Psychological Test and Assessment Modeling, Volume 56, 2014 (3), 255-274

Optimal item pool design for computerized adaptive tests with polytomous items using GPCM

Xuechun Zhou[1] *& Mark D. Reckase*[2]

Abstract

Computerized adaptive testing (CAT) is a testing procedure with advantages in improving measurement precision and increasing test efficiency. An item pool with optimal characteristics is the foundation for a CAT program to achieve those desirable psychometric features. This study proposed a method to design an optimal item pool for tests with polytomous items using the generalized partial credit model (G-PCM). It extended a method for approximating optimality with polytomous items being described succinctly for the purpose of pool design. Optimal item pools were generated using CAT simulations with and without practical constraints of content balancing and item exposure control. The performances of the item pools were evaluated against an operational item pool. The results indicated that the item pools designed with stratification based on discrimination parameters performed well with an efficient use of the less discriminative items within the target accuracy levels. The implications for developing item pools are also discussed.

Keywords: computerized adaptive testing, item pool design, G-PCM, *p*-optimality method

[1] *Correspondence concerning this article should be addressed to:* Xuechun Zhou, PhD, 19500 Bulverde Road, San Antonio, TX 78259, USA; email: Xuechun.Zhou@pearson.com

[2] Michigan State University

Introduction

The advantages of computerized adaptive testing (CAT) have been widely acknowledged. A majority of the current CAT applications were developed for tests consisting of dichotomous items, but applying CAT to tests including polytomous items are evaluated for educational and medical assessment as well (Gorin, Dodd, Fitzpatrick, & Koch, 2005). While a substantial amount of research has been conducted on item selection and ability estimation methods for CATs using various polytomous item response theory (IRT) models, few studies have investigated optimal pool characteristics for polytomous CAT implementations.

Since developing an item pool with the desired qualities is an arduous effort with respect to the cost and time spent on writing, revising and pretesting items, it is crucial to understand pool characteristics at the beginning. This goal can be achieved through optimal item pool design using simulations. The main product, an optimal blueprint, summarizes the item and pool attributes such as item distributions and pool size. The blueprint is used to build and manage item pools on a continuous basis.

There were two major methods to address the issue of optimal item pool design. One is the *p*-optimality method used in the current study, which will be introduced in detail later. Another used the integer programming method (Veldkamp & van der Linden, 2000). This method formulates the test assembly as a constrained optimization problem with all the desired qualitative and quantitative qualities being expressed as constraints (Veldkamp & van der Linden, 2000). This approach is flexible in simulating constrained CATs under any IRT model with the objective function varying as desired. However, its application focused more on optimal test assembly assuming an item pool already existed (van der Linden, Adelaide, & Veldkamp, 2006; Ariel, Veldkamp, & Breithaupt, 2006). The resulting item pool blueprint thus might not represent desired optimal characteristics. In addition, this approach did not calculate the number of items needed for a pool (Veldkamp & van der Linden, 2000).

Using the generalized partial credit model (G-PCM) (Muraki, 1992), this study aims to develop an optimal item pool design method for tests consisting of polytomous items by extending the *p*-optimality method (Reckase, 2007). Specifically, the research questions to be addressed are: 1) How do practical constraints such as content balancing and item exposure control affect the optimal item pool design and their performance? 2) For each combination of the constraints, what does the blueprint show with regard to the characteristics and distribution of the items, item pool information distribution, and pool size for a modeled CAT procedure?

Literature review

G-PCM and its information function

G-PCM is an extension of the partial credit model proposed by Masters (1982) by adding the discrimination parameter *a*. With the G-PCM, for item *j* with (m_j +1) possible categories, the probability for a given θ to receive a score k is denoted in function (1) (Muraki, 1992), and its information function is shown in (2) (Donoghue, 1994):

$$P_{jk}(\theta) = \frac{\exp\left[\sum_{v=1}^{k} D\mathrm{a}_{\mathrm{j}}\left(\theta - b_j + d_{jv}\right)\right]}{\sum_{c=1}^{m} \exp\left[\sum_{v=1}^{c} D\mathrm{a}_{\mathrm{j}}\left(\theta - b_j + d_{jv}\right)\right]} \quad (1)$$

$$I_j(\theta) = D^2 a_j^2 \left[\sum_{k=0}^{m_j} k^2 P_{jk}(\theta) - (\sum_{k=0}^{m_j} k P_{jk}(\theta))^2\right] \quad (2)$$

where a_j is a slope parameter that “indicates the degree to which categorical responses vary among items as θ levels change” (see Muraki, 1992). b_j is the location parameter indicating the overall difficulty of the item, and d_{jv} is a threshold parameter that “is interpreted as the relative difficulty of step k in comparing other steps within item j” (see Muraki, 1992). d_0 is defined as 0 arbitrarily because it is cancelled as a common factor. D is the constant 1.7.

The variations relevant to item information curve shape include the magnitude of a-parameter, the distance between the first and last threshold parameters, the ordering of threshold parameters, and the proximity of two adjacent threshold parameters (Akkermans & Muraki, 1997; Dodd & Koch, 1987).

p-Optimality method for optimal item pool design

An optimal item pool is defined as one that can always provide optimal items that satisfy the desired characteristics of a CAT program during implementation (Reckase, 2007). To achieve this, an item pool must have a sufficient number of items and a distribution matching the target population (Boyd, Dodd, & Choi, 2010; Veldkamp & van der Linden, 2000). Throughout this study, item pools are deemed optimal as one of an adequate number of items resulting from CAT simulations with all predetermined psychometric, statistical and practical specifications satisfied.

Perfectly, there should always be an informative item available for each ability estimate, $\hat{\theta}$, using a specified item selection method. Because θ is defined on a continuous scale, even when θs differ as slightly as .001, different items are needed in an absolute sense of optimality. On the other hand, the desired characteristics of items such as item information provided by two items across a short θ interval might vary quite negligibly. Including in an item pool a large quantity of items that function similarly is impractical as it greatly increases financial cost yet barely improves measurement precision.

