The impact of differing field-testing procedures on the accuracy of item calibrations in a computerized adaptive test

G. Gage Kingsbury
Northwest Evaluation Association

April 13, 2005

Paper presented to the annual meeting of the American Educational Research Association, Montreal, Canada.

This research was sponsored by the NCLEX Joint Research Committee

The impact of differing field-testing procedures on the accuracy of item calibrations in a computerized adaptive test

G. Gage Kingsbury
NWEA

Item Response Theory (IRT: Lord and Novick, 1968; Lord, 1980) derives part of its appeal from the fact that it allows the creation of measurement scales that are independent of the particular sample of individuals or test questions used to create the scales, and invariant when applied to particular groups of individuals within the population of interest. This invariance property is exceptionally valuable, because it provides us with the capacity to build measurement scales that can be expected to maintain their measurement characteristics even though we modify test forms or implement a computerized adaptive test (Weiss, 1982; Lord, 1980).

Given the assumptions of IRT, item parameters will be invariant, but item parameter **estimates** will not be invariant. Estimates will vary due to a number of factors that have been researched for at least twenty years. These include sampling fluctuation (Swaminathan and Gifford, 1983), departures from unidimensionality (Bejar, 1980), and other characteristics of the calibration design including item context (Yen, 1980).

Adaptive testing adds another level of complexity which may impact parameter estimation. Four characteristics of adaptive testing that contribute to this complexity are the following:

- Each test taker sees a different set of test questions.
- Test takers see sets of questions with different difficulty.
- Each test taker sees a set of questions targeted to his or her trait level.
- The adaptive test reacts dynamically to the performance of the test taker.

These characteristics change the role of the person taking the test from that of a passive test taker to an active participant in test design. This, in turn, changes the test characteristics, including the distribution of item difficulty, and item difficulty as a function of position of the item within the test. While the impact of test taker as test designer has not been well researched, it is clear that it has an impact on common psychometric exercises, such as identifying person fit, identifying DIF, and, more pertinently, estimating item parameters.

Research concerning item-parameter estimation in adaptive testing settings has two areas of focus. The first focus relates to the **calibration procedure**, in which the research asks the question "How do we calibrate items, given the data from adaptive tests". A variety of researchers have tackled this question synthesized by Ban, Hanson, Wang, Yi, and Harris (2001) and a number of approaches ranging from marginal maximum likelihood to

empirical Bayesian approaches have been established. The second focus relates to the **field-testing process**, and asks the question "How do we assign field-test items to test takers to calibrate items within an adaptive test". Some of the work done by Van der Linden and his associates (Van der Linden & Glas, 2000) concerning optimal calibration design is related to this aspect of research. The current study addresses this second focus of research concerning calibration of items within the context of an adaptive test.

This study is designed to investigate two attributes of item calibration within the context of a specific, operational adaptive test. The first attribute is the impact of sample size on the accuracy of item calibrations. The second is whether the calibration procedure could be modified by assigning specific field-test items to specific test takers in order to reduce the number of candidates needed to calibrate a given field-test item, and to assure that each item is calibrated to a desired level of precision. The processes for attaching these kinds of constraints to the operational items in adaptive tests have been well researched (Kingsbury and Zara, 1991; Stocking and Swanson, 1993; Van der Linden and Pashley, 2000).

The operational test that serves as the model for this study is the NCLEX-RN test. This adaptive test is used as a portion of the licensure procedure for registered nurses throughout the United States. While the NCLEX test is very high-volume, the field-testing process limits the number of items that can be added to the item pools, and dictates the number of field-test items that must be given to each candidate. Items developed for use in NCLEX RN tests are currently field tested by being administered to 400 candidates. Candidates take field-test items with unknown difficulties, and the number of candidates who take each item in the field trial is fixed.

The managers of operational adaptive tests need to have information about the quality of their current calibration processes. In addition, they need to know the impact of other possible alterations in the calibration process. Since the number of individuals taking a certification test is somewhat fixed, the number of items that can be added to the item pool in a period of time is also fixed. If the item calibration process could be streamlined by using less item responses, managers would have more flexibility in the use of its item trial slots. This could be used to reduce test length, or to allow the field testing of more items without substantial additional cost.

This simulation study will investigate two possible modifications to field-testing procedures that might result in a more streamlined process. The first modification studied will be the use of provisional calibrations to seed field-test items into the operational tests. The second modification will involve administering field-test items to as many candidates as necessary to provide parameter estimates with a prespecified level of stability. The study will also include an examination of the impact of combining the two modifications. Finally, the study will examine the expected impact of the recent change from calibrating with 500 field-test responses to calibrating with 400 field-test responses.

