Item pool design for computerized adaptive tests

Mark D. Reckase
Michigan State University[1]

Although item pools are critical to the proper functioning of computerized adaptive tests (CATs), there is very little in the research literature that indicates the desired features of an item pool. Most research articles use an existing item pool that was developed for other purposes. For example, Pastor, Dodd and Chang (2002) used items from NAEP; Wang and Kolen (2001) used an item pool from paper-and-pencil ACT Mathematics test forms. The standard texts on adaptive testing give little guidance. Flaugher (2000) indicates that "the item pool from which items are selected must contain high-quality items for many different levels of proficiency" (p. 38). Later in the chapter he provides an example with 200 items in the pool. Veldkamp and van der Linden (2000) give detailed guidance about designing an item pool once specifications have been developed, but they do not tell how to produce the specifications in the first place. Stocking (1994) suggested that item pools should contain six or seven times as many items as the length of the adaptive test, but this figure was based mostly on concerns over test security rather than the number and distribution of items needed to measure effectively. The information function for the item pool was of concern, but the criterion for information was based on the characteristics of previous paper-and-pencil forms. No specific procedures have been identified for developing the specifications for a computerized adaptive testing item pool.

The purpose of this paper is to describe some initial research designed to address the issue of the design of the ideal item pool for an adaptive test. This will be far from the definitive work on the subject because the implementation of adaptive testing is very complex, and it is becoming more complex every day as new methods for exposure control and content balancing are designed and implemented. The research reported here will begin with the very simple cases of item pool design where exposure control and content balancing are not considered. In fact, for the sake of beginning with a simple example, the first adaptive test to be considered is based on the Rasch model with maximum information item selection and maximum likelihood ability estimation. A fixed step is used to estimate ability until finite estimates can be obtained using the maximum likelihood procedure. Also, to keep the example simple, a fixed test length is used for the adaptive test. For this relatively simple situation, an ideal item pool is determined.

---

[1] Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL, April 2003.

The Definition of an Ideal Item Pool

The definition of an ideal item pool that is used here is that the pool contains every item that is requested by the item selection procedure. The following example shows how the ideal item pool is defined. Suppose that an examinee is randomly sampled from a distribution with mean 0 and standard deviation 1. That examinee has an ability of -1.15. The adaptive test starts with an initial ability estimate of 0.0 so the first item desired is one with a *b*-parameter of 0.0 because for the Rasch model the information is maximized when the item difficulty parameter matches the ability. Even though the examinee had a low ability, the answered the first item correctly so the ability estimate was increased by a fixed step of .7 to get the second ability estimate. Therefore the next item that was desired was one with a *b*-parameter of .7. That item was answered incorrectly and the first maximum likelihood estimate is computed as .35. Then of course, the next item desired is one with a *b*-parameter of .35. This process continues for all of the items on the test. All of the *b*-parameters required so that the item difficulty always exactly matched the current ability estimate are given in Table 1. This would be an ideal item pool for the estimation of ability of this first person on the test.

Insert Table 1 about here

Notice that as the test continues, the items selected for the ideal item pool have very similar *b*-parameter values. In fact, there are six items that have *b*-values that round to -.86 even though they have uniquely different values. It is doubtful that these items are really functioning differently and the specifications for the item pool for this person could simply say "six items with *b*-parameter estimates of -.86." Perhaps even those between -.8 and -.9 would be indistinguishable in practice because the amount of information they provide is negligibly different. Also, in reality, the *b*-values in an item pool are estimates and the estimation error may be greater than these slight differences. Therefore before going further to define the ideal item pool, the range of *b*-values that is indistinguishable in practice will be considered.

Consider the information function for a Rasch calibrated item. For convenience, the *b*-parameter for the item is assumed to be 0.0. The form of the item information function is exactly the same for any other item. The information function for the item is given in Figure 1.

