Psychological Test and Assessment Modeling, Volume 52, 2010 (2), 127-141

Designing item pools to optimize the functioning of a computerized adaptive test

Mark D. Reckase[1]

Abstract

Computerized adaptive testing (CAT) is a testing procedure that can result in improved precision for a specified test length or reduced test length with no loss of precision. However, these attractive psychometric features of CATs are only achieved if appropriate test items are available for administration. This set of test items is commonly called an "item pool." This paper discusses the optimal characteristics for an item pool that will lead to the desired properties for a CAT. Then, a procedure is described for designing the statistical characteristics of the item parameters for an optimal item pool within an item response theory framework. Because true optimality is impractical, methods for achieving practical approximations to optimality are described. The results of this approach are shown for an operational testing program including comparisons to the results from the item pool currently used in that testing program.

Key words: computerized adaptive testing, item pool design, item response theory, conditional standard error of measurement, item pool size

[1] *Correspondence concerning this article should be addressed to:* Mark D. Reckase, PhD, Michigan State University, 461 Erickson Hall, East Lansing, MI 48824, USA; email: reckase@msu.edu

Computerized adaptive testing (CAT) has been an operational option for measuring the level of an examinee on a construct since the 1980s. For example, during that time the MicroCAT testing system was marketed (Assessment Systems Corporation, 1984) and programs such as the Computerized Adaptive Screening Test (see Sands, Gade, and Knapp (1997) for a summary of development) were implemented. Since that time there have been numerous applications of computerized adaptive testing. The books by Drasgow and Olson-Buchanan (1999) and van der Linden and Glas (2000) provide summaries of some of the developments. With over 20 years of practical implementation, CAT is a mature technology.

The basic component parts of a CAT are so well known (see, for example, Wainer, Dorans, Eignor, Flaugher, Green, Mislevy, Steinberg, & Thissen (2000)) that only a brief summary will be given here. The major components of a CAT include: an item response theory model[2], a calibrated item pool, an item selection algorithm, a statistical method for locating the examinee on the construct, and a rule for stopping the test. Recent CAT procedures that are used for high-stakes testing may also include content balancing (e.g., Cheng & Chang, 2009) and special security provisions such as exposure control (e.g., Sympson & Hetter, 1985).

While there is a sizable literature on CAT, one component of the CAT is notable because of the limited guidance given to it. This component is the item pool. For example, in the classic work on CAT by Wainer, et al. (2000), Chapter 3 by Ronald Flaugher on item pools indicates that "To realize many of the measurement advantages of adaptive testing, the item pool from which items are selected must contain high-quality items for many different levels of proficiency" (p. 38). But the only guidance given for developing the item pool that meets this criterion is to "Create sufficient numbers of items in each content category, based on the test specifications established previously" (p. 39). Unfortunately, there is nothing in the book that indicates how to produce test specifications. There is an example of a CAT in Chapter 3 that uses an item pool of 200 test items. It is difficult to use the limited information in that chapter to design an item pool.

Recent work (see Belov and Armstrong (2009) and van der Linden, Ariel and Veldkamp (2006) for example) do address item pool design, but both do so assuming that a large set of items called a "master pool" already exists and the task is to select operational pools from it. Another approach by Veldkamp and van der Linden (2000) uses the shadow test approach to designing an item pool, but that approach used the characteristics of an existing item pool as a starting point.

Because there is limited guidance on the required characteristics of the item pool for a CAT, ad hoc procedures are often used to develop an item pool. For example, when developing a CAT to measure the effect of headaches on people, Ware, Bjorner, and Kosinski (2000) collected together the items from four existing measures and supplemented that with a few items written to enlarge the item pool. The result was a 53 item pool that focused on the lower end of the construct.

---

[2] Non-IRT CATs can also be developed, such as those based on the sequential probability test, but this article is limited to CAT procedures that use IRT.

The purpose of this article is to provide some preliminary guidance and methodology for the design of the characteristics of item pools for CATs. A conceptual framework is first provided for item pool design and then that framework is used to design the item pools for CATs that have specific measurement goals.

An optimal item pool

Before providing details about a process for designing an item pool for a CAT procedure, it is useful to consider the best possible item pool. The definition for the best possible, or optimal, item pool that is used here is that whenever the CAT item selection algorithm is searching for a test item to administer, exactly the item that is desired is available in the item pool. If the desired item is always available for every item selection, than the item pool can be considered to be optimal. A simple example may clarify this issue.