Modification I: Using provisional calibrations

Field-test items are currently selected randomly for candidates from a large pool of items. This may result in less able candidates encountering very high-difficult items, and more

able candidates encountering very low-difficulty items. These candidates do not provide as much information about item characteristics as we would gain from candidates seeing challenging but not frustrating field-test items.

By asking content experts to rate the difficulty of items, we should be able to give them provisional calibrations that bare some relationship to actual calibrations. If we select field-test items according to these provisional calibrations and the current trait level estimate for the candidate, we should be able to increase the amount of information obtained from each candidate taking the field-test item. As a side benefit, this process should also reduce the amount of consternation caused by an item of inappropriate difficulty appearing during an adaptive test.

Modification II: Specifying a level of calibration stability

Items that are included in the field tests vary substantially in difficulty. Currently, these items are administered to a consistent number of candidates. For items that tap very difficult content, or very elementary content, calibration estimates may have more error than those for items in the middle of the difficulty distribution. This modification halts calibration for an item when the calibration reaches a desired level of stability.

For this study, the level of calibration stability will be measured by calibrating the item following each 25 item administrations, after an initial minimum number of administrations. If the calibrations obtained for an item on two successive calibrations differ by less than a prespecified level, the calibration process for the item is complete.

Procedure

Using NCLEX-RN candidate, item, and test characteristics, a series of simulations were conducted to determine whether either modification or a combination of the two modifications results in a calibration process that could improve the one now used.

Instruments

The NCLEX-RN test (Zara, 1992) is used as a portion of the certification process for every new registered nurse practicing in the United States. Items developed for use in this test are field tested by being administered to 400 randomly chosen test takers during operational testing. The NCLEX uses the one-parameter logistic model for scaling.

For this simulation, the item parameters for operational items were from an operational item pools for the spring, 2003 NCLEX-RN. The field-test item parameters were randomly generated to have the same mean and standard deviation as the operational item pool.

Simulations

In each simulation, simulated candidates were administered an adaptive test according to the item presentation rules used in the operational NCLEX examinations. The distribution of simulated test takers was set to match that of the spring 2003 NCLEX-RN sample. Field-test items were simulated to mimic the distribution of difficulty in the

operational item pools. To simplify the simulation, tests were fixed at 75 items in length.

Calibration sample size. Three sample sizes were used as the point at which to calibrate the field-test items, as follows:

- **400** -- In this condition, an item received its final calibration after it had been administered to 400 simulated candidates
- **500** -- In this condition, an item received its final calibration after it had been administered to 500 simulated candidates
- **N** -- In this condition, each item is given to enough candidates that the estimated calibration stabilizes (with a predefined stabilization parameter -- delta) For this study, delta was set to .005, and calibration stability was estimated after each 25 administrations of an item

Provisional calibrations. Each of the sample calibration sample sizes was simulated

- **PC -- With provisional calibrations** -- In this condition, a set of provisional item calibrations was used for item selection. Field-test items were administered as if the provisional calibration were the actual calibration for the item. Only item selection was affected. The calibration process did not use the provisional calibration as a starting point or prior.
- **NP -- With no provisional calibrations** -- In this condition, items were selected randomly from the field-test items for administration. This is the current item selection process.

Provisional calibration accuracy. Since the degree to which item calibrations can be estimated from item characteristics varies greatly from one situation to another, four conditions representing varying validity were simulated, as follows:

- **.00 correlation** -- In this condition, provisional calibrations had no relationship to true item calibrations, except that the mean and the standard deviation were the same.
- **.40 correlation** -- In this condition, provisional calibrations had the same mean and standard deviation as the true calibrations, and the correlation was .40, approximating the low end of the range of calibrations between estimated and observed item difficulties reported in the literature.
- **.60 correlation** -- In this condition, provisional calibrations had the same mean and standard deviation as the true calibrations, and the correlation was .60.
- **.80 correlation** -- In this condition, provisional calibrations had the same mean and standard deviation as the true calibrations, and the correlation was .80, approximating the high end of the range of calibrations between estimated and observed item difficulties reported in the literature.

In the figures shown below, each condition is designated by the number of candidates, the presence of provisional calibrations, and the correlation of the provisional calibrations with actual calibrations. Therefore, condition 400PC4 indicates that a sample size of 400 was used to calibrate each item, provisional calibrations were used, and the correlation between the provisional and actual calibrations was .40.

Design and replications. All of the conditions above were completely crossed. Although the correlation between true and provisional calibrations should not influence the accuracy of calibrations when provisional calibrations are not used, the completely crossed design allows the direct comparison of all conditions. Each particular set of 100 field-test items was used with each condition, to assure comparability. Each condition was replicated 20 times.

