Educational and Psychological Measurement
http://epm.sagepub.com/

# Comparison of Exposure Controls, Item Pool Characteristics, and Population Distributions for CAT Using the Partial Credit Model

HwaYoung Lee and Barbara G. Dodd

*Educational and Psychological Measurement* 2012 72: 159 originally published online 24 June 2011
DOI: 10.1177/0013164411411296

Educational and Psychological Measurement
72(1) 159–175

DOI: 10.1177/0013164411411296
http://epm.sagepub.com

# Comparison of Exposure Controls, Item Pool Characteristics, and Population Distributions for CAT Using the Partial Credit Model

**HwaYoung Lee[1] and Barbara G. Dodd[1]**

**Abstract**

This study investigated item exposure control procedures under various combinations of item pool characteristics and ability distributions in computerized adaptive testing based on the partial credit model. Three variables were manipulated: item pool characteristics (120 items for each of easy, medium, and hard item pools), two ability distributions (normally distributed and negatively skewed data), and three exposure control procedures (randomesque procedure, progressive–restricted procedure, and maximum information procedure). A number of measurement precision indexes such as descriptive statistics, correlations between known and estimated ability levels, bias, root mean squared error, and average absolute difference, exposure rates, item usage, and item overlap were computed to assess the impact of matched or nonmatched item pool and ability distributions on the accuracy of ability estimation and the performance of exposure control procedures. As expected, the medium item pool produced better precision of measurement than both the easy and hard item pools. The progressive–restricted procedure performed better in terms of maximum exposure rates, item average overlap, and pool utilization than both the randomesque procedure and the maximum information procedure. The easy item pool with the negatively skewed data as a mismatched condition produced the worst performance.

**Keywords**

exposure control procedure, computerized adaptive testing, item response theory

[1]University of Texas at Austin, Austin, TX, USA

**Corresponding Author:**
HwaYoung Lee, University of Texas at Austin, Department of Educational Psychology, 1 University Station D5800, Austin, TX 78712, USA
Email: hwayoung@mail.utexas.edu

## Introduction

Computerized adaptive testing (CAT) has become widely used as an alternative to paper-and-pencil testing. The main concept behind CAT is the administration of items that are most appropriate to each examinee's current ability. For example, if an examinee answers an item correctly, a more difficult item will be administered as a next item. However, if the examinee answers an item incorrectly, a less difficult item will be administered as a next item. One of the main advantages of computerized adaptive tests over paper-and-pencil testing is test efficiency, because the CAT yields scores with equal or greater reliability while administering fewer questions (Lord, 1970; Urry, 1971; Vale & Weiss, 1975; Weiss, 1982). However, test security is an issue in such high-stakes testing even though improvement of test security is known as one of the main advantages of CAT. Basically, CAT algorithms choose the most informative items based on the examinee's estimated ability level, so more popular items can be administered frequently, whereas some items may be administered less frequently, or never be administered at all. If examinees with similar abilities share in the sets of items they were given with examinees who have yet to take the test, test security is threatened. Therefore, research investigating ways to improve test security for CAT is necessary.

Many strategies have been developed to control overexposed items in an effort to improve test security. A review of the literature indicated that exposure control procedures under polytomous models did not perform differently based on the measurement precision indexes as measured by correlations between known and estimated ability levels, bias and root mean squared error (RMSE). However, exposure properties, such as maximum exposure rate, pool utilization, and item overlap rates of the polytomous items, were different across exposure control procedures (Boyd, Dodd, & Choi, 2010). For example, the randomesque procedure and the modified within .10-logits procedure performed better than the Sympson–Hetter procedure under polytomous models even though the Sympson–Hetter procedure provided lower maximum exposure rate (Davis, 2002, 2004). The a-Stratified procedure and the Sympson–Hetter procedure did not perform well under polytomous models, yielding a high percentage of items not administered (Davis, 2004; Davis & Dodd, 2003; Johnson, 2007; Pastor, Dodd, & Chang, 2002). The progressive–restricted procedure controlled the maximum exposure rate and used the entire pool and produced small item overlap rates (Boyd, 2003; Grady & Dodd, 2009; McClarty et al., 2006). However, most research that investigated item exposure control procedures have used item pools, which contained a majority of items with medium difficulty as well as a normal ability distribution. Research of item exposure control procedures should be extended to various conditions, such as mismatched item pool characteristics and examinees' ability distributions. In addition, research has to be conducted in matched conditions that do not focus on examinees with medium ability level, such as gifted children who have high abilities.

Research (Gorin, Dodd, Fitzpatrick, & Shieh, 2005) has investigated the mismatched condition of item pool and ability distribution under the partial credit model.

Overall, there was not much difference in terms of measurement precision between a matched condition, using normally distributed item pool and normal ability distribution, and a mismatched condition, using normally distributed item pool with negatively skewed distribution. However, the authors did not include exposure control procedures in this study, so it is questionable whether exposure control procedures perform well under some conditions, such as a mismatched condition of an item pool with an ability distribution without reducing measurement precision. Therefore, the purpose of this study is to investigate the effect of exposure control procedures under various combinations of item pool characteristics and ability distributions in CAT systems based on the partial credit model.