Insert Figure 1 about here

The maximum information for this item is .7225 because the model is includes $D = 1.7$ so that it will be comparable to later work done with the three-parameter logistic model. The use of this constant makes the model similar to the normal ogive model. The range of values over which this item has roughly 90% or more of the maximum information is from -.4 to .4 or a width of .8 θ units. It might be argued that the item is equally effective in that range. The range that covers 80% or the maximum is roughly from -.6 to .6 or a width of 1.2. A range from -.3 to .3, or .6 wide covers 95% of the maximum. The reason for considering these values is to determine the width over which the item will not have a serious reduction in the information it provides for estimating the ability of examinees. The range of .6 certainly would not result in any serious drop in information and it is possible that the range of .8 might work nearly as well. Both of these ranges will be considered when developing an ideal item pool.

Rather than requiring that items match the ability estimate exactly, the requirement is that the item be within .3, or .4, of the value. This is operationalized by defining "bins" that are either .6 wide or .8 wide to store the required items. Reconsidering the example given in Table 1, the items are sorted into bins that are either .6 wide or .8 wide. Table 2 shows the number of items needed for the examinee according to the bins.

The allocation of items to bins shows that most of the items are in the -1.2 to -.61 bin when the .6 bin width is used, or the -1.6 to -.81 bin when the .8 bin width is used. This is the bin that corresponds to the true ability of the examinee, -1.15. The reason for bringing in the bin concept is to accumulate the number of items needed in the item pool over multiple examinees, the next step in the process of ideal item pool design.

When more than one person takes an adaptive test, they can use the items that have been selected for another individual. If a person is near -1.15 on the θ-scale, they can use many of the items from the first examinee. For example, suppose that two examinees have taken the adaptive test based on the Rasch model and there abilities are -.43 and -1.67. Because they start with

ability estimate 0.0, there are items that are in common to the items selected for these two examinees. The items that are needed to measure each examinee well are the union of the set of items for each examinee. The 25 items selected for the first examinee are shown in Figure 2.

Insert Figure 2 about here

The first examinee has most items in the bins for -.6 to -.01 and -1.2 to -.61. The second examinee has most items in the bins from -1.21 to -2.4. The second examinee used three of the items that were selected for the first examinee. Therefore, the pool needed for all three examinees required only 47 items instead of the 50 items that would be required if the unique items for each examinee were required to be put in the pool. The full pool used for the two examinees is shown in Figure 3.

Insert Figure 3 about here

The process for deriving the ideal pool for an adaptive test is to randomly select true θ-values from a hypothesized population of examinees and identify the ideal set of items needed for each examinee. The ideal pool is then the union of the sets of items for the examinees. The example given above shows that successive examinees can use items selected for previous examinees. The number of new items that need to be added for each examinee diminishes as the number of examinees increases. Ultimately, the number in the ideal item pool should asymptote to some value such that the full set meets the requirements of virtually all sampled examinees.

To show this result, suppose that examinees are randomly sampled from a normal distribution with mean 0.0 and standard deviation 1.0. Items are selected to be optimal for each examinee assuming the Rasch model and the union of the sets of items is formed to identify the ideal item pool for that examinee population. Figure 4 shows how the item pool increases in size as the number of sampled examinees increases.

Insert Figure 4 about here

The item pool size starts at 25 because that is the number of items needed to test one person. As the number of tested examinees increases, the required item pool increases until the item pool reaches 216 items. As more examinees are tested after that, they all use items selected for previous examinees. The set of items needed for the full set of examinees is shown in Figure 5. This distribution is not very smooth because it is based on a random sample of 200 examinees. However, it is clear that the distribution is not normal. It tends to be closer to a uniform distribution. To smooth out the distribution, the process for developing the distribution was replicated five times and the results were averaged. That result is given in Figure 6.