Calibration procedure. In each simulation, an unconditional maximum-likelihood calibration procedure, analogous to that used in WINSTEPS (Linacre, 2003) was used for calibration. This matches the operational system, since WINSTEPS is used for calibration in the NCLEX program. Field trials with WINSTEPS, using the settings normally used for calibration of field-test items indicate that the two procedures produce the same calibrations to two decimal places. The analogue procedure was used in place of WINSTEPS to allow the calibration to be imbedded in the simulation.

Method of field-test item selection. As in the actual NCLEX-RN, the simulation included fifteen field-test items embedded within sixty operational items. The field-test items were presented at random points during the test, with two limitations. No field-test items appeared during the first ten items administered to a simulated candidate, and no more than two field-test items were presented consecutively.

During the simulation, assumptions were made concerning the activity of the testing system when administering field-test items. The first assumption involves the manner in which field-test items are chosen for a candidate during active testing. Since it is probable that field-test items will be put out for field testing in groups, we need to know how the items in a group will be treated, as they reach the desired goal for calibration. Three approaches were considered, as follows:

- **natural selection** – all field-test items are considered eligible for presentation to any candidate, regardless of the number of times they have been presented in the past.
- **partially-constrained selection** -- All field-test items are considered eligible for presentation until they reach the desired calibration termination condition, at which point they become available for use only if no non-terminated items appear in the list of items to be chosen from randomly.
- **fully-constrained selection** – all field-test items are considered eligible for presentation until they reach the desired calibration termination condition, at which point they become unavailable for use. Field-test items are chosen from the remaining items.

Each approach has merit. The natural-selection approach is a logical choice if items are to be added to the field test interactively, and withdrawn from use when calibrated. It also uses the most candidates to bring all items in a group of field-test items to the calibration criterion. The fully-constrained approach reduces to a minimum the number of candidates needed to bring all items in a group of field-test items to the calibration criterion. It tends to administer items that are inappropriate at the end of a field-testing cycle. The partially-constrained approach requires more candidates than the fully-constrained approach, but assures that items will be administered to the appropriate

candidates. In this simulation, the partially-constrained approach was used in all conditions.

Analysis

Each condition was evaluated by the following criterion measures:

1. Overall calibration accuracy
2. Overall calibration bias
3. Overall number of candidates for calibration
4. Conditional calibration accuracy

Each criterion measure was examined using the average of the 20 simulation replications. Conditional calibration accuracy was calculated conditional on true item difficulty. For this calculation, items were blocked into 24 theta blocks (-3.00>=b<-2.75,..., 2.75<b<=3.00). Each block value was then computed using an item-weighted average across replications, within theta block.

Results

Calibration Accuracy and Bias

Table 1 shows the average calibration accuracy and bias for all calibration sample conditions (400, 500, and Variable), correlation levels (.00, .40, .60, and .80), with and without the use of provisional calibrations (PC and NP).

Several trends may be noted. First, as expected, the level of the correlation of the provisional calibrations with true item difficulties has no noticeable impact on the accuracy of calibrations obtained not using the provisional calibrations (the NP column). Second, within any one correlation level the variable termination condition resulted in the greatest inaccuracy, the 400-subject condition resulted in the next lower inaccuracy, and the 500 subject condition resulted in the most accurate difficulty estimates. Third, for the two lowest correlation levels, the condition without provisional calibrations resulted in the most accurate difficulty estimates. For the two highest correlation levels, the use of provisional calibrations resulted in the most accurate difficulty estimates.

It is useful to note that across all conditions, the use of 400 test takers rather than 500 added approximately .1 theta units to the average inaccuracy. This is an increase in error of approximately 10 percent.

Table 1. Average absolute difference between true item difficulty and estimated item difficulty for each simulated condition (averaged across all replications)

Num	Corr	NP	PC
400	0.00	0.114	0.121
500	0.00	0.095	0.113
Variable	0.00	0.151	0.172
400	0.40	0.108	0.110
500	0.40	0.096	0.097
Variable	0.40	0.155	0.151
400	0.60	0.108	0.090
500	0.60	0.098	0.088
Variable	0.60	0.150	0.144
400	0.80	0.116	0.079
500	0.80	0.100	0.070
Variable	0.80	0.156	0.121

Table 2 shows the average difference between the true and estimated item difficulties, averaged across all replications, for each simulated condition. The results of this bias analysis indicate that levels of bias were consistently less than .01 theta units, across all conditions. No pattern of bias is noticeable among the conditions.

