used with CR items, in addition to MC items. The other model is called the Bug-DINO–CR model, and is a modification of the Bug-DINO model that can be used with CR items, in addition to MC items.

The DINA–CR and Bug-DINO–CR models

When using the DINA–CR and Bug-DINO–CR models to analyse tests with MC and CR items, the responses to MC and CR items are scored as 1s/0s. The response to a CR item is scored as 1 when all the problem-solving steps for the CR item have been completed successfully, and as 0 otherwise. If a student's response to a CR item is scored as 1, we can know the student possesses each measured skill and does not possess each measured misconception associated with each of the problem-solving steps. This is why the accuracy of the determination of skills or misconceptions provided by the DINA–CR or the Bug-DINO–CR model is higher than that provided by the DINA or the Bug-DINO model, respectively.

The mathematical formulas of the DINA–CR and Bug-DINO–CR models are equal to those of the DINA and Bug-DINO models. What is new is the concept of scoring CR items as either 0's or 1's. Because CR items do not allow guessing, when using the DINA–CR or Bug-DINO–CR model to analyse CR item responses, the guessing parameters of these models are not estimated. Rather, they are fixed at zeros during the estimation process. Nevertheless, slipping is still be possible with CR items. Therefore, the slip parameters of the DINA–CR and Bug-DINO–CR models remain to be estimated. Note that there can be one or more CR items included in the test when using these models. However, CR items have a high cost of construction and scoring. In practice, there should not be too many CR items on a test. The optimal proportion of MC and CR items on the test is another issue, and is not discussed in this paper. This is issue that can be investigated in the future.

Usually, the problem-solving process of the CR item includes multiple steps. Each of the steps measures a skill or misconception. If a student incorrectly answers the CR item, s/he may lack the measured skills or possess the measured misconceptions of some steps, and thus, failed to successfully complete all these steps. However, we do not know which measured skill or misconception the student lacks or possesses if his/her answer to a CR item is wrong. For each problem-solving step, a correct step response can indicate that the student certainly possesses or does not possess the skill or misconception associated with that step. Therefore, for a CR item, students' problem-solving step responses can provide more information than only correct or incorrect responses. For modelling the problem-solving step responses of a CR item, two additional modified DINA and Bug-DINO models called the DINA–Step and Bug-DINO–Step models, are proposed.

The DINA–Step and Bug-DINO–Step models

When using the DINA–Step and Bug-DINO–Step models to analyse tests with MC and CR items, the responses to MC items and the problem-solving steps of CR items are scored as 1s/0s. As mentioned above, a CR item usually includes multiple problem-solving steps, and each of them, typically but not always, measures a skill or misconception. For the CR item, both the DINA–Step and Bug-DINO–Step models assume that the probability of successfully performing each problem-solving step does not influence the probabilities of successfully performing the other steps. Based on this assumption, each problem-solving step of a CR item can be modelled as a separate item in which a skill or misconception is measured, and the step response is scored as 1/0.

In a CR item, if each step is scored with a categorical score (i.e. 0 or 1), then the number of categorical scores for the CR item increases. For example, if a CR item measures three skills, then there are three categorical scores for this item. This increased number of categorical scores for the CR item is the reason the accuracy in diagnosing skills or misconceptions provided by the DINA–Step or Bug-DINO–Step model tends to be higher than that provided by the DINA–CR or Bug-DINO–CR model. However, increasing the number of categorical scores for a CR item also increases the loads of graders. As mentioned earlier, in practice, computerised scoring is a direction to be considered and recommended to address this issue.

Like the DINA–CR and Bug-DINO–CR models, when using the DINA–Step and Bug-DINO–Step models to analyse the problem-solving step responses of CR items, the guessing parameters of these models are not estimated but rather fixed at zeros during the estimation process. Nevertheless, slipping is still possible for each step. Therefore, the slip parameters of the DINA–Step and Bug-DINO–Step models are still estimated.

The second purpose of this paper is to illustrate how these models can be used to analyse MC and CR data for diagnosing students' skills and misconceptions. The empirical example below is used to illustrate the proposed models.

Methods

An empirical study was conducted to demonstrate the application of CDMs to tests with MC and CR items in diagnosing students' skills and misconceptions. An example of primary students' solving mathematics problems in the direct proportion unit was used in the empirical study. The direct proportion items were a subset of the original items used by Lu (2014) to which student responses had been collected. Methods of analysing real data using the models are explained. The parameter estimation method for the models and the evaluation criteria used for comparing the diagnostic results were also introduced.

Participants

The real data used in this study were responses of 497 grade-six students (268 boys and 229 girls), who were randomly selected from 19 classes in four Taiwanese primary schools, to eight direct proportion items. Six primary school mathematics teachers were employed as raters to grade the students' CR item responses. All of them worked in public primary schools, and had taught the direct proportion unit for many years.

Instruments

Items

Eight direct proportion items were used in this study (see Appendix 4). Each item included MC and CR parts. An example is shown in Table 3, in which students were asked to select one of four provided options (MC response) for Item 8, and to write down their problem-solving processes (CR response).

