

Abstract

Exposure control research with polytomous item pools has determined that randomization procedures can be very effective for controlling test security in computerized adaptive testing (CAT). The current study investigated the performance of four procedures for controlling item exposure in a CAT under the partial credit model. In addition to a no exposure control baseline condition, the Kingsbury-Zara, modified-within-.10-logits, Simpson-Hetter, and conditional Simpson-Hetter procedures were implemented to control exposure rates. The Kingsbury-Zara and the modified-within-.10-logits procedures were implemented with 3 and 6 item candidate conditions. The results show that the Kingsbury-Zara and modified-within-.10-logits procedures with 6 item candidates performed as well as the conditional Simpson-Hetter in terms of exposure rates, overlap rates, and pool utilization. These two procedures are strongly recommended for use with partial credit CATs due to their simplicity and strength of their results.

Strategies for Controlling Item Exposure in Computerized Adaptive Testing with the Partial Credit Model

Introduction

As recently as the early 1990s, computerized adaptive testing (CAT) was heralded as a tool for revolutionizing the world of measurement. Among its many laudable features, CAT would allow for increased efficiency in testing, improved test security through the presentation of individually tailored test forms, and the immediate scoring and feedback to examinees (Wainer, 1990). In addition, the use of the computer as the medium for test delivery offered many exciting possibilities for reduction in the reliance on multiple choice assessment through the introduction of new interactive, simulation, and constructed response item types. In the enthusiasm to embrace this emerging paradigm, however, certain operational details of CAT implementation were often overlooked (Wainer & Eignor, 2000).

Reckase (1989) discussed the four major components of a CAT: 1) the item pool, 2) the item selection method, 3) the trait estimation method, and 4) the stopping rule. Notably absent from this list is exposure control—an element of CAT that has now become an integral part of any operational CAT program. Exposure control focuses on preserving the integrity of the item pool by protecting items from over-exposure to examinees, and consequently, the compromise of items in the pool. The absence of exposure control from this list does not represent an omission (methods of controlling item exposure had been available since the mid-1980s), but rather the fact that the impact of a compromised item pool in terms of actual cost and public perception was not widely recognized. That began to change when, in 1994, Kaplan Educational Centers, one of the country's largest test preparation companies, was able to successively steal a significant portion of the GRE CAT item pool with fairly minimal effort (Wainer & Eignor,

2000). The high profile nature of this and other cases has made item pool security a chief concern for CAT programs today.

CAT creates special circumstances that can lead to the over-exposure of certain items within the pool to examinees. The CAT item selection algorithm itself, designed to maximize measurement precision, can constitute a threat to item security. Under maximum information item selection, certain items will be administered to almost all examinees and a small proportion of items available in the pool can account for a large proportion of items actually administered. Frequently exposed items will cease to be a valid measure of the trait because they may have been compromised. Any time an examinee has prior knowledge of an item, his/her response will not be an accurate measure of his/her true trait level. As such, research has produced a class of algorithms known as exposure control procedures which are designed to intercede in the item selection process and ensure that items are not over-exposed.

Exposure Control Procedures

Exposure control procedures run the gamut from the relatively simple to the more complex. Early attempts to control item exposure focused on the incorporation of a random component to item selection and have been called randomization procedures (Way, 1998). The Kingsbury-Zara (Kingsbury & Zara, 1989) procedure (also known as the “randomesque” procedure) uses maximum information to select a prespecified number of items from the item pool (the most informative item, the second most informative item, etc.). The next item to be administered is then randomly selected from this group of item candidates. Lunz and Stahl (1998) developed a procedure which similarly assembles a group of item candidates; however, rather than selecting a prespecified number, all items within 0.10 logits of the target item difficulty are identified and the next item to be administered is randomly chosen from among them.

While procedures that rely on random elements to control exposure rates are appealing for their simple logic and uncomplicated implementation, research with dichotomous item pools has produced mixed results in terms of their ability to affect sufficiently tight security for an item pool. Morrison, Subhiyah, & Nungester (1995) and Eignor, Stocking, Way, and Steffan (1993) both evaluated the Kingsbury-Zara procedure and concluded that it was not effective for reducing item exposure. However, Revuelta and Ponsoda (1998) found an increase in pool utilization and a decrease in maximum exposure rate when using the Kingsbury-Zara procedure. Bergstrom and Lunz (1999) demonstrated that the within-.10-logits procedure was able to control maximum exposure rates to less than 30% (item would be seen by less than 30% of test taking population) for most items; however, there were a handful of items with maximum exposure rates exceeding this value.