Table 2. Average difference (bias) between true item difficulty and estimated item difficulty for each simulated condition (averaged across all replications)

Num	Corr	NP	PC
400	0.00	0.005	0.003
500	0.00	0.007	-0.002
Variable	0.00	-0.003	0.005
400	0.40	-0.002	0.006
500	0.40	0.005	0.004
Variable	0.40	0.003	-0.003
400	0.60	-0.003	-0.001
500	0.60	-0.006	0.005
Variable	0.60	0.003	-0.002
400	0.80	0.005	0.006
500	0.80	-0.004	-0.005
Variable	0.80	-0.003	0.004

Conditional Accuracy

Figures 1 through 4 show the average absolute difference between calibration estimates and true item difficulties, for all items, across all replications, as a function of item difficulty. The trends noticeable here mirror the overall accuracy results.

Observing Figure 4, in which the correlation between actual and provisional item difficulties is at its highest level (.80), clear patterns are observed. For almost all difficulty levels, the condition with 500 responses using provisional calibrations (500pc8) results in the most accurate difficulty estimates. The conditions with provisional calibrations tend to consistently result in the most accurate estimates at virtually all difficulty levels. The stability-based termination tends to result in the least accurate estimates when compared to similar conditions.

One additional trend is that the differences in accuracy tend to be slightly greater for the most extreme items. The conditions that do not use provisional calibrations tend to have greater errors for the highest and lowest difficulty items within the item pools.

Figure 1. Average absolute difference between true item difficulty and estimated item difficulty for each simulated condition (averaged across all replications)
Correlation between true and provisional calibrations = .00