## *Polytomous Item Response Theory (IRT) Models and Partial Credit Model*

Polytomous IRT models allow for the scoring of items when multiple response categories are allowed, such as assessing attitudes using Likert-type scales, essay scoring, or partial credit scoring. Polytomous models, which can easily be extended from the dichotomous models, use multiple parameters to represent the probability of responding in a given category rather than a single parameter. Unlike dichotomous models that have a single-item characteristic curve, polytomous models have multiple category characteristic curves, which represent the relationship between ability level and the probability of responding in each category.

As one of polytomous models, the partial credit model (Masters, 1982) is an extension of the one-parameter logistic model and is appropriate for items that are scored into more than two categories. For each item $i$, a person's item score is categorized in one of $m_i + 1$ category scores, ranging from 0 to $m_i$. A category score represents the number of steps that are successfully completed. In other words, the category scores for item $i$ are the successive integers, denoted $x$, that take on the values 0, 1, . . ., $m_i$. Therefore, the probability of an examinee receiving a given category score on item $i$ can be expressed as follows:

$$P_{ix}(\theta) = \frac{\text{EXP} \sum_{k=0}^{x} (\theta - b_{ik})}{\sum_{h=0}^{m} \text{EXP} \sum_{k=0}^{h} (\theta - b_{ik})}, \quad x = 0, 1, \ldots, m_i, \tag{1}$$

where $b_{ik}$ represents the step difficulty of transitioning from one category of $m_i$ to the next category. The step difficulty parameters can be interpreted as the category thresholds, which are the intersections of two consecutive category response curves. So there would be four-step difficulty parameters in the case of an item with five categories. For the partial credit model, the step difficulties within an item do not necessarily need to be in order, but the steps within an item have to be completed in order (Dodd & Koch, 1987). For example, in the following math item, which has four categories with three step difficulties, $[(5.5/.5) - 2]^3$, an examinee cannot receive credit for the third step (i.e., $[\ ]^3$) before completing both the first and second steps.

However, the first step can be harder than the second step (11-2), meaning that intersection parameters (step difficulties) are not ordered (Dodd & Koch, 1987). In the partial credit model, all items are assumed to have equal discrimination power (Masters, 1982).

## Item and Test Information for Polytomous IRT Models

Samejima (1969) developed a general formula for calculating information for polytomous models. Item information in polytomous IRT models is calculated for the response categories, as well as for an item. The information for a given item is denoted as follows:

$$I_i(\theta) = \sum_{x=0}^{m_i} \frac{[P'_{ix}(\theta)]^2}{P_{ix}(\theta)}, \tag{2}$$

where $P_{ix}(\theta)$ is the probability of obtaining a category score of $x$ on item $i$, and $P'_{ix}(\theta)$ is the first derivative of $P_{ix}(\theta)$. As with information functions in dichotomous IRT, the test information function in polytomous IRT models is simply the sum of the item information functions, and that test information is inversely related to the precision with which ability is estimated. Samejima (1969) noted that polytomous scoring of an item provides more information than dichotomous scoring of the item, and it is true of all the polytomous models (Dodd, De Ayala, & Koch, 1995).

## Computerized Adaptive Testing

Reckase (1989) listed four major components of a CAT: item pool, item selection method, trait estimation method, and stopping rule. Exposure control and content balancing are also included in CAT components for the purpose of practical consideration in high-stakes testing conditions (Boyd, 2003). Because performance of exposure control procedures using the partial credit model is the main focus of this current study, exposure control procedures are discussed.

## Overview of Exposure Control Procedures

Georgiadou, Triantafillou, and Economides (2007) summarized exposure control procedures conducted from 1985 to 2005 and categorized exposure control procedures into five categories: randomization strategies, condition selection strategies, stratification strategies, multiple stages adaptive test designs, and combined strategies. When using randomization strategies, the next item is randomly selected from a candidate group that contains the most optimal items, such as the randomesque procedure (Kingbury & Zara, 1989) and the progressive procedure (Revuelta & Ponsoda, 1998). Conditional selection strategies constrain the probability of administering an item to target exposure rate, such as the Sympson–Hetter procedure (Sympson & Hetter,

1985) and the restricted procedure (Revuelta & Ponsoda, 1998). When using stratified strategies, the selection of the items is constrained on the basis of stratified discrimination parameters, such as the α-stratified procedure (Chang & Ying, 1999). For a multiple stages test (multiple stages test designs), the exposure control goal is achieved by preconstructing adaptive test forms, such as the computerized adaptive sequential test (Luecht & Nungester, 1998). The combined selection strategies were proposed to resolve both item overexposure and item underexposure such as the progressive–restricted procedure, which combines the progressive procedure with the restricted procedure (Revuelta & Ponsoda, 1998). Among several exposure control procedures, only randomesque and progressive–restricted procedures are described in detail for the purpose of the current study.

*Randomesque procedure.* The randomesque procedure (RA) proposed by Kingsbury and Zara (1989) randomly selects the next item for administration from a group of the most informative items rather than selecting the single most informative item. It is similar to the 5-4-3-2-1 procedure, but it repeatedly selects the same number from the most informative item group (e.g., 2, 3, 4, . . . , 10), then one is randomly selected from the most informative item group. In addition, this procedure does not switch to maximum information selection after the initial few items but, employing a random component, continues throughout testing. This procedure can control overexposed items throughout the test as well as at the beginning of the test (Morrison, Subhiyah, & Nungester, 1995; Stocking, 1992).

