

Running Head: CAT ITEM POOL CHARACTERISTICS

Optimizing Item Pool Characteristics to Control Item Exposure
in a Computerized Adaptive Test

Cherdsak Iramaneerat, M.D., M.H.P.E.

Department of Educational Psychology, College of Education

University of Illinois at Chicago

John Stahl, Ph.D.

Pearson VUE

Paper presented at the annual American Educational Research Association Meeting, Chicago, Illinois, 2007

Author for correspondence:
Cherdsak Iramaneerat, M.D., M.H.P.E.
1407 South Indiana Ave. Chicago, IL 60605
Tel: 312-945-3363
Email: cirama1@uic.edu

OPTIMIZING ITEM POOL CHARACTERISTICS TO CONTROL ITEM EXPOSURE
IN A COMPUTERIZED ADAPTIVE TEST

Abstract

We examined the relationship between item pool characteristics (pool size and item difficulty distribution) and item exposure rates in a Rasch model-based variable-length fixed standard error of measurement computerized adaptive test. Item utilization was most effective when item difficulty distribution in the pool was appropriate for the targeted probability of correct response (using pools that had mean item difficulty of 5.00, 4.595, and 4.193 logits for the tests that selected items with .5, .6, and .7 probability of correct response, respectively, for a group of examinees with mean ability measures of 5 logits). Appropriate pool sizes for tests that selected items with .5, .6, and .7 probability of correct response were 400-600 items, 400-600 items, and 600 items, respectively.

OPTIMIZING ITEM POOL CHARACTERISTICS TO CONTROL ITEM EXPOSURE IN A COMPUTERIZED ADAPTIVE TEST

A computerized adaptive test (CAT) is a test that combines the advanced information processing capabilities of computers with the understanding of cognitive ability testing through item response theory (IRT) to produce a test that is individually tailored to each examinee based on the examinee's responses to previous items in order to accurately estimate that examinee's ability (Parshall, Spray, Kalohn, & Davey, 2002a). A CAT offers many laudable features, including improved examinees' testing experience, increased testing efficiency, elimination of paper test books and answer sheets, improved test security, convenient test scheduling, and immediate scoring and feedback to examinees (Wainer, 2000). However, in providing examinees a continuous opportunity to test, a CAT creates a special circumstance that can lead to the overexposure of certain items within the pool to examinees, causing a threat to test security (Davis & Dodd, 2005). Thus, a good CAT program requires an effective system to maintain and renew an inventory of items in the pool to control item exposure (Way, Steffen, & Anderson, 2002).

A common strategy employed by large CAT testing programs in maintaining the security of their item inventory is to create several item pools and constantly rotate these pools over time. Large CAT testing programs generally keep their items in an item vat (a.k.a. item bank) and pull some portions of an item vat to develop multiple item pools, which are used to tailor tests for individual examinees (Way & Steffen, 1997; Way et al., 2002). Maintaining test security of a CAT program requires good planning on how to select items from an item vat to create item pools of appropriate size and difficulty distribution.

Researchers have studied the relationship between item pool sizes and item exposure in some operational CATs. Stocking (1994) demonstrated that a fixed-length CAT required an item pool size about six to eight times of parallel linear tests (or about 12 times the length of the CAT). Stahl and Lunz (1993) studied the degree of item overlap in five variable-length Rasch model-based CATs, where the examination length ranged from 50 to 100 items, employing a confidence interval stopping rule that stopped the test when examinee ability estimates were 1.65 times the SEM above or below the criterion-referenced passing standard. They suggested that the minimum item pool size to control item overlap should be 400-500 items, or 600–800 items if there was discrepancy between examinee ability and item difficulty distributions.

Besides stopping a CAT when the test reaches a specified number of items (a fixed-length CAT) or when the test reaches a point where it can tell that an examinee ability estimate is above or below the passing standard with a specified confidence interval (a confidence interval stopping rule), we can also stop the test when it can estimate examinee ability with a pre-specified level of precision (a fixed standard error of measurement (SEM) stopping rule), which guarantees that an ability measure of each examinee will be as precise as a test designer wants. This provides score estimates that conform to the *equal measurement error variance* assumption of some statistical analyses (Thissen & Mislevy, 2000). However, an appropriate item pool size for a variable-length CAT that employs a fixed SEM stopping rule has not yet been studied.

Determining an item pool size for a variable-length fixed SEM CAT is complicated because item utilization depends on the interaction between examinee characteristics, item characteristics, and the algorithm used in administering the test. One of the key variables in test administration algorithm that controls item utilization is the targeted probability of correct response. Although administering items that target at .5 probability of correct response yields

maximal information, resulting in the most effective test, sometimes a testing program does not select items that target at .5 probability of correct response to improve examinees' testing experience (i.e., letting them have correct responses more than incorrect responses), and to improve item pool utilization of a pool that does not have many difficult items (Bergstrom & Lunz, 1999). Thus, it is also practically important to examine item pool utilization in a setting where items are selected to target at probability of correct response higher than .5.

Besides the targeted probability of correct response, there are many factors that influence the item utilization in a variable-length fixed SEM CAT, including the desired precision level of examinee ability estimates, the content balancing, and item exposure control (Bergstrom & Lunz, 1999; Thissen & Mislevy, 2000). Determining an appropriate item pool characteristics to optimize item utilization requires the consideration of these factors and how they interact with examinees.

The precision of the examinee ability estimate is determined by the standard error of measurement (SEM). A CAT starts without the knowledge of an examinee's ability. A CAT then uses information obtained from the examinee's responses to estimate that examinee ability measure. At the beginning of the test, this examinee ability estimate has a large SEM. As the test progresses, the computer collects more information from the responses of an examinee and gradually reduces the SEM associated with the ability measure. The more precise the ability estimate is needed, the more items are required to be administered. The precision of examinee ability estimate is directly related to the internal consistency reliability of the test. The more precise the ability estimate, the more reliable the test result (Thissen, 2000). In other words, if we get a very precise ability estimate, we will be very confident that the test result will be the same if we repeat the test under the same condition (at the expense of test items used to obtain the

information). How reliable a test should be depends on the stakes of the test and the amount of test development resources we have. Most educational measurement professionals suggest that high-stake assessments should have internal consistency reliability higher than 0.9 (Downing, 2004).