The p-optimality method was introduced to approximate an optimal pool of smaller size with little loss of specified characteristic (Reckase, 2007). More specifically, items included in a pool are classified using a p-optimality criterion: p-proportional of a desired characteristic with p representing an accepted level. For instance, an item providing 98 % or more of the maximum information is deemed as 98-optimal. Furthermore, such items are considered equally optimal for $\hat{\theta}$ within a defined interval, and they are classified into item sets defined as "bin" units. An additional item is added into a final item pool unless there is no item in a bin or items in the bins reached their predetermined exposure criterion.

With items described in an acceptable *p*-optimal criterion, an optimal pool is generated under predetermined CAT features. A blueprint summarizes the resulting pool with respect to item characteristics and distribution, pool information distribution, and pool size. This method has been successfully applied in designing item pools under various CAT situations using 1-PL model (Rasch, 1960) and 3-PL model (Lord, 1980; Gu, 2007; He, 2010; Reckase, 2007).

Item selection and ability estimation methods

Maximum information is the most widely used method for item selection. Many previous studies on polytomous CAT using the G-PCM recommended maximum information selection method for its performance and ease of computation (Ho, 2010; Van Rijn, Eggen, Hemker, & Sanders, 2002; Veldkamp, 2003).

Content balancing and item exposure control are often used to address psychometric concerns in practice. As the content subtests being modeled are evenly distributed with unidimensionality assumed, the rotation method was used in this study (Boyd et al., 2010; Segall, Moreno, & Hetter, 1997). With the rotation method, items are selected in a fixed order from content-specific subsets (Segall et al., 1997).

In the previous studies of the item exposure control methods in polytomous CAT using the G-PCM, the *a*-stratified method was found to be effective in reducing item exposure and overlap rate and increasing pool utilization with little loss in measurement accuracy (Pastor, Dodd, & Chang, 2002; Yi & Chang, 2003; Yi, Wang, & Wang, 2003). This method controls item exposure rate by 1) dividing the item pool into *K* strata in ascending order of discrimination parameters, *a*, 2) dividing the test into *K* stages accordingly, and 3) selecting items from the corresponding *k*th stratum for administration at each stage until a stopping rule is satisfied (Chang & Ying, 1999). The number of strata is determined by the structure of the item pool such as the variation of the discrimination parameters and item pool size (Hau, Wen, & Chang, 2002). This approach was adopted along with the maximum exposure control.

Weighted likelihood estimation (Warm, 1989) derived from the maximum likelihood estimation (MLE) was used because of its performance in the fixed-length CAT using the G-PCM (Wang & Wang, 2001). Weighted likelihood estimate is obtained by solving the function below

$$\frac{\partial l(u\,|\,\theta)}{\partial \theta} - Bias\left(MLE\left(\theta\right)\right)*I\left(\theta\right) = 0 \tag{3}$$

where $l\left(U|\theta\right)$ = In $L\left(U|\theta\right)$ is the log-likelihood function, and $I\left(\theta\right)$ is the test information function.

For the G-PCM, the maximum likelihood estimate bias function is shown in function (4) (Samejima, 1993; Wang & Wang, 2001)

$$Bias\left(MLE\left(\theta\right)\right) = -\frac{1}{2\left[I\left(\theta\right)\right]^2}\sum_{j=1}^{n}\sum_{k_j} D^3 a_j^3 P_{jk}\left(\theta\right)\left(k-\sum_{c=0}^{m_j} cP_{jk}\left(\theta\right)\right)$$

$$*\left[k^2 - 2k\sum_{c=0}^{m_j} cP_{jk}\left(\theta\right) + 2(\sum_{c=0}^{m_j} cP_{jk}\left(\theta\right)^2 - \sum_{c=0}^{m_j} c^2 P_{jk}\left(\theta\right)\right] \tag{4}$$

where the notations were the same as defined above.

Method

To extend the *p*-optimality method to the G-PCM, strategies to define the bin unit and simulate an item pool with the constraints were proposed. The programming was implemented using the software MATLAB.

Item Sets: *aθ*-Bin

When the maximum information selection method is used, items are optimal in the sense that item information is maximized at a particular θ point. That is, during a simulation to design an item pool, given an initial θ or interim $\hat{\theta}_s$, a set of item parameters needs to be generated based on their mathematical relationship with the maximum information.

Because the polytomous items included in the operational achievement test were scored on a 7-point scale with the detailed rubric, items with six response categories were modeled in this study to mirror the operational items. With the G-PCM, the analytic solution of θ that maximizes $I(\theta)$ is unavailable. Furthermore, with seven parameters determining the amount of information, defining the bin unit using item parameters cannot describe item pool characteristics explicitly as with the applications of 1-PL and 3-PL model. On the other hand, given item parameters, there always exists a unique θ corresponding to where an item reaches its maximum information. Describing items based on the θ value captures a critical item characteristic without referring to location and threshold parameters directly. More importantly, it enables polytomous items to be graphically presented in a succinct manner, which is also essential for interpreting final blueprints. The "bin" concept was thus extended as *aθ*-bin with the θ representing where an item's information is maximized.

In this study, *aθ*-bin was defined by the range of *a*-bin and *θ*-bin. *a*-Bins stratify an item pool based on the variation of ***a***-parameter value (Chang & Ying, 1999; Yi et al., 2003). Upon examining the ***a***-parameter values of the operational items, three strata with the boundary set of (0.55, 0.75), (0.75, 0.95) and (0.95, 1.1) were used. *θ*-Bins centering on zero with a fixed width of 0.8 except at both ends were defined, resulting in 11 *θ*-bins in total.

p-Optimal item pool design

Without the analytic solution between θ and maximum $I(\theta)$, items cannot be generated simultaneously as $\hat{\theta}$ is updated during CAT simulations as applications with 1-PL and 3-PL model did. Instead, a loosely defined bootstrapping approach was adopted to design the final item pools. That is, supposing there were a master pool that has a large quantity of number of items more than a CAT program requires, resampling items with replacement for simulees during the simulation provides a close approximation of the desired optimal item pool. Because the predetermined CAT characteristics are satisfied during the simulation process, the resulting item pool blueprint delineates characteristics and optimal item pool needs.