More complicated procedures, such as the Sympton-Hetter procedure (Sympton & Hetter, 1985), constrain the maximum exposure rate to a predetermined target value by assigning an exposure control parameter to each item. These parameters are derived through a series of simulations in which the frequency with which an item is administered is recorded. Items which are frequently administered will have a low exposure control parameter, whereas items which are infrequently administered will have a high exposure control parameter. This is similar to the idea of handicapping a golfer—good golfers carry low handicaps and less proficient golfers carry higher handicaps. The exposure control parameter is a way of leveling the playing field so that all items in the pool have a reasonable chance of being administered. Once the exposure control parameters are derived, an item can only be administered if the value of its exposure control parameter exceeds that of a random number. The lower the exposure control parameter (a more frequently administered item), the less likely this is to happen.

While the Simpson-Hetter does guarantee that the global maximum exposure rate will be constrained to a given level, it does not take into account exposure rates conditional on trait level. Because item selection in CAT is based on the current trait estimate, examinees with similar trait estimates will tend to see the same items. Using the Simpson-Hetter, the overall global exposure rate of an item may be fairly low, but the exposure rate for examinees at a given trait level may be quite high.

The conditional Simpson-Hetter (Stocking & Lewis, 1998) addresses this concern by computing a matrix of exposure control parameters for items conditional on trait level. The parameters are computed in the same fashion as done in the Simpson-Hetter with the exception that the frequency of item administrations is tallied separately for each of 'm' discrete theta levels, resulting in each item having 'm' different exposure control parameters. An item can only be administered if the value of its exposure control parameter at the current theta estimate exceeds that of a random number.

Polytomous Item Response Theory

Polytomous item response theory (IRT) models can be used to describe the interaction of an examinee with an item which has multiple score points or categories such as an essay, short answer, or other constructed response item. These models are extended from the dichotomous models, but differ in that they use multiple parameters to represent the probability of responding in each category rather than a single item difficulty parameter representing the probability of answering the item correctly. These parameters may be called step values, step difficulties, or category boundaries depending on the particular model chosen. Rather than having a single item characteristic curve (ICC) to represent the relationship between trait level and the probability of a correct response, polytomous models have multiple category characteristic curves (CCCs) which represent the relationship between trait level and the probability of responding in a given

category. An examinee with a given trait level will be most likely to respond in the category whose curve is highest for their trait level.

The Partial Credit Model

While there are many different polytomous IRT models with different derivations and parameterizations, Masters (1982) partial credit model is one of the more commonly used polytomous models. The partial credit model is the generalized form of the one-parameter logistic model (1PL) and simplifies to the 1PL when used with two-category items (0/1 response). The probability that an examinee will score in category x on item i can be computed as:

$$P_{ix}(\theta) = \frac{\exp\left[\sum_{k=0}^x (\theta - b_{ik})\right]}{\sum_{h=0}^{m_i} \exp\left[\sum_{k=0}^h (\theta - b_{ik})\right]}, \quad (1)$$

where θ is the examinee's trait level, b_{ik} is the difficulty parameter (or step value) associated with score category x , and m_i is the number of score categories minus one. The partial credit model is a divide-by-total model, meaning that the probability of responding in a given category is calculated by normalizing the probability space to sum to 1.0. This is accomplished by dividing the numerator which represents a response in a given category by the denominator which represents each possible response. The partial credit model also allows for step values to be unordered such as would be the case when a later step in completing an item is easier than an earlier step. One example of this would be a multi-step math problem requiring examinees to first perform division (a more difficult skill) and later perform addition (an easier skill).

Exposure Control Research with the Partial Credit Model

Concerns over item exposure can be magnified when dealing with the context of polytomous items where item pools tend to be smaller and items require a constructed response

which may be more memorable to examinees. While there are not currently any operational CATs which incorporate polytomous scoring of constructed response items, the advent of automated algorithms for partial credit scoring (Bejar, 1991; Burstein, Kukich, Wolff, Lu, & Chodorow, 1998; Clauser, Margolis, Clyman, & Ross, 1997) will make this feasible in the not too distant future. As such technologies emerge, it becomes necessary to consider the topic of item exposure control with polytomous item pools. Polytomously scored items can be a richer source of information than dichotomous items with information functions which spread to cover a larger span of the theta scale rather than peaking at a more narrow range of theta (Koch & Dodd, 1989).