*Progressive–restricted procedure.* The progressive–restricted procedure (PR) proposed by Revuelta and Ponsoda (1998) combined the restricted maximum information procedure as a conditional component and the progressive procedure as a random component. Before administration of a CAT, the restricted maximum information procedure determines the available items, so items that have already achieved a predetermined exposure rate would not be allowed to be selected for administration; then by using the progressive procedure, the remaining items are weighted and selected based on a formula including the random and information components:

$$W_i = (1 - s)R_i + sI_i, \tag{3}$$

where $s$ represents the serial position in the test (how many items have been administered divided by the total test length), $I$ represents the item information provided by the item at the current estimated ability level, and $R$ represents a random uniform number. Item information is weighted by the serial position in the test, so it is not important in the beginning of the test. However, the contribution of serial position increases as the test progresses. On the other hand, a random component is weighted by one minus the serial position in the test, so it is important in the beginning of the test. However, the contribution of a random component decreases as the test progresses. The progressive–restricted procedure performed well in terms of measurement precision, item exposure control, and pool utilization (Boyd, 2003; Grady & Dodd, 2009; McClarty et al., 2006).

# Method

## Overview of Design

Three variables were manipulated in this study: item pool characteristics, ability distributions, and item exposure control procedures. Three item pools for this study were designed to provide information at different levels of the ability (easy, medium, and hard item pools). Two ability distributions were used to generate simulees (normally distributed and negatively skewed). A positively skewed distribution could be the mirror image of the negatively skewed distribution, so the implication for the mismatched conditions of item pool characteristics and ability distribution would be the same. Therefore, only a negatively skewed distribution was included in the current study. Three exposure control procedures were examined: a maximum information item selection procedure (MI), the RA (with six items), and the PR (with .30 exposure rate). In addition, 10 replications of the CAT procedures were conducted. Therefore, $3 \times 2 \times 3$ with 10 replications yielded 180 conditions.

## Item Pool Characteristics

Known item parameters for this study were obtained by duplicating item parameters used by Koch and Dodd (1989). The three CAT item pools (easy, medium, and hard) each containing 120 items were constructed for the purpose of this study, and each item in the three pools has three step difficulties. The item pools were (a) peaked at the low end of the latent trait distribution (easy item pool), (b) peaked in the middle of the latent trait distribution (medium item pool), and (c) peaked at the high end of the latent trait distribution (hard item pool).

## Data Generation

For the calibration sample, 10,000 simulees' known trait levels were drawn from a normal distribution with a mean of 0 and standard deviation of 1. Response data for 20 CAT data sets (10 for the normal CAT conditions and 10 for the skewed CAT conditions) were generated using the IRTGEN SAS macro (Whittaker, Fitzpatrick, Williams, & Dodd, 2003) for the partial credit model. Under the *normal* distribution condition, 1,000 simulees were randomly drawn from a normal distribution with a mean of 0 and standard deviation of 1. Under the skewed distribution condition, 1,000 simulees were randomly drawn from beta distribution with $\alpha = 5.0$ and $\beta = 1.8$, yielding a distribution with a mean of 0.74, a standard deviation of 0.16, and a skewness of –0.73. The known theta values for both the normally distributed data and negatively skewed data were standardized with a mean of 0 and a standard deviation of 1 prior to generating item responses. This procedure of sampling from a negatively skewed distribution was used in Gorin et al. (2005).

## Item Parameter Estimation

Using the calibration sample of 10,000 simulees, the three step difficulty parameters for the 360 four-category items were estimated using the PARSCALE (Muraki & Bock, 1993). The program used a marginal maximum likelihood estimation procedure with 30 normally distributed quadrature points to estimate the item parameters.

## CAT Simulation

A SAS program for CAT simulations (Boyd, 2003; Chen, Hou, & Dodd, 1998) was modified to satisfy each CAT condition for the present study. This simulation began with the initial θ of 0, and the maximum item information was used for initial item selection with one of the exposure control procedures. The examinees' abilities were estimated using expected a posteriori (EAP) estimation after each item is administered. For the normal data, EAP with a normal prior were used, and EAP with a skewed prior ($\alpha = 5.0$ and $\beta = 1.8$) was used for the skewed data. The test was terminated after 20 items were given. Each CAT condition (three item pool characteristics × two ability distributions × three exposure control procedures) were repeated for the 10 data sets.

## Data Analyses

Descriptive statistics, correlations between known and estimated thetas, bias, RMSE, and average absolute difference (AAD) were calculated to evaluate the various CAT conditions.

$$\text{BIAS} = \frac{\sum_{i=1}^{n}(\hat{\theta}_i - \theta_i)}{n}, \tag{4}$$

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(\hat{\theta}_i - \theta_i)^2}{n}}, \tag{5}$$

$$\text{AAD} = \frac{\sum_{i=1}^{n}\left|\hat{\theta}_i - \theta_i\right|}{n}, \tag{6}$$

In Formulas (4), (5), and (6), $\hat{\theta}_i$, $\theta_i$ represents the final ability estimate and known ability of examinee *i*, respectively, and *n* is the total number of simulated examinees in each condition. In addition, item exposure rates were computed by dividing the number of times an item is administered by the total number of simulees. The frequency distribution, mean, standard deviation, and maximum of the exposure rates were computed and summarized across conditions. The proportion of items within the pool never administered was also calculated as an indicator of pool utilization, and item overlaps, which are the average number of items shared by two examinees, were computed.