The students' responses of the eight direct proportion items were analysed to diagnose skills and misconceptions. The DINA, DINA–CR and DINA–Step were used to diagnose skills, whereas the Bug-DINO, Bug-DINO–CR and Bug-DINO–Step were used to diagnose misconceptions. As introduced in earlier sections, in CDMs, a Q-matrix is used to describe the relationships between the items, and the skills and misconceptions.

Q-matrices

Three skills and four misconceptions were measured by the eight direct proportion items. Two Q-matrices were constructed, one to describe the relationship between the direct proportion items and skills, and another to describe the relationship between the items and misconceptions. Table 4 provides the Q-matrix for the skills. The three skills are defined as S1 (equivalent ratios can be expressed in the direct proportion), S2 (solving proportion problems using equivalent ratios) and S3

Table 3. Students' problem-solving process and their selected answer choice.

Item	Intended measured skills and misconception
8. (2) There are two circles, one big and one small, and the ratio of their radius is 3:2. The length of diameter of the big circle is 56.52 cm *What is the length of diameter of the small circle?* This is the MC part of the item ① 84.78 cm ② 37.68 cm ③ 18.84 cm ④ 9.42 cm *Please write down your problem solving process:* This is the CR part of the item 56.52÷3.14=18 18÷2=9 3:2=9:□ =9÷3=3 12×3.14=37.68 □=3×2=6 6×2=12	*Skills to be measured*: S1, S2 *Misconceptions to be measured*: B1, B2, B3, B4

Note: S1 = Expressing equivalent ratios as a direct proportion; S2 = Solving proportion problems using equivalent ratios; B1 = Misplacing terms of ratios converted from fraction; B2 = Misplacing terms of ratios in the proportion; B3 = Both terms of ratios are multiplied or divided by different numbers; B4 = Misunderstanding ratios of outer and inner terms in the proportion are equivalent.

Table 4. The Q-matrix for skills.

	Skills		
Item	S1	S2	S3
1	0	0	1
2	1	1	0
3	1	1	0
4	1	0	1
5	1	0	1
6	1	0	1
7	1	0	1
8	1	1	0

(solving proportion problems using cross products). Table 5 provides the Q-matrix for misconceptions. The four misconceptions are defined as B1 (terms of ratios converted from fraction are misplaced), B2 (terms of ratios in the proportion are misplaced), B3 (both terms of ratios are multiplied or divided by different numbers) and B4 (misunderstanding that the ratios of the outer and inner terms in the proportion are equivalent).

Table 5. The Q-matrix for misconceptions.

	Misconceptions			
Item	B1	B2	B3	B4
1	0	0	0	1
2	1	1	1	1
3	1	1	1	1
4	1	1	0	1
5	1	1	0	1
6	1	1	0	1
7	1	1	0	1
8	1	1	1	1

Illustrations of analyses of real data using the different CDMs

This section illustrates the methods involved in using CDMs to analyse real data. As mentioned above, each of the direct proportion items had MC and CR responses. In this study, three data-sets obtained from the eight direct proportion items were analysed via the models. The first data-set represents the data for the test with MC items only, and it was analysed using the DINA or Bug-DINO model for diagnosing skills or misconceptions. The second data-set represents the data for the test with both MC and CR items, and the responses to the CR items are simply deemed correct and incorrect responses. This data-set was analysed using the DINA–CR or the Bug-DINO–CR model for diagnosing skills or misconceptions. The third data-set also represents the data for the test with both MC and CR items, but responses of CR items are problem-solving step responses. This data-set was analysed using the DINA–Step or the Bug-DINO–Step model for diagnosing skills or misconceptions. Note that the three data-sets were simply examples explaining how to use these models to analyse tests with MC and CR items. Researchers/teachers do not need to collect three sets of data when they want to use these models.

Analyses using the DINA and Bug-DINO models

The purpose of this first analysis was to use the DINA and Bug-DINO models to analyse the first data-set and thus diagnose skills and misconceptions, respectively. The first real data-set was composed of dichotomous MC responses obtained from direct proportion Items 1–8. The skill and misconception Q-matrices for the data that were analysed using the DINA and Bug-DINO models are shown in Tables 4 and 5, respectively.

Analyses using the DINA–CR and Bug-DINO–CR models

The purpose of this second analysis was to use the DINA–CR and Bug-DINO–CR models to analyse the second data-set for diagnosing skills and misconceptions, respectively. The second data-set was composed of dichotomous MC responses obtained from direct proportion Items 1 through 7 and dichotomous CR responses obtained from direct proportion Item 8. The skill and misconception Q-matrices for the data that were analysed using the DINA–CR and Bug-DINO–CR models are equal to those for the DINA and Bug-DINO models.