Based on operational item parameters, a master pool consisting of 3150 items was generated for the simulations conducted in this study, which was approximately 15 times larger than the biggest item pool used in the previous research. For all simulations, the prior ability distribution was assumed to be normal. Maximum item exposure rate was set to 0.20. Ability estimate after the last item was the final $\hat{\theta}$. The general modeled CAT procedure was summarized in Table 1.

There were four simulation conditions for designing the item pools: with content balancing and *a*-stratified constraints, with content balancing control, with *a*-stratified constraint, and unconstrained CAT. They are referred to Conditions 1 - 4 thereafter.

Table 1:
Summary of Modeled CAT Simulation Design

CAT Component		Simulation Procedure
Item Pool	Pool size	Master item pool: 1596 in Content 1, 1554 in Content 2
	Item parameters	Simulated parameters
Item Selection	Initial selection	*U*(-0.4, 0.4)
	Interim selection	Maximum information
	Content balancing*	Rotation
	Exposure control*	Restricted maximum exposure rate of 0.20 *a*-Stratified
Ability Estimation	Initial, interim, and final ability estimation	Weighted likelihood estimation and variable step size
Stopping Rule	Fixed length	Maximum number of items: 12

Note. Content balancing and exposure control apply as needed.

Extending operational item pool

16 polytomous items with six response categories administered in a large-scale achievement test formed the operational pool. They were calibrated using PARSCALE (Muraki & Bock, 1999). This pool was expanded to be of a similar size as the simulated pools using the item parameter replication (IPR) method (Raju, Fortmann-Johnson, Kim, Morris, Nering, & Oshima, 2009).

Evaluating item pool performance

Simulations were conducted to evaluate the performances of the simulated optimal pools (SOP) and extended operational pool (EOP). Two types of distribution were used: 1) 6000 simulees randomly sampled from $N(0,1)$ to evaluate the pool performance in general, and 2) 500 simulees at each of the 41 θ points from -4 to 4 in increments of 0.2 for an evaluation at a conditional level.

The evaluation criteria for ability estimation included Pearson product-moment correlation between the true θ and $\hat{\theta}$, bias, and root mean squared error (RMSE).

$$Bias = \sum_{i=1}^{\mathrm{n}} \left(\hat{\theta}_i - \theta_i \right) / n \tag{5}$$

$$\mathrm{RMSE} = \sqrt{\sum_{i=1}^{\mathrm{n}} \left(\hat{\theta}_i - \theta_i \right)^2 / n} \tag{6}$$

where n is the sample size.

In addition, classification accuracy at where θ equals 0.7 was obtained. This was measured by the percentage of the correct classification, the false-positive (FP) errors and the false-negative (FN) errors.

For item pool utilization, the criteria were overall pool usage, item exposure rate, percentage of items with varying exposure rate and test overlap rate. The discrepancy between the observed and expected item exposure rate follows χ^2 distribution (Chang & Ying, 1999) and is denoted as

$$\chi^2 = \sum_{j=1}^{N} \frac{\left(r_j - L / N \right)^2}{L / N} \tag{7}$$

where r_j is the observed exposure rate for item j, L is the test length, N is the number of items in the item pool.

Item exposure rate is the ratio of the number of item administrations to the total number of examinees. Test overlap rate is defined as the average proportion of items that two randomly selected simulees have in common (Way, 1998). For a fixed-length CAT, it is obtained by the formula below (Chen, Ankenmann, & Spray, 2003)

$$\mathrm{T}_{\mathrm{overlap}} = \frac{n\sum_{j=1}^{\mathrm{N}} \mathrm{r}_{\mathrm{j}}^{2}}{k(n-1)} - \frac{1}{n-1} \tag{8}$$

where k is the number of items in the test and the others as defined earlier.

Results

Item pool characteristics

When the maximum item exposure rate of 0.20 was applied, the practical constraints did not affect pool size much. The resulting SOPs contained 144, 147, 150, and 151 items respectively. Figure 1 plots the pool information curves. The impact of the practical constraints on the pool information was: 1) the pools designed with the *a*-stratified were less informative than those without it, and 2) when the *a*-stratified control was applied, the pool with the content balancing was slightly less informative than the one without it.

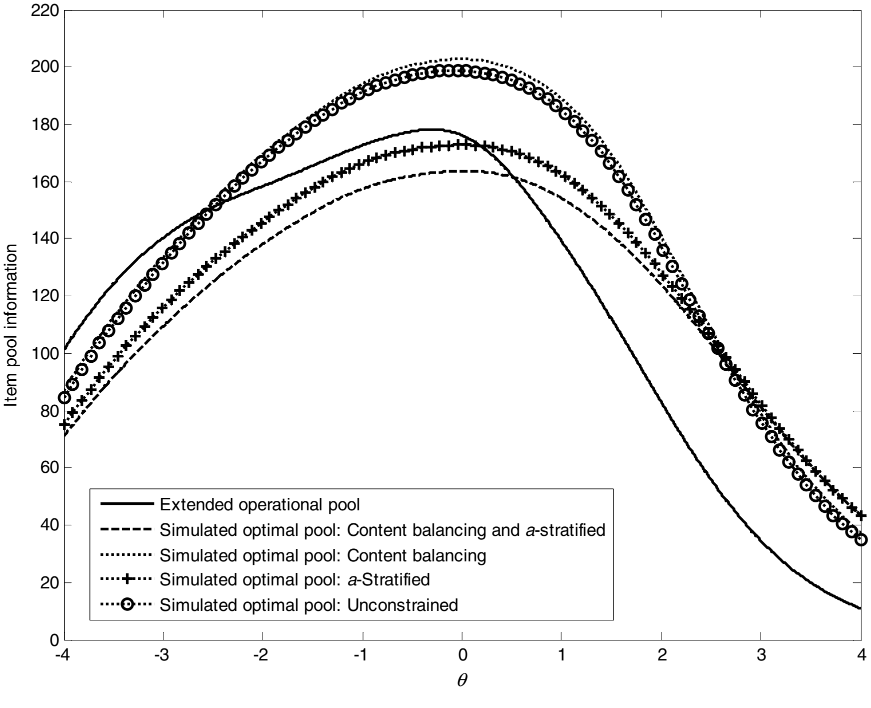


Figure 1:
Pool information curves of the EOP and SOPs

Because results yielded under the Conditions 3 and 4 are similar to Conditions 1 and 2, they were not presented here to save the space. Table 2 shows the descriptive statistics of the ***a***- and ***b***-parameter and the average maximum information the items provided. The distributions of the ***a***-parameter in the SOPs with the *a*-stratified constraint and the EOP were similar. The ***b***-parameter in the SOPs displayed larger mean and variations than those in the EOP with the inclusion of easy and difficult items.