Suppose that the CAT procedure has the following components: it is based on the Rasch model for dichotomous items; it uses maximum information item selection; it uses maximum likelihood estimation for the examinee location on the construct; it administers a fixed number of test items; and there are no security procedures such as exposure control. For this type of CAT procedure, an optimal item is one that has its Rasch difficulty parameter (*b*-parameter), exactly equal to the current estimate on the construct, θ. This relationship between the *b*-parameter and θ yields the maximum possible information from a test item because, for the Rasch model

$$P(u_{ij}=1\,|\,\theta_j,b_i)=\frac{e^{(\theta_j-b_i)}}{1+e^{(\theta_j-b_i)}}\,, \tag{1}$$

where u_{ij} is the response to Item *i* by Person *j*, the information function is given by

$$I_i(\theta)=P(u_{ij}=1\,|\,\theta_j,b_i)P(u_{ij}=0\,|\,\theta_j,b_i)\,. \tag{2}$$

This function is maximum when the probabilities of correct and incorrect responses are equal. This occurs when the probability of each is 0.5 and that occurs when the item difficulty is equal to the estimated value of θ.

An optimal item pool for this CAT procedure is one that has an item in the pool that has a *b*-parameter exactly equal to the current θ estimate for every item selection. Suppose that all examinees get the same first item, then the θ-estimate is updated with two possible values, one for correct and one for incorrect responses. For the optimal two item CAT, three items would be needed – the first item and the two possible second items. After the second item is answered, four items would be needed to match the possible estimates of θ and the total pool for a three item CAT would be 1 + 2 + 4 = 7. In general, the size of the item pool with items matching every θ-estimate is $2^n - 1$ where *n* is the number of items administered to an examinee. For a CAT test of length 20, the optimal item pool would need to contain 1,048,575 items, a number of items that is clearly impractical for any real test. However, if such an item pool existed, it would yield a CAT

for this particular implementation of the procedure that would give the best possible result.

An optimal item pool when the CAT is based on a two-parameter logistic model,

$$P(u_{ij}=1\,|\,\theta_j,a_i,b_i)=\frac{e^{a_i(\theta_j-b_i)}}{1+e^{a_i(\theta_j-b_i)}}, \tag{3}$$

where a_i is the item discrimination parameter, has interesting properties because the best item to administer when the current estimate is θ is an item with an infinite *a*-parameter and a *b*-parameter equal to θ. This item has infinite information at that θ-value because the information for this type of item is given by

$$I_i(\theta)=a_i^2P(u_{ij}=1\,|\,\theta_j,a_i,b_i)P(u_{ij}=0\,|\,\theta_j,a_i,b_i)\,. \tag{4}$$

Items of this type form a Guttman scale (Guttman, 1944). The same number of items would be needed in the item pool as for the Rasch model, but the accuracy of the estimates of location for the examinees would be much greater (i.e., standard error of approximately .000005). However, developing an item pool with these characteristics is even more impossible than creating one with more than 1,000,000 items for use with the Rasch model because it is impossible to produce items with infinite discrimination, so the concept of an optimal item pool is interesting, but not very useful.

p-optimal and *r*-optimal item pools

To make the concept of an optimal item pool useful for the design of practical item pools, it is important to consider the differences in characteristics of test items that are similar, but not exactly the same. That is, how much difference in information about an examinee's location on the construct is obtained from an item that exactly matches the current estimate of θ to one that has a *b*-parameter that differs from the optimal value by some small value, Δ.? One way to consider this issue is to determine the difference in the amount of information provided by a test item about θ when there is an exact match between θ and *b* and when these values are slightly different.

Figure 1 shows the information function for a test item whose functioning is accurately described by the Rasch model. The horizontal scale is $\theta - b$ so that the results generalize to all values of θ. When $\theta - b = 0$, that is, $\theta = b$, the information takes on a maximum value of 0.25. However, it might be acceptable from an item pool development perspective if the item that is available for selection has information that is within 90% of the maximum possible. The horizontal line in Figure 1 shows this level of information. If the selected item is within about .55 of exactly matching the θ-value, it will meet this criterion. If an item pool meets the criterion of always having items available for selection that are 90% or more of the maximum possible information, than the item pool can be said to be .9-optimal. This way of describing the design of an item pool is called *p*-optimal for proportion of maximum optimality.

