Graduate
Management
Admission
Council®

The Design of *p*-Optimal Item Pools for Computerized Adaptive Tests

Mark D. Reckase
Michigan State University

Keynote Address Presented June 7, 2007

2007 GMAC® Conference on Computerized Adaptive Testing

Abstract

CAT procedures do not function as expected unless they have an item pool that provides appropriate items for the item selection criteria. This paper describes a methodology for describing the characteristics of an item pool that is needed to support a CAT procedure. The methodology is called *p*-optimal because it does not require the specific item that matches a current ability estimate, but only one that provides at least *p*-proportion of the maximum information from the item. Optimal item pools are very large in size. The use of *p*-optimality allows the item pools to be reasonable in size while still providing an item pool that allows the CAT to function as planned. An example of the item pool design process is given using a simple example based on the Rasch model.

Acknowledgment

Presentation of this paper at the 2007 Conference on Computerized Adaptive Testing was supported in part with funds from GMAC®.

Copyright © 2007 by the Authors

All rights reserved. Permission is granted for non-commercial use.

Citation

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 [date] from www.psych.umn.edu/psylabs/CATCentral/

Author Contact

Mark D. Reckase, Michigan State University, 461 Erickson Hall, East Lansing, MI 48824.
Email: reckase@msu.edu

The Design of *p*-Optimal Item Banks for Computerized Adaptive Tests

Computerized adaptive tests (CATs) are methods for assessing the level of a target hypothetical construct for a person. The methods select test items during the process of test administration to match the characteristics of the person and optimize some measurement criterion, usually the amount of information provided about the location of the person on the construct. There is a substantial literature on the advantages and disadvantages of CATs, and there are number of textbooks on the subject including the classic work by Wainer et al. (2000) that is now in its second edition.

One of the components of a CAT that is often underappreciated is the set of items that are available to the selection algorithm. This set of items is usually called an "item bank" or "item pool." Although selection algorithms are designed to maximize some measure of information or minimize a measure of error, these algorithms cannot yield good estimates of location on the construct in an efficient way unless appropriate items are available for selection. Although the item bank for a CAT is a major component of the procedure, there is little in the research literature that provides guidance about the desired qualities of an item bank. Instead, the literature gives methods for selecting items for an item bank once specifications have been developed (e.g., Veldkamp & van der Linden, 2000) or methods for dividing a set of items into parallel banks to improve test security (e.g., Chang & Ying, 1999). The purpose of this presentation is to provide some guidance about item bank design for CATs. In particular, procedures will be described for determining the size of an item bank that is needed for a CAT to function properly and the distribution of item characteristics that is optimal for a particular implementation of a CAT. This work builds on earlier work by Patience and Reckase (1980) and Reckase (2001).

Along with providing a general model for approaching the item bank design problem, a number of specific examples will be provided. These examples deal with a variety of simple cases that highlight the characteristics of the item bank design methodology. Gu & Reckase (2007) extend this methodology to a more complex CAT model.

Definition of an Optimal Item Bank

The typical report of CAT research describes the characteristics of the item bank used in that research. Often that item bank consists of a set of readily available items, or the parameters of a set of items that are from existing forms of tests in a testing program. The item bank is typically summarized by distributions of the item parameters for the bank. However, there is seldom any evaluation of the quality of that item bank or an evaluation of whether the item bank is well matched to the requirements for the CAT. In this paper, an optimal item bank is defined as one that always has an item available for selection that matches the desired characteristics specified by the item selection routine for the CAT. For a simple example, suppose that a CAT is based on the Rasch model and the selection rule is to pick items that maximize the information provided by the items at the current estimate of trait level (θ). For the Rasch model, information is maximized when the difficulty parameter for the item is equal to the current estimate of θ.

Therefore an optimal item bank is one that always has an item available that has a *b* parameter that is equal to the current estimate of θ.