Analyses using the DINA–Step and Bug-DINO–Step models

The purpose of this third analysis is to use the DINA–Step and Bug-DINO–Step models to analyse the third data-set and thus diagnose skills and misconceptions, respectively. The third data-set was composed of dichotomous MC responses obtained from direct proportion Items 1–7 and dichotomous problem-solving step responses obtained from direct proportion Item 8. Because direct proportion Item 8 measures Skills 1 and 2 and Misconceptions 1–4, the Q-matrix for the DINA–Step contained nine items and the Q-matrix for the Bug-DINO–Step contained eleven items (see Tables 6 and 7).

Note that because, in this example, only the CR of Item 8 is used as an illustration, the Q matrices in Table 4 and Table 6 differ only in terms of the specifications regarding the skills needed for this item. The rows of the Q matrices in Tables 4 and 6 corresponding to the other items are the same. Likewise, the Q matrices in Table 5 and Table 7 differ only in terms of the specifications regarding the misconceptions for this item. The rows of the Q matrices in Tables 5 and 7 corresponding to the other items are also the same.

Table 6. The Q-matrix for the DINA–Step.

	Skills		
Item	S1	S2	S3
1	0	0	1
2	1	1	0
3	1	1	0
4	1	0	1
5	1	0	1
6	1	0	1
7	1	0	1
8a	1	0	0
8b	0	1	0

Note: Items 1–7 are MC items. Items 8a and 8b are one-skill CR items.

Table 7. The Q-matrix for the Bug-DINO–Step.

	Misconceptions			
Item	B1	B2	B3	B4
1	0	0	0	1
2	1	1	1	1
3	1	1	1	1
4	1	1	0	1
5	1	1	0	1
6	1	1	0	1
7	1	1	0	1
8a	1	0	0	0
8b	0	1	0	0
8c	0	0	1	0
8d	0	0	0	1

Note: Items 1–7 are MC items. Items 8a-8d are one-misconception CR items.

Parameter estimation

The nature of a CR item is such that the probability of correctly guessing the answer is very low, probably close to zero. To model CR item responses more properly, the guessing parameters of the DINA–CR, DINA–Step, Bug-DINO–CR and Bug-DINO–Step models can be set to zero; only the slip parameters are estimated in the estimation process. To estimate the parameters of the proposed models, an expectation-maximisation (EM) algorithm is adopted from de la Torre (2009b); its implementation codes were written in MATLAB R2012a. The details regarding the estimation process are given in Appendix 5.

Evaluation criteria

In this study, two evaluation criteria, correct attribute agreement rate (CAAR) and correct pattern agreement rate (CPAR; Chen, Xin, Wang, & Chang, 2012), were used to compare and evaluate the performances of the models. They were used to compare the skills and misconceptions diagnosed via the CDMs with the human classification results produced by expert graders for real data. The human grading results in this study were students' skills and misconceptions as they were scored based on the CR parts of all the eight direct proportion Items. Because these two criteria are rates, their values can range from 0 to 1 and larger values means higher diagnostic agreement. Their formulas can be found in Appendix 6.

Results

Results of diagnosing skills using the DINA, DINA–CR and DINA–Step models

Table 8 summarises the CAAR and CPARs for skills by comparing the skills diagnosed using the DINA, DINA–CR and DINA–Step models with those arrived at by expert raters. As shown in Table 8, both the mean of the CAARs for the skills and the CPAR provided by the DINA model were smaller than those provided by the DINA–CR model. The CAARs for Skills S1 and S2 provided by the DINA–CR model were obviously larger than those provided by the DINA model. Similarly, both the mean of the CAARs for the skills and the CPAR provided by DINA–CR model were smaller than those provided by the DINA–Step model. The CAARs for Skills S1 and S2 provided by the DINA–Step model were larger than those provided by the DINA–CR model, especially for Skill S1. These findings indicate that the diagnostic agreement for Skills S1 and S2 were significantly increased due to the effect of the CR responses obtained from direct proportion Item 8.

Table 8. Diagnostic agreement for each skills and entire pattern using the DINA, DINA–CR and DINA–Step models for real data.

	CAAR				
Fitted model	S1	S2	S3	Mean	CPAR
DINA	.8632	.7062	.7062	.7586	.4306
DINA–CR	.8974	.7445	.7183	.7867	.4628
DINA–Step	.9356	.7525	.7103	.7995	.4829

Note: CAAR = correct attribute agreement rate. CPAR = correct pattern agreement rate.

Results of diagnosing misconceptions using the Bug-DINO, Bug-DINO–CR and Bug-DINO–Step models

Table 9 summarises the CAAR and CPARs for misconceptions by comparing the misconceptions diagnosed using the Bug-DINO, Bug-DINO–CR and Bug-DINO–Step models with those arrived at by expert raters. As shown in Table 9, both the mean of the CAARs for the misconceptions and the CPAR provided by the Bug-DINO model were smaller than those provided by the Bug-DINO–CR model. The CAARs for Misconceptions B1–B4 provided by the Bug-DINO–CR model were obviously larger than those provided by the Bug-DINO model. Similarly, both the mean of the CAARs for the misconceptions and the CPAR provided by the Bug-DINO–CR model were smaller than those provided by the Bug-DINO–Step model. The CAARs for Misconceptions B1 to B4 provided by the Bug-DINO–Step model were larger than those provided by the Bug-DINO–CR model, especially for Misconceptions B1, B2 and B4. These findings indicate that the diagnostic agreement for Misconceptions B1–B4 were also significantly increased due to the effect of the CR responses that were obtained from direct proportion Item 8.