## Results

### Measurement Precision

Table 1 shows the grand means of the theta estimates and standard errors. The average correlation, RMSE, AAD, and Bias across 10 replications are also displayed in Table 1. Regarding item pool characteristics, measurement precisions did not differ substantially even though the medium item pool slightly performed better than both the easy and hard item pools. For example, bias was functionally zero and RMSEs were from 0.22 to 0.26 across item pool characteristics. AADs were around 0.20 across item pool characteristics. In addition, correlations between known and estimated simulee ability levels were similar, yielding a correlation more than 0.96 across item pool characteristics. For exposure control procedures, measurement precisions for both the RA and the PR procedures were similar to the MI(no exposure control condition), meaning that these procedures performed equally well in terms of measurement precision relative to the MI procedure.

The conditional plots to examine the accuracy of the estimates demonstrated that the ranges of the best accuracy of estimation differed across item pools as a function of matched items to examinees' ability. The medium item pool provided the best measurement precision across the entire ranges of ability, except in both extreme theta values, such as $\theta < -2.75$ or $\theta < 2.75$. The easy item pool yielded the smallest standard error in the ability ranges of –2.75 to 0.75. In addition, the hard item pool estimated ability well in the range of –0.75 to 1.75. The matched condition of the hard item pool and the skewed data set yielded the worst measurement for simulees at the low end of ability levels. It should be noted that there were few low ability simulees. In this sense, the conditional bias was also expected, in that the medium item pool with the normal data had a small discrepancy between known theta and estimated theta, except in extreme theta values. The easy pool in the low range and hard item pool in the high range of ability had a small discrepancy between known theta and estimated theta. However, within each item pool, the pattern was consistent across ability distributions and across exposure control procedures (see Figures 1 and 2).

### Maximum Exposure Rate

Exposure rates and pool utilization for each of the conditions are displayed in Table 2. The maximum exposure rate with the MI procedure for all three item pools was 1. For the RA procedure, the maximum exposure rate that ranged from 0.58 to 0.78 was slightly higher for both the easy and the hard item pools relative to the medium item pool with an exposure rate of 0.57 regardless of ability distributions. Because of the nature of the matched condition of the item pools to the ability distributions, the

**Table 1.** Overall Descriptive Statistics for the Exposure Control Procedures With Different Combinations of the Item Pool Characteristics and the Ability Distributions

| Item pool characteristics | Trait distribution | Exposure control | Grand mean $\hat{\theta}$ | Grand mean *SE* | Mean correlation | Mean Bias | Mean RMSE | Mean AAD |
|---|---|---|---|---|---|---|---|---|
| Easy | Normal | MI | −0.01 | 0.25 | 0.97 | 0.00 | 0.24 | 0.19 |
| | | RA-6 | −0.02 | 0.25 | 0.97 | −0.00 | 0.25 | 0.19 |
| | | PR-30 | 0.00 | 0.27 | 0.96 | −0.01 | 0.26 | 0.20 |
| | Negatively skewed | MI | 0.00 | 0.25 | 0.97 | 0.00 | 0.24 | 0.19 |
| | | RA-6 | −0.00 | 0.25 | 0.97 | 0.00 | 0.24 | 0.19 |
| | | PR-30 | 0.00 | 0.27 | 0.96 | −0.00 | 0.26 | 0.20 |
| Medium | Normal | MI | −0.01 | 0.22 | 0.97 | 0.00 | 0.23 | 0.18 |
| | | RA-6 | −0.01 | 0.23 | 0.97 | 0.00 | 0.23 | 0.18 |
| | | PR-30 | −0.01 | 0.24 | 0.97 | 0.00 | 0.23 | 0.18 |
| | Negatively skewed | MI | 0.00 | 0.23 | 0.98 | 0.00 | 0.22 | 0.18 |
| | | RA-6 | −0.00 | 0.23 | 0.98 | 0.00 | 0.22 | 0.18 |
| | | PR-30 | −0.00 | 0.24 | 0.97 | 0.00 | 0.23 | 0.18 |
| Hard | Normal | MI | −0.00 | 0.25 | 0.97 | −0.00 | 0.24 | 0.19 |
| | | RA-6 | −0.00 | 0.25 | 0.97 | 0.00 | 0.24 | 0.19 |
| | | PR-30 | −0.00 | 0.26 | 0.96 | −0.00 | 0.26 | 0.20 |
| | Negatively skewed | MI | −0.00 | 0.25 | 0.97 | 0.00 | 0.25 | 0.19 |
| | | RA-6 | −0.01 | 0.26 | 0.97 | 0.00 | 0.25 | 0.19 |
| | | PR-30 | −0.00 | 0.26 | 0.97 | 0.00 | 0.26 | 0.20 |

*Note*. MI = maximum information; RA-6 = randomesque with 6 items; PR-30 = progressive–restricted with .30 exposure rate; *SE* = standard error; RMSE = root mean squared error; AAD = average absolute difference. All statistics were averaged across 10 replications; each replication contained 1,000 observations.