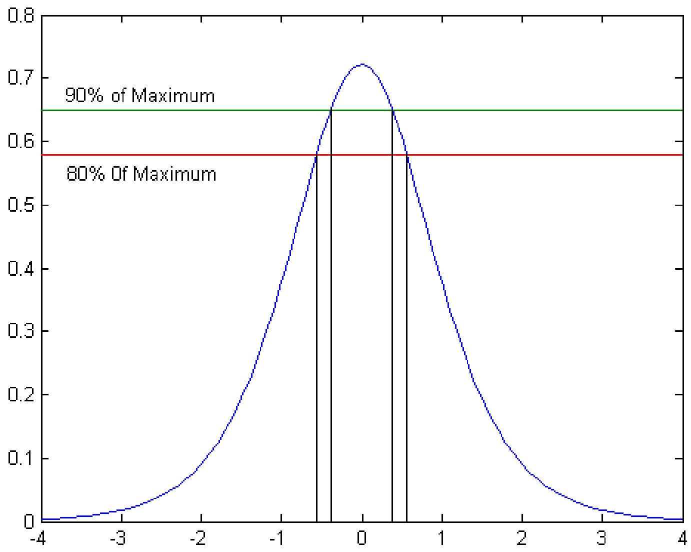
P-Optimality

This definition of optimal is somewhat unrealistic because θs are on a continuous scale. To implement this definition of optimality as stated, if estimates of θ were 1.225 and 1.227, two different items with *b*-parameters equal to 1.225 and 1.227 would be needed, even though the difference in those values probably is less than the error in estimation of the item parameters or the θ estimates. Requiring items to exactly match the θ estimates would also require unrealistically large item banks. To address these issues, the definition of optimal is relaxed somewhat. A purist will say that anything less than optimal is not optimal, which is true, but it is more realistic to consider the definition of optimal to be within the level of accuracy of item parameter estimation and meaningful differences in the functioning of test items. For actual test items, the difference in information provided by items with *b* parameters that are close to each other is small. Therefore, optimal is defined here as always having an item that is within a specified range of the characteristics of the item requested by the item selection algorithm. To make clear the deviation from true optimality, the design criterion is labeled "*p*-optimal," with the proportion reduction from optimality indicated by the value of *p*.

For the case of selecting items to maximize the information provided by the item, the definition of *p*-optimal that is used here is that the amount of information is *p* proportion or more of the maximum information for the items. That is, each item has an information function and the goal for selecting items is to use an item that has *p*-proportion or more of that maximum. For example, when seeking to have .9-optimal item pools, the goal is to use items that provide at least .9 of the maximum information for the item. For the Rasch model, this is fairly easy to determine because all items have the same form for the item information function. Figure 1 provides the item information function for an item with $b = 0$, along with intervals that show when the item provides at least .9 of the maximum or .8 of the maximum. The range that is within .9 of the maximum is roughly from −.4 to .4. Therefore, an item would be considered acceptable for .9-optimal item selection by the CAT maximum information algorithm if it were within .4 of the current θ estimate. The .8-optimal range is within .6 of the current estimate of θ. If .95 of the maximum is used as the criterion, the selected item should be within .3 of the estimated θ.

This definition of *p*-optimal results in a smaller number of required items for a bank than requiring an item that has a *b* parameter that exactly matches the current θ estimate; any item within a specified range can be selected for administration by the algorithm. To help determine the number of items required to meet the requirements of a CAT, the range of θs that are likely to be observed can be divided into regions called "bins". An item that has a *b* parameter within a bin is considered to be a *p*-optimal selection for a θ that falls within the limits of the bin. There is a small amount of inconsistency with this operational definition of *p*-optimality. If θ is located at one end of a bin, an item might be selected from the other end of the bin, resulting in a selection that is more than the desired distance away from the θ value. This argues for using bin sizes that are slightly smaller than that determined from the specified proportion of maximum information.

Figure 1. Information for a Rasch Item With *b* = 0



An Example

An example may help clarify some of these definitions. Suppose that the true θ for an individual is -1 and they are administered a CAT that uses items calibrated using the Rasch model. Item selection is maximum information and θ estimation is maximum likelihood. Until a correct and incorrect response is available, the θ estimate is increased by .7 after a correct response and decreased by .7 after an incorrect response. An initial estimate of 0 is used to start the CAT procedure. After each θ estimate, it is assumed that an item with a b parameter exactly equal to the current θ estimate is available for administration because that is the item that will have maximum information for that point on the score scale. The CAT is fixed length at 20 items.