Discussion and conclusions

Most CDMs have been developed for tests with MC items. Although these CDMs can utilise diagnostic information provided by MC items, they cannot handle the more useful diagnostic information provided by CR items. To handle the diagnostic information provided by CR items, this paper proposes new CDMs and illustrates methods of using these models to analyse MC and CR data and thus diagnose students' skills and misconceptions.

The empirical study was conducted to demonstrate the application of CDMs for tests with MC and CR items. Three data-sets created using eight direct proportion items were analysed via the models. The results showed that the diagnostic agreement for skills and misconceptions provided by the DINA–CR/Bug-DINO–CR models was better than that provided by the DINA/Bug-DINO models. This is because that CR items provide less uncertainty information than MC items. Furthermore, diagnostic agreement for skills and misconceptions provided by the DINA–Step/Bug-DINO–Step was better than that provided by the DINA–CR/Bug-DINO–CR. This is because the problem-solving step responses to the CR items provide more information than the simple correct and incorrect responses.

This study provides solid evidence that the proposed CDMs are effective than traditional CDMs in diagnosing skills and misconceptions. In practice, if a test

Table 9. Diagnostic agreement for each misconceptions and entire pattern using the Bug-DINO, Bug-DINO–CR and Bug-DINO–Step models for real data.

	CAAR					
Fitted model	B1	B2	B3	B4	Mean	CPAR
Bug-DINO	.8048	.7807	.8310	.8290	.8114	.5835
Bug-DINO–CR	.8813	.8531	.9135	.8954	.8858	.7525
Bug-DINO–Step	.9799	.8893	.9276	.9416	.9346	.7686

Note: CAAR = correct attribute agreement rate. CPAR = correct pattern agreement rate.

contains CR items, the DINA–CR/Bug-DINO–CR and DINA–Step/Bug-DINO–Step models can provide better diagnostic performance than the DINA/Bug-DINO model. However, the training process that allows expert graders to score CR items is costly for paper-and-pencil testing. Recent years have seen a growing interest in the study of diagnostic feedback in computer-based learning environments (Golke et al., 2015; Lee, 2016; Timmers et al., 2015; Van der Kleij et al., 2015). In current computer-based testing, students can input their CR responses through well-designed user interfaces, and the computerised scoring of CR items can be performed using recent developments in computer technologies such as an automated scoring mechanism (Yang et al., 2011). Therefore, to alleviate graders' load, CR items should be administered in computer-based testing with automated scoring mechanism rather than paper-and-pencil testing.

This article introduces five new models (i.e. DINA–CR, DINA–Step, Bug-DINO, Bug-DINO–CR and Bug-DINO–Step models) by building on two existing models, namely, the DINA and DINO models. All the models were used to analyse real data except for the DINO model. Regarding these models, in practice, when a paper-and-pencil test or a computer-based test contains MC items only, researchers/teachers should use the DINA or the Bug-DINO model for data analysis when diagnosing skills or misconceptions. When the paper-and-pencil/computer-based test contains both MC and CR items and the CR items are short-answer items that only require students to write answers, researchers/teachers should use the DINA–CR or Bug-DINO–CR model. When the CR items of the paper-and-pencil/computer-based test involve multiple problem-solving steps, researchers/teachers should use the DINA–Step or the Bug-DINO–Step model.

Because scoring CR items is costly, and writing out the problem-solving process used to solve CR items can be time-consuming, a paper-and-pencil/computer-based test should not contains too many CR items. As mentioned above, a computer-based test is more suitable for administering CR items than a paper-and-pencil test. Therefore, the DINA–CR/Bug-DINO–CR and the DINA–Step/Bug-DINO–Step should be applied to computer-based testing. The implementation codes for the models used in this study were written in MATLAB, though the code for the DINA model (de la Torre & Lee, 2010) was written in Ox (Doornik, 2002). The codes for these models can be made available upon request.

The proposed approached assumed that each step in CR items can be considered a separate item. This has the potential to violate of the conditional independence assumption needed in estimating the model parameters. Future studies should examine the extent to which this violation can impact the quality of item parameter estimates and, more importantly, attribute classifications. However, it can be noted that, with the current application, despite the potential violation of the conditional independence assumption, the proposed models still resulted in higher agreement rates with human graders compared to models that did not violate the assumption (i.e. models that used the original 0/1 responses).

Finally, this paper advocates the use of computers to efficiently administer and score CDAs. With computer-based CDAs, a natural extension is to employ computerised adaptive testing, which has been shown to be more efficient than linear test administration. However, a full implementation of cognitive diagnosis computerised adaptive testing would require developing a larger pool of items, as well as, item-selection algorithms appropriate for the proposed models.

Funding

Ministry of Science and Technology, Taiwan10.13039/501100004663 [grant number 102-2511-S-142-008-MY3].

