Modeling Partial Knowledge on Multiple-Choice Items Using Elimination Testing

Qian Wu, Tinne De Laet, and Rianne Janssen

KU Leuven

Abstract

Single-best answers to multiple-choice items are commonly dichotomized into correct and incorrect responses, and modeled using either a dichotomous item response theory (IRT) model or a polytomous one if differences among all response options are to be retained. The present study presents an alternative IRT-based modeling approach to multiple-choice items administered with the procedure of elimination testing, which asks test-takers to eliminate all the response options they consider to be incorrect. The partial credit model is derived for the obtained responses. By extracting more information pertaining to test-takers' partial knowledge on the items, the proposed approach has the advantage of providing more accurate estimation of the latent ability. In addition, it may shed some light on the possible answering processes of test-takers on the items. As an illustration, the proposed approach is applied to a classroom examination of an undergraduate course in engineering science.

Keywords: multiple-choice items, elimination testing, partial credit model, scoring group analysis, IRT

Modeling Partial Knowledge on Multiple-Choice Items Using Elimination Testing

Multiple-choice (MC) tests are widely used in achievement testing at various educational levels because of their objectivity in evaluation and simplicity in scoring. They are commonly administered with the instructions for test-takers to select one response option as the correct answer. Responses are then scored into correct and incorrect with or without a penalty for incorrect responses as in formula scoring or number right scoring, respectively. In terms of psychometric modeling, item response theory (IRT) is one generally used approach to estimate test-takers' ability that underlies the responses. Binary responses are often modeled using the one-, two-, or three-parameter logistic (1-, 2-, 3PL) model. If differences between all response options of a MC item are to be retained, the nominal response model (NRM; Bock, 1972) was extended for MC items by Samejima (1972) and by Thissen and Steinberg (1984). These models specify the probability of choosing each particular response option of an item to gain more information for the estimation of the latent ability.

However, MC tests with the single-best answer instructions have some disadvantages, among which susceptibility to guessing and insensibility to partial knowledge are the two major concerns. Various methods have been developed with the attempt to reduce guessing and extract more information regarding partial knowledge on MC items, such as changing the response method and the associated (dichotomous) scoring rule (Ben-Simon, Budescu, & Nevo, 1997). Previous studies have shown that alternative response methods with partial credit scoring rules can improve the psychometric properties of a test in terms of reliability and validity (e.g., Ben-Simon et al, 1997; Frary, 1982).

Yet, responses from the partial credit testing methods differ from the selection of a single response option, and thus, cannot be directly modeled by the IRT models. Therefore, the purpose

of the present study is to develop an alternative IRT-based modeling approach to MC items administered with partial credit testing procedures. The proposed approach employs the procedure of elimination testing (Coombs, Milholland, & Womer, 1956), which asks test-takers to eliminate all the response options they consider to be incorrect to capture their partial knowledge on MC items, and derives the partial credit model (PCM; Masters, 1982) for the obtained responses.

In the following, a brief summary of commonly used testing methods and modeling approaches to MC items is presented. After a review of the elimination testing procedure, we present our derivation of the PCM for elimination testing approach. Next, the proposed modeling approach is illustrated with a classroom examination of an undergraduate course in engineering science. Finally, the practical implications and potential limitations of the proposed approach are discussed.

Testing and Modeling Approaches to MC Items

Conventional Testing Procedures: the Single-Best Answer Instructions

MC items are commonly administered with the instructions to select a single best answer among the response options. When responses are dichotomized on correctness, an IRT model for binary data is used, such as the 1-, 2-, 3PL model. In particular, in addition to a difficulty and a discrimination parameter, the 3PL model (Birnbbaum, 1968) contains a guessing parameter. This parameter is assessed by a non-zero lower asymptote on the left side on the ability continuum, indicating that the very lower abilities have a probability of selecting the correct answer by guessing. Yet, the 3PL model may not be an optimal approach to MC items. First, there are two types of guessing that can take place on MC items: blind guesses and informed guesses. The

former are as a result of complete ignorance and purely dependent on luck, as expressed in the guessing parameter of the 3PL model, whereas the latter are deliberate choices made by test-takers after they are able to rule out one or more incorrect response options using their partial knowledge on the item. Therefore, the guessing parameter of the 3PL model does not fully address the two types of guessing. Second, dichotomization of responses may result in some loss of information regarding the incorrect responses, as they are treated as equivalent and indistinguishable in the 3PL modeling.

In order to extract more information from the incorrect responses, Bock (1972) developed the NRM specifying the probability of selecting each of the response options of a MC item as a function of the ability and characteristics of each response option. When the NRM is applied to MC items, one disadvantage is that all test-takers with a lower ability will select the very same incorrect response option rather than guess randomly. Samejima (1972) and Thissen and Steinberg (1984) extended Bock's NRM by introducing an additional latent response category of "no recognition" for those who do not know and guess. While Samejima's model assumes a fixed coefficient (the reciprocal of the number of response options of an item) for this "no recognition" category, Thissen and Steinberg's model allows this coefficient to be freely estimated to account for the differential plausibility of individual response options. As more information regarding the selections of distractors is incorporated into the modeling, these polytomous IRT models have the potential advantage of producing increased precision of the estimated latent ability than the dichotomous models, particularly at the lower end of the ability continuum (Bock, 1972).

Apart from the IRT modeling, another distinctive line of research on modeling MC items with single-best answers is within the framework of cognitive diagnostic models (CDMs; Rupp,

Templin, & Henson, 2010), such as the CDM for cognitively based MC options (de la Torre, 2009a), the generalized diagnostic classification models for MC option-based scoring (DiBello, Henson, & Stout, 2015), and using Bayesian networks to identify specific misconceptions associated with particular distractors (Lee & Corter, 2011). One key feature of the CDMs for MC items is that they specify a so-called Q-matrix indicating for each item response option whether a specific pre-determined cognitive attribute or sub-skill is present or required for that response option. The models provide a classification of test-takers with the same (sub)set of attributes. Compared to the IRT modeling, the CDM approach to MC items has the advantage of providing a more differentiated profile for individual test-takers including their specific cognitive (mis)understanding, so that more diagnostic information and remedial instructions are available for learning processes. On the other hand, the relative complex specification of the Q-matrix and the unique pattern of Bayesian network connectivity between specific misconception and items may hinder the implementation of the CDM approach to MC items (de la Torre, 2009b).

Alternative Testing Procedures: Different Response Methods with Partial Credit Scoring

To overcome the disadvantages of guessing and absence of partial knowledge in the single-best answer instructions, researchers have proposed alternative testing procedures with different response methods. Two frequently studied procedures are probability testing and elimination testing.

Probability testing, initially proposed by De Finetti (1965), enables test-takers to assign any numbers ranging from 0 to 1 to represent their subjective probability of each response option being correct. One simplified variant of probability testing is certainty-based marking, sometimes also referred to as confidence-based marking. It requires test-takers only to rate their

degree of certainty on the response option they have chosen to be the correct answer. The response is then scored on the correctness of the chosen option and the degree of certainty the test-taker has in his/her choice. These procedures allow the demonstration of partial knowledge in an elaborate way, but their flexibility in answering patterns also results in the complexity in scoring for standardized testing (Ben-Simon et al., 1997).