Content coverage of the test is always a major concern in test development. Besides selecting items to be administered to an examinee based on item difficulty level, the CAT administration algorithm also has to balance items from various content areas to make sure that the test does not have a problem of construct underrepresentation (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999). How items in the pool are used depends not only on their difficulty levels, but also on their content areas. In a test that is supposed to assess knowledge in all content areas equally using a CAT pool that has unbalanced number of items among different content areas, items from the content area that has less items in the pool will be used more often than those from the content area that has more items.

If a CAT selects test items based solely on their difficulty levels, certain items tend to be administered to nearly every examinee, while many items in the pool are never used. Overexposing certain items to examinees over a period of time is a threat to test security. On the other hand, underutilization of items can lead to inefficiency and unnecessary increased cost (Buyske, 2005). CAT programs generally employ an item exposure control mechanism to make a balanced use of the entire item pool rather than using only a small subset of items in the pool (Davey & Nering, 2002). The combination of an item exposure control procedure used in test administration and the acceptable level of item exposure rates can have a significant influence on how the items in the pool are utilized.

Among the many factors that can influence item pool utilization in a variable-length fixed CAT, this study focused on how the changes in the targeted probability of correct response affect item exposure rate of item pools of various characteristics in term of the pool size and the item difficulty distribution in the pool, while holding the desired precision level of examinee ability estimates, the content balance control, and the rules for item exposure control constant. In other words, we studied the item exposure rates in a variable-length fixed SEM Rasch model-based CAT as a function of (1) the targeted probability of correct response (.5, .6, and .7), (2) item pool size (400, 600, and 800 items), and (3) the difficulty distribution of items in the pool (normal distribution with means of 5.00, 4.60, and 4.15, all with a SD of 1.00).

Generally, the studies of item exposure in CAT focused mainly on the issue of test security resulting from overexposing items in the pool. In this study, we expanded the scope of discussion to address not only the problem of item overexposure, but also the problem of item underutilization. Development of CAT items is expensive. Underutilizing items is a waste of time and effort spent crafting those items that have never been used or rarely been used. Developing an item pool to avoid overexposing items without considering the issue of item underutilization can be costly. Thus, we carried out this study to find ways to optimize item pool characteristics to minimize both item overexposure and item underutilization.

We hypothesized that item exposure rates would be lowest when we administered the test from the largest item pool that had item difficulty distribution matched properly with the targeted probability of correct response in test administration algorithm. On the other hand, item exposure rates would be highest when we administered the test from the small item pool that had item difficulty distribution inappropriate for the targeted probability of correct response in test administration algorithm. The appropriate item pool characteristics that optimize item utilization

would be different for the three targeted probabilities of correct response. The findings from this study would help identify which combination of item difficulty distribution and item pool size led to the best item utilization (minimizing both item underutilization and item overexposure) for various targeted probability of correct response. This would help a CAT program manager in the development and maintenance of item pools in a fixed SEM CAT testing environment to assure test security with efficient use of resources.

Methods

We conducted a series of computerized adaptive testing simulation studies, using the Promissor CAT simulator (Becker, 2006). The test was administered and analyzed with the basic Rasch model for dichotomous responses (Rasch, 1960). The basic Rasch model is a probabilistic model that describes a dichotomous response as a function of the ability of an examinee and the difficulty of an item, using the following mathematical formula:

$$\ln\left(\frac{P_{ni}}{1 - P_{ni}}\right) = B_n - D_i, \quad (1)$$

where P_{ni} is the probability that an examinee n will answer an item i correctly, B_n is the ability level of an examinee n , and D_i is the difficulty level of an item i .

Test administration algorithm

We examined item exposure rates in 27 simulations of Rasch model-based CAT. These simulations employed three item pool sizes (400, 600, and 800 items), each had three item difficulty distributions (normal distributions with means of 5.00, 4.595, and 4.153 logits, all with a SD of 1.00 logit), and selected items to target three different probabilities of correct response (.5, .6, and .7) in testing a group of 1,000 simulated examinees, with a mean ability of 5.00 logits

and a SD of 1.00 logit. We chose this distribution (instead of a typical distribution with a mean of 0 logit and a SD of 1 logit) to avoid negative values in the analysis. The first simulation examined exposure rates of items in a 400-item pool that had mean item difficulty of 5.00 logits and a SD of 1.00 logit, using an item selection algorithm that selected items to target the probability of correct response of .5. However, to avoid item overexposure, the simulator randomly selected an item within 0.10 logit randomization interval of the target difficulty level (Bergstrom & Lunz, 1999). We limited the difficulty of the first ten items within +/- 0.20 logits of the previously administered item. The test continued until the examinee's ability estimate had a SEM equaled to or lesser than 0.3162, which corresponded to a value of the information function of 10, or a reliability of .9 (Thissen, 2000; Thissen & Mislevy, 2000).

Each item pool had four content areas; each contributed 25% of items in the pool. We administered items from all four content areas equally (25% each). To ensure the content coverage of the test, we did not allow a test to be shorter than 40 items. We limited the maximum number of items in each test at 100 because of the anticipated time constraints in a real testing situation.

To study the influence of the mismatch between item difficulty distribution and examinees' ability distribution, we repeated our simulation to the same group of examinees, with the change in item difficulty distribution to the ones with a mean difficulty of 4.595 and 4.153 logits, both with a SD of 1.00 logit. According to equation (1), these two distributions were appropriate for a CAT that administered items with expected probability of correct response of .6 and .7, respectively ($4.595 = 5.0 - \ln(0.6/0.4)$; and $4.153 = 5.0 - \ln(0.7/0.3)$). Thus, this study demonstrated how the discrepancy of .1 and .2 between the targeted probability of correct response and the appropriate probability for the pool would influence item exposure rates.