Exposure control research with the partial credit model has thus far yielded results which seem to stand in contrast to the earlier research findings with dichotomous item pools (Pastor, Chiang, Dodd, & Yockey, 1999; Davis, Pastor, Dodd, Chiang, & Fitzpatrick, 2003; Davis & Dodd, 2003; Boyd, Dodd, and Fitzpatrick, 2003). Contrary to conventional thinking regarding the necessary trade-off between test security and measurement precision, the implementation of exposure control algorithms with the partial credit model has resulted in a fairly minimal impact to trait estimation. In addition, procedures which control exposure through randomization seem to function fairly well in CATs based on the partial credit model, whereas procedures that control exposure through probabilistically determined exposure control parameters have shown themselves to be problematic.

Studies by Pastor, et al. (1999) and Davis, et al. (2003) evaluated the Simpson-Hetter in isolation and in conjunction with a simple content balancing mechanism with item pools ranging in size from 60 to 240. The results revealed two surprising findings—first that use of the Simpson-Hetter did not negatively impact measurement precision and second that it was relatively ineffective in controlling exposure rates for smaller item pools and equally ineffective

in promoting pool utilization in larger item pools. In addition, the authors reported problems in obtaining convergence of the exposure control parameters to the target exposure rate in the smaller item pools.

Davis and Dodd (2003) experimented with the use of a randomization method for controlling exposure in an item pool composed of testlets from the Verbal Reasoning section of the MCAT (Medical College Admission Test) scored with the partial credit model. The original Lunz and Stahl (1998) within-.10-logits procedure was modified to work with the partial credit items. The multiple step values (b-parameters) of each partial credit item precluded the identification of a single target difficulty. Instead the item candidates were selected by maximum information item selection from within a range of the current target trait estimate. Results with this new procedure indicated that while measurement precision was somewhat affected, the procedure was surprisingly successful in reducing exposure and overlap rates and increasing pool utilization. A second study (Boyd, Dodd, & Fitzpatrick, 2003) compared the modified-within-.10-logits method to the Simpson-Hetter procedure (target exposure rates: $r=.19$ and $r=.29$) with this same partial credit item pool. Their results showed that the modified-within-.10-logits procedure outperformed the Simpson-Hetter, yielding lower exposure and overlap rates and higher rates of pool utilization.

The results of these previous studies seem to indicate that exposure control strategies such as the Simpson-Hetter that have been traditionally preferred with dichotomous item pools may not provide the best option for controlling exposure with the partial credit model. However, more research is needed to determine the generalizability of these findings and whether other, as yet unexamined, exposure control procedures may be preferable. While Davis (2004) conducted a more comprehensive comparison of exposure control procedures with the generalized partial credit model, no such comparison has yet been made for the partial credit model.

Method

Overview of Techniques

The current study investigated the performance of four procedures for controlling item exposure in a computerized adaptive test (CAT) under the partial credit model (Masters, 1982). In addition to a no exposure control baseline condition, the Kingsbury-Zara, modified-within-.10-logits, Sympon-Hetter, and conditional Sympon-Hetter procedures were implemented to control exposure rates. For the Kingsbury-Zara and modified-within-.10-logits procedures, it was necessary to stipulate how large a group of item candidates would be formed from which the next item to be administered would be randomly selected. Two item candidate group sizes (3 and 6 items) for each procedure were evaluated in this study.

Data Generation

Live student response data from a large scale verbal reasoning test were used to obtain the known item parameters for an item pool of 157 polytomously scored items. Four sets of simulated data were generated from these known item parameters. Conventional methods for generating polytomous item response data were used (Dodd & Koch, 1987).

- 1) *Calibration sample.* To obtain estimated item parameters, item responses from 7,500 simulees, $N(0,1)$, to the 157 items were generated and submitted to Parscale (Muraki & Bock, 1993) for calibration according to the partial credit model. These resulting estimated item parameters were used in the seven CAT conditions.
- 2) *Sympon-Hetter sample.* In order to set the exposure control parameters for the Sympon-Hetter procedure, item responses from 8,000 simulees, $N(0,1)$, to the 157 items were generated. This sample will be discussed further under the topic of *Exposure Control*.

- 3) Conditional Sympson-Hetter sample. A uniformly distributed sample was generated for use in computing the exposure control parameters for the conditional Sympson-Hetter. One thousand simulee response strings at each of fifteen discrete levels of theta (-3.0 to 3.0 in increments of 0.5 logits) were generated for a total sample size of 15,000 simulees. This sample will be discussed further under the topic of *Exposure Control*.
- 4) CAT sample. A sample of 1,000 simulee responses, $N(0,1)$, were generated for evaluation with the seven CAT conditions.