Elimination testing (Coombs et al., 1956) asks test-takers to eliminate all the response options they consider to be incorrect, and responses are scored on the number of correction and incorrect eliminations. This is, in principle, equivalent to the subset selection testing by Dressel & Schmid (1953), which requires test-takers to indicate the smallest subset of response options that they believe contains the correct answer. In both the elimination and subset selection testing procedures, test-takers are not forced to end up with only one response option as the correct answer in case they are not confident in doing so. While these two procedures seem to be complementary to each other, studies regarding test-takers' answering behaviors on MC tests showed that responses under the two procedures do not fully match because of different strategies test-takers use and the framing effect (that test-takers perceive themselves in a gaining or losing domain) on their decision making under uncertainty (Bereby-Meyer, Meyer, & Budescu, 2003; Yaniv & Schul, 1997).

MC items can also contain multiple correct answer options with the number of correct options known or unknown to test-takers. Test-takers are asked to select response options they consider to be correct. Bo, Lewis, and Budescu (2015) proposed an option-based partial credit response model for MC items with multiple correct options, using a scoring rule that transforms test-takers' selection(s) of response options into a vector consisting of 1 for the option(s) selected and 0 otherwise. Responses are then scored by counting the number of matches between the

response vector and the key vector. To model the responses, they first specify the probability of a test-taker giving a correct response (selection or not) to each item option, and the product of these probabilities on individual item options gives the probability of a particular response pattern. Next, dividing this product probability by the sum of the probabilities of all permissible response patterns gives the unconditional likelihood of a response pattern on a MC item for a given test-taker.

Although the option-based partial credit response model is one of the few IRT models developed for partial knowledge testing on MC items, some caveats need to be noted. First, in its first modeling step, the model is formulated on an option local independence assumption for mathematical convenience, although it was not the authors' intention to make such an assumption (Bo et al., 2015). However, this condition can hardly be met in practice for MC items. Second, if there is no restriction on how many response options test-takers have to select, it is not possible to determine the nature of non-endorsement on an option. When test-takers do not select a response option, it may be because they indeed identify that option to be incorrect or because they just choose not to respond, i.e., omit that response option. This would clearly affect the estimation of the latent ability. Finally, the response format does not fully and explicitly identify all the knowledge levels that can possibly occur on MC items, such as full misconception and absence of knowledge (see sections below).

The Present Study

Despite the increasing interest in promoting partial credit testing procedures on MC tests (Dressel & Schmid, 1953; Gardner-Medwin, 2006; Lesage, Valcke, & Sabbe, 2013; Lindquist & Hoover, 2015; Vanderoost, Janssen, Eggermont, Callens, & De Laet, 2018), IRT models for

analyzing responses obtained from these partial credit testing procedures have not been as extensively studied as they are under the conventional testing procedures. The present study aims to propose an IRT-based modeling approach to MC items by employing the elimination testing procedure (Coombs et al., 1956) and the PCM (Masters, 1982). After a short review of elimination testing, the proposed PCM approach for elimination testing is derived.

Elimination Testing

Response patterns. Consider a MC item i with m_i response options of which test-takers know only one is correct. Test-takers are asked to eliminate as many incorrect response options as they possibly can. When they have eliminated $m_i - 1$ response options, the response corresponds to choosing a particular response option as in the single-best answer instructions. Using the classification of knowledge levels by Ben-Simon et al., (1997), elimination testing is able to distinguish all possible levels of knowledge: full knowledge (indicated by the elimination of all distractors), partial knowledge (eliminating a subset of distractors), partial misconception (eliminating the correct answer and a subset of distractors), full misconception (eliminating only the correct answer), and absence of knowledge (omission). Elimination testing can also be applied to MC items with r_i multiple correct answers ($0 < r_i < m_i$), with the number of the correct answers r_i known or unknown to test-takers. Table 1 presents the possible knowledge levels and response patterns for a MC item with four response options of which the first one is the correct answer. Note that the response pattern of eliminating of all response options (XXXX) is considered to be irrational in the current study as test-takers know that one of the response options is correct.

Table 1

Possible Knowledge Levels, Response Patterns, Score Rules, and the PCM modeling under Elimination Testing for a Multiple-Choice Item with Four Alternatives of Which the First One Is Correct

Eliminating	Knowledge Level	Response Pattern	Scoring		PCM	
			Coombs et al. (1956)	Arnold and Arnold's (1970)	Sum of binary sub-items	Score level
All distractors	Full Knowledge (FK)	OXXX	1	1	4	6
A subset of distractors	Partial Knowledge 2 (PK2)	OXXO	$\frac{2}{3}$	$\frac{1}{3}$	3	5
		OXOX				
		OOXX				
		OXOO	$\frac{1}{3}$	$\frac{1}{9}$	2	4
	Partial Knowledge 1 (PK1)	OOXO				
		OOOX				
		OOOO	0	0	1	3
		(XXXX*)				
The correct answer and a subset of distractors	Partial Misconception 2 (PM2)	XXXO	$-\frac{1}{3}$	$-\frac{1}{3}$	2	2
		XXOX				
		XOXX				
		XXOO	$-\frac{2}{3}$	$-\frac{1}{3}$	1	1
	Partial Misconception 1 (PM1)	XOXO				
		XOOX				
		XOOO	-1	$-\frac{1}{3}$	0	0
Only the correct answer	Misconception (MIS)					

Note. X: elimination; O: non-elimination; *: the response pattern of eliminating of all response options (XXXX) is considered to be irrational in the current study.

Scoring rules. Responses to an item with elimination testing are scored based on the number of distractors eliminated and whether the correct answer is among the eliminations. Depending on the rationale used, different scoring rules can be applied to the obtained responses. In the original scoring rule of Coombs et al., (1956), each correct elimination of a distractor is rewarded 1 point and a penalty of $m_i - 1$ points is given if the correct answer option is eliminated. Thus, the possible item scores range from $-(m_i - 1)$ to $(m_i - 1)$ and the expected score of random eliminations is zero. An example of the Coombs et al.'s scoring is given for the MC item in Table 1, which is a linear transformation of the original scoring by multiplying a factor of $\frac{1}{3}$. Arnold and Arnold (1970) developed a variant scoring rule based on the game theory. In their scoring rule, a penalty P is also applied to the elimination of the correct answer option to make sure the expected gain due to guessing is zero. Different from Coombs et al.'s scoring rule, each elimination of a distractor is rewarded with a different proportion of partial credit, more specifically, $C_d = \frac{d}{m_i - d} (-P)$ where C_d is the credit rewarded when the d number of distractors eliminated, and no partial credit is given as long as the correct option is eliminated. Hence, for the same MC item in Table 1, if the maximum item score is 1, the penalty and the scores for eliminating zero, one, two, three distractors are $-\frac{1}{3}, 0, \frac{1}{9}, \frac{1}{3}, 1$, respectively.