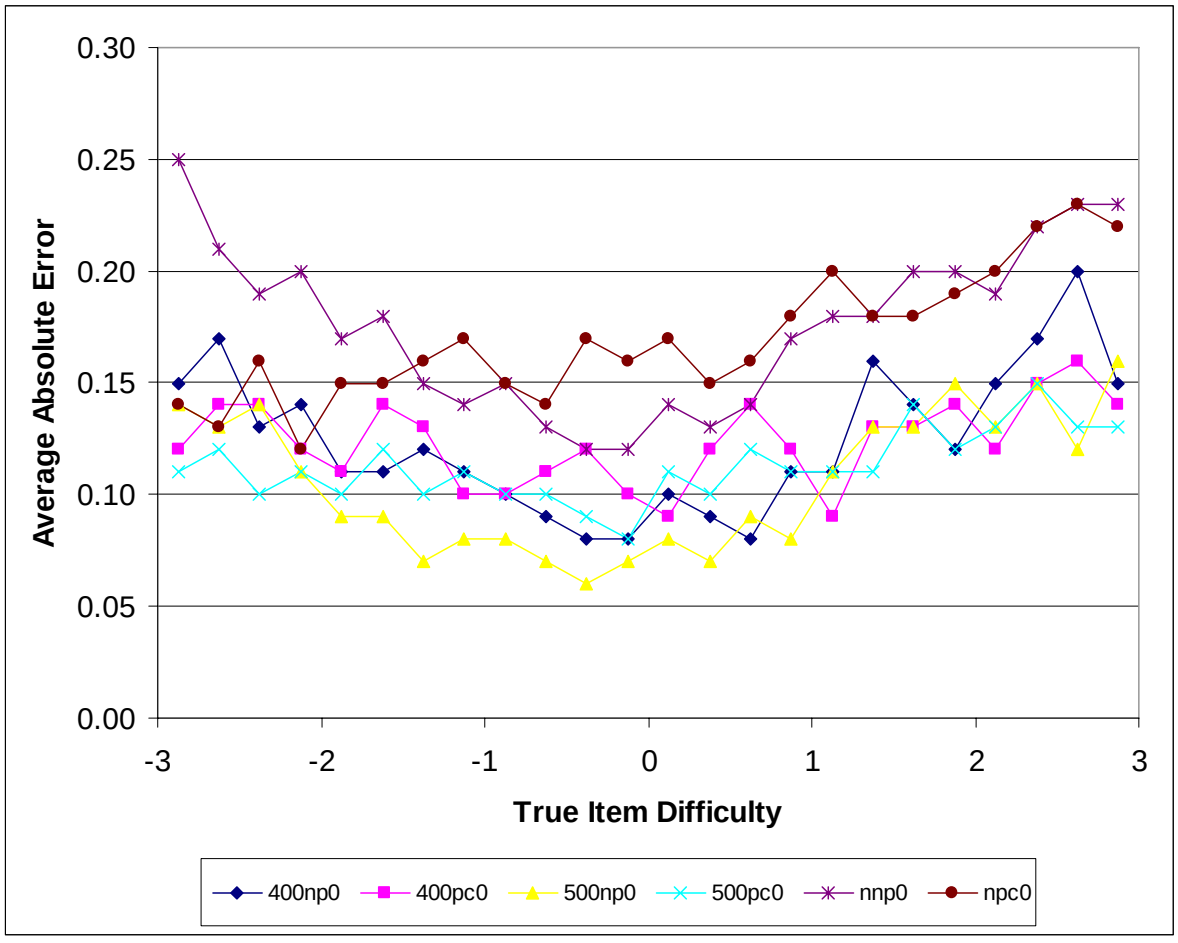


Figure 2. Average absolute difference between true item difficulty and estimated item difficulty for each simulated condition (averaged across all replications)
Correlation between true and provisional calibrations = .40

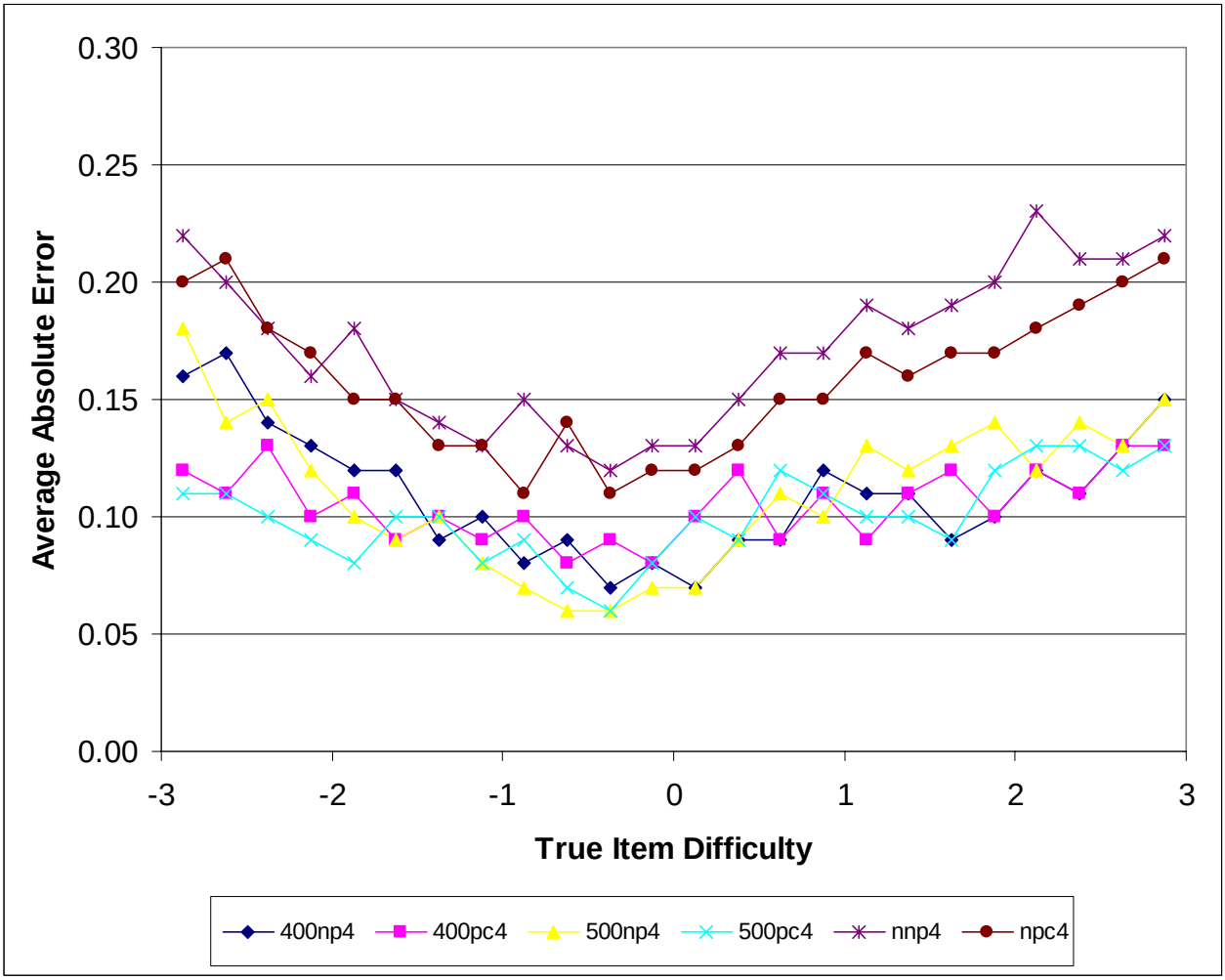


Figure 3. Average absolute difference between true item difficulty and estimated item difficulty for each simulated condition (averaged across all replications)
Correlation between true and provisional calibrations = .60

