Graduate
Management
Admission
Council®

Adaptive Item Calibration: A Process for Estimating Item Parameters Within a Computerized Adaptive Test

G. Gage Kingsbury
Northwest Evaluation Association

Presented at the Real-Time Analysis Paper Session, June 2, 2009

Abstract

The characteristics of an adaptive test change the characteristics of the field testing that is necessary to add items to an existing measurement scale. The process used to add field-test items to the adaptive test might lead to scale drift or disrupt the test by administering items of inappropriate difficulty. The current study makes use of the transitivity of examinee and item in item response theory to describe a process for adaptive item calibration. In this process an item is successively administered to examinees whose ability levels match the performance of a given field-test item. By treating the item as if it were taking an adaptive test, examinees can be selected who provide the most information about the item at its momentary difficulty level. This should provide a more efficient procedure for estimating item parameters. The process is described within the context of the one-parameter logistic IRT model. The process is then simulated to identify whether it can be more accurate and efficient than random presentation of field-test items to examinees. Results indicated that adaptive item calibration might provide a viable approach to item calibration within the context of an adaptive test. It might be most useful for expanding item pools in settings with small sample sizes or needs for large numbers of items.

Acknowledgment

Presentation of this paper at the 2009 Conference on Computerized Adaptive Testing was supported in part with funds from GMAC®.

Copyright © 2009 by the Author

Citation

Kingsbury, G. G. (2009). Adaptive item calibration: A process for estimating item parameters within a computerized adaptive test. In D. J. Weiss (Ed.), ***Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing.*** **Retrieved [date] from www.psych.umn.edu/psylabs/CATCentral/**

Author Contact

**G. Gage Kingsbury, Northwest Evaluation Association, 121 NW Everett St.
Portland, OR 97209, U.S.A. Email: Gage.Kingsbury@nwea.org**

Adaptive Item Calibration: A Process for Estimating Item Parameters Within a Computerized Adaptive Test

Adaptive testing (Weiss, 1982) provides us with a very efficient measurement technology that has been applied in a broad variety of measurement venues. Existing uses range from personnel testing (Sands, Waters, and McBride, 1997) to certification and licensure testing (Zara, 1992) to educational achievement testing (Kingsbury & Houser, 1999). Adaptive testing provides us with a variety of opportunities to improve testing practice, but occasionally it also presents unexpected challenges. One of these challenges arises when we wish to add new items to an existing measurement scale using item response theory (IRT; Lord & Novick, 1968; Lord, 1980).

In the forty years since the development of item response theory (IRT), researchers have created a variety of techniques to apply the theory in order to create measurement scales to measure a wide variety of constructs. Central to each of these techniques is the need to estimate item parameters by connecting item response information to these scales (Swaminathan & Gifford, 1980). Commonly, a group of examinees responds to a group of items, and then the item parameters are estimated from a full data structure or from a data structure with data missing by design. This approach allows consistent estimation of item parameters with known levels of accuracy.

Adaptive testing adds a level of complexity to item parameter estimation. Four characteristics of adaptive testing that contribute to this complexity are the following:

1. Each examinee sees a different set of test questions.
2. Examinees see sets of questions with different difficulty.
3. Each examinee sees a set of questions targeted to his or her trait level.
4. The adaptive test reacts dynamically to the performance of the examinee.

These characteristics change the role of the person taking the test from that of a passive examinee 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 examinee 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 differential item functioning, and, more pertinently, estimating item parameters.

Research concerning item parameter estimation in adaptive testing settings has two aspects. The first aspect 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 addressed this question, including a synthesis by Ban, Hanson, Wang, Yi, and Harris (2001). In general, those processes that work well in traditional tests also work with information from adaptive tests.

The second aspect of parameter estimation within an adaptive test relates to the *field-testing process,* and asks the question "How do we assign field-test items to examinees to calibrate items within an adaptive test?". The work done (Buyske, 1998; Van der Linden & Glas, 2000; Holman & Berger, 2001) concerning optimal calibration design is related to this aspect of research, and applies to traditional tests and adaptive tests as well. This aspect of item calibration has been more challenging, since the four aspects of adaptive testing mentioned above each impact the data that are obtained for use in item calibration. This is the aspect of parameter estimation that was addressed in this study.

This study introduces a process for assigning examinees to particular field-test items to allow efficient item calibration. The process performs in a manner similar to the process that is used to assign operational items to examinees in an adaptive test. Since the process is similar to adaptive testing, we will call it *adaptive item calibration***.** This study describes the adaptive item calibration process and provides an example. It then discusses some of the unique elements in the process that will need to be managed, and presents a simulation of the process to identify its operating characteristics.

Adaptive Item Calibration

When items are written for inclusion in an adaptive testing item pool, we know their content characteristics, but not their measurement characteristics. To discover an item's measurement qualities we need to administer it to a reasonable group of examinees. Commonly, we administer a set of field-test items to a randomly assigned group of examinees. Although this administration can be done as a standalone field test, it is often accomplished by seeding field-test items into an operational test.

Random assignment is a very reasonable process, and it will definitely provide information concerning each item's measurement characteristics that can be used for IRT analysis. At the same time, random assignment has two problems.