This CAT is simulated by assigning a score of 0 or 1 for an item depending on whether a uniform random number is greater than or less than the probability of correct response for the administered item for the true θ. The procedure begins by selecting an item with a b parameter equal to 0, the item with maximum information at the initial θ estimate. That item has a probability of correct response of .1545 for a person with θ equal to -1. If a uniform random number of .3764 is generated, a response of 0 is assigned to the item, because the random number is greater than the probability of correct response. Because scores of both 0 and 1 are not available, the maximum likelihood estimate is not defined and the estimate of θ is reduced by .7 to $-.7$. The next item is selected to have a b parameter of $-.7$ and the process continues as above. When the response string contains both a 0 and a 1, maximum likelihood is used to estimate θ. The CAT terminates when 20 items have been administered.

Table 1 contains the b parameters for the items selected for administration and the item response for each item. Because the requested item with b parameter equal to the θ estimate is assumed to be available, the b parameter for the next item is the same as the θ estimate after the previous item. The final θ estimate for this adaptive test is -1.1263. This is the estimate after the 20th item on the test. It is slightly higher than the last *b* parameter because the last item score was a 1 for a correct response.

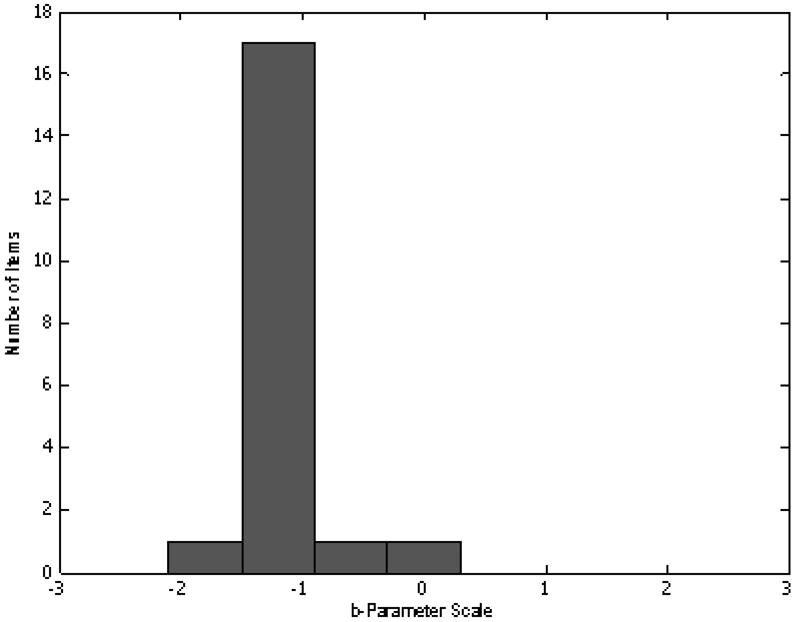
Table 1. *b* Parameters (and θ Estimates) and Item Scores for a Simulated CAT Based on the Rasch Model with True $\theta = -1$

Item Number	*b* Parameter	Item Score
1	0	0
2	−0.7000	0
3	−1.4000	1
4	−1.1986	0
5	−1.5924	1
6	−1.2914	1
7	−1.0607	0
8	−1.2537	1
9	−1.0878	0
10	−1.2332	1
11	−1.1037	0
12	−1.2204	1
13	−1.1142	0
14	−1.2116	1
15	−1.1217	0
16	−1.2052	1
17	−1.1272	1
18	−1.0538	0
19	−1.1229	0
20	−1.1880	1

A quick scan of the item parameters shows that some of them are very similar. From a practical perspective, the differences in difficulty of those items are of no practical significance. Using the concept of bins, the number of items for each range of the θ scale can be counted to give the number of items needed in each range for the *p*-optimal measurement of this examinee. For this example, a bin width of .6 is used, based on a .8-optimal criterion. The results are presented in a bar graph in Figure 2.