To study the influence of the change in item pool size, we studied these simulations on three different item pool sizes: 400, 600, and 800 items. We selected these pool sizes based on prior studies of item pool sizes in fixed-length CATs (Stocking, 1994) and fixed confidence interval CATs (Stahl & Lunz, 1993).

To study the influence of changes in targeted probability of correct response, we repeated our simulation on each set of item difficulty distribution and item pool size, with the change in item selection criteria to the ones that targeted at a probability of correct response of .6 and .7.

Item exposure

We monitored percent exposure rate of each item, which is defined as the percentage of computerized exams in which that item appears. In other words, item exposure rate is the proportion of examinees that saw the test item. We categorized exposure rates into five categories.

(1) *Overexposed items* were items that have been used more than the acceptable maximum percent exposure. Testing programs generally set the maximum percent exposure control at 15% to 25% (Davey & Nering, 2002; Muckle, Bergstrom, Becker, & Stahl, 2005; Parshall, Harmes, & Kromrey, 2000; Parshall, Spray, Kalohn, & Davey, 2002b). Thus, we used 20% as our acceptable maximum percent exposure.

(2) *Frequently used items* were items that have been used more than the optimal exposure rates, but did not exceed the 20% maximum limit.

(3) *Properly used items* were items that have their exposure rates within the optimal range of 5 – 15% (Way, 1998).

(4) *Underutilized items* were items that were used less than 5% (Parshall et al., 2002b).

(5) *Unused items* were items that had exposure rate of 0%.

We then examined the scatterplot of item difficulty by percent exposure of each item pool to see the ranges of item difficulty where overexposure and underutilization occurred. This provided diagnostic information that could help guide the direction for item development to replenish the pool (Parshall, 2002).

Finally, we compared average item exposure rates with the existing guideline on acceptable exposure rates (Way, 1998) to determine how well each administration did in controlling item exposure.

Data source

We generated all data using Microsoft (2002) Excel. These generated data included a group of 1,000 examinees (mean ability = 5.00 logits, SD = 1.00 logit), and nine sets of item pools, which were combinations of three pool sizes (400, 600, and 800 items) and three item difficulty distributions (mean difficulty = 5.00, 4.595, and 4.153 logits, SD = 1.00 logit). Each pool contained items of four content areas; each contributed 25% of items in the pool. Each content area had an item difficulty distribution similar to that of the pool that their items belonged to.

Results

Examinees

Examinee ability estimates that were generated by Microsoft (2002) Excel ranged from 2.01 to 8.27 logits with a mean of 5.00 logits and a standard deviation of 0.98.

Item pools

Microsoft (2002) Excel generated nine item pools that were combinations of three item difficulty distributions and three different pool sizes. Descriptive statistics of difficulty measures of items in these nine pools were shown in Table 1.

[INSERT TABLE 1 ABOUT HERE]

Frequency distributions of item percent exposure rates

We summarized frequency distributions of item percent exposure rates of item pools that contained 400, 600, and 800 items in Tables 2, 3, and 4, respectively. Item pool utilization was most effective when the difficulty distribution of items in the pool was appropriate for the targeted probability of correct response used for test administration (e.g., selecting items with expected probability of correct response of .5 from a pool that had mean item difficulty of 5.00), resulting in the highest proportions of properly used items. When selecting items to target the probability of correct response that mismatched with the item difficulty distribution in the pool, we observed more underutilization and overexposure. When the pool size is larger than appropriate, we saw higher proportions of unused and underutilized items. On the other hand, when the pool size is smaller than appropriate, we saw higher proportions of frequently used and overexposed items.

If a CAT program manager is planning for administering a CAT that targets .5 probability of correct response, the most effective item pool is the one that has 600 items with a normally distributed item difficulty distribution with a mean equals to mean examinees' ability. This pool yielded the highest proportion of properly used items (87%), with no overexposure

problem and only 1% of unused items. With resource limitation, one can also choose a 400-item pool with a normally distributed item difficulty distribution with a mean equals to mean examinees' ability. This pool still yielded a good proportion of properly used items (82%), with only 1% overexposed items and 1% unused items.

If a CAT program manager is planning for administering a CAT that targets .6 probability of correct response, the item pool that optimizes item utilization is the one with 600 items with a normally distributed item difficulty distribution with a mean 0.405 logits lower than mean examinees' ability. This pool yielded the highest proportion of properly used items (90%) with no overexposed items and only 1% of unused items. If one has a limited number of items, opting for a 400-item pool with a mean item difficulty 0.405 logits lower than mean examinees' ability is a good alternative. It still had up to 84% of properly used items, with only 2% of overexposed items.

In a setting that administers a CAT that targets .7 probability of correct response, a 600-item pool with normally distributed item difficulty measures with mean 0.847 logits lower than mean examinee's ability provides the best item utilization. Ninety-four percent of items in the pool were used properly, with no overexposed items and only 1% of unused items. However, opting for a 400-item pool in this situation is not a viable alternative. Even when item difficulty distribution was appropriate for the targeted probability of correct response, only 70% of items were properly used, and up to 6% of items were overexposed.

[INSERT TABLES 2, 3, AND 4 ABOUT HERE]

Scatterplots of item difficulty by percent exposure

Figures 1, 2, and 3 showed scatterplots of 400-item, 600-item, and 800-item pools, respectively. Scatterplots in each row represent item exposure of the same item pool when targeting three different probabilities of correct response. Scatterplots in each column represent item exposure resulting from using the same targeted probabilities of correct response in item selection algorithm that selects item from three item pools of different difficulty distributions. These scatterplots illustrated a consistent pattern of overexposure and underutilization.