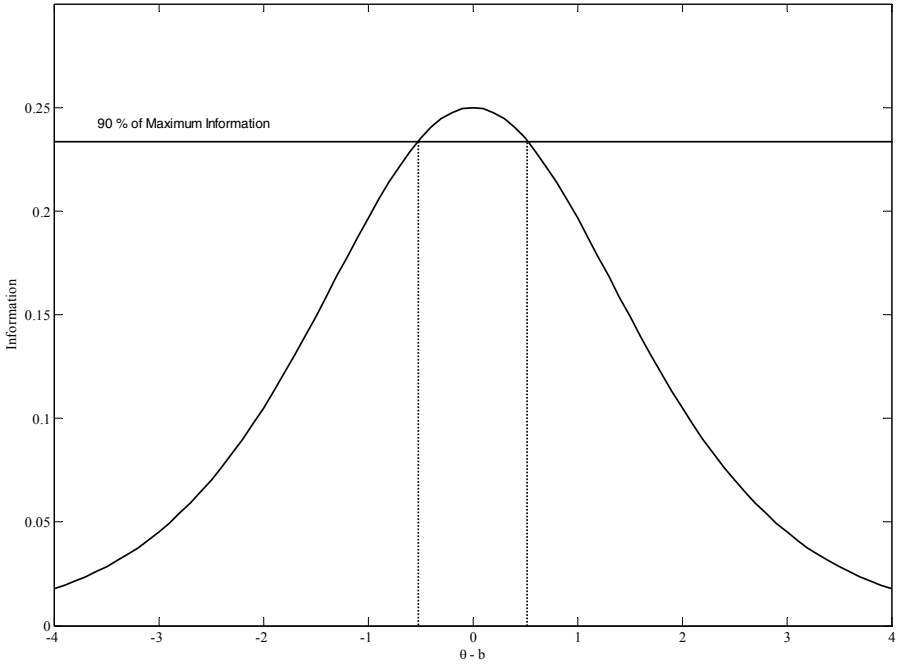

Figure 1:
Information Function for a Test Item Fit by the Rasch Model

If the value of *p* is closer to 1.0, the width of the range in which items will be acceptable is narrower. For .95-optimal using the Rasch model, the selected item should be within about .35 of the *b*-parameter that exactly matches the current estimate of θ. This means that there is a range on the *b*-parameter scale of .7 with the current estimate of θ in the middle. For a .8-optimal pool, the item selected would have to be within about .85 θ-units of the desired item, or the full range of acceptable items is 1.7. An alternative way of describing how close an item pool is to optimal is to specify the how close on the θ-scale the items have to be to the desired item in terms of the *b*-parameters. In the case of 90% of maximum information described above, the *b*-parameter must be within .55 of the θ estimate. This specifies a range and the criterion can be called range optimality and the goal is to be *r*-optimal (e.g. .55-optimal). Note that .55-optimal means that there is a range of width 1.1 on the *b*-parameter scale. In some cases it is more convenient to use *p*-optimality and in other cases *r*-optimality. In this paper, the specification of a value of *p* for *p*-optimality is used to determine the value of *r* for *r*-optimality. In general, *r*-optimality is easier to use when designing item pools for an operational CAT procedure.

Designing an item pool

The process for designing an item pool for a CAT using optimality criteria is fairly straight forward when the Rasch model is the basis for the CAT algorithm. Suppose, for example, that a CAT uses the Rasch model with maximum likelihood estimation of the construct, maximum information selection of the item, and a fixed test length of 20 items per examinee. For this type of CAT, maximum likelihood estimates do not have finite values until both correct and incorrect responses have been obtained from the examinee. Until that occurs, construct estimates are computed by adding .7 to the previous estimate after a correct response and subtracting .7 from the previous estimate after an incorrect response. All examinees begin this CAT with an estimate of level on the construct of 0.0. Given this CAT design, what is a practical item pool design that will support the proper functioning of the CAT?

Suppose that a single person who has a true θ of -1 is administered this CAT test with an infinite item pool that contains every possible *b* parameter. The first item administered has a *b*-parameter of 0 and after each update to the estimate of θ, an item with *b*-parameter that maximizes the information is available for administration. Table 1 shows the estimate of θ and the *b*-parameter after each response to the test. The set of *b*-parameters gives the optimal item set for this examinee for this CAT design.

Table 1:
Item Responses, θ-estimates, and *b*-parameters for a 20-item Rasch Model Based CAT

Item Number	*b*-parameter/θ-estimate	Item Score
1	0	0
2	-0.70	1
3	-0.35	0
4	-1.06	0
5	-1.66	1
6	-1.19	1
7	-0.83	1
8	-0.52	0
9	-0.79	0
10	-1.02	0
11	-1.23	1
12	-1.04	0
13	-1.22	0
14	-1.38	1
15	-1.23	1
16	-1.09	1
17	-0.96	0
18	-1.08	1
19	-0.97	0
20	-1.08	1

If the goal is to produce a *p*-optimal pool with criterion .95 of maximum information, the range for the corresponding *r*-optimal definition (.35-optimal) is .7. This means that instead of needing items at every θ estimate, the *b*-parameter scale can be divided into intervals of width .7 and the number of items needed in each interval can be tallied. In this case, intervals are marked off with the first one centered on the 0 point on the scale (-.35 to .35) and stepped off in either direction. These ranges on the scale are called "item bins" and the number of items needed in each bin is tallied. For the items selected for the CAT administration shown in Table 1, the distribution of items over bins with *p*-optimality of .95 is given in Figure 2.

This figure shows that one item would be needed in the range from -.35 to .35, but nine items would be needed in the range from -1.05 to -.35 and ten items are needed in the range from -1.75 to -1.05. This is reasonable because the true θ is -1 and after getting the general range of the person parameter, the adaptive test administers items in this range. The figure shows the *r*-optimal set of items for this examinee given the criterion for optimality. To measure this person well, the item pool should contain the number of items specified in the ranges given.