When comparing these two scoring rules for elimination testing, one can notice that Coombs et al.'s rule gives more credit to partial knowledge and is able to identify all possible knowledge levels, whereas Arnold and Arnold's rule rewards less credit to partial knowledge and does not make further distinctions among misinformation. As Frary (1980) stated that when test makers are favoring either of the scoring rules, the content and purpose of the test should be of concern. In the later application section of the current study, Coombs et al.'s scoring rule was used, because it was one of the objectives to maintain the differentiation among all possible

knowledge levels to facilitate the interpretation of results. Yet, Arnold and Arnold's scoring rule can be particularly useful when a large score inflation due to awarding partial credit is not desired (Vanderoost et al., 2018). In addition, because the highest partial credit does not exceed half of the maximum item score, situations can be avoided that test-takers may be able to pass an exam with only partial knowledge on all items.

Note that the ordering of knowledge levels based on the elimination scoring rules purely represents the ordering of performance levels. It does not refer to the ordering of the cognitive learning process. For example, partial misconception always receives negative scores and thus is associated with lower score levels, but this does not necessarily mean that test-takers with partial misconception have less knowledge than those who do not know. On the contrary, test-takers may start out with ignorance at the lowest level and later develop misconception as they proceed.

Elimination testing in practice. Studies on elimination testing have shown that this testing procedure reduced the amount of full credit responses that may be attributed to lucky guesses (Bradbard, Parker, & Stone, 2004; Chang, Lin, & Lin, 2007; Wu, De Laet, & Janssen, 2018). Vanderoost et al. (2018) did an in-depth between-subject comparison between elimination testing with Arnold and Arnold's scoring rule and formula scoring for a high-stakes exam among students of medicine. They found that scoring methods did not produce a significant difference on the expected test scores, but the amounts of omission responses and guessing were largely reduced in elimination testing. Students in their study preferred elimination testing to formula scoring as the former reduced their test anxiety and improved test satisfaction. In the study of Bush (2001), students also reported a lower level of stress with elimination testing compared to other conventional testing procedures.

Note that elimination testing also gives penalties for eliminating the correct answer. Concerns may arise about the effect of negative penalties on test-takers due to risk aversion (e.g., Budescu & Bar-Hillel, 1993; Budescu & Bo, 2015; Espinosa & Gardeazabal, 2010). However, with elimination testing, test-takers are offered the opportunity to explicitly express both their certainty and uncertainty. In a simulation study, Wu et al., (2018) compared the expected answering patterns under correction for guessing and elimination testing, and showed that risk aversion has a bigger impact for test-takers with partial knowledge, but elimination testing helps to reduce the effect of risk aversion in comparison to correction for guessing. As females are normally considered to be more risk averse, empirical studies (Bond et al., 2013; Bush, 2001; Vanderoost et al., 2018) showed that elimination testing reduced the difference between female and male students in terms of the amount of omission responses and test performance, and therefore is less disadvantageous to risk averse test-takers.

Derivation of the PCM for Elimination Testing

The PCM. In elimination scoring, responses to each of the MC item options can be coded as 1 for each correct response, i.e., a correct elimination of a distractor or a non-elimination of the correct answer, and 0 otherwise. In this way, item i is considered as a testlet with m_i response options as its binary sub-items. These binary sub-items can be summed for item i , resulting in a score varying from 0 to m_i , as is illustrated in the second-to-last column of Table 1.

Huynh (1994) showed that a set of locally independent binary Rasch items is equivalent to a partial credit item with step difficulties in an increasing order. When dependency is created by clustering Rasch items around a common theme, their sum score follows the PCM. Thus, for

the binary sub-items of a MC item, it can be derived that the resulting sum s_i for a person with ability θ follows the PCM as:

$$P(s_i|\theta) = \frac{\exp(s_i\theta - \eta_{is})}{\sum_{j=0}^{m_i} \exp(j\theta - \eta_{ij})} = \frac{\exp(s_i\theta - \sum_{h=0}^s \delta_{ih})}{\sum_{j=0}^{m_i} \exp(j\theta - \sum_{h=0}^j \delta_{ih})} \quad (1).$$

Equation 1 gives two different parameterizations of the PCM for the binary sub-item sum score.

The η_{ij} are the category parameters that represent the difficulty of reaching the sum score s_i on the item, and the δ_{ij} parameters denote the individual step difficulty of scoring from category $j - 1$ to category j . The δ_{ij} parameters also correspond to the point on the latent continuum where the probability of observing category j equals the probability of observing category $j - 1$ on item i . Huynh (1994) pointed out that if the δ_{ij} parameters do not follow an increasing order, it is an indication of dependence among the binary items that compose the sum. In case $\delta_{ij} > \delta_{i,j+1}$, category j is nowhere along the ability continuum the modal response category. As an extension of the Rasch model that assumes equal discrimination across items, applying the PCM to binary sub-items also assumes that all response options have comparable discriminatory power both within and between MC items.

Although the sum of binary sub-items is an easy and evident way to obtain an a priori order of the elimination responses, it is not in line with the constructed classification of knowledge levels. For example, in Table 1 a sum score (of binary sub-items) of 2 can be obtained either by a test-taker with Partial Knowledge 1 eliminating one distractor or by one with Partial Misconception 2 eliminating two distractors and also the correct answer. Therefore, using the sum of binary sub-items does not appear to be an optimal way to classify elimination responses into ordered categories for the PCM.

The ordered partition model (OPM). In elimination testing, each response option of a MC item can be at either one of the two states: eliminated or not eliminated, and responses to a MC item i with m_i response options can yield 2^{m_i} possible answering patterns (cf. response patterns in Table 1). Since different response patterns can lead to an equivalent performance level on an item, a second approach to modeling responses in elimination testing is the OPM by Wilson (1992), which is an IRT model for analyzing responses in partially ordered categories. The advantage of the OPM is that it breaks the one-to-one correspondence between response patterns and score levels, allowing more than one response pattern to be scored onto the same level within an item. For the MC item i with m_i response options, the k_{th} ($k = 1, 2, \dots, 2^{m_i}$) response pattern can be assigned to one of the possible knowledge levels l ($l = 0, 1, \dots, L_i$) using a scoring function $B_i(k)$, so $B_i(k) = l$. Then the OPM can be applied to model the probability of a person with ability θ responding in a particular response pattern y_{ik} at knowledge level l on the item as:

$$P(y_{ik}|\theta) = \frac{\exp(B_i(k) \theta - \eta_{ik})}{\sum_{j=1}^{2^{m_i}} \exp(B_i(j) \theta - \eta_{ij})} \quad (2).$$

The $B_i(k)$ is an integer-valued, a priori defined score function, reflecting the partial ordering of the response patterns, and the η_{ik} are parameters associated with a particular response pattern k . Note that although $B_i(k)$ is equal for all response patterns at the same knowledge level, the η_{ik} parameters are specific for each of the response patterns within that level.