Insert Figures 5 and 6 about here

The ideal item distribution given in Figure 6 is somewhat more rounded and symmetric than the distribution in Figure 5. With more replications, that distribution can be expected to become even more regular. This distribution has approximately 23 items in each of the middle categories and then drops to about 20 items more than a standard deviation from the mean of the distribution. Outside two standard deviations from the mean, the number of items drops further with only a few items with *b*-parameters beyond +4 and -4. This ideal pool has 221 items.

The same type of analysis can be done for a CAT using the three-parameter logistic model, but it is a little more complicated because optimal values for the *a*- and *c*-parameters can not be defined. These values would really be positive infinity and zero, respectively, but those values are not realistic. Instead, the values for the *a*- and *c*-parameters can be sampled from realistic distributions and then the ideal *b*-parameter is the one that yields maximum information given the other parameters. There is a further complication because the *a*- and *b*-parameters tend to be correlated. That relationship could be modeled to derive a realistic item pool. An example of an item pool for a CAT using the three-parameter logistic model is shown in Figure 7. The *b*-parameters for this pool are reported on the standard score scale for the test. The bin width for this pool is .25 on the usual metric corresponding to 96% of the maximum information for a typical item.

Insert Figure 7 about here

This ideal item pool has approximately 600 items in it and it is slightly non-symmetric because of the effects of the lower asymptote parameter on the item information. The larger pool size is due to the smaller bin width. If larger bins are used, the pool size will drop.

Discussion and Conclusions

The purposes of this paper are to identify an area of research in adaptive testing that has received little attention and to suggest a methodology for approaching the problem. The area is the design of item pools for computerized adaptive tests. An example of item pool design is given for the simple case of a test based on the Rasch model with maximum information item selection, maximum likelihood ability estimation, and a fixed test length. The results show that a pool of approximately 200 items that are distributed relatively evenly over the range from -2.5 to 2.5 is appropriate for examinees sampled from a standard normal distribution.

The methodology that is proposed is very general. Other forms of examinee distributions can be used and different forms of IRT models can be the basis for the method. Substantial work needs to be done in this area, especially when exposure control and content balancing methods are part of the adaptive testing procedure. No doubt, those added features will require that the item pools be larger. The amount of increase in size is not known at this time. This is a very rich area for future research.

References

Flaugher, R. (2000). Item pools. In Wainer, H. (2000). *Computerized adaptive testing: a primer.* Mahwah, NJ: Lawrence Erlbaum Associates.

Pastor, D. A., Dodd, B. G., & Chang, H. (2002). A comparison of item selection techniques and exposure control mechanisms in CATs using the generalized partial credit model. *Applied Psychological Measurement, 26*(2), 133-146.

Stocking, M. L. (1994, February). *Three practical issues for modern adaptive testing item pools (RR-94-5).* Princeton, NJ: Educational Testing Service.

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: Kluwer.

Wang, T. & Kolen, M. J. (2002). Evaluating comparability in computerized adaptive testing: issues, criteria and an example. *Journal of Educational Measurement, 38*(1), 19-49.

Table 1

b-Parameters for the Ideal Item Pool
for an Examinee with θ = -1.15

Item Number	*b*-parameter
1	0
2	0.7000
3	0.3500
4	-0.0808
5	-0.4449
6	-0.1609
7	-0.3879
8	-0.5953
9	-0.7914
10	-0.9809
11	-1.1667
12	-1.0049
13	-0.8731
14	-0.7597
15	-0.8630
16	-0.7679
17	-0.8558
18	-0.9395
19	-0.8613
20	-0.7893
21	-0.8571
22	-0.9222
23	-0.8605
24	-0.9191
25	-0.8633

Table 2
Item Allocation to Bins .6 Wide or .8 Wide

Bin Boundaries .6	Number of Items	Bin Boundaries .8	Number of Items
-4.2 - -3.61		-4.0 - -3.21	
-3.6 - -3.01		-3.2 - -2.41	
-3.0 - -2.41		-2.4 - -1.61	
-2.4 - -1.81		-1.6 - -0.81	13
-1.8 - -1.21		-0.8 - -0.01	9
-1.2 - -0.61	17	0 – 0.79	3
-0.6 - -0.01	5	0.8 – 1.59	
0 – 0.59	2	1.6 – 2.39	
0.6 – 1.19	1	2.4 – 3.19	
1.2 – 1.79		3.2 – 4.00	
1.8 – 2.39			
2.4 – 2.99			
3.0 – 3.59			
3.6 – 4.20			