Now, suppose that a second examinee with a true θ of .5 is administered the CAT. Figure 3 shows the distributions of items over bins for this examinee. This set of items is somewhat higher on the *b*-parameter scale than the set in Figure 2 because of the higher θ-level of the examinee, but there is also a lot of overlap in the sets of items. If the items

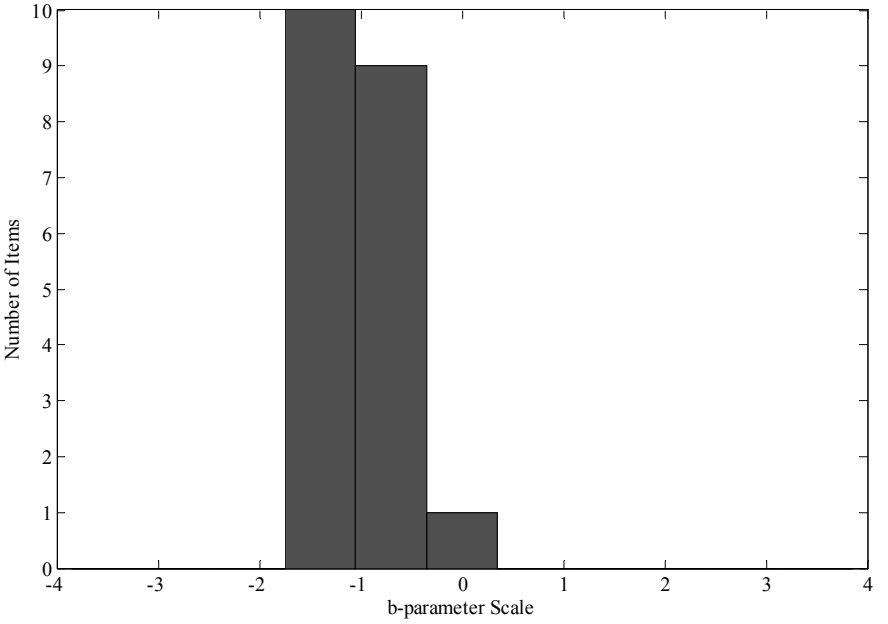

Figure 2:
Distribution of Items over Bins of Width .7 for True θ = -1

selected for the first examinee can also be used for the second examinee, then to have the *r*-optimal sets for both examinees requires only 33 items instead of the 40 that would be required if the items were not reused. The combined set is the union of the two sets of items. The union of the two sets is shown in Figure 4.