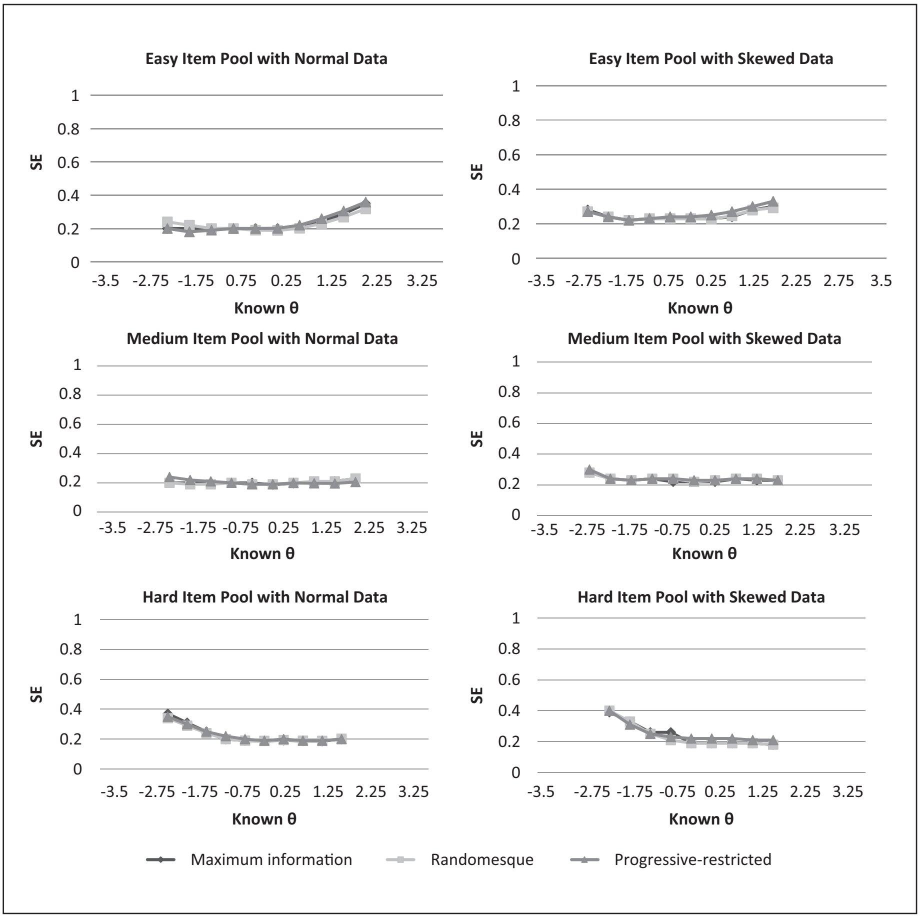


**Figure 1.** Grand mean standard error for the exposure control procedures with different combinations of the item pool characteristics and the ability distributions

maximum exposure rates of the RA procedure, in both the medium item pool with the normal data and the hard item pool with the skewed data, were lower than those of the RA procedure in the mismatched conditions, such as 0.57 for the medium item pool with the normal data and 0.58 for the hard item pool with the skewed data. However, the maximum exposure rate of the RA procedure in a big mismatched condition, such as easy item pool with the skewed data, was highest with exposure rate 0.78. On the other hand, the PR procedure constrained the exposure rate with .3, so the maximum exposure rate across all conditions was .3.

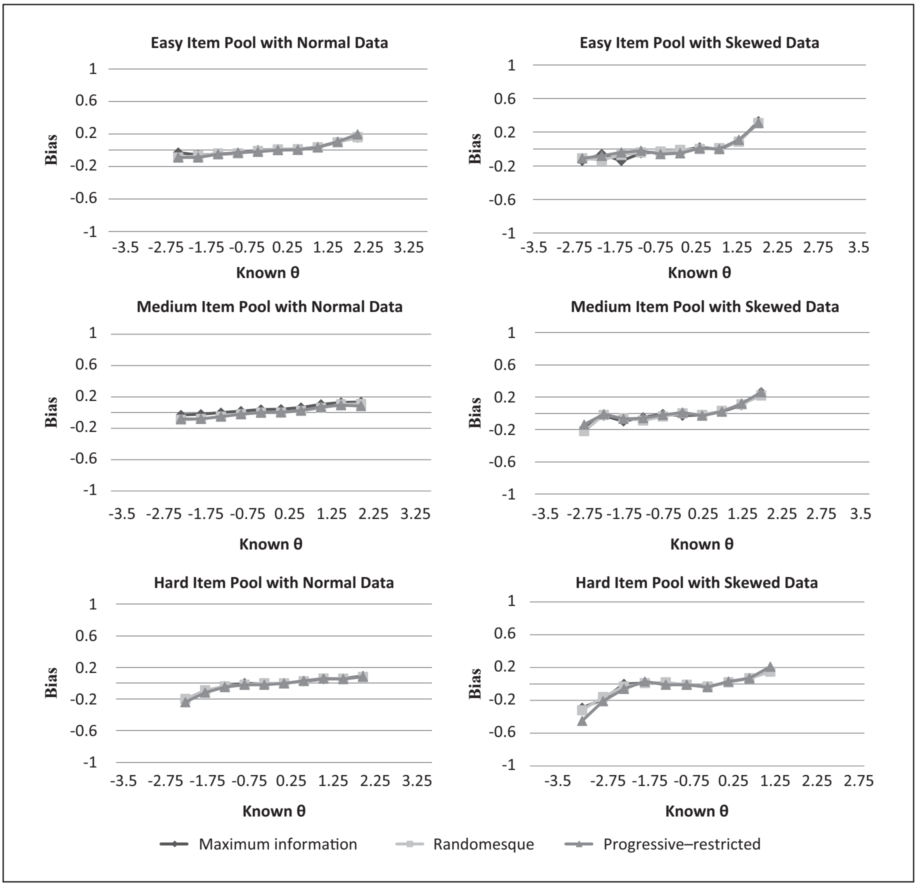


**Figure 2.** Grand mean standard bias for the exposure control procedures with different combinations of the item pool characteristics and the ability distributions