1. It might not be the most efficient process. When creating large item banks for adaptive tests, efficiency might be a high priority. A process that provides useful measurement information with smaller sample sizes would be very useful in operational testing.

2. Random assignment of items will make them less targeted to the examinees. This might not sound like a problem, but if a struggling examinee encounters a very difficult field-test item, it might hurt their motivation, and it will certainly stand out from the items that surround it. We know little about the impact of this type of item on the examinee, but it could easily change the way in which the individual interacts with the item, and with the rest of the test.

In order to address these two issues, the adaptive item calibration process has been developed. This process assigns examinees to an item in much the same manner that an adaptive tests assigns items to an examinee. The process has a number of interesting operational characteristics and tuning parameters that we will address below. First, though, it is useful to explicate the algorithm itself:

Adaptive Item Calibration Algorithm

1. A pool of field-test items is established. The items in this pool have no existing item responses.

2. Each item is given an initial provisional difficulty estimate. (This can be done using content experts as judges to assign a difficulty estimate, or by using the mean difficulty estimate from items with similar content.)

3. Rules for positioning field-test items within a test are established. These rules can vary in complexity from "The 20th and 21st items in each test will be field-test items" to "Three

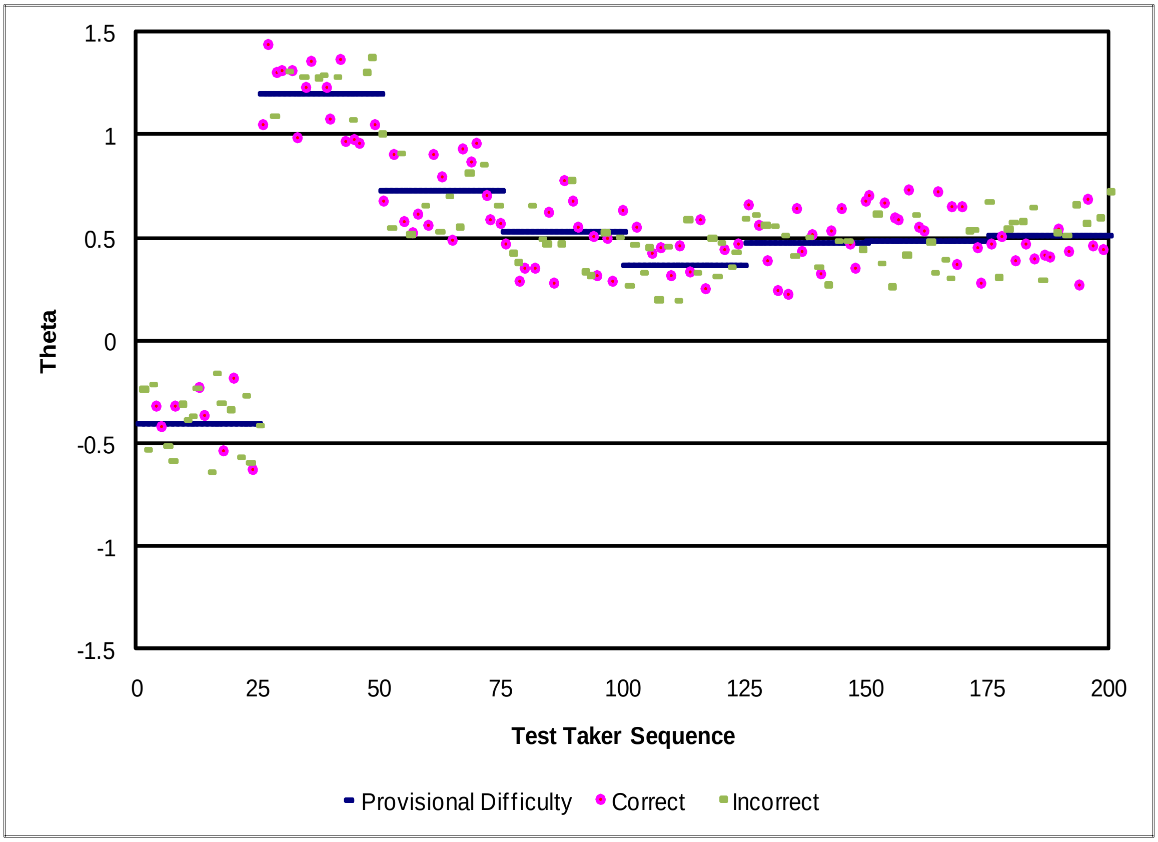
field-test items will be administered at randomly identified positions within the first 30 items on the test, with no more than one field-test item being administered consecutively and no field-test items appearing within the first five items."