Figure 1
Information for a Rasch Item

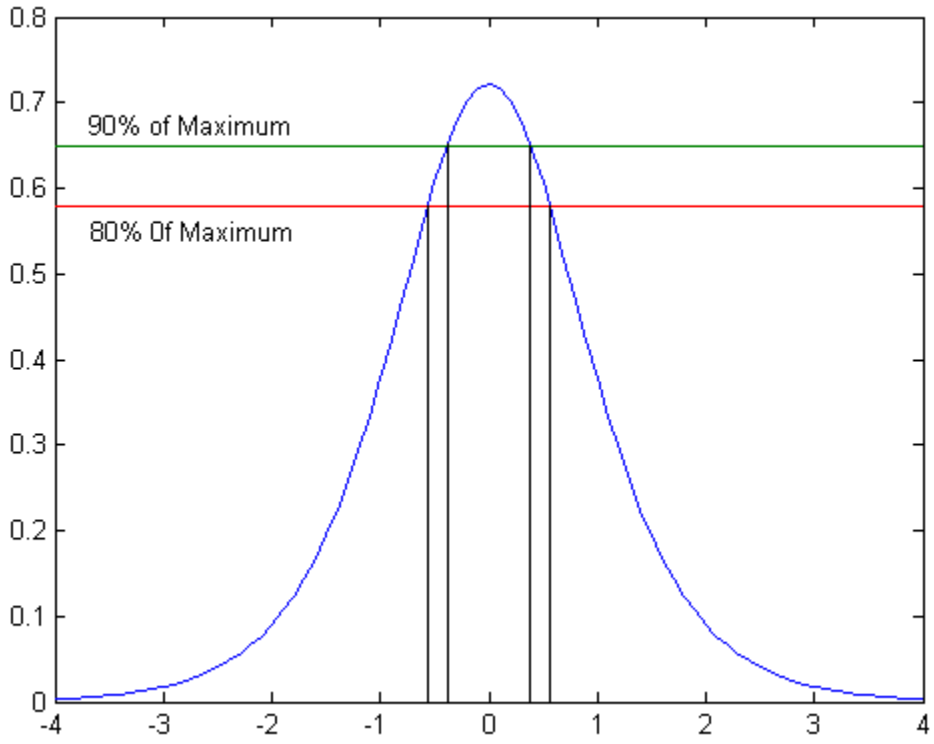

Figure 2
Items for Examinee with θ = -.43

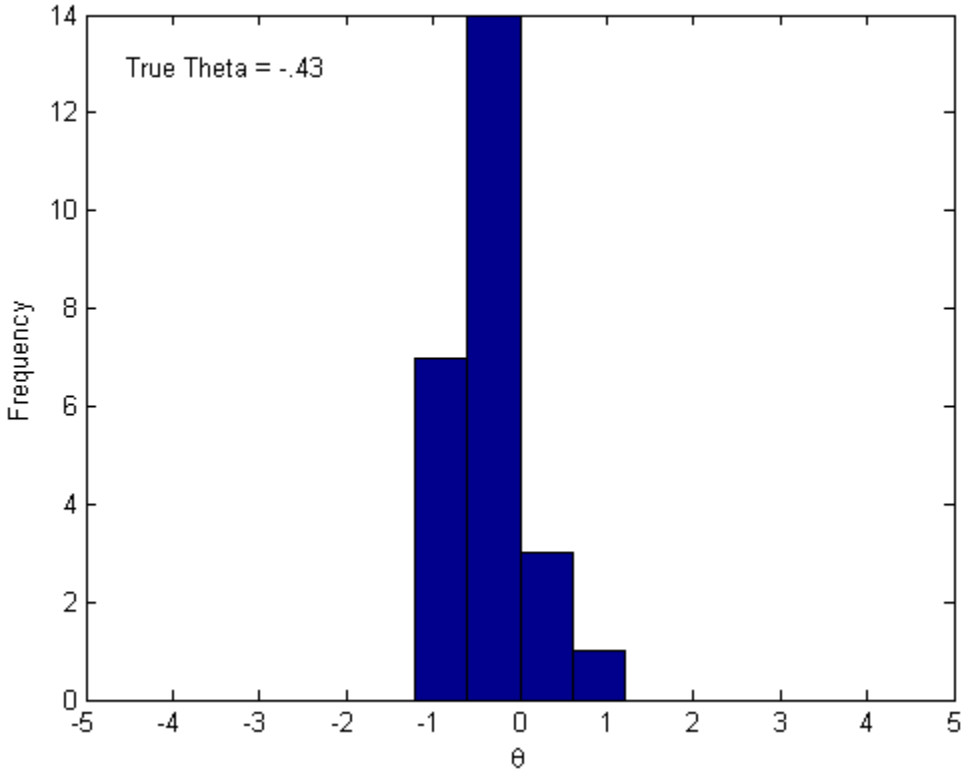