Modeling the probability of all possible response patterns using the OPM can help to investigate differences among response patterns within the same performance level on a MC item. However, a potential drawback is that it may yield a great number of parameters to be

estimated. Say a MC item i with $m_i \geq 5$ response options, there will be at least 32 (2^5) possible response patterns and η_{ik} parameters to be estimated.

The PCM for elimination testing. According to Wilson (1992), the OPM and the PCM are consistent with each other in the way that the PCM estimates only one parameter η_{is} for all response patterns that are scored on the same level s_i , while the OPM has a separate η_{ik} parameter for each of the possible response patterns within the same level s_i . Summing the probabilities of the response patterns y_{ik} of the same level s_i in the OPM equals the probability of getting a sum s_i in the PCM as

$$\eta_{is} = \sum_{\sum y_{ik}=s_i} \eta_{ik} \quad (3).$$

Therefore, the PCM can be seen as special case of the OPM, with responses being modeled in terms of their ordered partition levels.

Built on this equivalence, we derive the PCM approach to elimination testing using the ordered knowledge levels. When a MC item is administered with elimination testing, using the corresponding scoring rule as the scoring function $B_i(k)$, the 2^{m_i} possible response patterns can be partially ordered into $L_i + 1$ score levels $(0, 1, \dots, L_i)$ that are in line with the pre-defined knowledge levels. Instead of modeling the probability of each individual response pattern y_{ik} using the OPM, the probability of responding in a score level l_i (i.e., summing over the response patterns within a knowledge level) is then modeled using the PCM. In order to have integer-valued scores that are necessary to apply the PCM, the values of actual elimination scores should be translated to categories ranging from 0 and L_i (as can be seen in the last column of Table 1).

The generalized partial credit model (GPCM). The PCM assumes equal discrimination power both within and across MC items. If this condition is not met, the generalized partial credit model (GPCM; Muraki, 1992) can be used, which extends the PCM by incorporating a varying slope parameter α_i across items:

$$P(s_i|\theta) = \frac{\exp[a_i(s_i\theta - \eta_{is})]}{\sum_{j=0}^{m_i} \exp[a_i(j\theta - \eta_{ij})]} \quad (4).$$

Thus, in the GPCM the discrimination power of an item is a combination of the slope parameter α_i and the set of category parameters η_{ij} of the item (Muraki, 1992).

Application

As an illustration, the proposed PCM for elimination testing modeling approach is applied to a MC test in the context of a classroom examination in engineering science.

Data

Two MC tests were administered with the elimination testing procedure to 425 undergraduates of engineering science following the course *Electrical Circuits*. A trial test was first carried out in the middle of the semester so that students could familiarize themselves with the new testing procedure. Students completed afterwards a questionnaire regarding their opinions about elimination testing. At the end of the semester, the second MC test was administered as the final examination, of which the data were analyzed in the current study. The final exam consisted of 24 MC questions that were based on the course content throughout the semester. Each MC question had four response options, of which only one was the correct answer. For each response option, there were two choices—“possible” and “impossible”. If students knew the correct answer, they were asked to mark the correct response option as

possible and all the incorrect ones as impossible; if they did not know the correct answer, they should mark the response option(s) that they could eliminate with certainty as impossible and the response option(s) they thought might be correct as possible. For each correct elimination of a distractor, $+\frac{1}{3}$ was awarded, whereas a penalty of -1 point was given if the correct answer was eliminated, following the linear transformation of the original Coombs et al.'s (1956) scoring rule shown in Table 1.

Data Analysis

Score group analysis. The response data were first examined using the score group analysis (Eggen & Sanders, 1993). In this analysis, students were first classified into four score groups using the quartiles of the total test scores based on the eliminating scoring rule mentioned above. For each score group, the proportions of students (a) who considered each of the response options as a possible correct answer (i.e., not eliminating) and (b) who scored on each of the knowledge levels on the item were calculated. These proportions were plotted to give a visualization of (a) the attractiveness of each response option and (b) the relative frequency of each knowledge level within each score group on a MC item.

The score group analysis presents some psychometric characteristics of the classic testing theory. The proportion of Full Knowledge (FK) equals the item difficulty (p -value) and the slope of FK gives an indication of the item-total correlation. To enable a better visualization, Partial Knowledge 1 and 2 (PK1 and PK2) were merged into partial knowledge (PK), and Partial Misconception 1 and 2 (PM1 and PM2) and Misconception (MIS) into misconception (MI).

Psychometric modeling. Following the PCM for elimination testing, responses on each item were scored into seven ordered performance (knowledge) levels ranging from 0 to 6 (cf., the last column in Table 1). However, Level 0 (only eliminating the correct answer; MIS) was not observed on 10 items. In line with Wilson and Masters (1993), this level was collapsed by downcoding the levels above it. Thus, for those 10 items the score levels varied from 0 (PM1) to 5 (FK). The responses from elimination testing were analyzed using the PCM and the GPCM with the marginal maximum likelihood estimation. The step parameters δ_{ij} were estimated. All analyses were performed using the *mirt* package (Chalmers, 2012) in R.

Model fit. According to Muraki (1997), the goodness-of-fit of polytomous IRT models can be tested item by item. Thus, the fit of the model to each item was tested by means of the Orland and Thissen's chi-square statistics (Chalmers, 2012) comparing the observed and expected response frequencies from the model in aggregated ability groups. In case of misfit, the empirical plot was examined in which the observed proportions of each response category within each group were plotted against the modeled response curve of that category to identify the location(s) of misfit.

Comparison with the dichotomous modeling. Additionally, in order to compare the current approach with the binary score modeling approach on the precision of ability estimation, the responses of elimination testing were dichotomized. FK responses were considered as correct and all the misconception responses (PM1, PM2, and MIS) as incorrect. Three different categorizations of the partial knowledge responses (PK1 and PK2) were performed by considering (1) both to be correct; (2) both to be incorrect; and (3) PK1 to be incorrect and PK2

to be correct. The Rasch model was used to estimate person abilities. The average standard errors of the ability estimates and the information from the Rasch model and the PCM for elimination testing were compared.

Results

Descriptive Statistics

Figure 1 presents the percentages of the responses in each knowledge level on the 24 MC items. It shows that for most of the items the dominant response pattern was to eliminate three distractors showing FK, except for Item 8, 19, 20, and 21. On Item 8, the percentages of the responses in each knowledge level were very similar, and about one third of the students showed partial knowledge (PK1 or PK2). On Item 19, 20 and 21, omission (NK) was the most frequently observed response among students. Correspondingly, these four items had relatively higher item difficulties in terms of the proportion of a full credit response.

 Insert Figure 1 about here

Figure 1. Percentages of the elimination responses on each knowledge levels on the MC items.

Score group analysis. The score group analysis revealed several different patterns across items, varying in item difficulties and slopes. As an illustration, two relatively extreme examples are presented in Figure 2 and 3.