If a pool had the distribution of item difficulty that was appropriate for the targeted probability of correct response (the three pools in the diagonal line of each figure (plots A, E, and I)), the peak exposure occurred around the mean item difficulty level 5.0, 4.6, and 4.1 logits in plots A, E, and I, respectively). This peak exposure occurred because there were many examinees who had their ability close to the mean examinees' ability (5.0 logits), requiring items in these ranges. There were also two secondary peaks about 1-2 logits higher and lower than the primary overexposure area. This was due to the use of a fixed SEM stopping rule, which administered more items to examinees with very high or very low ability level than those with average ability level.

If a pool had item difficulty distribution in a range higher than the range where item selection algorithm generally chose items (the three pools on the upper right corner of each figure (plots B, C, and F)), overexposure occurred with easy items, while underutilization occurred with difficult items. On the other hand, if a pool had item difficulty distribution in a range lower than the range where item selection algorithm generally chose items (the three pools on the lower left corner of each figure (plots D, G, and H)), overexposure occurred with difficult items, while underutilization occurred with easy items.

[INSERT FIGURES 1, 2, AND 3 ABOUT HERE]

Average item exposure rates

Table 5 summarized minimum, maximum, mean, and SD of item exposure rates. A general recommendation for average item exposure rate in a CAT for high stakes admission was 8-12% (Way, 1998). Thus, for a test that administered items with .5 or .6 probability of correct response, an appropriate pool size was 400 items. For a test that administered items with .7 probability of correct response, an appropriate pool size was 600 items. A pool size of 800 items seemed to be underutilized for all the three targeted probabilities of correct response. One noteworthy finding is that all but three simulations (out of 27 simulations) have unused items (as indicated by the exposure rates of 0). This demonstrated that there almost always were some items in the pool that were never administered to examinees.

[INSERT TABLE 5 ABOUT HERE]

Another interesting finding from this table is the dispersion of item exposure rates, as indicated by their standard deviations. We observed two trends in the dispersion of item exposure rates. First, a standard deviation was always lowest when we administered the test at the targeted probability of correct response that was appropriate for item difficulty distribution of the pool. As the mismatch between the targeted probability of correct response and the item difficulty distribution of the pool increased, a standard deviation went up. Second, a standard deviation depends on item pool size as well. A large item pool had relatively less dispersion of item exposure rates.

Discussion

This study illustrated how the changes in item characteristics (pool size and item difficulty distribution) affected item exposure rates in a variable-length fixed SEM CAT that was administered with the Rasch model. The findings supported the hypothesis that the average item exposure rate would be lowest when we administered the test from a large pool (an 800-item pool) that had item difficulty distribution matched properly with the targeted probability of correct response (selecting items that targeted .5 probability of correct response from a pool with a mean item difficulty of 5.00). The highest average item exposure rate was also found when we used a small item pool (a 400-item pool). However, contrary to our hypothesis, the highest average item exposure rate was found when we targeted the probability of correct response of .7, using the pool that has a mean item difficulty of 4.10, which was supposed to be appropriate for that item selection algorithm. This could be due to test information function. Administering items that target at .5 probability of correct response provide the maximal information for the estimation of examinee's ability, producing the shortest and most effective test for that examinee (Bergstrom & Lunz, 1999; Gershon, 2004). Thus, in opting for a test administration algorithm that selects items with .7 probability of correct response, we did not obtain the maximal information, resulting in a longer test. In our study, we found that when using the same item pool, a test administration algorithm that targets .5 probability of correct response requires about 7 – 8 items less than a test administration algorithm that targets .7 probability of correct response to achieve the same measurement precision of examinees' ability. By administering a longer test, a test that targets at .7 probability of correct response (using 50 – 52 items per test) exposed more items than a test that targets at .5 probability of correct response (using 42 – 45 items per test), even with appropriate item difficulty distribution in the pool.

However, focusing only on average item exposure rates did not provide a complete picture of item utilization. Although employing a CAT that targets .5 probability of correct response yielded the lowest average item exposure rates, regardless of the appropriateness of item difficulty distribution in item pool, it did not always optimize item utilization.

Administering a CAT to target .5 probability of correct response from a pool that has item difficulty distribution appropriate for .6 or .7 probability of correct response resulted in overexposure of difficult items and underutilization of easy items (as demonstrated in plots D and G in Figures 1, 2, and 3). Low average item exposure rates in these simulations occurred because the low exposure rates of underutilized items canceled out high exposure rates of overexposed items. Careful consideration of item exposure scatterplots and the standard deviations of item exposure rates revealed that the match between item difficulty distribution in item pool and the targeted probability of correct response is a critical factor that dictates item utilization. Using an item pool that has item difficulty distribution appropriate for the targeted probability of correct response resulted in the most optimal item utilization with low variation in item exposure rates of items in all ranges of difficulty (as demonstrated by having a low SD of item exposure rates and a flat item exposure scatterplot). Mismatch of item difficulty distribution in an item pool with the targeted probability of correct response resulted in underutilization of items in one range of difficulty, while overexposing items in another range of difficulty.

The findings from this study corresponds with the theoretical framework that underlies a computer adaptive test. Administering a test that targets at .5 probability of correct response produces the most effective test, achieving the desired precision level of examinees' ability estimates using the minimum number of test items. Developing an item pool of appropriate size with item difficulty distribution that matches with the targeted probability of correct response

can maximize the proportion of properly used items, while minimizing overexposed and underutilized items. A manager of a variable-length fixed SEM CAT program can use the findings from this study to guide the development of item pools that have proper characteristics in terms of pool size and difficulty distribution to optimize item utilization.