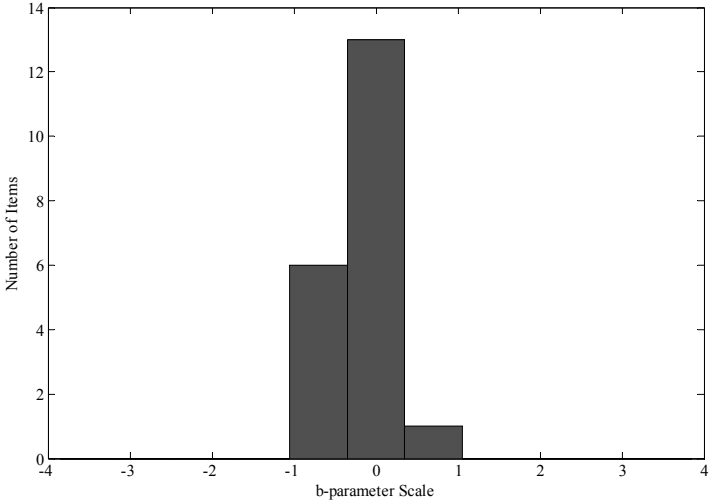

Figure 3:
Distributions over Bins of Size .7 for True θ = .5

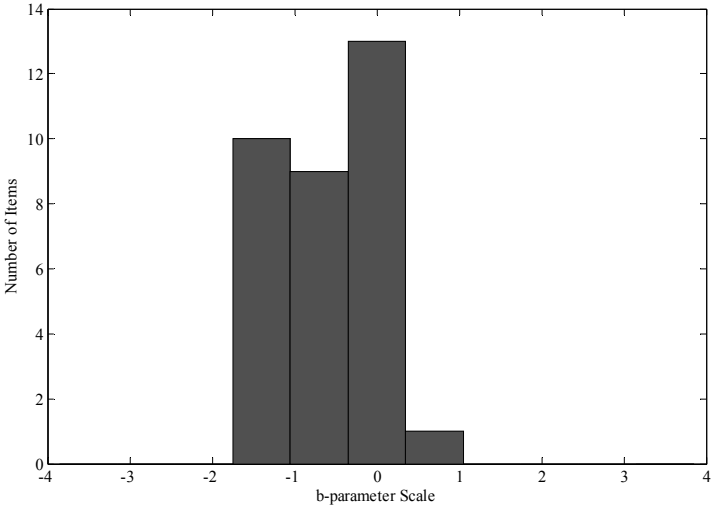

Figure 4:
Distribution over Bins of Size .7 for the Items for Two Examinees

The process of determining the *r*-optimal set of items for each examinee and then finding the minimum common set by taking the union of the individual item sets provides a means for determining the size and desired characteristics of the item pool needed for a target population of examinees. The technique is to randomly sample an examinee from the target population, determine the optimal item set for that examinee and allocate the items to bins according to the optimality criterion, then sample another examinee and do the same. After each examinee is sampled and the items are assigned to bins, the union of the item sets is formed. Then the next set of items is selected and the union is formed with the previous set defined by the previous union. The process continues until the numbers of items in the union reaches an asymptote. The distribution of item parameters in the union of item sets when the asymptote is reached is the *r*-optimal item pool for the target examinee distribution and the CAT design.

The CAT design used for the examples in Figures 2, 3, and 4 can be used to demonstrate the process of determining item pool size and the *r*-optimal item pool distribution. Suppose that the target population of examinees is normally distributed with mean 0 and standard deviation of 1 and the CAT uses the Rasch model with maximum information item selection, maximum likelihood estimation, and a fixed length of 20 items for each examinee's test. The number of items contained in the union of the items for examinee *n* and the items for the previous *n*-1 examinees is shown in Figure 5 for one implementation of the item pool design process. This graph shows that the item pool size for the

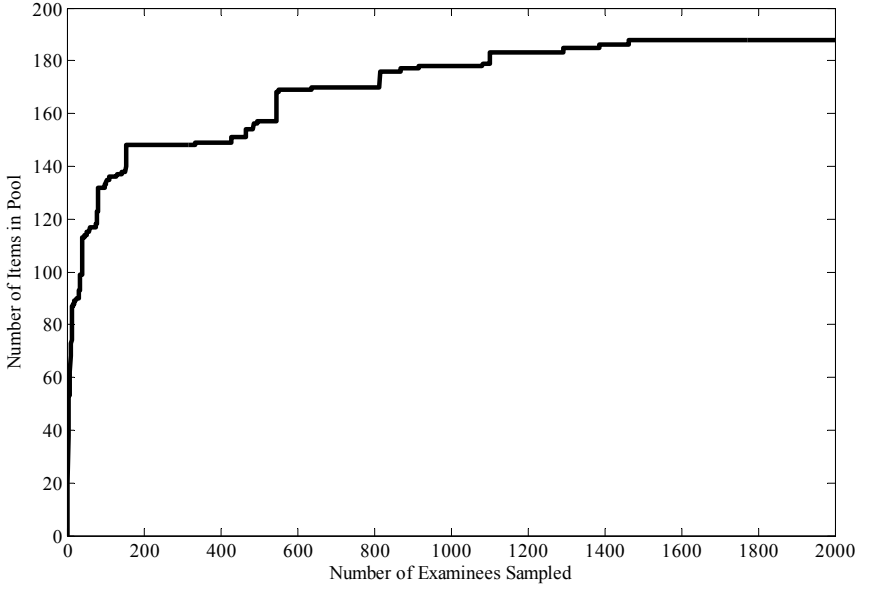

Figure 5:
Number of Items Needed in the *r*-optimal Item Pool for *r* = .7 and a Rasch-based CAT of Fixed Length of 20 Items

r-optimal pool with bin size of .7 asymptotes to 188 items. The distribution of *b*-parameters for the items in that pool is given in Figure 6.

The distribution of *b*-parameters is flatter than a normal distribution and it is spread over a range from -5 to 4.5. The distribution is not smooth because the selection process is a random simulation. Replicating the simulation gives slightly different results. The results can be stabilized by averaging several replications. Figure 7 shows the *b*-parameter distribution from averaging ten replications of the process. Note that the symmetry of the distribution is more evident than it is in the distribution from a single simulation shown in Figure 6.