4. When the rules indicate that an examinee should be administered a field-test item, the item to be administered is chosen from the pool by selecting the item that provides the most information based on its provisional item difficulty and the examinee's momentary trait level estimate. Additional constraints based on content, cognitive complexity, item type, and other considerations may also be applied here.
5. After a particular field-test item has been administered to a prespecified number of individuals (small enough to allow the item difficulty estimate to change quickly throughout the calibration process), the provisional item difficulty estimate is updated using the desired item calibration procedure. This new provisional difficulty estimate is then used to calculate the item information used for further item selection.
6. Steps 4 and 5 are repeated until a prespecified number of item responses have been obtained, or until the provisional item difficulty estimate has stabilized to a predetermined level.
7. For the item in question, the final provisional item difficulty estimate becomes the operational item difficulty estimate (provided the item survives other statistical requirements.)
8. The process continues until all items are calibrated. In an operational setting, it is likely that there would be a queue of items that would be introduced to the testing process periodically, so that it would be a continuous process, rather than a batch process.

Example

If we have an item with a true difficulty of 0.5 on the trait of interest, examinees will respond to this difficulty regardless of our difficulty estimate. This allows us to use the item responses to hone our initial difficulty estimates so that they become a better representation of the true difficulty. Figure 1 shows the progression of difficulty estimates for an item with true difficulty of 0.5, initial provisional difficulty estimate of –.5, and a calibration group size of 25 examinees.

Figure 1. Item Difficulty Estimate and Distribution of Examinees' Final Trait Level Estimates as a Function of the Number of Examinees Taking the Item



As this item is field tested, the first 25 examinees to encounter the item are chosen as if the item has a difficulty of –.5. One of the first things that can be seen in this figure is that the examinees are distributed along the θ continuum around the provisional difficulty. This is due to the standard error of measurement of each examinee's trait level estimate at the time at which they encounter the field-test item. If the field-test item occurs as the 25th item in the individual's test, the momentary standard error of measurement will be approximately .5 times the original standard deviation of the trait in the population (assuming a deep item pool with perfect fit to the 1-parameter logistic model). By the end of the test, we have a more precise estimate, the final trait level estimate (identified as Theta in Figure 1), which may be used as the anchor point in calibration. The difference between the momentary trait level estimate used to select the field-test item, and the final trait level estimate used as the anchor for item difficulty estimation, provides the distribution of individual achievement around the provisional difficulty estimate which is required to obtain a stable estimate.

After the first 25 individuals take the item, its provisional difficulty estimate is updated based on the 25 responses received. In this example the bulk of the responses were incorrect, indicating that the item was more difficult than originally estimated. The new calibration estimate after 25 responses (1.2 in this case) is then applied to the item and used to collect the next 25 responses.

Following the 50th administration, all 50 item responses are used to calculate the new provisional difficulty estimate. In this case, the estimate of 1.2 caused examinees to respond correctly more than expected, and so the new provisional difficulty estimate becomes somewhat lower (.73).

This process continues until the number of item responses desired for final calibration is obtained (200, in this case) and the item difficulty estimate is finalized (.51, in this case). In this example, the provisional difficulty estimate stabilizes near the true value after the first 75 individuals see the item. This example helps to highlight some of the interesting characteristics of the adaptive calibration procedure, which are described in detail below.

Interesting Aspects of the Process

Initial provisional item difficulty estimates. In step 2 of the calibration process described above, it was mentioned that the initial value for a provisional calibration could be established by a content expert. By asking content experts to rate the difficulty of items, we might be able to start the calibration process with provisional calibrations that have a positive correlation with actual item difficulties. However, this process will only function well if content experts can estimate difficulty reasonably well.

The literature around this issue suggests two approaches to identifying provisional item difficulty estimates. In the first approach, individuals are asked to rate or rank the item statistics associated with a set of items after being given instruction on the meaning of those statistics. Two examples of this type of research are Bejar (1983) and Impara and Plake (1998). The outcome of this type of study tends to be a low but positive correlation between ratings and observed item statistics. The correlations tend to range from .00 to .70, varying widely with the content of the assessment and the specifics of the tasks that the raters were performing.

In the second approach, the individuals (or computer models) that provide the estimates of item statistics use characteristics of the items to help determine their difficulty. Examples of this approach are seen in work done on generative tests with item families (Bejar, 2003), model-based measurement systems that use item length and complexity to judge item difficulty (Burdick, Stenner, & Kyngdon, 2010), and systems that use difficulties of existing items with similar content to determine estimates for new items (NWEA, 2003). This approach tends to result in higher correlations between estimated and observed item statistics, ranging from approximately .40 to .90

It is clear that provisional estimates of item difficulty can be obtained prior to collecting field-test data that will provide some information. It is not as clear what the quality of the information will be. The simulation for this study assumes the low end of the observed range (a correlation of 0.0 between the initial provisional estimate of difficulty and the true difficulty) to provide the best test possible for the new procedure.

Adaptive processes with two unknowns. The intent behind the development of this field-testing process is to obtain more accurate estimates of item difficulty using smaller sample sizes. The process differs slightly from that used commonly in adaptive testing, because we don’t know the trait level of the examinee with great precision when the examinee is chosen for use with the field-test item.

We can estimate the SEM for each examinee who will take the item, based on the provisional difficulty of the item, the position of the field-test slot within the operational test, and the characteristics of the item pool. This information allows us to estimate and control the variance of the individuals who will take the item. In general, the earlier in a test that a field-test item appears, the greater the variance of trait level estimates will be around the item's provisional difficulty estimate. Some ramifications of this control mechanism will be described in the discussion below.