ORCID

Chun-Hua Chen http://orcid.org/0000-0002-3441-1284

References

Arieli-Attali, M., & Liu, Y. (2015). Beyond correctness: Development and validation of concept-based categorical scoring rubrics for diagnostic purposes. *Educational Psychology*. Advance online publication. doi: 10.1080/01443410.2015.1031088

Ashlock, R. B. (1994). *Error patterns in computation* (6th ed.). Englewood Cliffs, NJ: Prentice Hall.

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-raterR V.2. *The Journal of Technology, Learning, and Assessment, 4*(3). Retrieved from http://napoleon.bc.edu/ojs/index.php/jtla/article/view/1650

Attali, Y., Powers, D., Freedman, M., Harrison, M., & Obetz, S. (2008). *Automated scoring of short-answer open-ended GRE subject test items* (GRE Board Research Rep. No GRE-04-02). Princeton, NJ: ETS.

Bradshaw, L., & Templin, J. (2014). Combining item response theory and diagnostic classification models: A psychometric model for scaling ability and diagnosing misconceptions. *Psychometrika, 79*, 403–425. doi:10.1007/s11336-013-9350-4

Carlin, B. P., & Louis, T. A. (2000). *Bayes and empirical Bayes methods for data analysis*. New York, NY: Chapman & Hall.

Chen, P., Xin, T., Wang, C., & Chang, H. H. (2012). Online calibration methods for the DINA model with independent attributes in CD-CAT. *Psychometrika, 77*, 201–222. doi:10.1007/s11336-012-9255-7

de la Torre, J. (2009a). A cognitive diagnosis model for cognitively based multiple-choice options. *Applied Psychological Measurement, 33*, 163–183. doi:10.1177/0146621608320523

de la Torre, J. (2009b). DINA model and parameter estimation: A didactic. *Journal of Educational and Behavioral Statistics, 34*, 115–130. doi:10.3102/1076998607309474

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika, 76*, 179–199. doi:10.1007/s11336-011-9207-7

de la Torre, J., & Douglas, J. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika, 69*, 333–353. doi:10.1007/BF02295640

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. doi:10.1111/j.1745-3984.2009.00102.x

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*, online first. doi:10.1177/0748175615569110.

Doornik, J. A. (2002). *Object-oriented matrix programming using Ox* (Version 3.1). [Computer software]. London: Timberlake Consultants Press.

Ercikan, K., Sehwarz, R. D., Julian, M. W., Burket, G. R., Weber, M. M., & Link, V. (1998). Calibration and scoring of tests with multiple-choice and constructed-response item types. *Journal of Educational Measurement, 35*, 137–154. doi:10.1111/j.1745-3984.1998.tb00531.x

Feasel, K., Henson, R., & Jones, L. (2004). *Analysis of the gambling research instrument (GRI)*. Unpublished manuscript.

Golke, S., Dörfler, T., & Artelt, C. (2015). The impact of elaborated feedback on text comprehension within a computer-based assessment. *Learning and Instruction, 39*, 123–136. doi:10.1016/j.learninstruc.2015.05.009

Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement, 26*, 301–321. doi:10.1111/j.1745-3984.1989.tb00336.x

Harks, B., Rakoczy, K., Hattie, J., Besser, M., & Klieme, E. (2014). The effects of feedback on achievement, interest and self-evaluation: The role of feedback's perceived usefulness. *Educational Psychology, 34*, 269–290. doi:10.1080/01443410.2013.785384

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*, 81–112.

Huang, T.-W., & Wu, P.-C. (2013). Classroom-based cognitive diagnostic model for a teacher-made fraction-decimal test. *Educational Technology & Society, 16*, 347–361.

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. doi:10.1177/01466210122032064

Ketterlin-Geller, L. R., & Yovanoff, P. (2009). Diagnostic assessments in mathematics to support instructional decision making. *Practical Assessment, Research and Evaluation, 14* (16), 1–11.

Lee, H. (2016). Which feedback is more effective for pursuing multiple goals of differing importance? The interaction effects of goal importance and performance feedback type on self-regulation and task achievement. *Educational Psychology, 36*, 297–322. doi:10.1080/01443410.2014.995596

Lesh, R., Post, T., & Behr, M. (1988). Proportional reasoning. In J. Hiebert & M. Behr (Eds.), *Number concepts and operations in the middle grades* (pp. 93–118). Reston, VA: Lawrence Erlbaum & National Council of Teachers of Mathematics.

Lord, F. M. (1980). *Application of item response theory to practical testing problems*. Hillside, NJ: Lawrence Erlbaum. doi:10.4324/9780203056615

Lu, C.-H. (2014). *The discussion of problem-solving strategies and math performance in the direct proportion unit of the grade school students* (Unpublished master's thesis). National Taichung University of Education, Taichung, Taiwan.

Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika, 64*, 187–212. doi:10.1007/BF02294535

Modestou, M., & Gagatsis, A. (2007). Students' improper proportional reasoning: A result of the epistemological obstacle of "linearity". *Educational Psychology, 27*, 75–92. doi:10.1080/01443410601061462

Neidorf, T. S., Binkley, M., Gattis, K., & Nohara, D. (2006). *Comparing mathematics content in the National Assessment of Educational Progress (NAEP), Trends in International Mathematics and Science Study (TIMSS), and Program for International Student Assessment (PISA) 2003 assessments* (NCES 2006-029). Washington, DC: National Center for Education Statistics, U.S. Department of Education.

Pekrun, R., Cusack, A., Murayama, K., Elliot, A. J., & Thomas, K. (2014). The power of anticipated feedback: Effects on students' achievement goals and achievement emotions. *Learning and Instruction, 29*, 115–124. doi:10.1016/j.learninstruc.2013.09.002

Roberts, M. R., Alves, C. B., Chu, M.-W., Thompson, M., Bahry, L. M., & Gotzmann, A. (2014). Testing expert-based versus student-based cognitive models for a grade 3 diagnostic mathematics assessment. *Applied Measurement in Education, 27*, 173–195. doi:10.1080/08957347.2014.905787

Rossi, G., Sloore, H., & Derksen, J. (2008). The adaptation of the MCMI-III in two non-English speaking countries: State of the art of the Dutch language version. In T. Millon & C. Bloom (Eds.), *The Millon inventories: A practitioner's guide to personalized clinical assessment* (2nd ed., pp. 369–386). New York, NY: Guilford Press.

Rupp, A., & Templin, J. (2008). The effects of q-matrix misspecification on parameter estimates and classification accuracy in the DINA model. *Educational and Psychological Measurement, 68*, 78–96. doi:10.1177/0013164407301545

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

Rust, C. (2007). Towards a scholarship of assessment. *Assessment and Evaluation in Higher Education, 32*, 229–237. doi:10.1080/02602930600805192

Sykes, R. C., & Hou, L. (2003). Weighting constructed-response items in IRT-based exams. *Applied Measurement in Education, 16*, 257–275. doi:10.1207/S15324818AME1604_1

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement, 20*, 345–354. doi:10.1111/j.1745-3984.1983.tb00212.x

Templin, J. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods, 11*, 287–305. doi:10.1037/1082-989X.11.3.287

Timmers, C. F., Walraven, A., & Veldkamp, B. P. (2015). The effect of regulation feedback in a computer-based formative assessment on information problem solving. *Computers & Education, 87*, 1–9. doi:10.1016/j.compedu.2015.03.012

Van der Kleij, F. M., Feskens, R. C. W., & Eggen, T. J. H. M. (2015). Effects of feedback in a computer-based learning environment on students' learning outcomes: A meta-analysis. *Review of Educational Research, 85*, 475–511. doi:10.3102/0034654314564881

Williamson, D. M., Bejar, I. I., & Hone, A. S. (1999). 'Mental model' comparison of automated and human scoring. *Journal of Educational Measurement, 36*, 158–184. doi:10.1111/j.1745-3984.1999.tb00552.x

Williamson, D. M., Bejar, I. I., & Sax, A. (2004). Automated tools for subject matter expert evaluation of automated scoring. *Applied Measurement in Education, 17*, 323–357. doi:10.1207/s15324818ame1704_1

Yang, C.-W., Kuo, B.-C., & Liao, C.-H. (2011). A ho-irt based diagnostic assessment system with constructed response items. *Turkish Online Journal of Educational Technology, 10*, 46–51.

Zenisky, A. L., & Sireci, S. G. (2002). Technological innovations in large-scale assessment. *Applied Measurement in Education, 15*, 337–362. doi:10.1207/S15324818AME1504_02

Appendix 1. The DINA model

The item response function of the DINA model is defined as

$$P(X_{ij} = 1|\alpha_i) = \left(1 - s_j\right)^{\eta_{ij}} g_j^{(1-\eta_{ij})}, \tag{A.1}$$

where η_{ij} is defined as $\eta_{ij} = \prod_{k=1}^{K} \alpha_{ik}^{q_{jk}}$. $\eta_{ij} = 1$ indicates that student i possesses all the skills required for item j, and $\eta_{ij} = 0$ otherwise. The item parameter $s_j = P(X_{ij} = 0 | \eta_{ij} = 1)$ is the probability of an incorrect response to item j when $\eta_{ij} = 1$. The item parameter $g_j = P(X_{ij} = 1 | \eta_{ij} = 0)$ is the probability of a correct response to item j when $\eta_{ij} = 0$. α_i represents the skill pattern of student i.

Appendix 2. The DINO model

The item response function of the DINO model is defined as

$$P(X_{ij} = 1|\alpha_i) = \left(1 - s_j\right)^{\xi_{ij}} g_j^{1-\xi_{ij}}, \tag{B.1}$$

Where ξ_{ij} is defined as $\xi_{ij} = 1 - \prod_{k=1}^{K} (1 - \alpha_{ik})^{q_{jk}}$. $\xi_{ij} = 1$ indicates that respondent i possesses at least one of the disorders required for item j, and $\xi_{ij} = 0$ otherwise. The item parameter $s_j = P(X_{ij} = 0 | \xi_{ij} = 1)$ is the probability of an negative response to item j when $\xi_{ij} = 1$. The item parameter $g_j = P(X_{ij} = 1 | \xi_{ij} = 0)$ is the probability of a positive response the item j when $\xi_{ij} = 0$. α_i represents the disorder pattern of student i.