## Pool Utilization

Table 2 contains the mean percentages of items not administered across the 10 replications of each CAT condition. Generally, the medium item pool produced better pool utilization for each exposure control procedure than both the easy and hard item pools. Regarding item exposure control procedures, the percentage of items not administered for the MI procedure was higher than those for both the RA and PR procedures. The MI procedure left between 13% and 33% of the item pool unused, depending on the combinations of item pool characteristics and ability distributions. The RA procedure left between 12% and 30% of the item pool unused. Even though the RA procedure performed better in terms of pool utilization than the MI procedure for the easy and

**Table 2.** Mean Maximum Exposure Rates and Pool Utilization Averaged Across Replications

| Item pool characteristic | | Normal data | | | Negatively skewed data | | |
|---|---|---|---|---|---|---|---|
| | | MI | RA-6 | PR-30 | MI | RA-6 | PR-30 |
| Easy | Maximum exposure rate | 1.00 | 0.72 | 0.30 | 1.00 | 0.78 | 0.30 |
| | Percentage of pool not administered | 17.67 | 13.83 | 0.00 | 28.33 | 17.00 | 0.00 |
| Medium | Maximum exposure rate | 1.00 | 0.57 | 0.30 | 1.00 | 0.57 | 0.30 |
| | Percentage of pool not administered | 13.00 | 11.67 | 0.00 | 22.17 | 16.17 | 0.00 |
| Hard | Maximum exposure rate | 1.00 | 0.63 | 0.30 | 1.00 | 0.58 | 0.30 |
| | Percentage of pool not administered | 19.17 | 18.33 | 0.00 | 32.83 | 30.15 | 0.00 |

*Note.* MI = maximum information; RA-6 = randomesque with 6 items; PR-30 = progressive–restricted with .30 exposure rate.

medium item pools, regardless of the item pool characteristics and the ability distributions, the RA procedure in the hard item pool was not substantially beneficial for pool utilization when compared with the MI procedure. On the other hand, the PR procedure was able to use the entire item pool, yielding 0% of the pool never being administered, meaning the PR procedure performed the best in terms of pool utilization.

## *Item Overlap*

To investigate overall mean item overlap, different abilities average overlap, and similar abilities overlap across 10 replications, each simulee's audit trail was compared with the audit trails of every other simulee. When examinees' known abilities differed by one logit or less, they are defined to have "similar" ability levels. However, the examinees are defined to have "different" ability levels when their known abilities differ by more than one logit. Table 3 shows the overall mean and minimum and maximum item overlap percentages averaged across 10 replications of each CAT condition. The overall mean overlap rates, in general, for both the easy and hard item pools, were slightly higher, regardless of ability distributions, than for those of the medium item pool. Specifically, the MI and RA procedures in the medium item pool with the normal data as a matched condition yielded 35% and 31% of overall mean overlap rates, respectively. On the other hand, these procedures in the easy item pool with skewed data as a big mismatched condition yielded 51% and 45% of overall mean overlap rates, respectively. However, the PR procedure was lowest overall among the mean overlap rates, ranging from 23% to 26% across all conditions.

**Table 3.** Mean Number of Overall Overlap Items Averaged Across 10 Replications

| | | Normal data | | | Negatively skewed data | | |
|---|---|---|---|---|---|---|---|
| Item pool characteristics | Exposure control | Grand mean (%) | Minimum (%) | Maximum (%) | Grand mean (%) | Minimum (%) | Maximum (%) |
| Easy pool | MI | 9.20 (46.00) | 1 (5.00) | 20 (100) | 10.13 (50.65) | 1 (5.00) | 20 (100) |
| | RA-6 | 8.20 (41.00) | 0 (0) | 20 (100) | 9.95 (45.25) | 0 (0) | 20 (100) |
| | PR-30 | 4.96 (24.80) | 0 (0) | 19.67 (98.35) | 4.88 (24.40) | 0 (0) | 19 (95.00) |
| Medium pool | MI | 6.93 (34.65) | 1 (5.00) | 20 (100) | 7.10 (35.50) | 1 (5.00) | 20 (100) |
| | RA-6 | 6.24 (31.20) | 0 (0) | 20 (100) | 6.32 (31.60) | 0 (0) | 20 (100) |
| | PR-30 | 4.54 (22.70) | 0 (0) | 18 (90.00) | 4.62 (23.10) | 1 (5.00) | 18.30 (91.50) |
| Hard pool | MI | 8.13 (40.65) | 1 (5.00) | 20 (100) | 7.64 (38.30) | 1 (5.00) | 20 (100) |
| | RA-6 | 7.41 (37.05) | 0 (0) | 20 (100) | 7.02 (35.10) | 0 (0) | 20 (100) |
| | PR-30 | 5.22 (26.10) | 0 (0) | 18.2 (91.00) | 5.04 (25.20) | 0 (0) | 18.5 (92.00) |

*Note*. MI = maximum information; RA-6 = randomesque with 6 items; PR-30 = progressive–restricted with .30 exposure rate.