Figures 2 to 4 demonstrate the item distribution of the EOP and SOPs from the Conditions 1 and 2. Because the operational test was not an adaptive one, Figure 2 illustrates that the items reached their maximum information in the middle of the ability scale. Additionally, the EOP contained highly informative items: 85 % of them having the ***a***-parameter larger than 0.75.

For the SOP under the most constrained condition, Figure 3 shows that it contained the items that were informative across the entire θ scale and they were distributed somehow symmetrical around the central θ-bin across the strata. Furthermore, the item distributions in two content areas were not the same. More items were included in the three θ-bins in the middle for all three strata. In content area one, there were fewer items in the first stratum than in other two strata. It should be noted that there were more items with low discrimination in the SOP compared with the EOP, 24 % versus 15 %.

As shown in Figure 4, the SOP under Condition 2 included a large proportion of highly discriminative items: 88 % with the ***a***-parameter larger than 0.95. The average item maximum information was larger as well, 1.46 versus 1.30. In addition, fewer items were needed for the first content area.

Table 2:
Statistics of Discrimination, Location Parameters, and Maximum Information

	a				***b***				**Maximum Information**			
Pool Size	**Mean**	***SD***	**Min**	**Max**	**Mean**	***SD***	**Min**	**Max**	**Mean**	***SD***	**Min**	**Max**
Extended Operational Pool												
144	0.89	0.14	0.64	1.14	-1.08	0.40	-1.83	-0.50	1.30	0.43	0.71	2.37
Simulated Optimal Pool: Condition 1[*]												
144	0.91	0.14	0.67	1.10	-0.25	1.05	-1.91	1.97	1.24	0.27	0.75	1.83
Simulated Optimal Pool: Condition 2[*]												
150	1.02	0.12	0.56	1.10	-042	0.98	-1.88	2.00	1.46	0.28	0.60	1.83

Note. Condition 1: Content balancing and *a*-stratified constraints. Condition 2: Content balancing control.

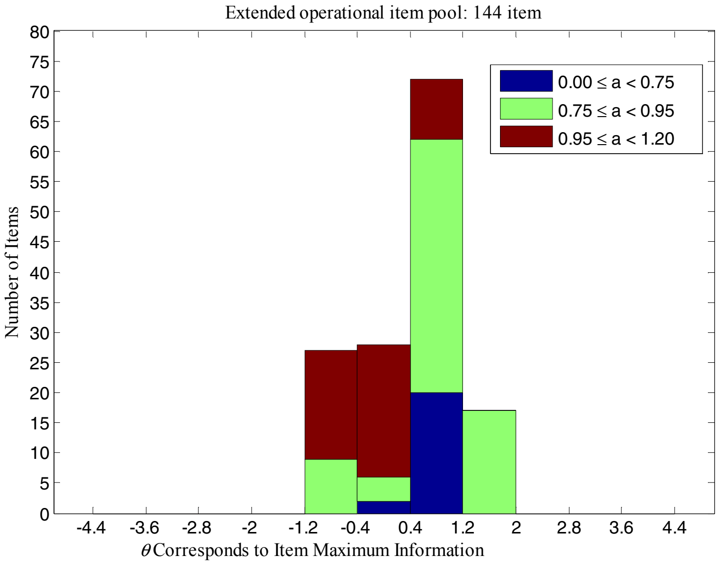


Figure 2:
Item distribution of the EOP

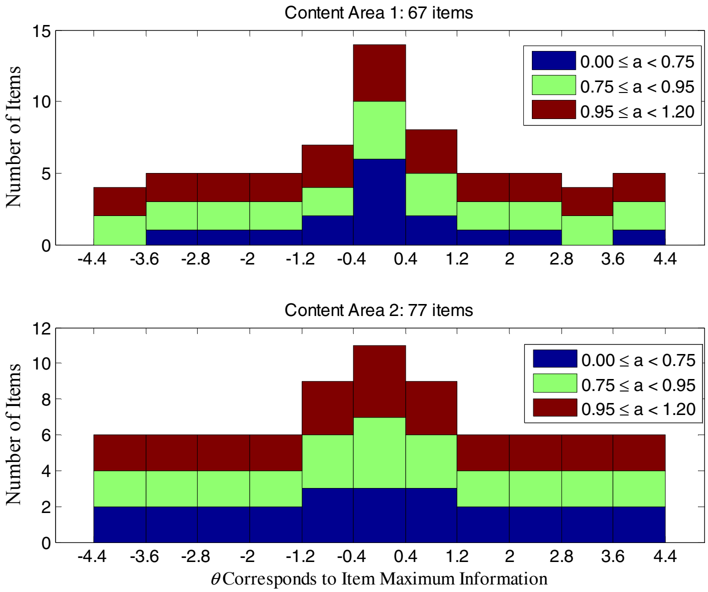


Figure 3:
Item distributions for the SOP under Condition 1

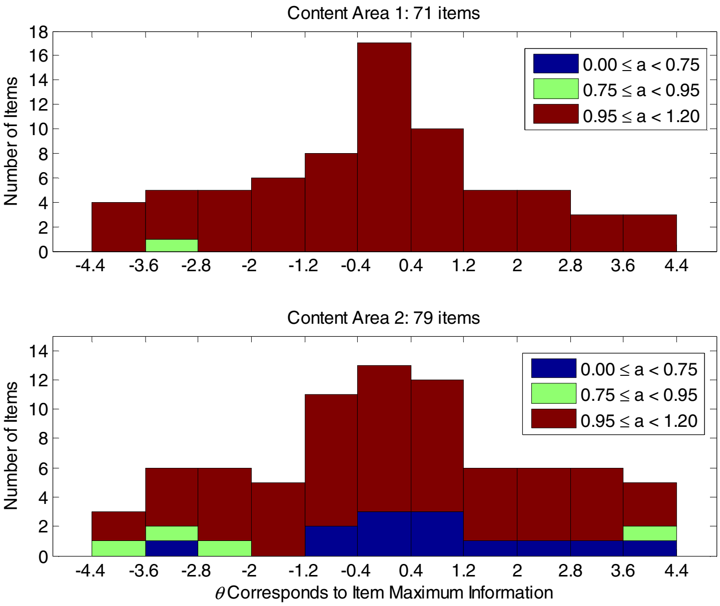


Figure 4:
Item distributions for the SOP under Condition 2

Item Pool Performances

To compare their performances, the EOP was split into different subsets based on the conditions the simulated pools were designed.