Appendix 3. The Bug-DINO model

The item response function of the Bug-DINO model is defined as

$$P(X_{ij} = 1|\beta_i) = (1 - s_j)^{1-r_{ij}} g_j^{r_{ij}}, \tag{C.1}$$

Where r_{ij} is defined as $r_{ij} = 1 - \prod_{l=1}^{L} (1 - \beta_{il})^{q_{jl}}$. $r_{ij} = 1$ indicates that student i possesses at least one of the misconceptions measured by item j, and $r_{ij} = 0$ otherwise. The item parameter $s_j = P(X_{ij} = 0 | \gamma_{ij} = 0)$ is the probability of an incorrect response to item j when $r_{ij} = 0$. The item parameter $g_j = P(X_{ij} = 1 | \gamma_{ij} = 1)$ is the probability of a correct response to item j when $r_{ij} = 1$. β_i represents the misconception pattern of student i. In contrast to the DINO model, possessing one of the disorders will result in a positive endorsement, whereas possessing a misconception will result in a negative endorsement.

Appendix 4. Eight direct proportion items for test

Item		Intended measured skills and misconception
1. (3) In the propotion 甲 : 40 = 0.8 : 0.5.		*Skills to be measured*: S3 *Misconceptions to be measured*: B4
what is the value of 甲 ? ① 10 ② 25 ③ 64 ④ 80	This is the MC part of the item.	
Please write down your problem solving process: 甲:40=0.8:0.5 甲=40×0.8÷0.5 =32÷0.5 =64	This is the CR part of the item.	
2. (4) Three bags of candies can be divided into twelve packages. Zong Tong has bought twenty-four bugs of candies.		*Skills to be measured*: S1, S2 *Misconceptions to be measured*: B1, B2, B3, B4
How many packages can he divide 24 bugs of candies into? ① $1\frac{1}{2}$ packages ② 6 packages ③ 8 packages ④ 96 pachages	This is the MC part of the item.	
Please write down your problem solving process: 3:12=24:X 24÷3=8 12×8=96	This is the CR part of the item.	
3. (1) An activity of points for cash is held by the department store. 100 points are equivalent to 3 dollars. In the activity, Pei Xuan has got 72 dollars.		*Skills to be measured*: S1, S2 *Misconceptions to be measured*: B1, B2, B3, B4
How many points does she use for cash? ① 2400 points ② 24 points ③ $4\frac{1}{6}$ points ④ $2\frac{4}{25}$ points	This is the MC part of the item.	

Please write down your problem solving process:

設折抵現金72元用30點

100:3=□:72

72÷3=24

100×24=2400

This is the CR part of the item.

4. (2) When the height of the triangle is fixed, the bottom and area of the teiangle are in direct proportion. If the length of the bottom is 12 cm, then the area is 72 cm^2.

Skills to be measured: S1, S3

Misconceptions to be measured: B1, B2, B4

What is the length of the bottom when the area is 48 cm^2?

① 6 cm ② 8 cm ③ 18 cm ④ 288 cm

This is the MC part of the item.

Please write down your problem solving process:

設底為□cm

12:72=□:48

48×12=576

576÷72=8

This is the CR part of the item.

5. (1) Two liters of green tea and twenty-one grams of sugars make the best to drink.

Skills to be measured: S1, S3

Misconceptions to be measured: B1, B2, B4

How may liters of green tea make and thirty grams of sugars make the best to drink?

① $2\frac{6}{7}$ liters ② $\frac{7}{10}$ liters ③ $1\frac{3}{7}$ liters ④ 315 liters

This is the MC part of the item.

Please write down your problem solving process:

2:21=□:30

$\square = 2\frac{6}{7}$

$2\frac{18}{21} = 2\frac{6}{7}$

$$\begin{array}{r} 2 \\ 21\overline{)60} \\ \underline{42} \\ 18 \end{array}$$

This is the CR part of the item.

6. (2) The height of a tree is 225 cm, the length of its shadow is 2.4

Skills to be measured: S1, S3

m. At the same time, the length of shadow of the flagpole is 2.4 m.		*Misconceptions to be measured*: B1, B2, B4
What is the height of the flagpole? ① 37.5 cm ② 150 cm ③ 192 cm ④ 337.5 cm	This is the MC part of the item.	
Please write down your problem solving process: 225:3.6 :□ = 2.4 225×2.4=540 540÷3.6=150	This is the CR part of the item.	
7. (2) A construction company wants to build a house. Three-fifths of the work has been completed in 45 days.		*Skills to be measured*: S1, S3 *Misconceptions to be measured*: B1, B2, B4
How many days does it need to complete the remaining work? ① 18 days ② 27 days ③ 30 days ④ 67.5 days	This is the MC part of the item.	
Please write down your problem solving process: 3:5=□:45 45×3=135 135÷5=27	This is the CR part of the item.	
8. (2) There are two circles, one big and one small, and the ratio of their radius is 3:2. The length of diameter of the big circle is 56.52 cm.		*Skills to be measured*: S1, S2 *Misconceptions to be measured*: B1, B2, B3, B4
What is the length of diameter of the small circle? ① 84.78 cm ② 37.68 cm ③ 18.84 cm ④ 9.42 cm	This is the MC part of the item.	
Please write down your problem solving process:	This is the CR part of the item.	