Table 4 contains mean and minimum and maximum overlap rates for simulees with similar and different ability levels averaged across 10 replications. For simulees of similar ability levels, both the RA and PR procedures reduced the mean overlap rates for simulees of similar abilities as well as different abilities, compared with the MI procedure. The MI procedure produced highest mean overlap rates for simulees of similar ability levels, ranging from 54% to 73%. On the other hand, the RA procedure produced lower mean overlap rates, ranging from 49% to 65%, than the MI procedure. However, the PR procedure produced considerably lower mean overlap rates, ranging from 32% to 34%, of item overlap rates than both the MI and RA procedures. The overlap rates for simulees of different abilities had similar patterns of those for simulees of similar abilities.

In summary, using the MI procedure yielded higher item overlap rates than using either the RA procedure or the PR procedure. The RA procedure controlled item exposure by randomly selecting an item from a group of six most informative items for the current ability estimate, but the overall exposure rate was not capped. The PR procedure, on the other hand, capped the maximum exposure rate and, therefore, provided the best control of item exposure rate.

## Conclusion and Discussion

In general, polytomous CAT based on the partial credit model was relatively robust under the different combinations between item pool characteristics and ability distributions in terms of measurement precision and exposure control properties. This is because polytomous items provide information across wider ranges of ability levels. Therefore, if an item pool is sufficiently large, CAT designs are relatively robust to changes in the underlying ability distribution (Keng, 2008). This result was consistent with several studies (Chen, Hou, Fitzpatrick, & Dodd, 1997; Gorin et al., 2005; Keng, 2008). The results of the current study, however, revealed that the interactions with item pool characteristics and ability distribution differentially affect the performance of exposure control procedures. For example, the MI procedure in the easy item pool with the negatively skewed data as a big mismatched condition had the worst performance in terms of item overlap rates. In addition, the RA procedure in mismatched conditions such as the easy item pool with the skewed data had higher maximum exposure rate and item overlap rates than in the matched condition such as the medium item pool with the normal data. However, the PR30 procedure performed equally well in terms of exposure properties even in the big mismatched condition. Thus, the current study showed the PR procedure outperformed the RA procedure in all conditions.

In summary, the medium item pool, in general, produced better results in terms of precision of measurement, maximum exposure rates, and pool utilization regardless of ability distributions. Regarding the comparison of exposure control procedures, the PR procedure performed much better in terms of item exposure rate, mean item overlap rates, and pool utilization across item pool characteristics than the MI procedure or

**Table 4.** Mean Number of Overlap Items for Similar and Different Abilities Averaged Across 10 Replications

| | | Normal data | | | | | | Negatively skewed data | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Similar abilities[a] | | | Different abilities[b] | | | Similar abilities[a] | | | Different abilities[b] | | |
| Item pool | Exposure control | Grand mean (%) | Minimum (%) | Maximum (%) | Grand mean (%) | Minimum (%) | Maximum (%) | Grand mean (%) | Minimum (%) | Maximum (%) | Grand mean (%) | Minimum (%) | Maximum (%) |
| Easy | MI | 13.16 (65.50) | 1 (5.00) | 20 (100) | 4.89 (24.30) | 0 (0) | 19 (95.00) | 14.66 (73.30) | 1 (5.00) | 20 (100) | 5.02 (25.10) | 1 (5.00) | 17.04 (85.20%) |
| | RA-6 | 11.17 (58.55) | 0 (0) | 20 (100) | 4.56 (22.80) | 0 (0) | 17 (85.00) | 13.01 (65.05) | 0 (0) | 20 (100) | 4.62 (23.10) | 0 (0) | 16.2 (81.00) |
| | PR-30 | 6.51 (32.55) | 0 (0) | 20 (100) | 3.27 (16.35) | 0 (0) | 16.71 (83.53) | 6.69 (33.45) | 0 (0) | 19 (95.00) | 2.83 (14.15) | 0 (0) | 15.6 (78.00) |
| Medium | MI | 10.86 (54.30) | 1 (5.00) | 20 (100) | 3.27 (13.35) | 1 (5.00) | 16.25 (81.25) | 11.09 (55.45) | 1 (5.00) | 20 (100) | 2.51 (12.55) | 1 (5.00) | 13.60 (68.00) |
| | RA-6 | 9.80 (49.00) | 0 (0) | 20 (100) | 2.36 (11.80) | 0 (0) | 13 (65.00) | 9.94 (49.70) | 0 (0) | 20 (100) | 2.15 (10.75) | 0 (0) | 12.40 (62.00) |
| | PR-30 | 7.07 (35.35) | 0 (0) | 18 (90.00) | 1.87 (9.35) | 0 (0) | 12.25 (61.25) | 7.13 (35.65) | 0 (0) | 18.40 (92.00) | 1.80 (9.00) | 0 (0) | 11/80 (59.00) |
| Hard | MI | 11.86 (59.30) | 1 (5.00) | 20 (100) | 3.99 (19.95) | 1 (5.00) | 17.75 (88.75) | 11.99 (55.95) | 1 (5.00) | 20 (100) | 3.70 (18.50) | 1 (5.00) | 19.20 (96.00) |
| | RA-6 | 10.76 (53.80) | 0 (0) | 20 (100) | 3.84 (19.20) | 0 (0) | 16.75 (83.75) | 10.08 (50.40) | 0 (0) | 20 (100) | 3.57 (17.85) | 0 (0) | 17.20 (86.00) |
| | PR-30 | 6.46 (32.30) | 0 (0) | 18.25 (91.25) | 3.43 (17.15) | 0 (0) | 15.50 (77.50) | 6.88 (34.40) | 0 (0) | 18.33 (91.65) | 3.41 (17.05) | 0 (0) | 16.33 (81.65) |