CAT simulations

A simulated CAT was conducted for each of the seven study conditions using an item pool of 157 polytomously scored items and a fixed length 20 item test. The initial theta estimate for each simulee was set to zero. Maximum information was used to select items (contingent upon exposure control and content constraints) and maximum likelihood estimation (MLE) was used to compute both provisional and final theta estimates. A variable step-size measure (Dodd & Koch, 1987) was used for provisional theta estimation early in the test until responses were made into two different categories (usually after the 2nd item administration) at which point MLE could begin. The variable step-size computes the theta estimate for a simulee by identifying the minimum and maximum step values in the pool and adding half the distance to the end of the pool to the previous theta estimate. The direction of movement (positive or negative) is determined by the simulee's response to the previous item (a low category response will cause the new theta estimate to be more negative, whereas a high category response will cause the new theta estimate to be more positive). In the current study, the variable step-size algorithm was modified slightly to accommodate content balancing such that rather than identifying the minimum and maximum step values in the item pool, the algorithm identified the minimum and maximum step values within the selected content area.

Content Balancing

Content balancing was accomplished using the Kingsbury and Zara (1989) constrained CAT (or CCAT) method. Items in the pool were classified as belonging to one of nine item types based on their content area (social sciences, natural sciences, humanities) and the number of categorical responses (3, 4 or 5). After each item administration, the total percent of items on a simulee's test from each of these nine groupings was computed and compared to predetermined targets. The item type whose percent administered differed the most from its target was then cued up as the next item type which would be administered. The targets for content balancing were determined by computing the percentage of each item type in the pool (see Table 1).

Exposure Control

Four different methods of exposure control were employed in the CAT conditions—Simpson-Hetter, conditional Simpson-Hetter, Kingsbury-Zara, and modified-within-10-logits. The first two methods required a two-phase implementation where the exposure control parameters were derived through simulation in the first phase and applied to actual CATs in the second phase. The second two methods could be implemented directly through programmatic changes to the CAT algorithm and did not require a simulation phase.

Simpson-Hetter

For the first phase of the Simpson-Hetter, 30 iterations were run with a target exposure rate (r) of 0.39 so that the maximum probability of administering an item would converge to 0.40 (Hetter & Simpson, 1997). Several initial attempts were made to set the target exposure rate to lower levels (0.19 and 0.29), but were unsuccessful in obtaining convergence of the exposure control parameters to the desired target with this item pool and CAT structure. Further attention is given to this issue in the *Discussion* section.

Each iteration of this phase consisted of administering CATs to the 8,000 simulees in the Sympon-Hetter sample and tallying up the number of times each item was selected for administration. This number was used to compute the probability of selection for each item, $P(S)$, by dividing the total number of times it was selected for administration by the total sample size (8,000). An item's exposure control parameter was assigned as a direct result of its $P(S)$ according to the following rules:

$$\text{If } P(S) > r, \text{ then } K_i = r/P(S),$$

$$\text{IF } P(S) \leq r, \text{ then } K_i = 1.0$$

where r is the target exposure rate and K_i is the item's exposure control parameter. An item could then only be administered if its K_i exceeds the value of a uniform random number. A higher K_i indicated that an item was more likely to be administered if selected. A lower K_i indicated that an item was less likely to be administered if selected. For the first iteration, K_i was set equal to 1.0 for all items so that all items would be administered if selected, thereby establishing a baseline for exposure rates. To ensure that there would always be at least one item available for administration, the twenty items with the largest K_i values at the end of each iteration had their K_i automatically reset to 1.0. These items were also selected with respect to the content requirements of the CAT. In this way, there would always be a sufficient number of items to allow for a full length CAT.

In the second phase of the Sympon-Hetter, the exposure control parameters resulting from the final iteration of the first phase were used to regulate the administration of items to simulees in the $N=1000$ CAT sample. An item was selected based upon maximum information item selection (and content balancing criterion), but its K_i was evaluated against a random number drawn from a uniform distribution before it could be administered. If an item's K_i exceeded the value of the uniform random number, it was administered. If not, it was blocked

from further selection for the current simulee and another item was selected via maximum information. This continued until an item was found that could be administered.

Conditional Simpson-Hetter

For the first phase of the conditional Simpson-Hetter, 30 iterations were run with a target exposure rate (r) of 0.39 at each of the 15 theta levels so that the conditional maximum probabilities of administering an item would converge to 0.40. Each iteration of this phase consisted of administering CATs to the 15,000 simulees in the conditional Simpson-Hetter sample and tallying up the number of times each item was selected for administration at each of the 15 theta levels. After each iteration, a vector of K_i s was computed for each item to reflect its $P(S)$ at each level of theta. As with the Simpson-Hetter, each vector of K_i s was set equal to 1.0 for the first iteration so that a baseline exposure rate could be determined.