Simulation of the Calibration Procedure

Method

Adaptive test characteristics. Each adaptive test consisted of 52 items drawn from a pool made up of 1,000 operational items and 100 field-test items. 50 of the items in each test were operational, and two were field-test items. All operational items in the test were chosen using a randomesque maximum information strategy with a pool of the 10 most informative items (Kingsbury & Zara, 1989). Field-test items were selected according to the adaptive calibration procedure and were administered as item 22 and 28 in each test. Tests were scored using maximum-likelihood scoring. All items were assumed to fit the one-parameter logistic (1PL) IRT model. All item difficulties were drawn from a normal distribution with a mean of 0.0 and a standard deviation of 1.0.

One aspect of the calibration process simulated here is that it is designed to be used in an ongoing field-testing process using a field-testing queue. This means that as one item completes its field testing, it is turned off, and another item begins field testing in its place. In order to simulate this continuous field-testing process, items that completed their field testing in the simulation were still available for selection, even though the information did not influence the calibration. (In earlier versions of the simulation, the items for which field testing was completed were made unavailable for selection. This resulted in the remaining field-test items having a wider range of examinees, which resulted in less accurate calibrations.)

Initial provisional calibration accuracy. Since the degree to which item calibrations can be estimated from item characteristics varies greatly from one situation to another, the simulation used the bottom of the range of correlations between true item difficulty and item difficulty commonly seen in research (a correlation of 0.0). The initial provisional calibrations, therefore, had no relationship to true item calibrations, except that the mean and the standard deviation were the same.

Simulated examinees. Simulated examinees (sims) had true trait levels that were drawn from at random from a normal distribution with a mean of 0.0 and a standard deviation of 1.0. Each sim responded to the items administered according to the 1PL model. The number of sims in each replication varied according to the number needed to fully calibrate each item.

Design and replications. The two field-testing conditions in the study were random calibration and adaptive calibration. Each condition was simulated with 10 sets of 100 field-test items. This 10-replication design allowed the identification of consistent differences between conditions. Each operational item pool and set of field-test items was used in both the random selection and the adaptive selection conditions, to assure comparability. Each replication used a unique set of field-test and operational items. Each replication continued until each field-test item had received 500 responses.

Calibration procedure. Conditional maximum-likelihood was used to estimate item difficulties at each stage of the process. The final trait-level estimate for each sim was used as a fixed anchor for the calibration process, to assure that all estimates were on a common measurement scale. Each item was calibrated following each 10-item administration and the difficulty estimate obtained was used as the new provisional calibration. If a perfect response vector occurred during calibration, the new provisional calibration was set to the old provisional calibration plus .5 (in the case of a vector with no correct responses) or the old provisional calibration minus .5 (in the case of a vector with all correct responses.)

Analysis. Random sampling item calibration and adaptive item calibration were compared using the following criterion measures:

1. Calibration accuracy as a function of sample size.
2. Calibration bias as a function of sample size.
3. Conditional calibration accuracy.