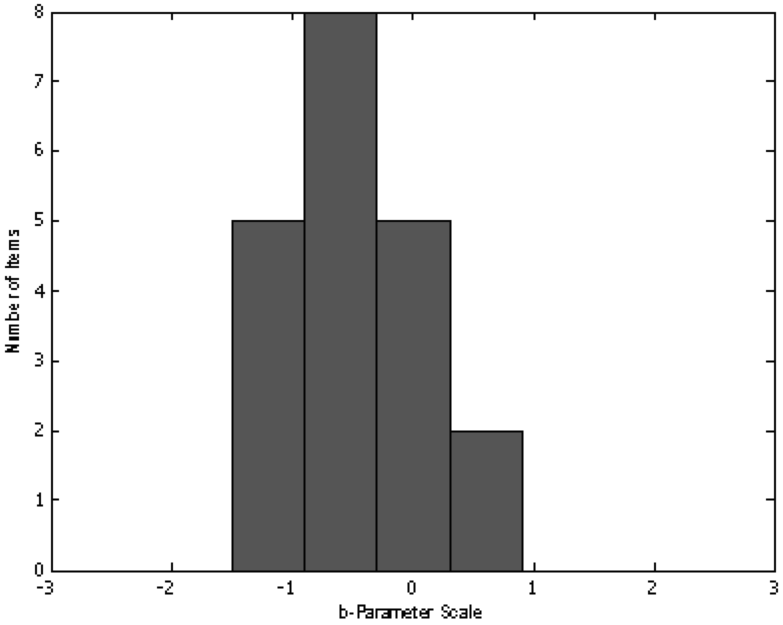
The results of the counts of the items in each of the bins shows that 1 item was needed from the range −.3 to .3, one item was needed from −.9 to −.3, 17 items were needed from −1.5 to −.9 (the range that included the true θ) one item was needed from −2.1 to −1.5. For this administration of the CAT, this would be the 20 items that would be needed for there to be an item available near every item that was requested. For one person, the CAT needs 20 items because items can not be reused for the same person.

Figure 2. Number of Items in Each Bin for a Test with True θ = −1



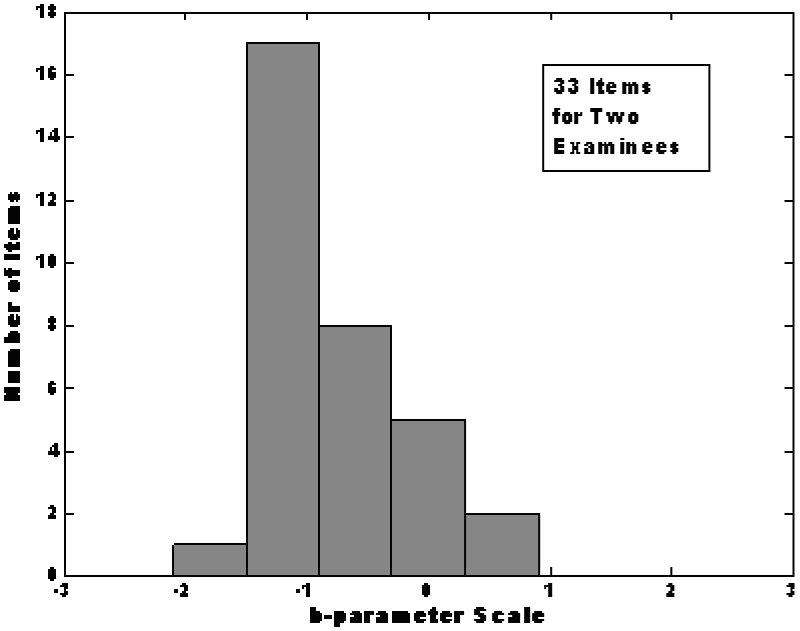
It will be possible to reuse some of the items if a second examinee is located relatively close to the first examinee. Suppose the CAT is administered to a second examinee with θ = −.5. The bar chart with counts of items in bins for that examinee is given in Figure 3. For this examinee, two items are needed from .3 to .6, five items are needed from −.3 to .3, eight items are needed from −.9 to −.3, and five items are needed from −1.5 to −.9.

Figure 3. Number of Items in Each Bin for a Test With True θ = −.5



If test security is not a concern, there is no need to have 20 unique items for the second examinee if the items for the first examinee are already in the item bank. The first examinee already needed 17 items between −1.5 and −.9, more than enough for the five items required by the second examinee. In general, the items needed for both examinees is the larger of the two counts for each bin—one item in the range −2.1 to −1.5, 17 in the range −1.5 to −.9, eight from −.9 to −.3, five from −.3 to .3, and two from .3 to .6. The total of items needed for both examinees is 33 rather than the 40 that would be needed if items could not be reused. This number is the union of the sets of items for the two examinees. The distribution of items in bins for the item bank for two examinees is shown in Figure 4.