In an ideal situation, where a CAT can target at .5 probability of correct response, a manager of the testing program should plan for an item pool that contains 600 items with a mean difficulty of items about the same with the mean ability estimates of examinees. This pool functions very well, with no overexposed items. Targeting at .5 probability of correct response leads to an effective test that does not need many items to arrive at the required precision level of examinees' ability estimates. Thus, an 800-item pool is unnecessarily large, resulting in underutilization problem. However, a 400-item pool is also a good alternative for a CAT program that has a limited number of items in the inventory. However, a 400-item pool has a little problem with overexposing about 1% of items and using 12% of items too often. The scatterplot (Figure 1A) showed that overexposure problem occurred in the 0.10 logits randomization interval around the mean item difficulty. So, if a testing program wants to use a 400-item pool, we suggest adding more items that have difficulty measures close to the mean item difficulty to alleviate overexposure problem.

In a CAT that targets .6 probability of correct response, a manager of the testing program should plan for an item pool that contains 600 items with a mean difficulty of items about 0.405 logits lower than mean ability estimates of examinees. Although the discrepancy between the mean examinees' ability and the mean item difficulty leads to a slightly longer test to arrive at the pre-specified precision, an 800-item pool is still too large and has too many underutilized items. A 400-item pool is still a good alternative in this situation if there are not many items

available in an item vat. A CAT program manager can address the potential overexposure problem by adding more items that have difficulty measures close to the mean item difficulty (as suggested by the item exposure scatterplot, Figure 1E).

In a CAT that targets .7 probability of correct response, an appropriate item pool is the one with 600 items with a mean difficulty 0.847 logits lower than mean ability examinees' ability estimates. A 400-item pool is not a viable alternative in this situation.

An interesting finding from item exposure scatterplots is the distribution of overexposed and underutilized items when the item difficulty distribution of the pool was not appropriate for the targeted probability of correct response. If a pool had mean item difficulty measures higher than what was appropriate for the targeted probability of correct response, easy items in the pool had a major problem with overexposure, while difficulty items in the pool were underutilized. On the other hand, if a pool had mean item difficulty measures lower than what was appropriate for the targeted probability of correct response, difficult items in the pool were overexposed, while easy items in the pool were underutilized. This finding has a practical significance for a CAT program that is being developed based on items used in paper-and-pencil examination. Item pools developed for a paper-and-pencil examination generally do not have sufficient number of difficult items to target .5 probability of correct response (Bergstrom & Lunz, 1999). Thus, attempting to provide the most efficient CAT by targeting .5 probability of correct response while having insufficient amount of difficult items in the pool can lead to overexposure of these items. In this situation, the CAT program manager should focus the effort in generating difficult items to optimize item utilization.

The limitation of this study is its simulation nature. The simulation was conducted based on many assumptions, including the characteristics of examinees, the characteristics of item

pools, and the testing algorithm. Thus, the findings have certain limitations in context generalizability. The results can only be generalized to a CAT program that uses the same test administration algorithm in testing a group of examinees with the same ability distribution, using items from the pools with similar difficulty distributions. The simulation also assumed that examinees would respond to items in accordance with the item characteristic curve of the basic Rasch model for dichotomous responses. In actual testing environment, examinees may not respond strictly in accordance with the model, exhibiting various types of testing behaviors (e.g., guessing, fatigue, and test anxiety) causing deviation from the model expectation. Thus, in actual testing, one should consider the evaluation of the fit of the examinees' responses to the measurement model as well.

Conclusion

This study examined the effect of changes in two item pool characteristics, including item difficulty distribution and item pool size, on exposure rates of items in a variable-length fixed SEM CAT that was administered and analyzed with the basic Rasch model for dichotomous responses that targeted the probability of correct response of .5, .6, and .7. Our simulations revealed that average item exposure rates depended on the targeted probability of correct response. The test that targeted .5 probability of correct response provided the shortest test, resulting in the lowest average item exposure rates for any given item pool size. However, appropriate utilization of items in the pool depended on item pool size and the item difficulty distribution in the pool. We achieved the best utilization of items when we used an item pool of appropriate size with item difficulty distribution that was appropriate for the targeted probability of correct response. The optimal item pool for tests that target the probability of correct response

of .5, .6, and .7 for a group of examinees with a mean ability measure of 5 logits are item pools with 400-600 items that have mean item difficulty of 5 logits, 400-600 items that have mean item difficulty of 4.595 logits, and 600 items that have mean item difficulty of 4.193 logits, respectively.

References

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Becker, K. (2006). Promissor CAT simulator (Version 1.0). Chicago.
- Bergstrom, B. A., & Lunz, M. E. (1999). CAT for certification and licensure. In F. Dragow & J. B. Olson-Buchanan (Eds.), *Innovations in computerized assessment* (pp. 67-91). Mahwah, NJ: Lawrence Erlbaum Associates.
- Buyske, S. (2005). Optimal design in educational testing. In M. P. F. Berger & W. K. Wong (Eds.), *Applied optimal designs* (pp. 1-19). Hoboken, NJ: John Wiley & Sons.
- Davey, T., & Nering, M. (2002). Controlling item exposure and maintaining item security. In C. N. Mills, M. T. Potenza, J. J. Fremer & W. C. Ward (Eds.), *Computer-based testing: Building the foundation for future assessments* (pp. 165-191). Mahwah, NJ: Lawrence Erlbaum Associates.
- Davis, L., & Dodd, B. (2005). *Strategies for controlling item exposure in computerized adaptive testing with the partial credit model* (PEM Research Report No. 05-01). Austin, TX: Pearson Educational Measurement.
- Downing, S. M. (2004). Reliability: On the reproducibility of assessment data. *Medical Education*, 38, 1006-1012.
- Gershon, R. C. (2004). Computer adaptive testing. In E. V. Smith, Jr & R. M. Smith (Eds.), *Introduction to Rasch measurement: Theory, models, and applications* (pp. 601-629). Maple Grove, MN: JAM Press.