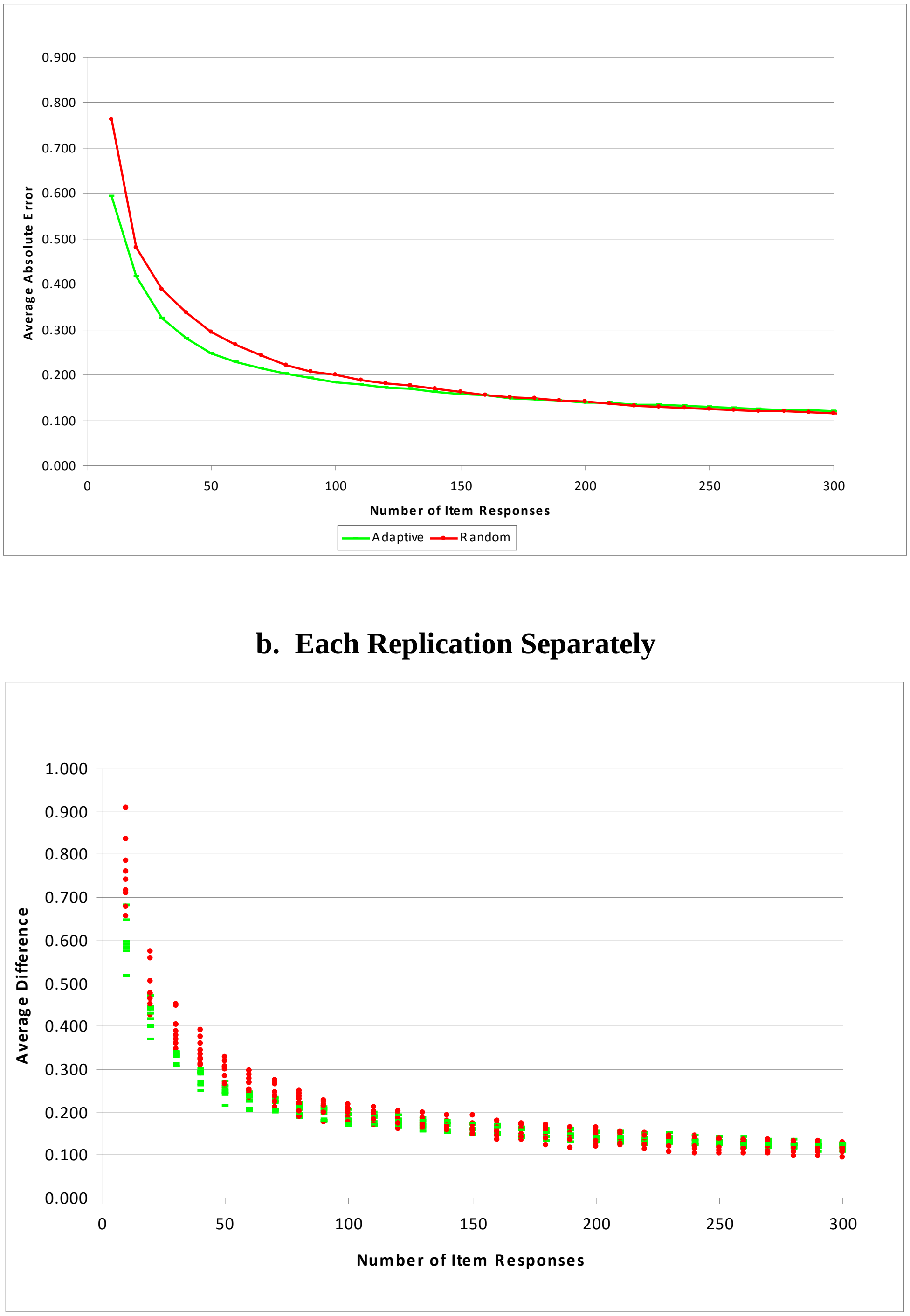
Simulation Results

Mean absolute differences. Figures 2a and 2b show the mean absolute difference between the observed and true item difficulties for random sampling and adaptive sampling as a function of the number of sims taking the item. Figure 2b shows each replication separately, while Figure 2a combines the results across all replications.

From Figure 2a it can be seen that the adaptive item calibration approach resulted in lower error for all sample sizes less than 150. For sample sizes greater than 150, the two approaches resulted in very similar levels of absolute error in item difficulty estimates. The largest difference in absolute error between the two approaches was observed for the smallest sample size (*N* = 10).

Figure 2b shows that the pattern of absolute error observed was fairly consistent across replications. For smaller sample sizes, each replication shows a smaller average absolute error in item difficulty estimates for the adaptive item calibration approach. For sample sizes greater than 100 there is less consistency in the differences observed. For sample sizes greater than 150, the average absolute errors for different replications from the two sampling approaches overlap completely.

Figure 2. Absolute Error in Item Difficulty Estimates as a Function of the Number of Item Responses