Table 3 presents the evaluation results of the ability estimates, classification, and pool utilization. Under Condition 1, the magnitudes of the bias were negligible for the SOP and EOP. RMSE from the SOP was slightly smaller than that of the EOP. The correlation between the true θ and $\hat{\theta}$ was the same, and both were significant. The average test information provided by both pools was nearly the same. The SOP yielded a slightly lower percentage of correct classification than the EOP, 66 % versus 69 %.

The χ^2 value of the item exposure rate of the SOP was slightly larger than that of the EOP. Examining the item usage in the EOP indicated a high percentage of the items that reached the preset maximum exposure rate. More specifically, 31 % of the items were fully used in the EOP and 25 % in the SOP. However, the EOP had a much larger proportion of items that was not used than the SOP, 30 % versus 15 %.

Table 3:
Summary Statistics of Item Pools Performance

Statistic	**Extended Operational Pool**	**Simulated Optimal Pool**	**Extended Operational Pool**	**Simulated Optimal Pool**
	Condition 1		**Condition 2**	
Bias	0.0024	-0.002	-0.02	-0.01
RMSE	0.37	0.36	0.27	0.31
Correlation	0.94	0.94	0.96	0.95
Test information	14.85	14.92	17.97	18.73
Correct classification	69 %	66 %	72 %	74 %
FP errors	16 %	20 %	14 %	12 %
FN errors	15 %	14 %	14 %	14 %
χ^2 of item exposure rate	10.05	10.66	6.86	9.10
Items with exposure rate equals 0.2	31 %	25 %	31 %	25 %
Items with exposure rate between .02 and 0.2	21 %	33 %	22 %	30 %
Items with exposure rate less than .02	18 %	27 %	22 %	28 %
Items that are not used	30 %	15 %	25 %	17 %
Test overlap rate	0.18	0.17	0.17	0.17
Pool size	144	144	144	150

Figure 5 plots the conditional test information, bias, RMSE, and test overlap rate under Condition 1. The SOP provided test information consistently above the target information level, 10.0, even at the extreme θ values. The EOP resulted in test information greater than 10.0 except for the θ values roughly larger than 2.0. For both pools, there was positive bias at the low θ levels and negative bias at the high end. The SOP had a consistently smaller test overlap rate.

Figures 6 and 7 illustrate the item exposure rate for 6000 simulees. As depicted in Figure 6, for the EOP, the distribution of the exposure rate for each stratum was similar for the two content areas. The fully exposed items spread approximately between -1.20 and 1.20, corresponding to the three θ-bins in the middle. Because the EOP had a small number of items in the first stratum, those items were well used. Many items in the second stratum that were informative for the middle ability levels were rarely or never administered. The items in the third stratum in the content area one were better used than those in the content area two.