Figure 3
Item Pool for Two Examinees

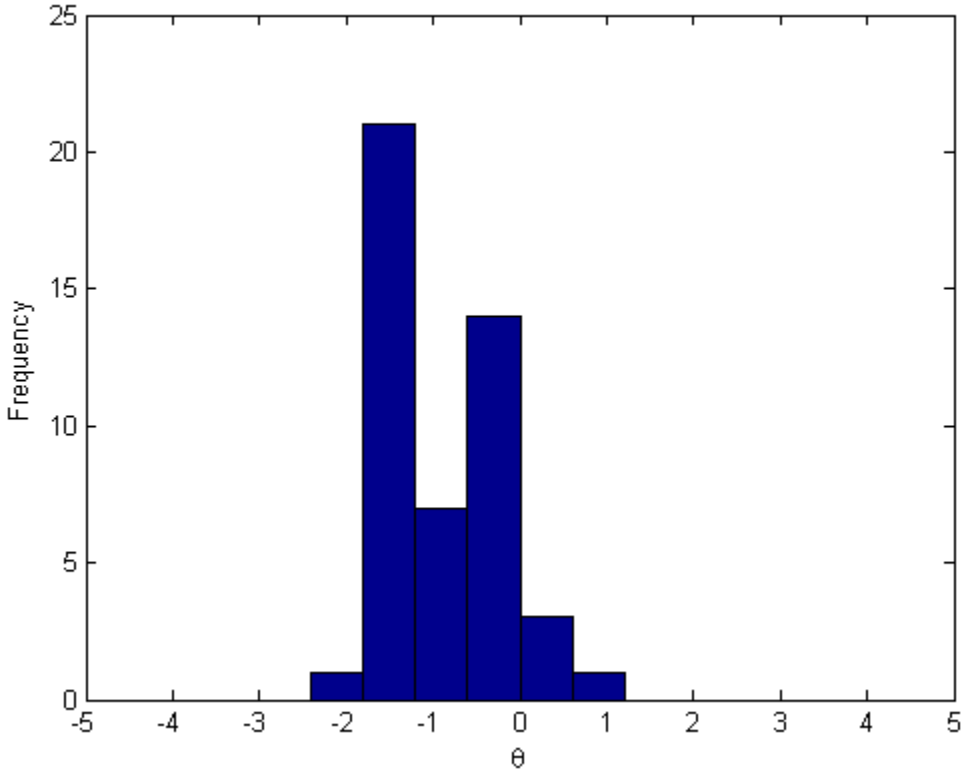

Figure 4
Increase in Pool Size as Number of Examinees Increases