Figure 2 gives students' responses on Item 14 (a) at the level of response options and (b) at the item level in terms of knowledge levels. Option C was the correct answer. The plots show that the proportions of students considering options A, B, and D being the possible correct

answer decreased as the abilities increased (Figure 2a), corresponding to the increased proportion of FK across score groups at the item level (Figure 2b). It is interesting to notice in Figure 2a that although students in lower ability groups were less able to single out option C as the correct answer, the percentage of them accepting it as a possible answer was still high. This pattern of three distractors being easily identified and eliminated suggested that Item 14 was a rather easy item in terms of the proportion of the correct response, evidenced by a mean item score of 0.76.

 Insert Figure 2 about here

Figure 2. Score group analysis of the responses on Item 14 at the response option level (not-eliminating) (a) and at the knowledge level (b). Option C was the correct answer. FK = Full Knowledge; PK = Partial Knowledge; NK = No Knowledge; MI = Misconception.

 Insert Figure 3 about here

Figure 3. Score group analysis of the responses on Item 20 at the response option level (not-eliminating) (a) and at the knowledge level (b). Option A was the correct answer. FK = Full Knowledge; PK = Partial Knowledge; NK = No Knowledge; MI = Misconception.

Figure 3 presents the corresponding plots of responses on Item 20 with Option A being the correct answer. At the response option level (Figure 3a), there was also a high percentage of non-elimination of the correct answer A, but for the lower three score groups the percentages of not eliminating the distractors were much higher as opposed to those on Item 14 in Figure 2a. This shows that for those students it was not very easy to identify options B, C, and D to be

incorrect and eliminate them, resulting in a dominant category of NK at the item level (Figure 3b). Only for the highest ability group, an increase in the elimination of the three distractors was observed in the left panel, leading to a higher percentage of FK on the item in the right panel. Correspondingly, Item 20 had a mean item score of 0.20.

PCM Modeling

Item category characteristic curves (ICCCs). The estimated partial credit step parameters δ_{ij} for each item are presented in Appendix A. From the results, it can be seen that the step parameters δ_{ij} were non-increasing for all items, suggesting that the elimination of each response option within a MC item was not independently made. When the modeled probabilities of each category are plotted along the ability scale, two main patterns can be discerned from the ICCCs.

The first pattern, observed on 14 of the 24 items, is illustrated in Figure 4a, which shows the ICCCs of Item 14. The breaking of the rank order of the step parameters was very pronounced going from Category 4 (PK1) to 5 (PK2) ($\delta_{i5} < \delta_{i4}$). Hence, Category 4 (PK1) was nowhere a modal response category along the ability continuum. Furthermore, the rank order of the step parameters was non-increasing from Category 5 (PK2) to 6 (FK) ($\delta_{i6} < \delta_{i5}$), and consequently Category 5 (PK2) was never the modal response category on the ability continuum, neither. This pattern implies that the three distractors of those items were not as attractive as the correct answer and could be easily identified by students. Once students were able to eliminate one distractor, tentatively reaching PK1, it was more likely for them to continue to eliminate at least one more distractor and end up in Category 6 (FK).

The second pattern is presented in Figure 4b, showing the ICCCs of Item 5. Compared to the pattern in Figure 4a, the reversing of the rank order of the step parameters was further observed at going from Category 3 (NK) to 4 (PK1) (consequently, $\delta_{i6} < \delta_{i5} < \delta_{i4} < \delta_{i3}$). This indicates that Category 3 (NK), 4 (PK1), and 5 (PK2) would not be the modal response categories on the ability continuum. Instead, Category 2 (PM2) and 6 (FK) were the two most probable response categories. Note that both response categories were associated with eliminating three response options on an item. Hence, this result implies that on those items when students have correctly eliminated two distractors, they tended to choose one particular response option from the remaining two as the answer, ending up either in Category 2 (PM2) when a distractor was chosen, or in Category 6 (FK) when the correct answer was chosen. The probability of ending up with PM2 versus FK depended on the latent ability as shown in Figure 4b.

 Insert Figure 4 about here

Figure 4. Response category curves of Items 14 (a) and Item 5 (b) from the PCM modeling for elimination testing.

Item fit. The item fit statistics of the PCM can be found in Appendix A. The PCM did not provide an adequate fit for five items of the test. To examine the misfit, Figure 5 presents the empirical plot for Item 6 as a typical example of the mis-fitted items. First, it can be seen that the model differentiated a larger range on the ability scale, while the students in the current test were centered in a relatively small range on the ability scale from -1 to 1 . Second, the plots show that most of the misfit was found in the categories of PK1 and PK2. Lower ability students seemed to

have a higher proportion of scoring in these two partial knowledge categories than the model predicted, whereas the opposite results were found for higher ability students. This confirms the results from the modeling that when having correctly eliminated at least one distractor, students tended to continue eliminating. Moreover, the plots further distinguished that it was higher ability students that chose to continue eliminating and obtained FK, supported by the overestimation of Category 5 (PK2) and the underestimation of Category 6 (FK) in the modeled ICCCs for them. On the other hand, lower ability students tended to withdraw their elimination and opted for PK 1 and PK 2, as these categories were under-represented by the model for these ability groups. In case they went on with elimination, they would be more likely to indicate a wrong response option as the correct answer (PM2), supported by a higher proportion of PM2 observed in the lower ability groups than the model predicted.

 Insert Figure 5 about here

Figure 5. The empirical plot of the observed proportions of the responses against the modeled response category curves in aggregated ability groups on Item 6.

GPCM Modeling

The parameter estimates and the item fit statistics of the GPCM can be found in Appendix B. The slope parameters α_i ranged from 0.19 on Item 8 to 1.21 on Item 2. The ICCCs showed the similar pattern as those from the PCM, but because almost all items, except Item 21, had a slope parameter estimate smaller than one, the ICCCs from the GPCM were flatter on the ability scale. It is also interesting to note that the GPCM indeed improved the fit statistics on the

items that had a poor fit in the PCM, but for two items (Item 13 and 20) the fit was less good in the GPCM than in the PCM.

Comparison with the Dichotomous Modeling

The partial credit responses in elimination testing were dichotomized in three ways, and modeled using the Rasch model. On average, the PCM for elimination testing provided more accurate estimation of the ability as opposed to the dichotomous modeling approach. The average standard error of the ability estimates in the PCM (0.15) was three times smaller than those in the three dichotomous Rasch models, namely, 0.48 for scoring all PK as correct, 0.51 for all PK as incorrect, and 0.48 for the split approach.

At the item level, the PCM for elimination testing yielded much higher information on all items in comparison to the dichotomous Rasch modeling. Figure 6 plots the information functions of (a) Item 9 and (b) Item 21 as two examples. The solid line depicts the information function from the PCM for elimination testing, and the dotted lines are those from the Rasch modeling for the recoded binary responses. The three information functions of the dichotomous modeling showed similar patterns, except that considering all PK responses as wrong was a bit more informative for the higher abilities. On the other hand, the information from the PCM for elimination testing was much higher than all three Rasch modeling approaches. In particular, Figure 6b shows that besides the maximum information peak around the ability level of 0.5, there was also a relatively larger amount of information on the curve around the ability level of -1 . This shows that the PCM for elimination testing also provided a bit more information in the lower range of the ability where few FK responses were observed.