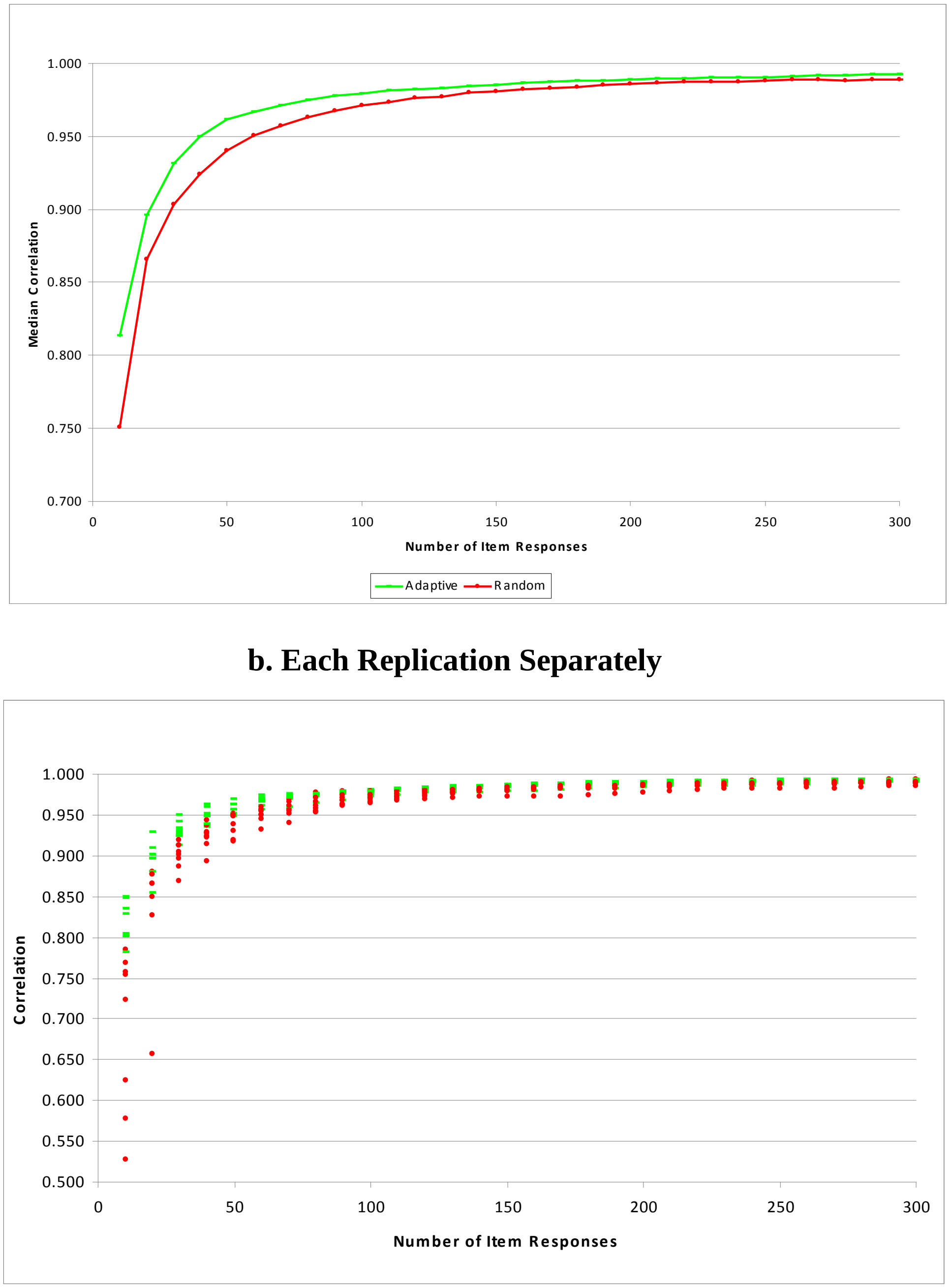
a. Average Across All Replications

b. Each Replication Separately

Fidelity coefficients. Figures 3a and 3b show the median correlation observed between true and estimated item difficulties as a function of sample size. Figure 3a shows the results across all replications, while Figure 3b shows the results from each individual replication.

Figure 3a shows that the adaptive item calibration approach resulted in a higher correlation between estimated and true item calibrations for all sample sizes. The largest differences in correlations were observed for the smaller calibration sample sizes but the direction of difference was consistent across all sample sizes studied. These differences may be interpreted in two ways, the absolute difference in correlations at a particular sample size, or the sample size needed to obtain the same correlation.

Figure 3. Correlation Between True and Estimated Item Difficulty as a Function of Number of Item Responses



a. Median Across All Replications

b. Each Replication Separately

For a

calibration sample size of 200 sims, the adaptive calibration approach resulted in a median correlation across replications of .989. The corresponding value for the random assignment approach was .986. This difference in correlations was .003, or 21% of the maximum difference possible from the lower value. The average percentage of the maximum difference possible across all sample sizes was approximately 27%, and varied little from the smallest sample sizes to the largest.

The adaptive calibration approach first reached a correlation of .989 at a sample size of 190 sims. The random assignment approach reached the same correlation at a sample size of 260 sims. Across all calibration sample sizes observed, the difference in needed calibration sample size needed to reach any given correlation ranged from 10 to 100 sims. The adaptive calibration approach required fewer sims at every correlational level. The difference in sample size required increased with the value of the correlation.