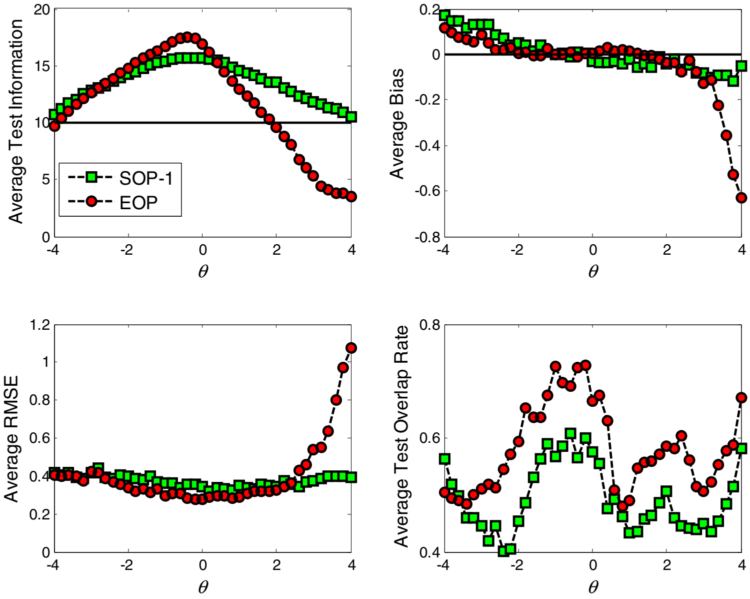


Figure 5:
Pool performance comparison under Condition 1

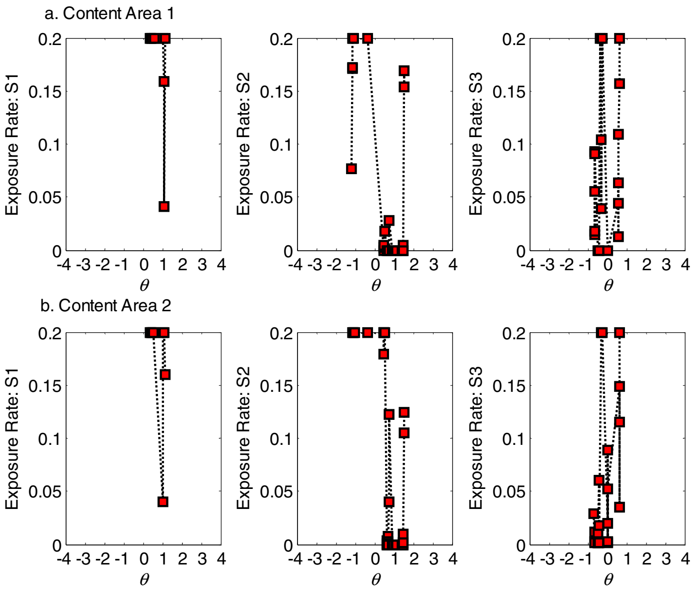
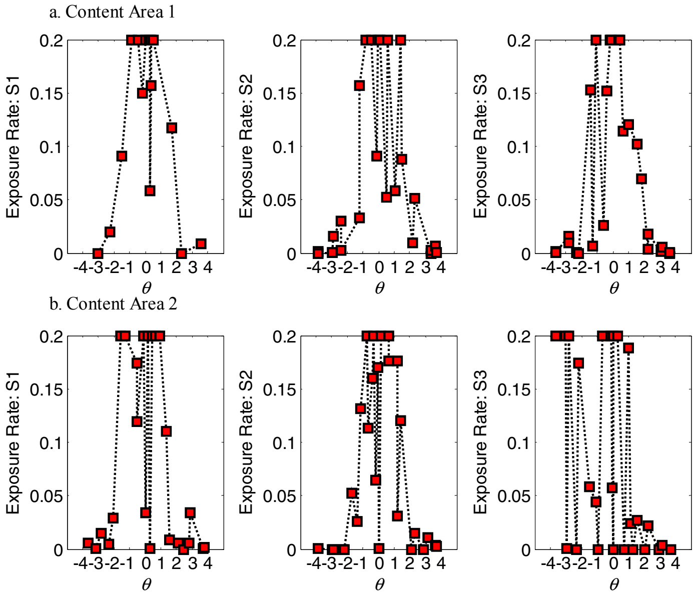


Figure 6:
Item exposure rate of the EOP under Condition 1



a. Content Area 1

b. Content Area 2

Figure 7:
Item exposure rate of the SOP under Condition 1

Figure 7 demonstrates that for the SOP, the distribution of the item exposure rate was similar at the first and second stratum, but differed at the third for two content areas. For both content areas, the items in the four θ-bins in the center were well used and the items informative for the extreme θ values tended to be under-exposed.

For Condition 2, Table 3 indicates that the ability estimates from both pools showed small negative bias. The correlation between the true θ and $\hat{\theta}$ for both pools was almost identical. The average test information from the SOP was higher than the EOP. Negligible differences in classification results were observed. In regards to the item pool usage, the χ^2 of the item exposure rate of the SOP was larger than that of the EOP. The SOP had a smaller percentage of items that were never administered.

As shown in Figure 8, without the a-stratified constraint, the SOP provided test information much higher than the target level across the θ scale. The EOP did not reach the desired level of 10.0 above the θ point 2.0. For both pools, there was positive bias at the left tail, and negative bias at the right.

Figure 9 indicates that the item exposure rate distribution of the two content areas in the EOP was similar. The items in the two center θ-bins were better used than the items at the ends.

Figure 10 shows that the items in the three θ-bins in the middle were well used for the SOP. Several informative items at the lower end were also highly exposed.