 Insert Figure 6 about here

Figure 6. Information functions of Items 9 (a) and Item 21 (b). The solid line depicts the information function from the PCM for elimination testing, and the dotted lines are those from the three dichotomous Rasch modeling.

Discussion

The application illustrates the feasibility and efficacy of the PCM for elimination testing in practice. The descriptive score group analysis and the psychometric modeling showed that the proposed approach improved the estimation of test-takers' ability, and also provided additional information on the response behaviors of test-takers of different ability levels.

On the other hand, the PCM modeling failed to yield a statistically good model fit for some items, with the misfit mostly occurring at the partial knowledge categories. This can be partly due to the fact that elimination testing was still new to the students (as this was their first few encounters with elimination testing) and they still held a predominant tendency of attempting to identify the correct answer as in the conventional testing instructions. If students are not fully familiar with the testing procedure, it will be less likely for them to make use of the advantages of elimination testing. Yet, previous studies have shown that despite elimination testing initially being new and more complex to them, students still preferred it to the conventional instructions (e.g., Bond et al., 2013; Bradbard et al., 2004; Bush, 2001; Coombs et al., 1956; Vanderoost et al., 2018). Therefore, more acquaintance and practice for students may be necessary when this new testing procedure is being implemented, so that students can get

more accustomed to it to demonstrate their partial knowledge when they doubt about the single-best answers.

Another possible reason for the partial misfit may be that the MC test consisted of a number of easy items. It can be seen from Figure 1 that the full credit response was the dominant category on most of the items. When students can easily solve a MC item, it is less likely to observe partial knowledge on the item. A closer examination of the ICCCs indeed showed that more able students were prone to choose one response option after having correctly eliminated two distractors, whereas less able students tended to be conservative in their elimination. Nevertheless, it can be considered as an added value of the proposed approach to be more informative and diagnostic for examiners to investigate the different answering behaviors of students with different levels of ability in the response process. Note that the differentiation across categories among students requires a large range of the ability scale, but normally the abilities of a certain group of students are only observed within a limited range. A test composed of items of a wide range of difficulties may provide more insights into the performance of the model.

Finally, the results that the GPCM did not lead to a fully fitting model suggest that the addition of a varying slope parameter did not necessarily lead to a more convenient summary of the response data.

General Discussion

The present study proposed a relatively simple and straightforward partial credit modeling approach to MC items administered with the elimination testing procedure. By capturing partial knowledge and incorporating information on distractors into the modeling, the

proposed PCM for elimination testing has several theoretical advantages as well as some limitations.

Advantages

The first advantage of the proposed approach is that more information is available for estimating the latent ability. By rewarding partial credit, a differentiation is made between knowledge levels that may not lead to the correct response, but nevertheless differ with respect to the level of understanding. Hence, a fine-grained measure and estimation of test-takers' latent ability can be achieved from the fixed set of MC items.

Second, a natural a priori ordering of the responses can be obtained even if the response options of a MC item are not ordered with respect to their quality. Although it is possible in principle to compose response options that correspond to increasing levels of (partial) knowledge, it mostly implies a heavy investment in item writing. This is in contrast to some of the CDMs for MC items (e.g., de la Torre, 2009a; DiBello et al., 2015) where a key element in the modeling is the definition and specification of cognitively based response options that are associated with specific desired attributes.¹ This requires a large amount of joint effort from experts of different disciplines. With elimination testing, responses are no longer a single selection of one option, but a vector consisting of responses to the individual options of the MC item. The associated scoring rules can then provide a meaningful way of partitioning the response patterns into partially ordered score levels that correspond to increasing performance levels on MC items.

Third, one of the popular response process theories for MC items with the conventional testing instructions is the attractiveness theory. It states that the probability of choosing an

response option depends on how attractive it is in comparison to the other options. This response theory is formalized in models such as the NRM and its extensions by modeling the probability of choosing a response option as the ratio of a term for the option of interest to the sum of such terms for all options. Our PCM for elimination testing follows another possible response process, which posits that test-takers evaluate each response option and eliminate the one(s) they recognize to be incorrect, followed by a possible comparison of the remaining ones (Ben-Simon et al., 1997; Frary, 1988; Lindquist & Hoover, 2015). This elimination process theory is also adopted in the Nedelsky IRT model (Bechger, Maris, Verstralen, & Verhelst, 2005). The conventional approach of scoring MC items dichotomously and applying the 1-, 2-, or 3PL model is not in line with either of these two covert response process theories for solving MC items. However, this does not mean the dichotomous modeling approach is invalid, but being consistent with a response strategy of how test-takers answer MC items can be considered as another attractive feature of the proposed approach.

Finally, when the proposed PCM for elimination testing is used to model the responses in terms of the order score (knowledge) levels, it is assumed that the different response patterns within the same score level are interchangeable. In contrast to the NRM approach, the response curves of each item option are not modeled separately, but the equivalent information can be obtained from the descriptive score group analysis.

Limitations

The first limitation of the proposed approach is that it is a model with strong assumptions. It expects test-takers to follow the process of distractor elimination and report their level of knowledge on each item. This process may not be prominent for MC items of all content

domains. For instance, compare choosing a synonym from a list of words versus choosing the correct numeric value after having performed a complex calculation. In the latter case, test-takers may still be able to eliminate certain distractors that contain implausible values given the problem. Also, when test-takers are not fully familiar with the testing procedure and therefore make less use of the advantage of this elimination strategy, or when a MC test consists of a large number of easy items, it is less likely to obtain sufficient observations of partial knowledge responses.

On the other hand, concerns may arise that elimination testing leads test-takers to use certain “tricks” to easily eliminate incorrect response options, for example, by looking for signifiers like “always” and contradiction/resemblance among options. Awarding partial credit to such eliminations may reduce test validity. These concerns indeed put a requisite on the quality of distractors, but to have attractive distractors is also a requisite for constructing good MC items in general. In the presence of an “easy” distractor, an explicit elimination would be more informative than an implicit exclusion and a potential guess on the remaining response options.

Furthermore, there may be some additional factors inherent in MC tests that are not fully captured by the proposed approach. The over- or underestimation of the occurrences of the PK levels in different ability groups as well as the result that some items had a worse fit in the GPCM may imply that students make an answering decision not only based on their ability, but also on other factors. Research shows that different scoring or testing procedures can to some extent affect the answering behaviors of test-takers with partial knowledge (e.g., Bereby-Meyer et al., 2003; Wu et al., 2018). In the current application the original Coombs et al.’s scoring rule was used, which was “generous” in rewarding partial knowledge, and it was possible that students could pass the test with only partial knowledge on all the items. It requires further

research on whether the proposed approach is robust under different scoring rules for elimination testing, such as Arnold and Arnold's. Yet, which scoring rule to use should be clearly communicated to test-takers beforehand, and it is not a decision that can be made in the modeling process afterwards.

Nevertheless, the proposed PCM for elimination testing approach provides an alternative way to score and model the responses to MC items. It shares with the dichotomous approach the relative simplicity of scoring as well as with the polytomous approach that more information about incorrect response options is taken into account for the latent ability estimation.