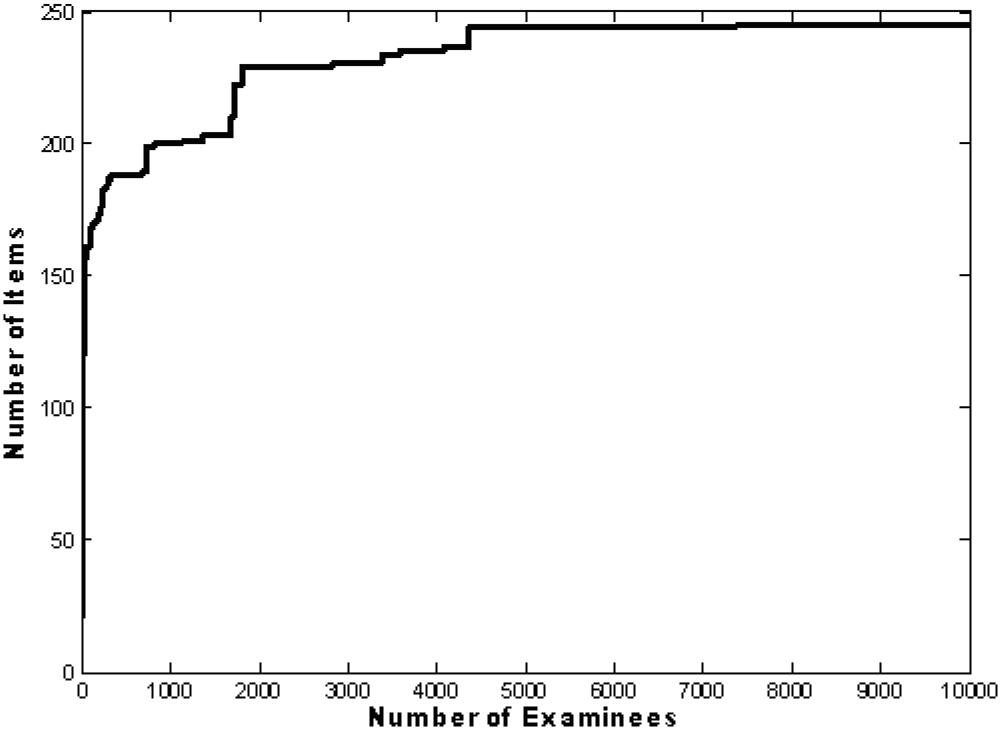
Figure 4. Number of Items in each Bin Needed for Two Examinees With θ = −1 and −.5



To determine the total number of items needed for a CAT, this process of selecting items can be continued for the number of examinees expected to be administered a test from the same item bank. The general process is to randomly select an examinee from the expected examinee population, determine the optimal set of items for that examinee, find the union of the items needed for that examinee with any previously selected items for examinees sampled earlier, and continue this process until the specified sample number of examinees is reached. The resulting distribution of items in bins gives the *p*-optimal item bank distribution for the specified population of examinees.

Suppose the hypothetical Rasch model based CAT is designed for a population of examinees that is distributed normally with mean 0 and standard deviation 1. Further, suppose that this low stakes CAT is expected to be administered to 10,000 examinees over a period of time. What is the .95-optimal distribution of item difficulty parameters for this application?

Figure 5. Increase in Number of Items Needed in the Item Pool With Increase in Number of Persons Tested