- Microsoft Corporation. (2002). Microsoft Excel (Version 10.65). Redmond, WA.
- Muckle, T., Bergstrom, B. A., Becker, K., & Stahl, J. A. (2005). *Impact of altering randomization intervals on precision of measurement and item exposure*. Paper presented at the 2005 Annual meeting of the American Educational Research Association, Montreal, Canada.
- Parshall, C. G. (2002). Item development and pretesting in a CBT environment. In C. N. Mills, M. T. Potenza, J. J. Fremer & W. C. Ward (Eds.), *Computer-based testing: Building the foundation for future assessments* (pp. 119-142). Mahwah, NJ: Lawrence Earlbaum Associates.
- Parshall, C. G., Harmes, J. C., & Kromrey, J. D. (2000). Item exposure control in computer-adaptive testing: The use of freezing to augment stratification. *Florida Journal of Educational Research*, 40(1), 28-52.
- Parshall, C. G., Spray, J. A., Kalohn, J. C., & Davey, T. (2002a). Computerized adaptive tests. In C. G. Parshall, J. A. Spray, J. C. Kalohn & T. Davey (Eds.), *Practical considerations in computer-based testing* (pp. 126-152). New York: Springer-Verlag.
- Parshall, C. G., Spray, J. A., Kalohn, J. C., & Davey, T. (2002b). Item pool evaluation and maintenance. In C. G. Parshall, J. A. Spray, J. C. Kalohn & T. Davey (Eds.), *Practical considerations in computer-based testing* (pp. 169-193). New York: Springer-Verlag.
- Rasch, G. (1960). *Probabilistic models for some intelligence and achievement tests*. Copenhagen: Danish Institute for Educational Research (Expanded edition, 1980. Chicago: University of Chicago Press).

- Stahl, J. A., & Lunz, M. E. (1993, April). *Assessing the extent of overlap of items among computerized adaptive test*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Atlanta, GA.
- Stocking, M. L. (1994). *Three practical issues for modern adaptive testing item pools* (ETS Research Report No. 94-5): Educational Testing Service, Princeton, NJ.
- Thissen, D. (2000). Reliability and measurement precision. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (2nd ed., pp. 159-184). Mahwah, NJ: Lawrence Erlbaum Associates.
- Thissen, D., & Mislevy, R. J. (2000). Testing algorithms. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (2nd ed., pp. 101-134). Mahwah, NJ: Lawrence Erlbaum Associates.
- Wainer, H. (2000). Introduction and history. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (2nd ed., pp. 1-21). Mahwah, NJ: Lawrence Erlbaum Associates.
- Way, W. D. (1998). Protecting the integrity of computerized testing item pools. *Educational Measurement: Issues and Practice*, 17(4), 17-27.
- Way, W. D., & Steffen, M. (1997, April). *Strategies for managing item pools to maximize item security*. Paper presented at the Annual meeting of the National Council on Measurement in Education, San Diego.
- Way, W. D., Steffen, M., & Anderson, G. S. (2002). Developing, maintaining, and renewing the item inventory to support CBT. In C. N. Mills, M. T. Potenza, J. J. Fremer & W. C. Ward (Eds.), *Computer-based testing: Building the foundation for future assessments* (pp. 143-164). Mahwah, NJ: Lawrence Earlbaum Associates.

Table 1

Descriptive Statistics of Difficulty Measures of Items from the Nine Simulated Pools

Pool	N	Minimum	Maximum	Mean	SD
1	400	2.29	8.10	5.02	1.00
2	400	1.90	7.67	4.62	0.98
3	400	1.50	6.45	4.10	0.93
4	600	1.57	7.81	5.02	1.02
5	600	1.57	7.80	4.61	1.03
6	600	0.35	6.58	4.07	1.01
7	800	2.33	8.46	5.01	0.96
8	800	1.55	7.51	4.65	0.99
9	800	0.82	7.40	4.14	1.00

Table 5

Descriptive Statistics of Percent Exposure Rates of Items from All 27 simulations

Simulation	Pool size	Mean difficulty	Target probability	Target			
				Minimum	Maximum	Mean	SD
1	400	5.02	.5	0	23.80	10.74	3.76
2	400	5.02	.6	0	31.70	11.13	5.65
3	400	5.02	.7	0	54.40	12.57	11.92
4	400	4.62	.5	0.10	31.20	10.78	5.59
5	400	4.62	.6	0.10	25.10	11.19	3.74
6	400	4.62	.7	0	34.70	12.69	7.32
7	400	4.10	.5	0	56.30	11.13	12.60
8	400	4.10	.6	0.10	37.40	11.50	7.23
9	400	4.10	.7	0	25.40	12.94	4.20
10	600	5.02	.5	0	14.30	7.12	2.24
11	600	5.02	.6	0	24.40	7.41	4.08
12	600	5.02	.7	0	43.50	8.44	8.35
13	600	4.61	.5	0	23.80	7.14	3.70
14	600	4.61	.6	0	14.80	7.42	2.26
15	600	4.61	.7	0	33.70	8.41	4.87
16	600	4.07	.5	0	42.50	7.27	7.75
17	600	4.07	.6	0	27.40	7.58	4.49
18	600	4.07	.7	0	17.40	8.58	2.46
19	800	5.01	.5	0	13.60	5.33	1.67
20	800	5.01	.6	0	18.70	5.56	2.92
21	800	5.01	.7	0	38.60	6.30	6.72
22	800	4.65	.5	0	19.00	5.34	2.57
23	800	4.65	.6	0	11.1	5.56	1.66
24	800	4.65	.7	0	27.40	6.37	4.36
25	800	4.14	.5	0	28.50	5.38	5.37
26	800	4.14	.6	0	17.40	5.59	2.80
27	800	4.14	.7	0	13.60	6.42	1.87