Acknowledgements

The authors wish to thank the editor George Engelhard, the associate editor Jonathan Templin, and the three anonymous reviewers for their helpful comments and suggestions on this work.

Notes

¹ One anonymous reviewer has pointed out that the diagnostic approach using Bayesian networks to diagnosing misconceptions from selections of distractors (e.g., Lee & Corter, 2011) does not bear such complexity. The Bayesian net inference method can also lead to a data-based estimated ordering of the response options as the PCM for elimination testing approach proposed in the current study.

Appendix A

Parameter Estimates and Item Fit Statistics from the PCM

Item (<i>i</i>)	Step Parameters (δ_{ij})						Item fit		
	1	2	3	4	5	6	χ^2	<i>df</i>	<i>p</i> -value
1	-2.86	-1.85	1.14	—	-1.96	-1.85	85.89	57	0.008*
2	—	-1.89	0.59	-0.77	-1.12	-2.54	54.54	42	0.093
3	-2.23	-0.72	-1.02	0.90	-1.77	-2.57	34.67	39	0.668
4	-2.32	-3.00	-0.05	0.46	0.87	-3.00	97.70	81	0.100
5	—	-3.17	1.95	-0.55	-1.99	-1.57	75.00	65	0.186
6	-1.32	-3.07	-0.07	0.53	-2.00	-2.14	80.10	49	0.003*
7	-2.48	-2.36	-0.95	0.42	0.16	-1.58	135.00	103	0.019
8	—	-2.54	-0.95	0.06	-0.18	-0.10	176.18	130	0.004*
9	—	-2.56	-1.07	0.10	0.06	-1.90	106.91	92	0.137
10	0.83	-3.87	-1.42	0.91	-0.06	-2.31	94.27	75	0.066
11	-1.63	-2.47	-2.27	2.90	-1.74	-1.53	91.43	62	0.009*
12	—	-0.47	-4.18	1.40	-0.80	-1.27	111.59	83	0.020
13	-1.23	-0.57	-3.09	1.60	0.05	-2.37	78.81	55	0.019
14	—	-2.86	-1.46	-0.28	-1.01	-1.75	73.92	65	0.210
15	—	-1.13	1.03	1.31	-3.82	-1.25	61.06	63	0.546
16	-1.83	-2.33	0.29	0.59	-2.23	-1.34	94.83	73	0.044
17	-0.91	-1.59	-1.54	3.18	-2.66	-2.52	52.82	50	0.366
18	—	-1.25	-0.78	1.19	-1.26	-1.01	118.10	97	0.072
19	-1.90	-1.23	-2.60	1.62	0.39	-2.02	118.51	69	0.000*
20	—	-1.94	-2.62	2.40	-0.36	-0.58	73.72	59	0.094
21	-3.85	-0.76	-1.82	3.99	-1.95	-0.54	98.60	74	0.030
22	—	-3.30	1.07	0.10	-0.18	-3.47	44.95	45	0.474
23	-2.63	-2.70	0.54	2.03	-2.23	-2.43	72.93	63	0.184
24	-0.57	-3.69	-1.77	1.34	-0.47	-2.13	96.13	67	0.011

Note. * $p < .01$.

Appendix B

Parameter Estimates and Item Fit Statistics from the GPCM

Item (<i>i</i>)	Discr (a_i)	Step Parameters (δ_{ij})						Item fit		
		1	2	3	4	5	6	χ^2	<i>df</i>	<i>p</i> -value
1	0.58	-5.07	-3.34	1.86	—	-3.45	-3.21	75.88	54	0.026
2	1.21	—	-2.47	-0.31	-1.34	-1.44	-2.26	31.56	36	0.680
3	0.57	-4.03	-1.41	-1.91	1.48	-3.17	-4.51	32.76	36	0.623
4	0.36	-5.87	-7.86	0.28	1.59	2.60	-8.43	96.71	85	0.181
5	0.20	—	-13.94	11.31	-1.51	-9.29	-7.92	94.65	73	0.045
6	0.76	-2.26	-4.52	-0.53	0.33	-2.86	-2.85	66.17	48	0.042
7	0.28	-7.74	-7.52	-2.63	2.06	0.83	-5.94	129.48	105	0.053
8	0.19	—	-11.53	-3.59	1.17	-0.77	-1.46	172.92	140	0.031
9	0.58	—	-4.57	-1.96	0.10	0.08	-3.25	101.34	88	0.157
10	0.70	0.79	-5.83	-2.31	1.08	-0.18	-3.20	82.40	69	0.129
11	0.63	-2.84	-4.16	-3.79	4.54	-2.76	-2.34	71.81	60	0.141
12	0.38	—	-0.78	-10.63	3.97	-1.95	-3.42	103.53	83	0.063
13	0.61	-2.22	-1.11	-5.21	2.53	0.07	-3.79	80.37	51	0.005*
14	0.30	—	-8.55	-3.99	-0.20	-2.92	-5.81	91.89	70	0.041
15	0.43	—	-2.38	2.64	3.24	-8.79	-2.89	59.35	62	0.572
16	0.31	-4.94	-6.70	1.66	2.48	-6.85	-4.32	103.93	78	0.027
17	0.56	-1.72	-2.93	-2.82	5.55	-4.73	-4.44	47.18	47	0.465
18	0.39	—	-2.84	-1.69	3.30	-3.13	-2.69	124.12	100	0.051
19	0.48	-3.82	-2.46	-5.33	3.46	0.87	-4.22	115.76	68	0.000*
20	0.31	—	-5.62	-8.04	8.05	-1.29	-2.59	98.63	59	0.001*
21	0.25	-13.91	-2.13	-6.67	16.15	-8.08	-3.30	96.48	81	0.115
22	0.44	—	-7.31	2.67	0.40	-0.29	-7.91	43.71	44	0.484
23	0.47	-5.38	-5.58	1.23	4.34	-4.61	-5.09	67.35	57	0.164
24	0.44	-0.95	-8.13	-3.80	3.16	-0.96	-4.78	98.37	69	0.012

Note. * $p < .01$.