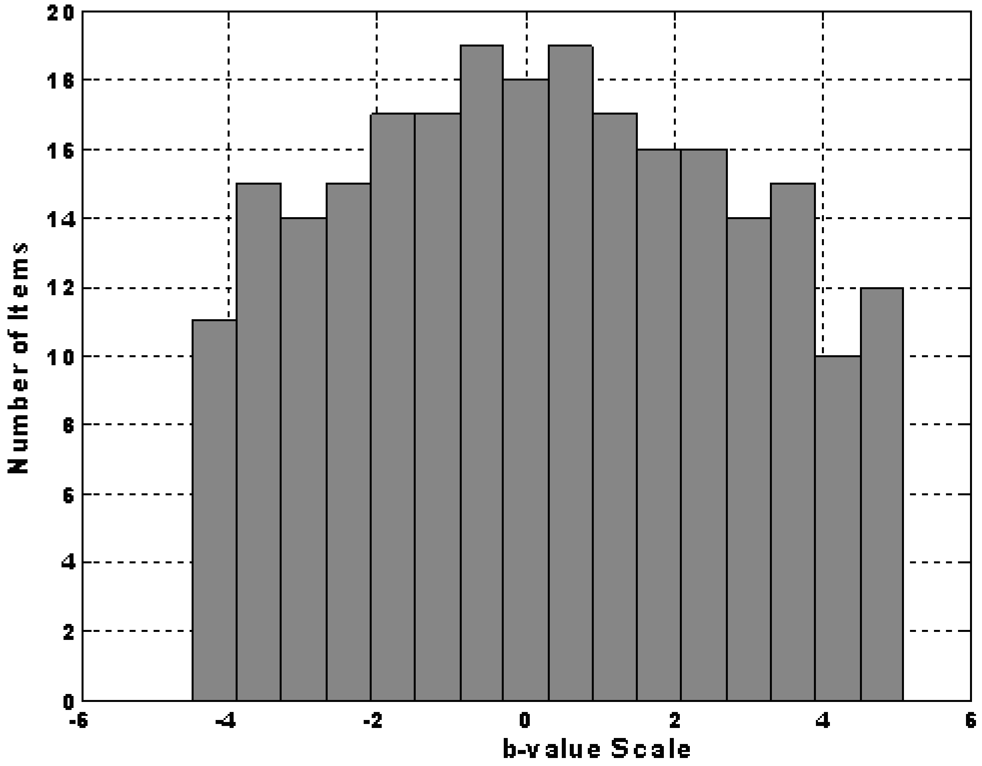
The number of items in the union of the item sets for the randomly selected examinees is graphed as a function of the number of examinees. That graph is presented in Figure 5. The graph shows that the number of items needed in the bank increases very quickly as the number of examinees increases, and then reaches an asymptote after about 5,000 examinees. The asymptote is at 218 items. Figure 6 shows the distribution of *b* parameters over the bins for the 218 items after the asymptote in item pool size was reached. Note that this distribution is not normal, but is flatter and does not drop to zero in the tails. This is because items are needed for good measurement of extreme abilities even if there is only one individual at the extremes. This suggests that selecting 218 items from tests designed for paper-and-pencil administration will not be optimal for computerized adaptive tests.

Summary and Conclusions

This paper describes a methodology for determining the desired characteristics of an item pool for a CAT. An example is provided based on a simple adaptive testing model using the Rasch model. The results show that the item pool size is dependent on the distribution of the examinee population and the number of persons who will take the CAT. The results also show that the form of the desired item pool is not a normal distribution of Rasch difficulty values, but rather one that is flattened with quite high frequencies at the tails of the distribution.

The item pool needed for a CAT is quite specific and is related to the design of the procedure and the characteristics of the examinee population. Other work in this area has also shown the effects of exposure control procedures on the desired characteristics of the item pool. The investigation of the desired characteristics of the item pools for CATs is clearly an area that is in need of further research. CATs work well only when they have item pools that support the methodology. The best CAT procedure may not work well when the item pool does not provide the kind of item that is requested by the procedure.

Figure 6. Distribution of Items Over Bins for the *p*-Optimal Item Pool of 218 Items



References

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

Gu, L. & Reckase, M. D. (2007*). Designing optimal item pools for computerized adaptive tests with Sympson-Hetter exposure control*. Paper Presented at the 2007 GMAC Conference on Computerized Adaptive Testing, Minneapolis, MN.

Patience, W. M. & Reckase, M. D. (1980, April*). Effects of program parameters and item pool characteristics on the bias of a three-parameter tailored testing procedure*. Paper presented at the meeting of the National Council on Measurement in Education, Boston, MA.

Reckase, M. D. (2001, September). *Item pool design for computerized adaptive tests.* Invited small group session at the 6th Conference of the European Association of Psychological Assessment, Aachen, Germany.

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.

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