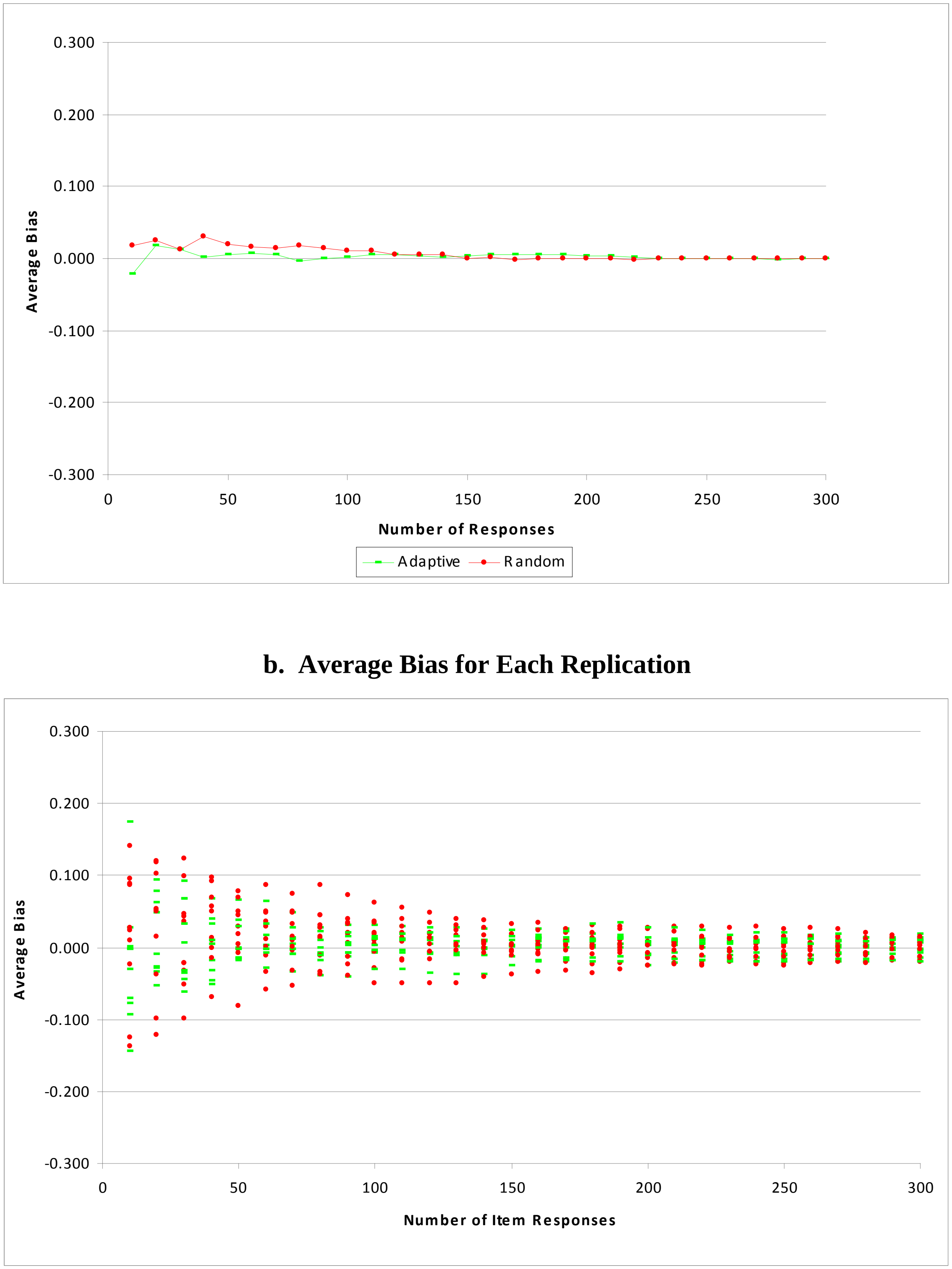
Figure 3b shows that in the individual replications, the pattern of correlations matched the pattern observed for the median across replications. Values observed for the random assignment approach tend to be somewhat more variable than those observed for the adaptive calibration approach, particularly for small calibration sample sizes. For all sample sizes, the adaptive calibration approach resulted in correlations that were more closely clustered and tended to be higher than those observed for the random assignment approach.

Bias. Figures 4a and 4b show the mean difference (bias) observed between the true and estimated item difficulties as a function of the calibration sample size across all replications and for each replication, respectively. Since a maximum-likelihood procedure was used to estimate item difficulties, substantial bias was not expected in the simulation and the results show little bias in the aggregate.

From Figure 4a, it can be seen that the overall bias across replications was less than .1 θ units, even with a sample size of 10 sims. The largest positive bias observed was .033, for the random assignment approach with 40 observations. The largest negative bias observed was .022, for the adaptive calibration approach with 10 observations. As the sample size increased above 100, neither approach resulted in overall bias greater than .010 θ units for any sample size.

Observing Figure 4b, it can be seen that the individual replications resulted in substantially more variability of bias. The largest bias value range observed was –.145 to .174, for the adaptive calibration approach with a calibration sample size of 10 sims. As the sample size increased, each approach resulted in less bias and more homogeneity of bias across replications. The two approaches had very similar patterns of bias across replications for all sample sizes.

Figure 4. Bias Between True and Estimated Item Difficulty as a Function of the Number of Responses