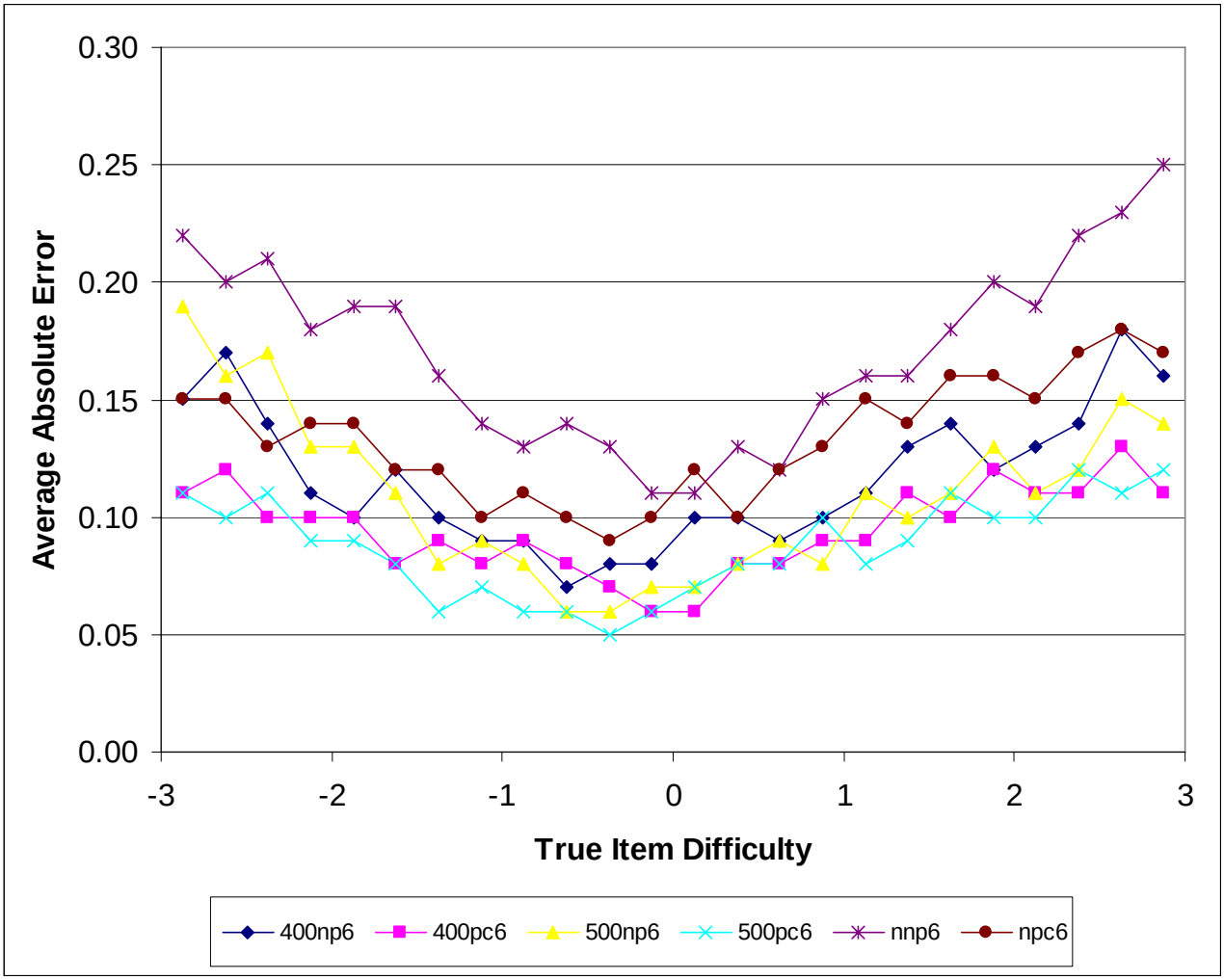
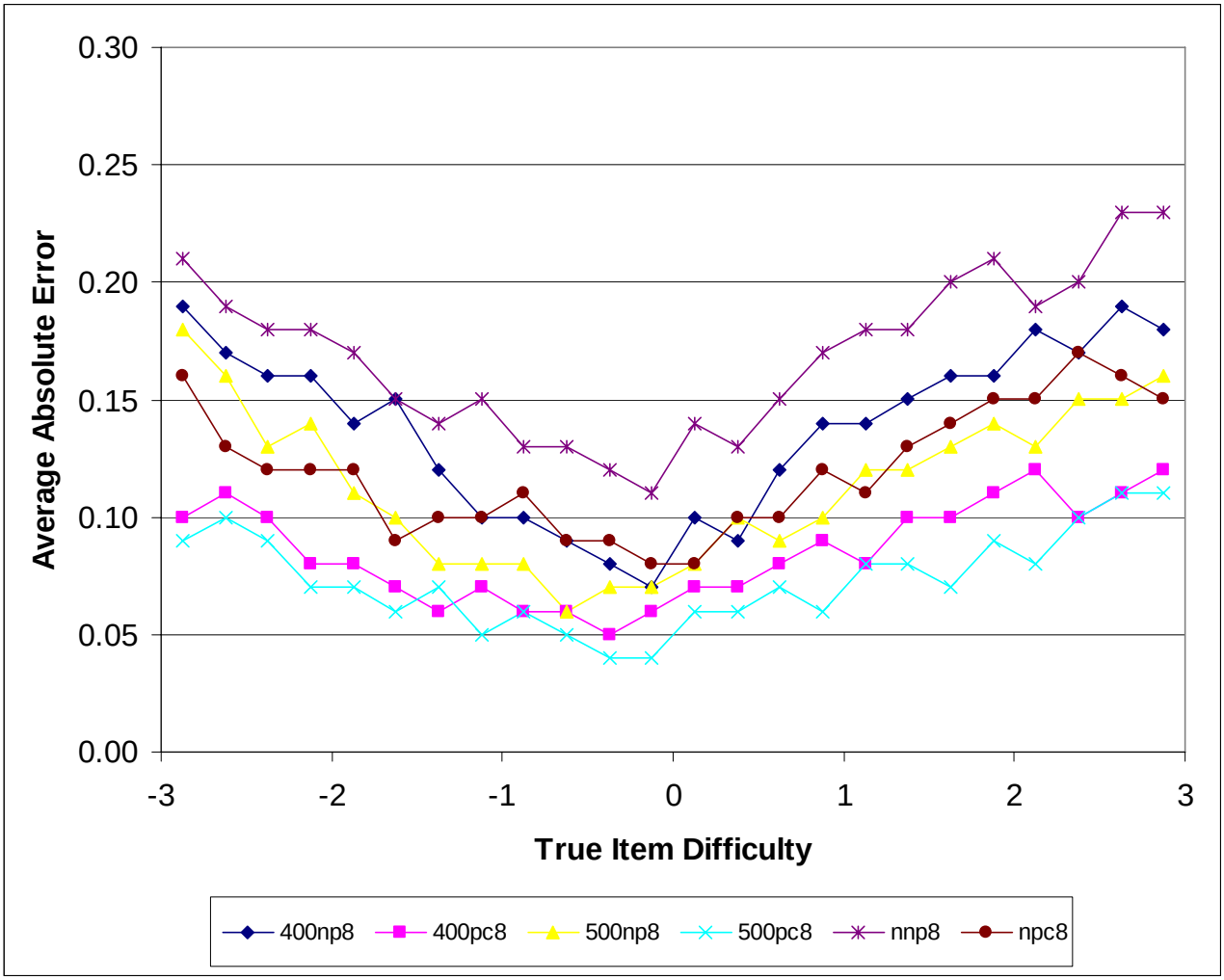