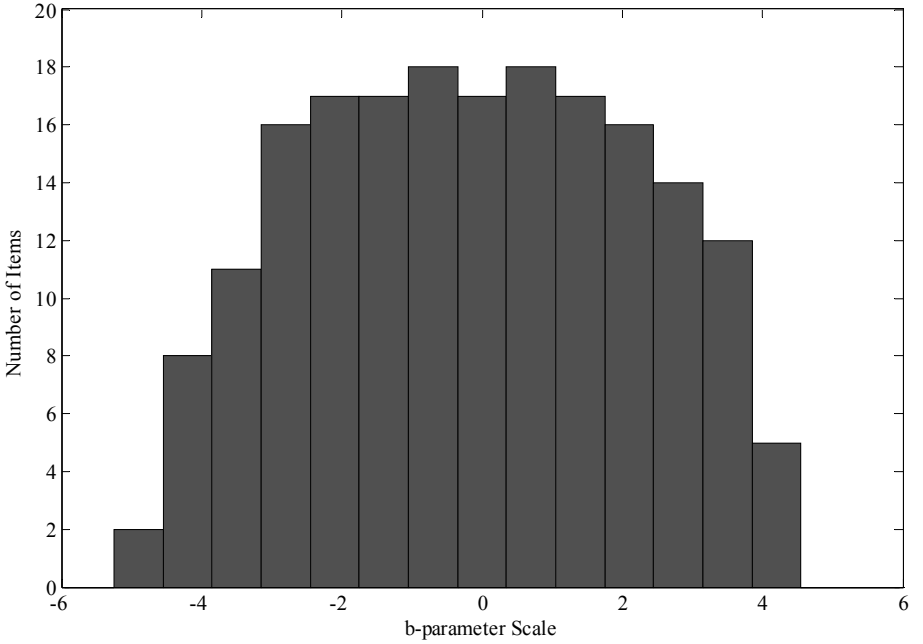

Figure 6:
Distribution of *b*-parameters for *r*-optimal Pool

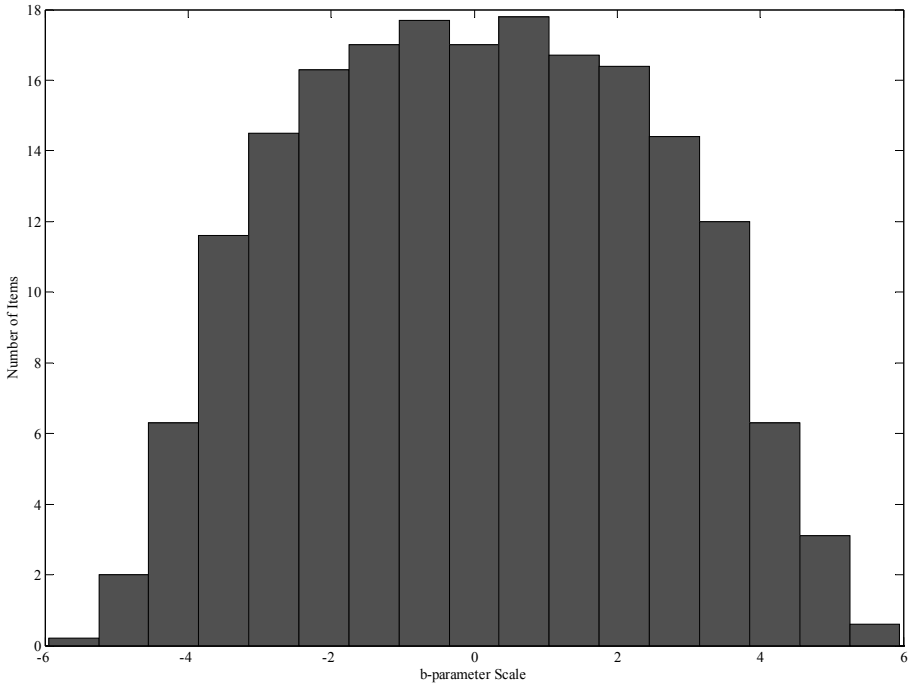

Figure 7:
Average Distribution of *b*-parameters for Ten Replications

Application to a Certification/Licensure Examination

To show the practicality of the procedures described in the previous section, they have been applied to the design of an item pool for a certification/licensure examination that is administered using CAT methodology. The IRT model that is used for this examination is the Rasch model. The item selection algorithm is maximum information, the same as used in the demonstration in the previous section. The proficiency estimation procedure is somewhat more complex. Until both a correct and an incorrect response is present in the response string for an examinee, Owen's Bayesian estimation procedure (Owen, 1975) is used to update the proficiency estimate. After both correct and incorrect responses are present in the response string, maximum likelihood estimation is used.

The testing algorithm also uses content balancing and exposure control. Initially, the content balancing cycles through the eight content areas on the test in a preset pattern. In later parts of the test, the content balancing procedure selects the next item to administer by determining the largest difference between the desired proportion of each content area and the observed proportion. The exposure control method is to select the 15 items that have *b*-parameters closest to the estimated θ and then randomly select one of those for

administration. The practical implications of this exposure control procedure for item pool design is that the *r*-optimal pool will need to have at least 15 items in the bin that contains the θ estimate so that an appropriate item will be selected.

The CAT for this application is variable length. All examinees must respond to at least 60 test items. After 60 items, the successive estimates are checked to determine if the confidence interval around the estimate contains the passing score for the test. If it does not, the test stops. If the confidence interval includes the passing score, another test item is selected and administered. The maximum test length is 250 items. If a pass/fail decision was not made after the 250th item, then if the estimate was above the passing score, a pass was recorded and otherwise the examinee failed the test.