References

- Arnold, J. C., & Arnold, P. L. (1970). On scoring multiple choice exams allowing for partial knowledge. *The Journal of Experimental Education*, 39, 8–13.
doi:10.1080/00220973.1970.11011223
- Bechger, T. M., Maris, G., Verstralen, H. H. F. M., & Verhelst, N. D. (2005). The Nedelsky model for multiple-choice items. In L. A. van der Ark, M. A. Croon, & K. Sijtsma (Eds.), *New developments in categorical data analysis for the social and behavioral sciences*. (pp. 187–206). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.
- Ben-Simon, A., Budescu, D. V., & Nevo, B. (1997). A comparative study of measures of partial knowledge in multiple-choice tests. *Applied Psychological Measurement*, 21, 65–88.
- Bereby-Meyer, Y., Meyer, J., & Budescu, D. V. (2003). Decision making under internal uncertainty: The case of multiple-choice tests with different scoring rules. *Acta Psychologica*, 112, 207–220. doi:10.1016/S0001-6918(02)00085-9
- Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord, M. R. Novick, & A. Birnbaum (Eds.), *Statistical theories of mental test scores* (pp. 374–472). Oxford, England: Addison-Wesley.
- Bo, Y., Lewis, C., & Budescu, D. V. (2015). An option-based partial credit item response model. In R. E. Millsap, D. M. Bolt, L. A. van der Ark, & W.-C. Wang (Eds.), *Quantitative psychology research. Springer proceedings in mathematics & statistics* (Vol. 89, pp. 45–72). Cham, Switzerland: Springer. doi:10.1007/978-3-319-07503-7_4
- Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, 37, 29–51. doi:10.1007/BF02291411

- Bond, A. E., Bodger, O., Skibinski, D. O. F., Jones, D. H., Restall, C. J., Dudley, E., & van Keulen, G. (2013). Negatively-marked MCQ assessments that reward partial knowledge do not introduce gender bias yet increase student performance and satisfaction and reduce anxiety. *PLoS ONE*, 8(2), e55956. doi:10.1371/journal.pone.0055956
- Bradbard, D. A., Parker, D. F., & Stone, G. L. (2004). An alternate multiple-choice scoring procedure in a macroeconomics course. *Decision Sciences Journal of Innovative Education*, 2, 11–26. doi:10.1111/j.0011-7315.2004.00016.x
- Budescu, D., & Bar-Hillel, M. (1993). To guess or not to guess: A decision-theoretic view of formula scoring. *Journal of Educational Measurement*, 30, 277–291. doi:10.1111/j.1745-3984.1993.tb00427.x
- Budescu, D. V., & Bo, Y. (2015). Analyzing test-taking behavior: Decision theory meets psychometric theory. *Psychometrika*, 80, 1105–1122. doi:10.1007/s11336-014-9425-x
- Bush, M. (2001). A multiple choice test that rewards partial knowledge. *Journal of Further and Higher Education*, 25, 157–163. doi:10.1080/03098770120050828
- Chalmers, R. P. (2012). mirt : A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, 48(6), 1–29. doi:10.18637/jss.v048.i06
- Chang, S.-H., Lin, P.-C., & Lin, Z.-C. (2007). Measures of partial knowledge and unexpected responses in multiple-choice tests. *Educational Technology & Society*, 10(4), 95–109.
- Coombs, C. H., Milholland, J. E., & Womer, F. B. (1956). The assessment of partial knowledge. *Educational and Psychological Measurement*, 16, 13–37. doi:10.1177/001316445601600102

- De Finetti, B. (1965). Methods for discriminating levels of partial knowledge concerning a test item. *British Journal of Mathematical and Statistical Psychology*, 18, 87–123.
doi:10.1111/j.2044-8317.1965.tb00695.x
- 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
- 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. doi:10.1177/0146621614561315
- Dressel, P. L., & Schmid, J. (1953). Some modifications of the multiple-choice item. *Educational and Psychological Measurement*, 13, 574–595.
doi:10.1177/001316445301300404
- Eggen, T., & Sanders, P. F. (1993). *Psychometrie in de praktijk* [Psychometrics in practice]. Arnhem, the Netherlands: Cito Instituut voor Toetsontwikkeling.
doi:10.2143/NK.25.1.2017369
- Espinosa, M. P., & Gardezabal, J. (2010). Optimal correction for guessing in multiple-choice tests. *Journal of Mathematical Psychology*, 54, 415–425. doi:10.1016/j.jmp.2010.06.001
- Frary, R. B. (1980). The effect of misinformation, partial information, and guessing on expected multiple-choice test item scores. *Applied Psychological Measurement*, 4, 79–90.
doi:10.1177/014662168000400109

- Frary, R. B. (1988). Formula scoring of multiple-choice tests (correction for guessing). *Educational Measurement: Issues and Practice*, 7(2), 33–38. doi:10.1111/j.1745-3992.1988.tb00434.x
- Gardner-Medwin, A. R. (2006). Confidence-based marking - Towards deeper learning and better exams. In C. Bryan & K. Clegg (Eds.), *Innovative assessment in higher education* (pp. 141–149). London, England: Routledge. Retrieved from <http://www.tmedwin.net/~ucgbarg/tea/innovass7.pdf>
- Huynh, H. (1994). On equivalence between a partial credit item and a set of independent Rasch binary items. *Psychometrika*, 59, 111–119. doi:10.1007/BF02294270
- Lee, J., & Corter, J. E. (2011). Diagnosis of subtraction bugs using Bayesian networks. *Applied Psychological Measurement*, 35, 27–47. doi:10.1177/0146621610377079
- Lesage, E., Valcke, M., & Sabbe, E. (2013). Scoring methods for multiple choice assessment in higher education – Is it still a matter of number right scoring or negative marking? *Studies in Educational Evaluation*, 39, 188–193. doi:10.1016/j.stueduc.2013.07.001
- Lindquist, E. F., & Hoover, H. D. (2015). Some notes on corrections for guessing and related problems. *Educational Measurement: Issues and Practice*, 34(2), 15–19. doi:10.1111/emip.12072
- Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47, 149–174. doi:10.1007/BF02296272
- Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, 16, 159–176. doi:10.1177/014662169201600206
- Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic measurement: theory, methods, and applications*. New York, NY: Guilford Press.

- Samejima, F. (1972). *A new family of models for the multiple choice item* (Research No. 79–4). Knoxville: University of Tennessee. Retrieved from <http://www.dtic.mil/dtic/tr/fulltext/u2/a080350.pdf>
- Thissen, D., & Steinberg, L. (1984). A response model for multiple choice items. *Psychometrika*, 49, 501–519. doi:10.1007/BF02302588
- Vanderoost, J., Janssen, R., Eggermont, J., Callens, R., & De Laet, T. (2018). Elimination testing with adapted scoring reduces guessing and anxiety in multiple-choice assessments, but does not increase grade average in comparison with negative marking. *PLoS ONE*, 13(10), e0203931. doi:10.1371/journal.pone.0203931
- Wilson, M. (1992). The ordered partition model: An extension of the partial credit model. *Applied Psychological Measurement*, 16, 309–325. doi:10.1177/014662169201600401
- Wilson, M., & Masters, G. N. (1993). The partial credit model and null categories. *Psychometrika*, 58, 87–99. doi:10.1007/BF02294473
- Wu, Q., De Laet, T., & Janssen, R. (2018). Elimination scoring versus correction for guessing: A simulation study. In M. Wiberg, S. Culpepper, R. Janssen, J. González, & D. Molenaar (Eds.), *Quantitative psychology. Springer proceedings in mathematics & statistics* (Vol. 233, pp. 183–193). Cham, Switzerland: Springer. doi:10.1007/978-3-319-77249-3_16
- Yaniv, I., & Schul, Y. (1997). Elimination and inclusion procedures in judgment. *Journal of Behavioral Decision Making*, 10(3), 211–220. doi:10.1002/(SICI)1099-0771(199709)10:3<211::AID-BDM250>3.0.CO;2-J