Figure 4. Average absolute difference between true item difficulty and estimated item difficulty for each simulated condition (averaged across all replications)
Correlation between true and provisional calibrations = .80



Number of Candidates Needed for Calibration

Table 3 shows the average number of candidates used to obtain the final calibrations, averaged across all replications. Within each level of correlation, the variable length termination resulted in the lowest number of candidates to complete calibration. The fixed, 400-candidate conditions used more candidates, while the 500-candidate conditions used the most candidates to calibrate an item. It should be noted that the amount of variability observed is related directly to the choice of partially-constrained item selection.

Table 3. Average number of candidates needed to complete calibration for an item in each simulated condition (averaged across all replications)

Num	Corr	NP	PC
400	0.00	2597.16	3251.82
500	0.00	3254.15	4070.67
Variable	0.00	1340.75	2069.17
400	0.40	2590.76	2900.44
500	0.40	3257.83	3591.32
Variable	0.40	1397.41	1776.78
400	0.60	2604.02	3141.78
500	0.60	3249.22	3922.52
Variable	0.60	1490.55	1973.60
400	0.80	2596.85	3296.73
500	0.80	3252.27	4065.30
Variable	0.80	1477.74	2038.32

Table 4 shows the average number of item responses needed to calibrate an average item in each condition. As expected, the fixed termination conditions use 400 or 500 responses. The variable termination conditions consistently use substantially fewer responses (approximately 60% of the responses used in the 500-response condition). The use of provisional calibrations increases the number of responses used by a small percentage, across conditions.