The process for designing the *r*-optimal pool for this case was the same as described in the previous section except that the items were tallied into bins for each content area and the optimal item set contained 15 items at each θ to account for the exposure control procedure. The distribution of *b*-parameters for items for one of the content areas is shown in Figure 8. This distribution uses a bin width of .8. In this figure there is a central peak that is at the passing θ-level for the test. Many more items are needed in that bin because examinees near the passing score take the tests that may reach 250 items. In other regions of the scale, a minimum of 15 items are needed to support the exposure control procedure. For this content area 233 items were needed for the *r*-optimal pool.

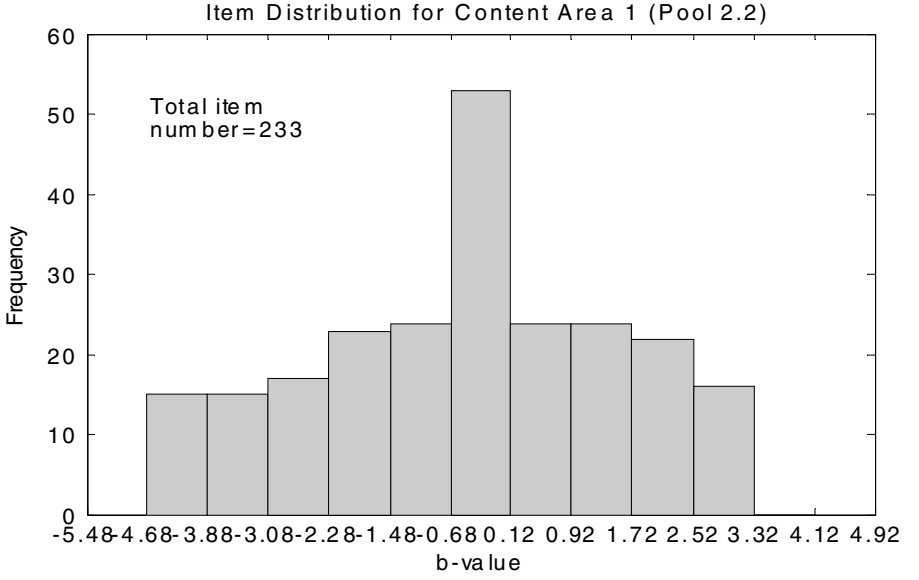

Figure 8:
Item Pool for One Content Area of the Licensure/Certification Test

The total size of the item pool over all of the content areas was roughly eight times the size of the single content area, 1602 items. The shape of the distribution of the items was proportional to that in Figure 8 so it is not shown here. The size of the pool may seem large, but it is much smaller than the operational item pool that was used for this testing program. The operational item pool size was 2000 items.

Because there was a substantial difference in the size of the item pool designed through this process and the actual operational pool, there was a concern that the CAT using the *r*-optimal pool might not function as well as the existing operational pool. To check this, the CAT was simulated using the two different *r*-optimal pools (they differed in bin size[3]) and the actual item pool and the conditional standard error of measurement was computed at equally spaced points along the θ scale. A graph of the conditional standard errors is given in Figure 9. This graph has an interesting shape. There is a dip in the middle of the conditional standard error curves near the passing score where the test length can sometimes be 250 items. In that region, the different item pools resulted in equal precision of estimates. However, as the distance from the center of the scale increased, the conditional standard errors for the *r*-optimal pools were fairly constant while those for the operational item pool increased. The operational item pool had more items

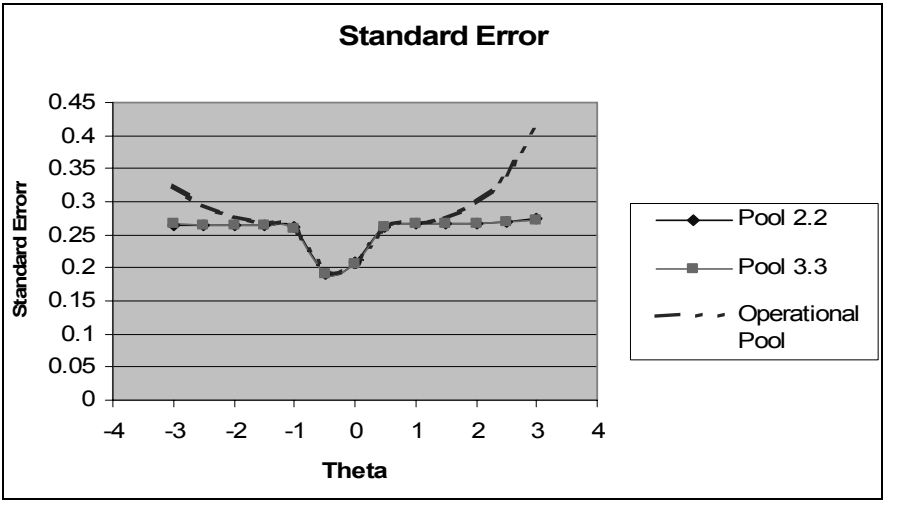

Figure 9:
Comparison of Conditional Standard Errors for *r*-optimal and Operational Item Pool

[3] The two item pools are labeled 2.2 and 3.3. These two pools were selected from several that were developed and the numbering refers to variations on the bin-width that was used for accumulating numbers of items.

than necessary near the passing score and had less spread in difficulty than the *r*-optimal pool. The design of the *r*-optimal pool showed that the operational item pool can be reduced in size without sacrificing precision of measurement.