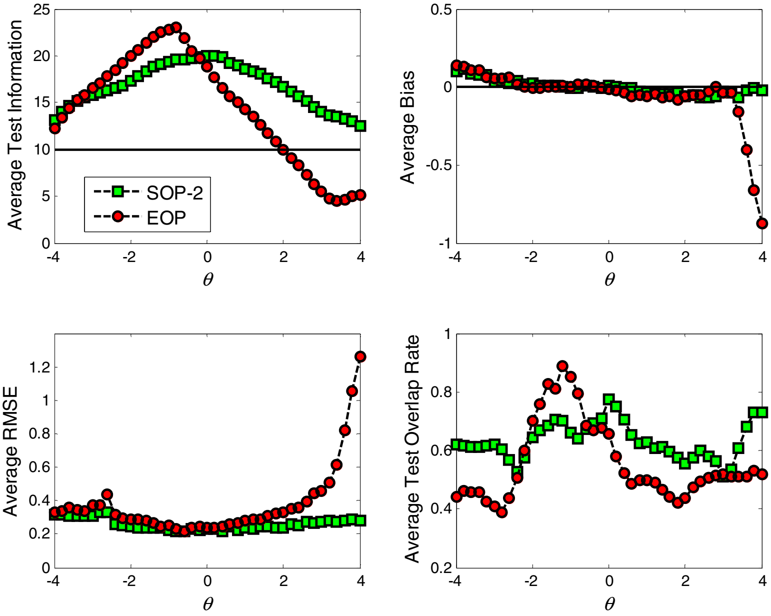


Figure 8:
Pool performance comparison under Condition 2

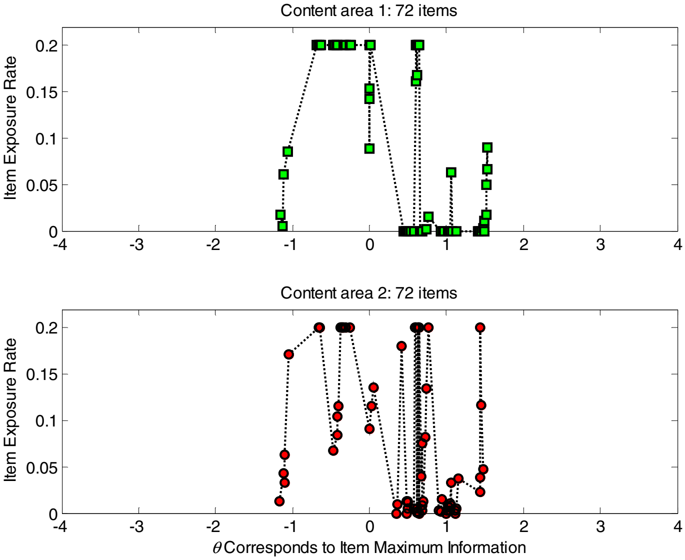


Figure 9:
Item exposure rate of the EOP under Condition 2

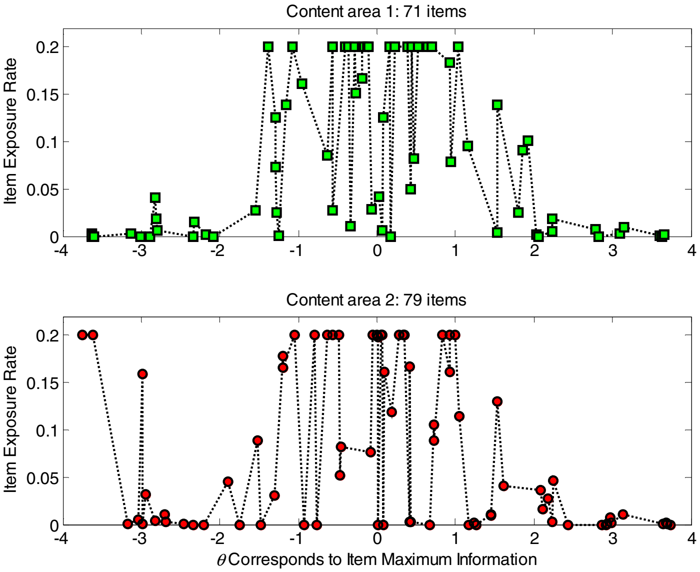


Figure 10:
Item exposure rate of the SOP under Condition 2

To summarize, the evaluation results of the pool performance are closely related to the pool characteristics. When the pools are designed with the *a*-stratified exposure control, the average test information was approximately 15.0, and it reached nearly 19.0 without the constraint. Accordingly, larger RMSE and smaller correlation coefficients were obtained. With regard to the item pool usage, percentage of items that are fully used, rarely used, and never used were quite comparable for the three pools designed with constraints. The SOPs had a smaller proportion of the items that were fully and never administered, but it also had a higher percentage of the items that were rarely used. Furthermore, compared with the EOP, when the *a*-stratified method applied, the conditional test overlap rate of the SOP was consistently lower.

Conclusions and implications

To conclude, the practical constraints of the *a*-stratified exposure and content balancing did not affect pool size to a large extent when the pools were designed under the G-PCM. It should be noted that the maximum item exposure rate of 0.20 was applied for all the conditions, which controlled the pool size to a certain degree.

Regarding the optimal item pool design, the *a*-stratified exposure control affected the pool characteristics to a great extent. First, the items included in the SOPs without the constraint had larger ***a***-parameter than those with the constraint. Correspondingly, the average maximum information those items provide was greater. Secondly, as depicted in

Figure 1, the pool information of the SOPs without the constraint was much larger than the SOPs with the constraint. On the other hand, the content balancing of rotation method had little impact on designing the pools. When the conditions differed only by the content balancing constraint, the distributions of the ***a***- and ***b***-parameter were quite similar.

The average test information differed greatly with and without the a-stratified exposure procedure. However, the difference in the average test information did not imply substantial measurement precision in practice. While the conditional SEM under both conditions was smaller than the desired level of 0.30, a small decrease in the SEM using a larger number of highly discriminative items might not be practical. This is consistent with what previous research has found: the a-stratified control resulted in small decreases in measurement precision in polytomous CAT (Davis, 2004; Pastor et al., 2002).

With regard to the item pool usage, the item usage was quite comparable for the three pools designed with constraint. Compared with the EOP, the SOPs are better utilized in the sense that there were fewer items that were fully exposed, fewer items that were not used, and more items that were well used. When the a-stratified method applied, the lower conditional test overlap rate also suggests a better item pool usage. In addition, the items that were informative at the ends were seldom used, suggesting the potentials to reduce pool sizes by increasing the width of the θ-bins at the ends.

To summarize the blueprint, when the a-stratified method was applied, the item distribution in three strata was similar for the resulting pools. In general, fewer items were needed in the first stratum. With the content balancing control, the number of items for each content area differed as well: fewer items were needed for the first content area. Without any constraints, the SOP yielded more precise measurement, but the item pool usage was not as good as the ones with the constraints.

Concerning item characteristics, following properties are observed from the SOPs: 1) The distributions of the ***a***- and ***b***-parameter were similar under all conditions. Due to the absence of the informative items at the tails in the EOP, the distribution of the ***b***-parameter in the SOPs showed a larger mean and standard deviation; 2) The threshold parameters, d_j, were orderly distributed from the easiest to the hardest. This is the property demonstrated in the operational items, which was also embedded in item generation. Because the sum of the threshold parameters was constrained to zero and polytomous items with six response categories were modeled, the threshold parameters were fairly symmetrical around zero.

For the information distribution at the item and item pool level, the SOPs consisted of items that were informative across the entire θ scale. For the item pools, the pool information curves of the SOPs were smooth and somewhat bell shaped compared with the EOP.

To sum up, the evaluation results showed that the p-optimal item pools designed in this study supported the polytomous CAT implementations with the anticipated measurement accuracy and pool usage. Furthermore, this was achieved with a small number of highly discriminating items. The SOPs designed with the proposed method show advantages in two ways: including items with the maximum information spread across the entire ability continuum and the ***a***-parameter spanning evenly when the *a*-stratified method applied. The practical implications of the results are that 1) item pools with approximately 150 items are sufficient for the polytomous CAT with the defined characteristics in the study; 2) while highly discriminating items are usually desired, including items with varying discriminating parameters can achieve the measurement accuracy as expected when the distribution of item difficulty matches the ability distribution of the target examinee population.

With respect to operational practice, the resulting blueprints can be used to help transforming current paper and pencil testing pool into a one for adaptive testing purposes as needed. For instance, items informative for the high θ levels need to be supplemented. Also, since many of the operational items that are informative for the middle range of θ scale were not efficiently used, they can be used in a rotating manner in CAT operations to maximize their usage and control item exposure.

Furthermore, because the inclusion of more items with low ***a***-parameter did not decrease measurement precision, they can be supplemented to reduce development cost. The item generation strategy can be adopted to generate item parameters with operational characteristics. Then, pool performance can be evaluated to update blueprints dynamically with necessary modifications. After desired performance is achieved, real items with desired characteristics, such as their parameters and content specifications, can be provided to item writers to help them understand item features for writing endeavors.