Note: S1 = Expressing equivalent ratios as a direct proportion; S2 = Solving proportion problems using equivalent ratios; S3 = Solving proportion problems using cross products; B1 = Misplacing terms of ratios converted from fraction; B2 = Misplacing terms of ratios in the proportion; B3 = Both terms of ratios are multiplied or divided by different numbers; B4 = Misunderstanding ratios of outer and inner terms in the proportion are equivalent.

Appendix 5. The EM algorithm

The steps of the EM algorithm for the DINA model are described as follows:
Iteration 0:
Step 1: The prior distribution of possible attribute patterns is assumed to follow the uniform distribution. Therefore, the prior probability of each possible attribute pattern α_l is set to $\frac{1}{2^K}$. K is the number of attributes.
Step 2: Draw the item parameters s_j and g_j randomly from $U(0, .1)$.
Iteration t:
Step 3: Compute the posterior probability of each possible attribute pattern α_l. The formula of posterior probability is defined as

$$P(\alpha_l|X_i) = L(X_i|\alpha_l)P(\alpha_l) / \sum_{l=1}^{L} L(X_i|\alpha_l)P(\alpha_l), \tag{E.1}$$

where X_i denotes the item responses of student i; $P(\alpha_l)$ denotes the prior probability and $L(X_i|\alpha_l)$ denotes the likelihood function. The formula of the likelihood function is defined as

$$L(X_i|\alpha_l) = \prod_{j=1}^{J} P_j(\alpha_l)^{X_{ij}}[1 - P_j(\alpha_l)]^{1-X_{ij}}, \tag{E.2}$$

where $P_j(\alpha_l)$ denotes the probability of a correct response to item j by the α_l.
Step 4: Compute the item parameters s_j and g_j. The formulas of them are defined as

$$\hat{s}_j = (I_{jl}^{(1)} - R_{jl}^{(1)})/I_{jl}^{(1)} \text{ and } \hat{g}_j = R_{jl}^{(0)}/I_{jl}^{(0)}, \tag{E.3}$$

where $I_j^{(1)} = \sum_{\alpha_l:\alpha_l' q_j = q_j' q_j} \sum_{i=1}^{I} P(\alpha_l|X_i)$ is the expected number of students who possess all the attributes measured by item j, $R_j^{(1)} = \sum_{\alpha_l:\alpha_l' q_j = q_j' q_j} \sum_{i=1}^{I} X_{ij}P(\alpha_l|X_i)$ is the number of the students among $I_j^{(1)}$ correctly answering item j. Similarly, $I_j^{(0)} = \sum_{\alpha_l:\alpha_l' q_j < q_j' q_j} \sum_{i=1}^{I} P(\alpha_l|X_i)$ and $R_j^{(0)} = \sum_{\alpha_l:\alpha_l' q_j < q_j' q_j} \sum_{i=1}^{I} X_{ij}P(\alpha_l|X_i)$, they belong to the students who lack at least one of the attributes measured by item j.Finally, Steps 3 and 4 are repeated until convergence. Implementation of the algorithm uses an empirical Bayes method (Carlin & Louis, 2000) updating the prior distribution of the latent classes after each iteration based on the posterior distributions of the students.
For the Bug-DINO model, $I_j^{(1)}$ is the expected number of students who possess at least one of the attributes measured by item j, and $R_j^{(0)}$ is the expected number of students who possess none of the attributes measured by item j. For the DINA–CR, Bug-DINO–CR, DINA–Step and Bug-DINO–Step models, because CR items do not allow guessing, hence the EM algorithms can be obtained from the algorithm described above by setting the item parameter $g_j = 0$ in step 2.

Appendix 6. The CAAR and CPAR

The formula of CAAR for the attribute k is defined as

$$\text{CAAR}_k = \sum_{i=1}^{N} I[\hat{\alpha}_{ik} = \alpha_{ik}]/N, \tag{F.1}$$

where I is the indicator function; $\hat{\alpha}_{ik}$ and α_{ik} are the estimated and true attribute k of student i, respectively. The formula of CPAR for the attribute pattern i is defined as

$$\text{CPAR}_i = \sum_{i=1}^{N} I[\hat{\alpha}_i = \alpha_i]/N, \tag{F.2}$$

where $\hat{\alpha}_i$ and α_i are the estimated and true attribute pattern of student i, respectively.