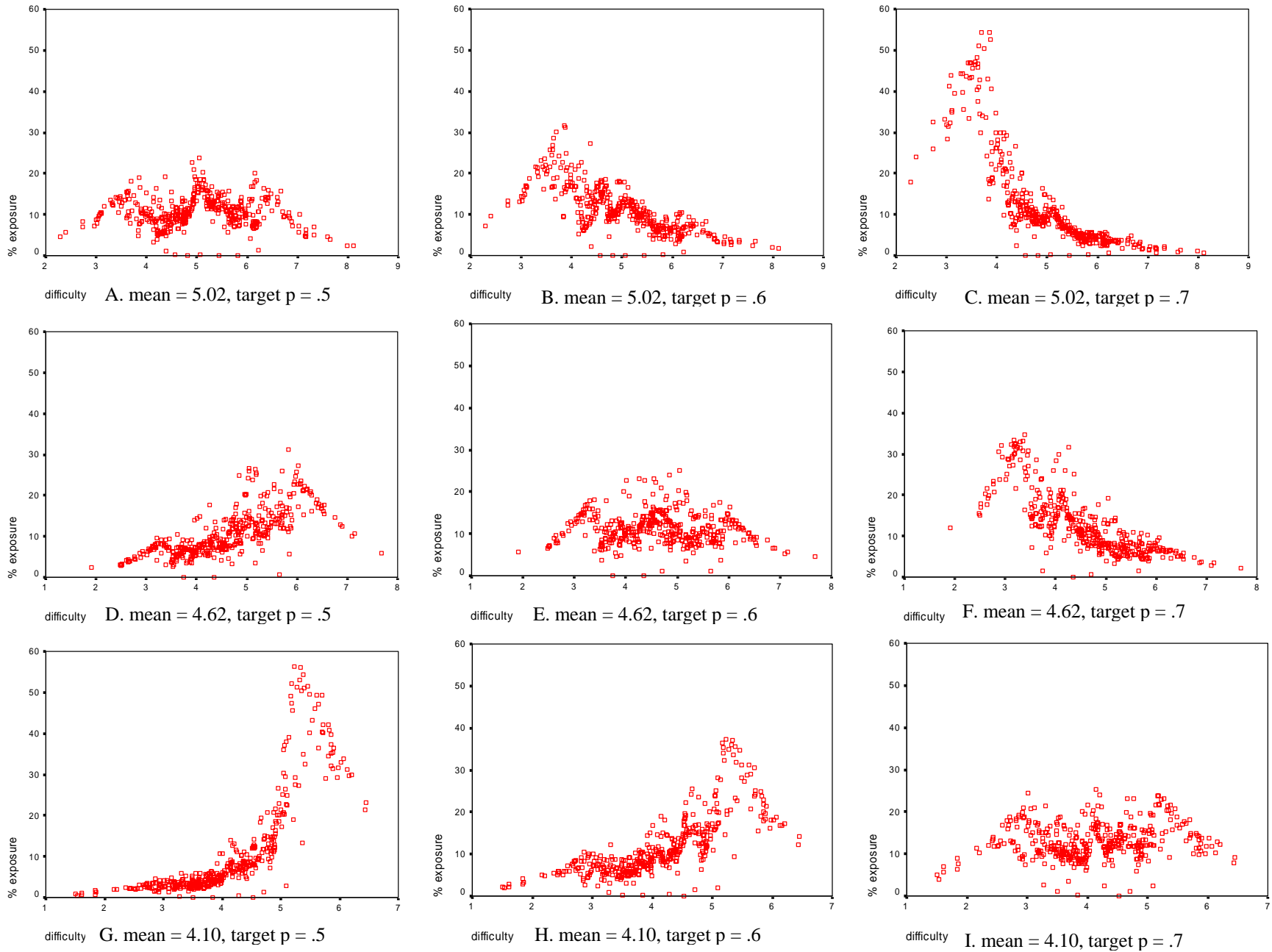


Figure 1 Item exposure scatterplots of 400-item pools.