Discussion

The purpose of this article is to describe a procedure for designing item pools for computerized adaptive tests and show that the use of the procedure results in efficient designs that do not sacrifice the advantages of the CAT. The procedure requires that it be possible to simulate the full method used for the CAT including content balancing, exposure control, and the item selection method or methods. Also, it is necessary to know something about the target examinee population. The simulations used real application of the procedures used the observed performance distribution of examinees as the target distribution for the simulation of the CAT. The bin size also makes a difference in the procedure. Larger bins result in less accurate targeting of the test items to the examinees estimated θ. Results from preliminary studies indicate that bin sizes in the range from .6 to .8 result in very small differences. Much bigger or smaller bin sizes result in important differences in the results.

The differences in the shapes of the distributions of *b*-parameters between the example and the real test highlight the effect of different CAT designs on the requirements for the item pool. The real example was focused on decision making at a specific point and it used a variable length test. The initial example had a fixed length test with a goal of estimating all examinees equally well. These design features of the CAT result in different requirements for the item pool. The practical implications of this result are that there is no correct answer to the question "How big should a CAT item pool be?" The examples here used 288 and 1602 items. The size and distribution of difficulty of the item pool is dependent on the design of the CAT and the distribution of performance of the target examinee population.

While this methodology has been applied in a number of practical situations and it works well, there are many challenges ahead to general implementation. An important one is to generalize the procedures to the cases of CATs based on the two- and three-parameter logistic models. In those cases, the optimal item has an infinite *a*-parameter. That is clearly impossible. Designing item pools with realistic distributions for the *a*-parameters is still a challenge, but work is being done in that area.

The critical point of this research is that the characteristics of the item pool for a CAT are important. The design of the item pool can affect the performance of the CAT. The results from the simulation procedure described here give targets for pool development. There is a remaining challenge of producing the test items that will have the desired distribution.

References

Assessment Systems Corporation (1984). *User's Manual for the MicroCAT Testing System*. St. Paul, MN: Author.

Belov, D. I. & Armstrong, R. D. (2009). Direct and inverse problems of item pool design for computerized adaptive testing. *Educational and Psychological Measurement, 69*(4), 53-547.

Cheng, Y., & Chang, H. (2009). The maximum priority index method for severely constrained item selection in computerized adaptive testing. *British Journal of Mathematical and Statistical Psychology, 62,* 369-383.

Drasgow, F., & Olson-Buchanan, J. B. (1999). *Innovations in Computerized Assessment*. Mahwah, NJ: Lawrence Erlbaum.

Guttman, L. (1944). A basis for scaling qualitative data. *American Sociological Review, 9,* 139-150.

Owen, R. J. (1975). A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of the American Statistical Association, 70,* 351-356.

Sands, W. A., Gade, P. A., & Knapp, D. J. (1997). The computerized adaptive screening test. In W. A. Sands, B. K. Waters & JU. R. McBride (Eds.), *Computerized adaptive testing: from inquiry to operation*. Washington DC: American Psychological Association.

Sympson J. B., & Hetter, R. D. (1985, October). Controlling item-exposure rates in computerized adaptive testing. *Proceedings of the 27th annual meeting of the Military Testing Association.* San Diego, CA: Navy Personnel Research and Development Center.

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., & van der Linden, W. J. (2000). Designing item pools for computerized adaptive testing. In van der Linden, W. J. & Glas, C. A. W. (Eds.) (2000). *Computerized Adaptive Testing: Theory and Practice*. Dordrecht: Kluwer.

van der Linden, W. J., & Glas, C. A. W. (2000). *Computerized Adaptive Testing: Theory and Practice*. Dordrecht: Kluwer.

Wainer, H., Dorans, N. J., Eignor, D., Flaugher, R., Green, B. F., Mislevy, R. J., Steinberg, L., & Thissen, D. (2000). *Computerized adaptive testing: a primer (2nd edition)*. Mahwah, NJ: Lawrence Erlbaum.

Ware, J. E. Jr., Bjorner, J. B., & Kosinski, M. (2000). Practical implications of item response theory and computerized adaptive testing: a brief summary of ongoing studies of widely used headache impact scales. *Medical Care, 38*(9). Supplement II, II-73 – II-82.