In the second phase of the conditional Simpson-Hetter, the 157 (item) X 15 (theta) matrix of K_i parameters was applied to constrain the exposure of items to the CAT sample. Following maximum information item selection for a simulee, the appropriate K_i parameter was found by looking up the row and column in the K_i matrix which represented the selected item and the theta level closest to the simulee's current estimated theta value. This K_i parameter was then compared to a random number drawn from a uniform distribution. If the K_i parameter was greater than the random number then the item was administered to the simulee; otherwise, the item was blocked from further selection for the current simulee and another item was selected via maximum information. This continued until an item was found that could be administered.

Kingsbury-Zara

Rather than selecting the single most informative item for a simulee's current theta estimate, the Kingsbury-Zara procedure selected the 3 or 6 (depending on the condition) most

informative items. From these item candidates, one item was randomly selected for administration and the other item candidates were returned to the pool.

Modified-within-.10-logits

Maximum information item selection was used to select item candidates within a prespecified range of a simulee's current theta estimate. Items were selected which maximized information at the current theta estimate, 0.10 logits below the current theta estimate, and 0.10 logits above the current theta estimate. For the 3-item candidate condition, one item was selected at each of these points. For the 6-item candidate condition, two items were selected at each of these points. From this group of item candidates, one item was randomly selected for administration and the other item candidates were returned to the pool.

Data Analyses

Given that exposure control procedures function by modifying optimal item selection, it is not surprising that there is an expected tradeoff between exposure control and theta estimation. Therefore, in any evaluation of an exposure control procedure it is important to look at variables which measure both test security and measurement precision. In the current study, measurement precision was evaluated for each condition with the following variables:

- *Number of nonconvergent cases.* After all 1,000 simulees had taken each CAT condition, the number of simulees for whom extreme theta estimates were obtained (greater than 4.0 or less than -4.0) was tallied.
- *Theta estimate.* The mean and standard deviation of the theta estimate was computed for each condition and compared to the mean and standard deviation of known thetas.

- *Standard error of measurement.* The standard error of measurement was computed for each simulee as the inverse of the square root of test information. The mean and standard deviation of the standard error was then determined for each condition.
- *Correlation between known and estimated thetas.* The mean correlation between known and estimated theta values was computed for each condition.
- *Bias and RMSE.* Bias and root mean squared error (RMSE) provide another summative statistic for comparing known and estimated theta. The equations to compute these statistics are:

$$Bias = \frac{\sum_{k=1}^n (\hat{\theta}_k - \theta_k)}{n}, \quad (2)$$

$$RMSE = \left[\frac{\sum_{k=1}^n (\hat{\theta}_k - \theta_k)^2}{n} \right]^{1/2}, \quad (3)$$

where $\hat{\theta}_k$ is the estimate of trait level for simulee k, θ_k is the known trait level for simulee k, and n is the total number of simulees.

In addition to the above measures, test security was evaluated for each condition with the following variables:

- *Exposure rate.* The exposure rate of an item is calculated as the number of times an item was administered divided by the total number of simulees in the CAT sample (1,000). A frequency distribution of exposure rates is provided for each condition. In addition, the mean, standard deviation, and maximum exposure rate was computed for each condition.
- *Pool utilization.* The number and percent of items not administered was calculated and provides an indication of the extent to which the item pool was utilized in each condition.

- *Test overlap.* The number of items in common between the tests of two simulees (overlap) was computed for every possible pairing of simulees. The mean test overlap between two simulees was then computed for each condition. A data file containing the number of items shared among the simulees as well as the difference between their known theta values was then created to obtain an index of test overlap conditional on theta. Simulees were defined to have “similar” trait levels when their known thetas differed by two logits or fewer and “different” trait levels when their known thetas differed by more than two logits (Davis & Dodd, 2003; Davis, et al., 2003, Davis, 2004).

Results

Item Pool

Table 1 provides descriptive statistics of the estimated item parameters for the total item pool and for each of the nine item types for which content balancing was implemented. As can be seen, the majority of the pool (63%) was comprised of 3-category items (two step values) with roughly 19% 4-category items (three step values) and an additional 19% 5-category items (four step values). Similarly, the item types with the largest numbers of items were humanities with 3-categories and social sciences with 3-categories. Figures 1 and 2 present the information function for the total pool and separately by item type, respectively. The information for the total pool peaks at a theta value of -0.5 logits. The information functions for humanities with 3 categories (H3) and social sciences with 3 categories (SS3) are somewhat larger than for the other item types due to the larger number of items in these classifications. However, all item types appear to be targeting the same relative area of the theta scale, peaking from -0.2 to -0.8 logits.