*Note.* MI = maximum information; RA-6 = randomesque with 6 items; PR-30 = progressive-restricted with .30 exposure rate.
a. Abilities within .1 logit.
b. Greater than .1 logit difference between abilities.

the RA procedure. Previous research found that the RA procedure, in general, performed better in terms of exposure control properties than the MI procedure (Boyd, 2003; Davis, 2002, 2004). However, the present study found that the RA procedure yielded item overlap rates and pool utilization that were only slightly better than the MI procedure. Thus, the future study might increase the informative item group from 6 to 10 items to achieve a lower item overlap rate.

To date, there are no studies investigating the combination of item pool characteristics with exposure control procedures in mismatched conditions under polytomous models. This study shows the performance of exposure control procedures for polytomous models depending on the alignment item pool information and the distribution of ability of those tested.

## Declaration of Conflicting Interests

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

## Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

## References

Boyd, A. M. (2003). *Strategies for controlling testlet exposure rates in computerized adaptive testing systems* (Unpublished doctoral dissertation). The University of Texas at Austin.

Boyd, A. M., Dodd, B. G., & Choi, S. W. (2010). Polytomous models in computerized adaptive testing. In M. L. Nering & R. Ostini (Eds.), *Handbook of polytomous item response theory models* (pp. 229-255). New York: Routledge.

Chen, S., Hou, L., & Dodd, B. G. (1998). A comparison of maximum likelihood estimation and expected a posteriori estimation in CAT using the partial credit model. *Educational and Psychological Measurement, 58*, 569-595.

Davis, L. L. (2002). *Strategies for controlling item exposure in computerized adaptive testing with polytomously scored items* (Unpublished doctoral dissertation). The University of Texas at Austin.

Davis, L. L. (2004). Strategies for controlling item exposure in computerized adaptive testing with the generalized partial credit model. *Applied Psychological Measurement*, *28*, 165-185.

Davis, L. L., & Dodd, B. G. (2003). Item exposure constraints for testlets in the verbal reasoning section of the MCAT. *Applied Psychological Measurement, 27*, 335-356.

Dodd, B. G., De Ayala, R. J., & Koch, W. R. (1995). Computerized adaptive testing with polytomous items. *Applied Psychological Measurement*. *19*, 5-22.

Dodd, B. G., & Koch, W. R. (1987). Effects of variations in item step values on item and test information in the Partial Credit Model. *Applied Psychological Measurement, 11*, 371-384.

Gorin, J. S., Dodd, B. G., Fitzpatrick, S. J., & Shieh, Y. Y. (2005). Computerized adaptive testing with the partial credit model: Estimation procedures, population distribution, and the item pool characteristics. *Applied Psychological Measurement*, *29*, 433-456.

Keng, L. (2008). *A comparison of the performance of testlet-based computer adaptive tests and multistage tests* (Unpublished doctoral dissertation). The University of Texas at Austin.

Kingsbury, G. G., & Zara, A. R. (1989). Procedures for selecting items for computerized adaptive testing using partial credit scoring. *Applied Measurement in Education, 2*, 359-375.

Koch, W. R., & Dodd, B. G. (1989). An investigation of procedures for computerized adaptive testing using partial credit scoring. *Applied Measurement in Education, 2*, 335-357.

Lord, F. M. (1970). Some test theory for tailored testing. In W. Holtzman (Ed.), *Computer-assisted instruction, testing, and guidance* (pp. 139-183). New York, NY: Harper & Row.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149-174.

McClarty, K. L., Sperling, R., & Dodd, B. G. (2006, April). *A variant of the progressive-restricted item exposure control procedure in computerized adaptive testing system based on the 3PL and partial credit models.* Paper presented at the American Educational Research Association, San Francisco, CA.

Morrison, C. A., Subhiyah, R. G., & Nungester, R. J. (1995, April). *Item exposure rates for unconstrained and content-balanced computerized adaptive tests*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Muraki, E., & Bock, R. D. (1993). PARSCALE computer program [Computer program]. Chicago, IL: Scientific Software International.

Pastor, D. A., Dodd, B. G., & Chang, H. H. (2002). A comparison of item selection techniques and exposure control mechanisms in CATs using the generalized partial credit model. *Applied Psychological Measurement*, *26*, 147-163.

Revuelta, J., & Ponsoda, V. (1998). A comparison of item control methods in computerized adaptive testing. *Journal of Educational Measurement, 35*, 311-317.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika, Monograph Supplement, 34* (4, Pt. 2), 100.

Stocking, M. L. (1992). *Controlling item exposure rates in a realistic adaptive testing paradigm* (Research Report No. 93-2). Princeton, NJ: Educational Testing Service.

Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement*, *6*, 473-492.

Whittaker, T. A., Fitzpatrick, S. J., Williams, N. J., & Dodd, B. G. (2003). IRTGEN: A SAS macro program to generate known trait scores and item responses for commonly used item response theory models. *Applied Psychological Measurement, 30*, 299-300.