Table 4. Average number of item responses needed to complete calibration for an item in each simulated condition (averaged across all replications)

Num	Corr	NP	PC
400	0.00	400	400
500	0.00	500	500
Variable	0.00	257.15	276.25
400	0.40	400	400
500	0.40	500	500
Variable	0.40	264.80	275.05
400	0.60	400	400
500	0.60	500	500
Variable	0.60	266.45	270.65
400	0.80	400	400
500	0.80	500	500
Variable	0.80	262.25	268.00

Discussion and Conclusions

It is clear from this study that the addition and use of reasonable accurate provisional calibrations (correlation with true calibrations >= .40) in the field testing process should enable more accurate calibration with a given field-test sample size. The gains are particularly noticeable for items at the extremes of the difficulty distribution. It is unclear whether provisional calibrations can be made with a level of accuracy high enough to improve calibration accuracy for this test, but it would probably be useful to investigate further.

Another clear finding is that the use of variable length stopping criterion for calibration reduced the sample size needed to calibrate items, but also resulted in a substantial increase in the error associated with the parameter estimate. More research is needed relative to the stopping rules before it is determined whether this process is advantageous for operational use.

The third major trend in the results was that the use of the smaller fixed calibration sample (400 responses) added 10 to 15 percent to the error in the calibration estimates compared to the larger fixed calibration sample (500 responses). Developers of operational adaptive tests need to consider difference in error as one element of their field-test design.

While the difficulties involved in calibrating items within the context of an adaptive test are well documented, less is known about the relative advantages of using different approaches toward calibration. This study was designed to add to our knowledge concerning the relative costs associated with using different calibration approaches within the context of an adaptive licensure examination.

In certification and licensure examination development, each item costs a great deal to develop, and each item has a limited life span. If an item is exposed to less test takers while it is being field tested, its useful life span will increase. The common approach of randomly assigning items to test takers fails to take advantage of the characteristics of an adaptive test. This study has examined the capacity of the adaptive test to give us more accurate calibrations for items through the use of our knowledge of the content area to create provisional calibrations. While this is only an initial study, the potential for use within operational adaptive tests is direct, and may be substantial.

References

Ban, J. C., Hanson. B. H., Wang, T., Yi, Q., & Harris, D. J. (2001). A comparative study of on-line pretest item calibration-scaling methods in computerized adaptive testing. Journal of Educational Measurement, 38, 191-212.

Bejar, I. I. (1980). A procedure for investigating the unidimensionality of achievement tests based on item parameter estimates. Journal of Educational Measurement, 17, 283-296.

Kingsbury, G. & Zara, A. (1991). A Comparison of Procedures for Content-Sensitive Item Selection in Computerized Adaptive Tests. *Applied Measurement in Education*, 4 (3) 241-261.

Linacre, J. M. (2003). Winsteps user manual.

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

Lord, F. M. & Novick, M. R. (1968). Statistical theories of mental test scores. Reading, MA: Addison-Wesley.

Stocking, M.L. & Swanson, L. (1993). A Method for Severely Constrained Item Selection in Adaptive Testing. *Applied Psychological Measurement*, 17(3), 277-292.

Swaminathan, H. & Gifford J. A. (1983). Estimation of parameters in the three parameter latent trait model. In D. J. Weiss (Ed.), New horizons in testing: Latent trait test theory and computerized adaptive testing. New York: Academic Press.

Van der Linden, W. J. & Pashley, P.. J. (2000). Item selection and ability estimation in adaptive testing with shadow tests. In W. J. van der Linden & C. A. W. Glas (eds.), *Computerized adaptive testing: Theory and practice* (pp.1-25). Boston: Kluwer.

Van der Linden, W. J. & Glas, C. A. W. (2000). Cross-validating item parameter estimation in adaptive testing. In A. Boorsma, M. A. J. van Duijn, & T. A. B. Snijders (Eds.), Essays on item response theory. New York: Springer

Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement, 6*, 473-492.

Yen, W. M. (1980). The extent, causes and importance of context effects on item parameters for two latent trait models. Journal of Educational Measurement, 17, 297-311.

Zara, A. R. (April, 1992). *A comparison of computerized adaptive and paper-and-pencil versions of the national registered nurse licensure examination*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.