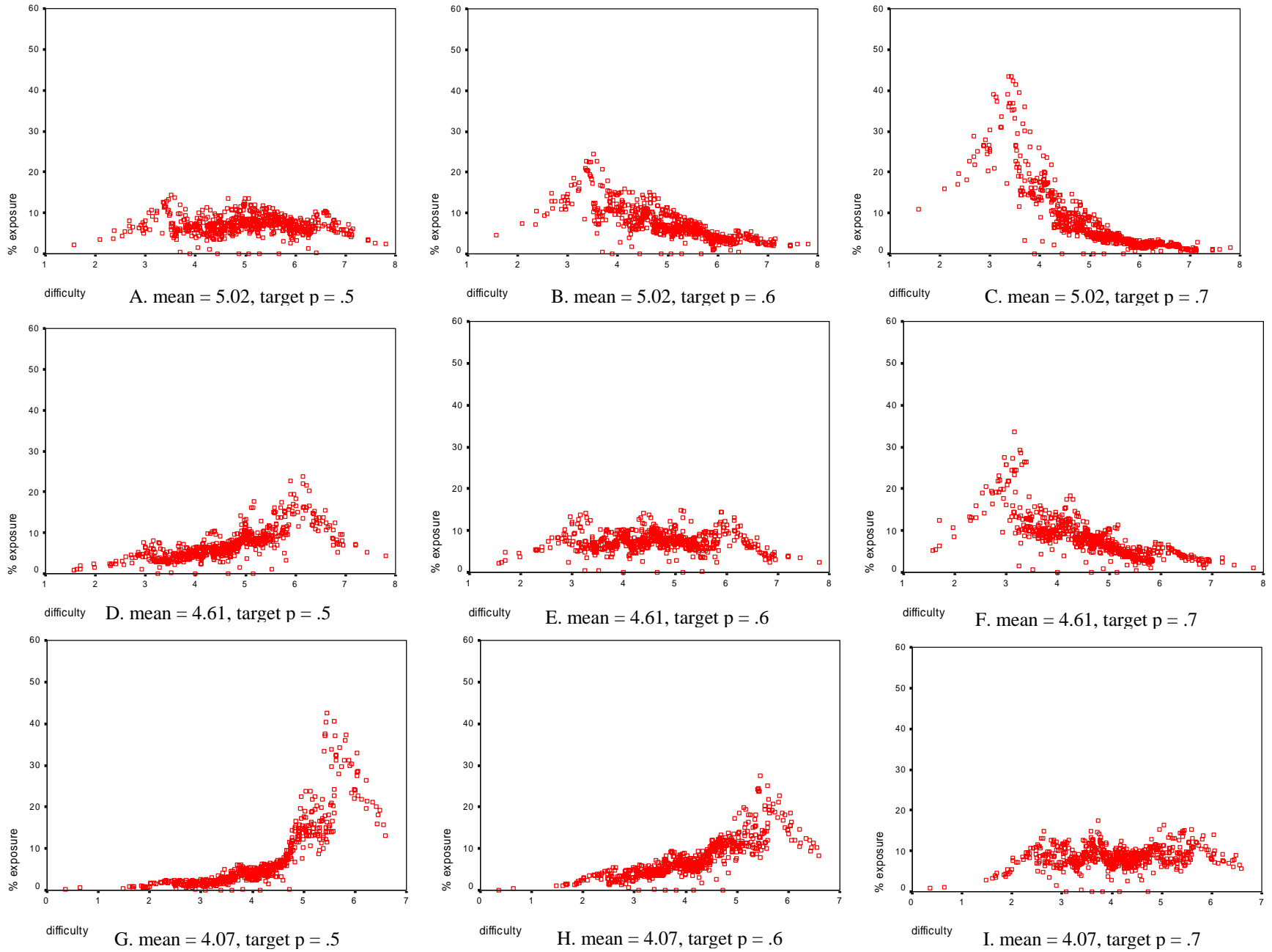


Figure 2 Item exposure scatterplots of 600-item pools.

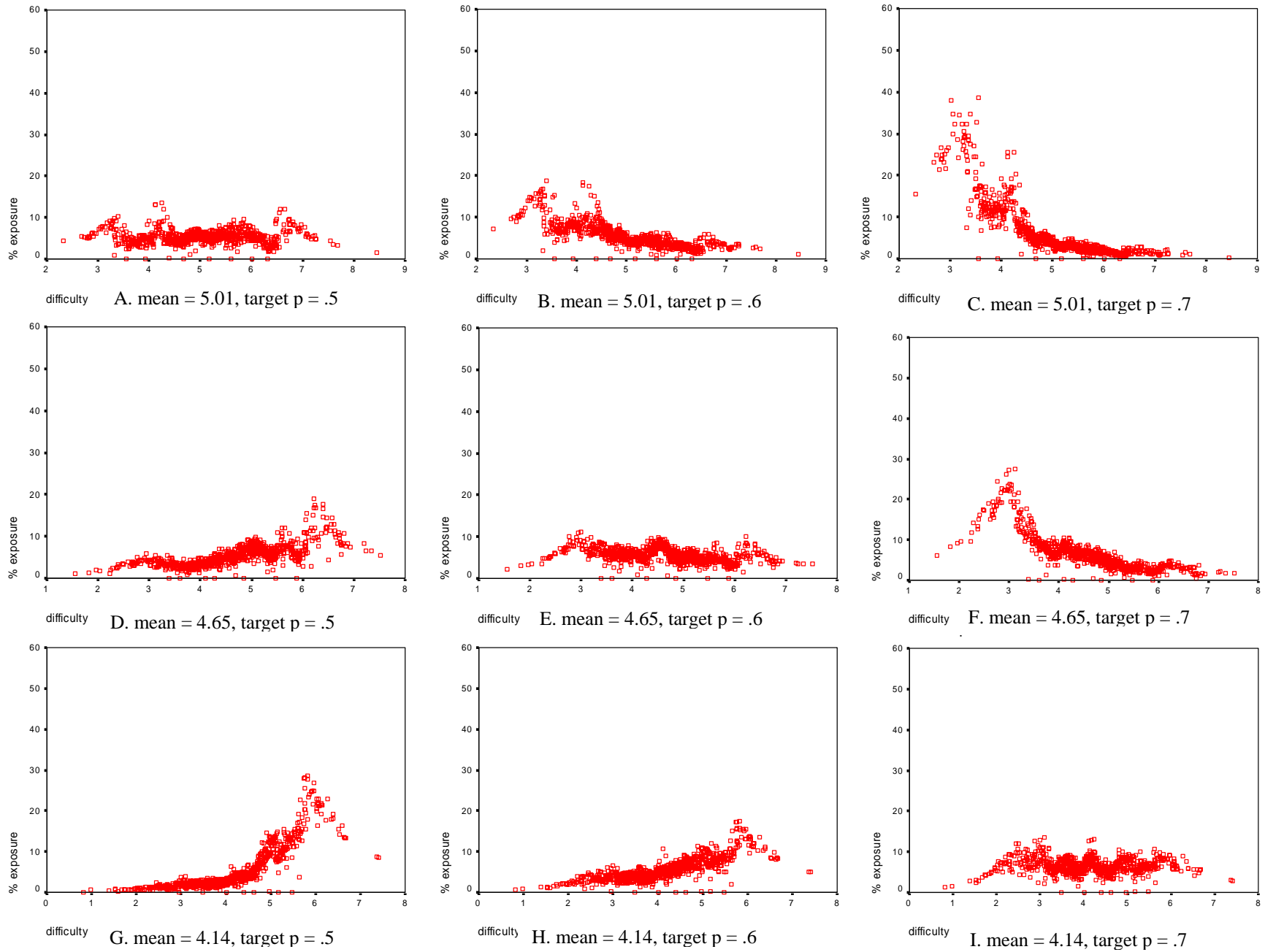


Figure 3 Item exposure scatterplots of 800-item pools.