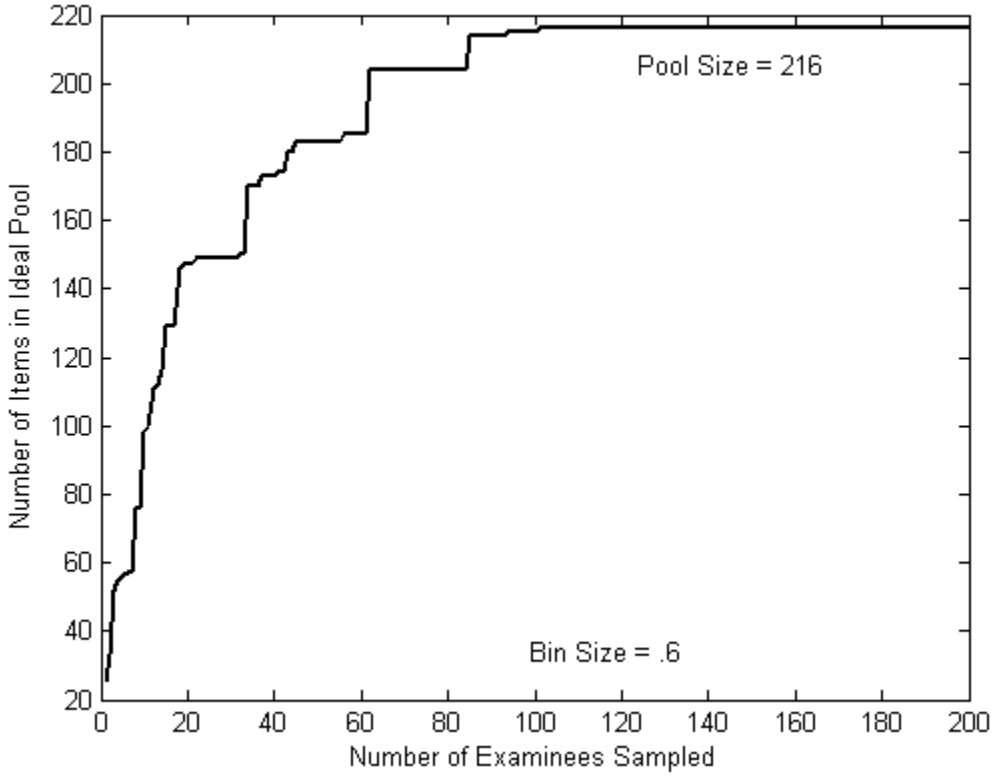

Figure 5
Ideal Item Pool Assuming a Standard Normal Distribution
and Bin Size .6

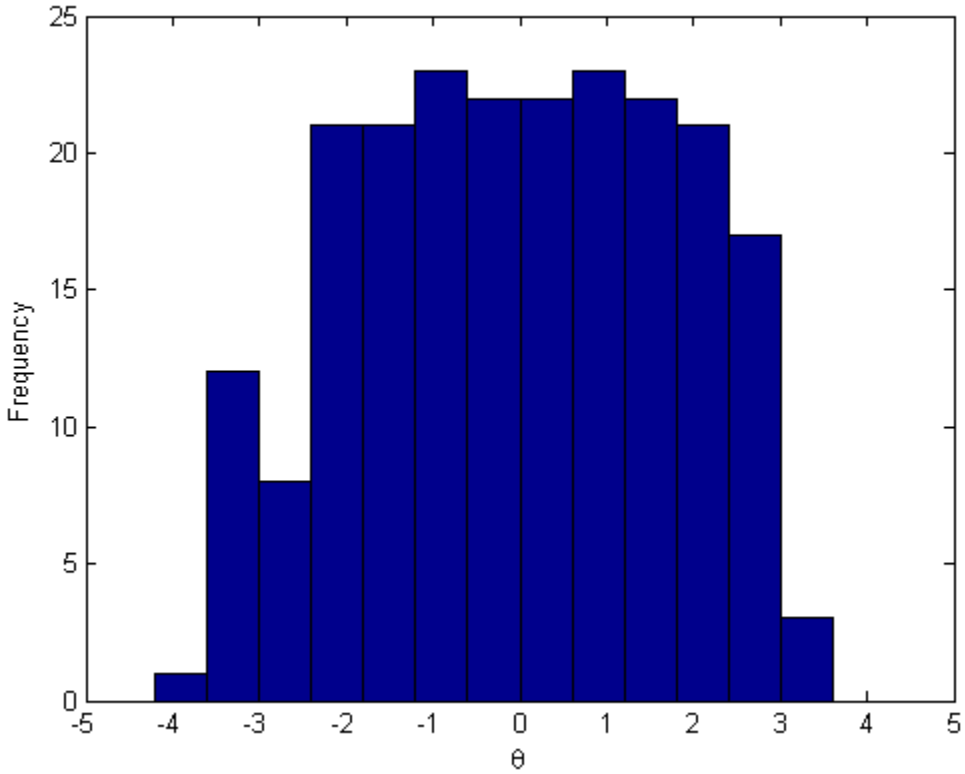

Figure 6