This study is restricted by the fact that the extended item pool reduplicated from a limited number of operational items may not represent a CAT pool in practice. Furthermore, a CAT program consisting of polytomous items exclusively does not reflect current practice. It is thus worthwhile to examine how the findings could be incorporated with those from the optimal item pool design research using dichotomous IRT models to design item pools for mixed-format CAT implementations. Future studies are also needed to investigate to what extent pool sizes can be reduced without losing measurement precision by adjusting the θ-bin width. A combination of narrow widths in the center and wide ones at the end might be a reasonable approach to approximate optimal pools of smaller sizes and improve item usage.

References

Ariel, A., Veldkamp, B. P., & Breithaupt, K. (2006). Optimal testlet pool assembly for multistage testing designs. *Applied Psychological Measurement*, *30*(3), 204-215.

Akkermans, W., & Muraki, E. (1997). Item information and discrimination functions for trinary PCM items. *Psychometrika, 62(4),* 569-578.

Boyd, A., Dodd, B., & Choi, S. (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.

Chang, H., & Ying, Z. (1999). *a*-stratified multistage computerized adaptive testing. *Applied Psychological Measurement, 23*(3)*,* 211-222.

Chen, S., Ankenmann, R. D., & Spray, J. (2003). The relationship between item exposure and test overlap in computerized adaptive testing. *Journal of Educational Measurement, 40(2),* 129-145.

Davis, L. L. (2004). Strategies for controlling item exposure in computerized adaptive testing with the generalized partial credit model. *Applied Psychological Measurement*, *28*(3), 165-185.

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*(4), 371-384.

Donoghue, J. R. (1994). An empirical examination of the IRT information of polytomously scored reading items under the generalized partial credit model. *Journal of Educational Measurement, 31(4),* 295-311.

Gorin, J. S., Dodd, B. G., Fitzpatrick, S. J., & Koch, W. R. (2005). Computerized Adaptive Testing With the Partial Credit Model: Estimation Procedures, Population Distributions, and Item Pool Characteristics. *Applied Psychological Measurement*, *29*(6), 433-456.

Gu, L. (2007). Designing optimal item pools for computerized adaptive tests with exposure controls. Unpublished doctoral dissertation. Michigan State University.

Hau, K.T., Wen, J.B., & Chang, H.H. (2002, April). *Optimum number of strata in the astratified computerized adaptive testing design.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

He, W. (2010). Optimal item pool design for a highly constrained computerized adaptive test. Unpublished doctoral dissertation. Michigan State University.

Ho, T. (2010). A Comparison of item selection procedures using different ability estimation methods in computerized adaptive testing based on the Generalized Partial Credit Model. Unpublished doctoral dissertation. University of Texas at Austin.

Lord, F. M. (1980). *Applications of item response theory to practical problems*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*(2), 149-174.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*(2), 159-176.

Muraki, E. & Bock, D. (1999). *PARSCALE: Parameter scaling of rating data (Version 4.1)* [Computer software]. Lincolnwood 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(2),* 147-163.

Raju, N. S., Fortmann,K. A., Kim, W., Morris, S. B., Nering, M. L., & Oshima, T. C. (2009). The item parameter replication method for detecting differential functioning in the polytomous DFIT framework. *Applied Psychological Measurement, 33(2),* 133-147.

Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. Denmarks Paedagogiske Institut, Copenhagen.

Reckase, M. D. (2007). The design of *p*-optimal item bank for computerized adaptive tests. In D. J. Weiss (Ed.). *Proceedings of the 2007 GMAC Conference on Computerized Adaptive Testing.* Retrieved from www.psych.umn.edu/psylabs/CATCentral/

Samejima, F. (1993). An approximation for the bias function of the maximum likelihood estimate of a latent variable for the general case where the item responses are discrete. *Psychometrika*, *58*(1), 119-138.

Segall, D. O., Moreno, K. E., & Hetter, D. H. (1997). Item pool development and evaluation. In W. A.Sands, B. K.Waters, & J. R. McBride (Eds.), *Computerized adaptive testing: From inquiry to operation* (pp. 117–130). Washington DC: American Psychological Association.

Smarter Balanced Assessment Consortium: Technology-enhanced items guidelines. (2012). Retrieved from http://www.psych.umn.edu/psylabs/CATCentral/www.smarterbalanced.org/

Snyman, J. A. (2005). *Practical mathematical optimization: An introduction to basic optimization theory and classical and new gradient-based algorithm.* New York, NY: Springer Science Business Media, Inc.

van Rijn, P. W., Eggen, T. J. H. M., Hemker, B. T., & Sanders, P. F. (2002). Evaluation of selection procedures for computerized adaptive testing with polytomous items. *Applied Psychological Measurement, 26(4),* 393-411.

van der Linden, W. J., Adelaide, A., & Veldkamp, B. P. (2006). Assembling a computerized adaptive testing item pool as a set of linear tests. *Journal of Educational and Behavioral Statistics, 31*(1), 81-100.

Veldkamp, B. P. (2003). Item selection in polytomous CAT. In H. Yanai, A. Okada, K. Shigemasu, Y Kano, & J. J. Meulman (Eds.). New developments in psychometrics (pp. 207-214). Tokyo, Japan: Springer-Verlag.

Veldkamp, B. P. & van der Linden, W. J. (2000). Designing item pools for computerized adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.) *Computerized adaptive testing: Theory and practice*. Dordrecht, the Netherlands: Kluwer.

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*(3), 427-450.

Wang, S., & Wang, T. (2001). Precision of Warm's weighted likelihood estimates for a polytomous model in computerized adaptive testing. *Applied Psychological Measurement*, *25*(4), 317-331.

Way, W. D. (1998). Protecting the integrity of computerized testing item pools. *Educational Measurement: Issues and Practice, 17,* 17-27.

Yi, Q., & Chang, H.H. (2003). a-Stratified CAT design with content blocking. *British Journal of Mathematical and Statistical Psychology, 56* (2), 359-378.

Yi, Q., Wang, T., & Wang, S. (2003). Implementing the *a*-Stratified method with *b* blocking in computerized adaptive testing with the Generalized Partial Credit Model. Paper presented at the annual meeting of American Educational Research Association, Chicago, IL.