Ideal Item Pool Based on 30 Replications of 200 Test Administrations

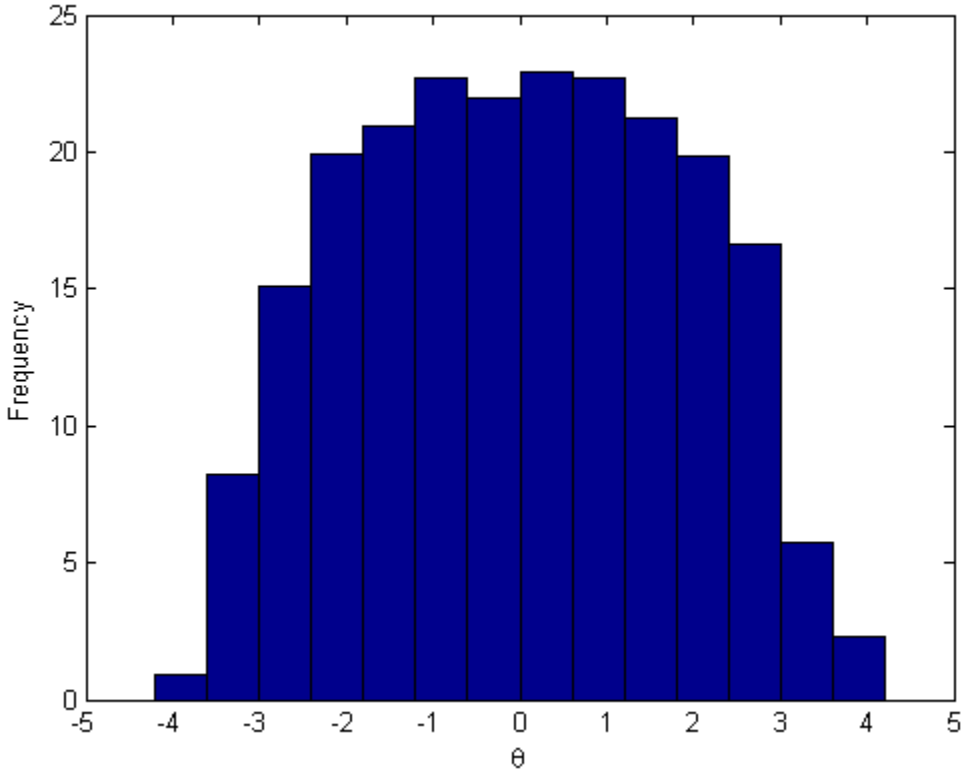

Figure 7
An Item Pool Based on the Three-Parameter Logistic Model