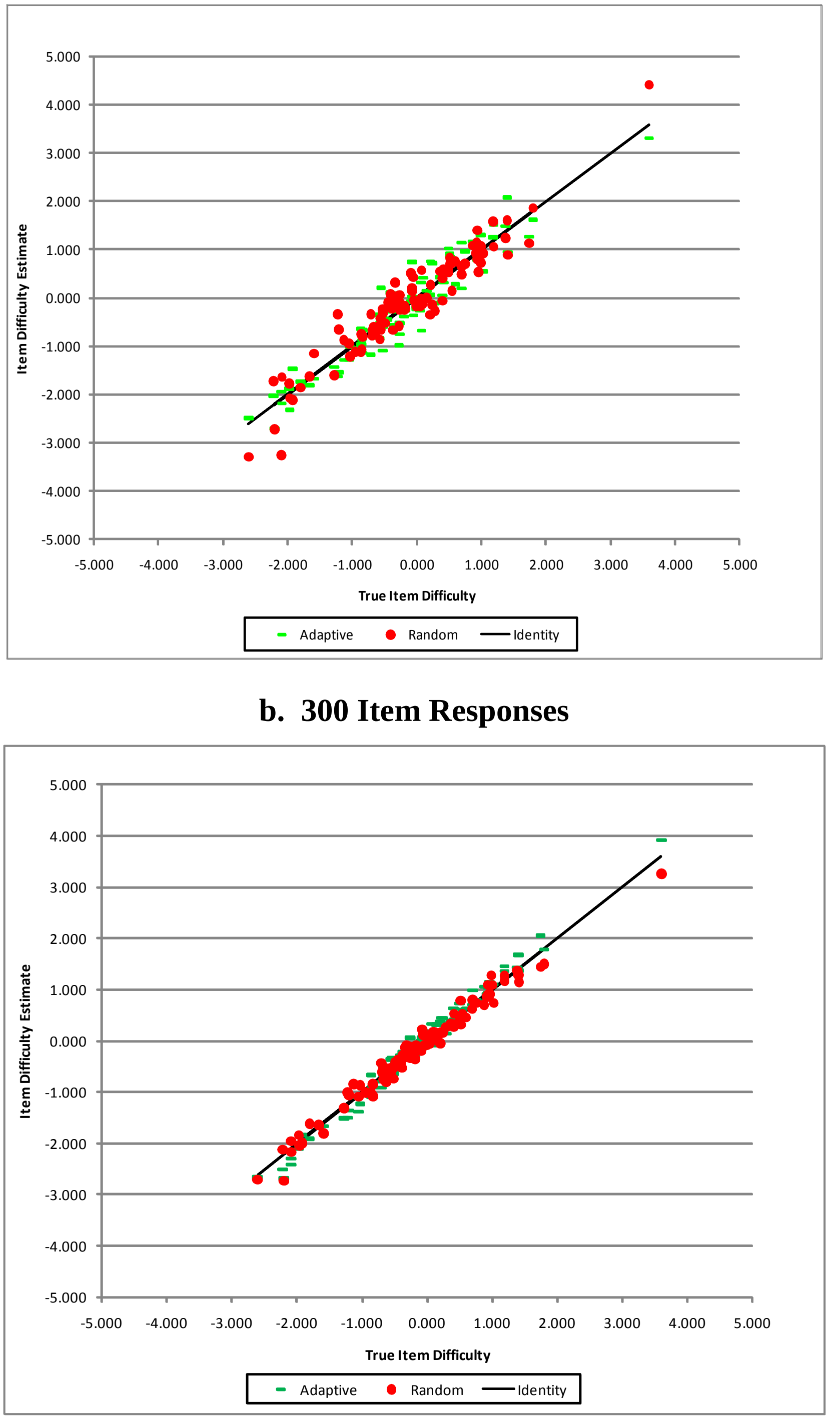


a. Average Across All Replications

b. Average Bias for Each Replication

Conditional accuracy. One characteristic of item difficulty estimation that is required is that the estimates be accurate at all observed trait levels. Figures 5a and 5b show the relationship between the true difficulty and the estimated difficulty for each item included in the first replication, after 50 responses and 300 responses respectively.

Figure 5. Relationship Among True and Estimated Item Difficulties for the First Replication



a. 50 Item Responses

b. 300 Item Responses

From Figure 5a, it can be seen that with a sample size of 50 sims, both the random assignment approach and the adaptive calibration approach resulted in modestly accurate item difficulty estimates across the range of the trait. The largest difference between true and estimated difficulty for the random assignment approach was −1.16 θ units for an item with a true difficulty of –2.09. The largest difference between true and estimated difficulty for the adaptive calibration approach was –.60 θ units for an item with a true difficulty of 0.01. While no substantial difference in the pattern of the relationship between true and estimated item difficulty was observed, the random assignment approach seemed to result in slightly more discrepant results for extreme trait levels.

Figure 5b shows that with a sample size of 300 sims, both the random assignment approach and the adaptive calibration approach resulted in more accurate item difficulty estimates across the range of the trait. The largest difference between true and estimated difficulty for the random assignment approach was –0.52 for an item with a true difficulty of –2.20. The largest difference between true and estimated difficulty for the adaptive calibration approach was –0.48 for an item with a true difficulty of –2.20 (the same item). For this sample size, no substantial difference in the pattern of the relationship between true and estimated item difficulty was observed for the two approaches.

Discussion

While the challenges 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 examinee selection in item calibration. This study was designed introduce a new procedure for examinee selection for item calibration. It was also designed to add to our information concerning the advantages and disadvantages associated with using different examinee selection approaches to facilitate item calibration within the context of an adaptive test.

In item pool development, each item costs a great deal to write and field test, and each item has a limited life span. If an item is exposed to fewer examinees while it is being field tested, its useful life span will increase. This is particularly true in high-stakes certification and licensure tests.

The common approach of randomly assigning items to examinees fails to take advantage of the characteristics of an adaptive test. An adaptive test is a very accurate method of assessing examinee capability with few items. The results here indicate that the use of an adaptive process to select examinees for specific items might enable us to assess the characteristics of an item with fewer people.

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 might be substantial.

This study suggests several lines of research concerning the selection of individuals to field-test items. The first is the obvious need for extension of the findings of this study into a live-data field trial. The second is a more thorough investigation of the characteristics and use of provisional item difficulties and their role in field testing. The third is an extension of the adaptive calibration model to allow items to be positioned within an adaptive test to take advantage of the conditional distribution of examinee trait levels. The use of provisional calibrations and the positioning of items according to the information needed warrants further discussion.

Several years ago, Mislevy (1988) suggested the use of collateral information for the Bayesian estimation of item parameters. If we can obtain reasonably accurate provisional item difficulty estimates (correlation with true calibrations of .40 or higher) and use these estimates within the adaptive calibration approach, this collateral information might enable more accurate calibration with a given field-test sample size. Gains might be 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.

The relationship between the position of a field-test item within the operational items on the test and the quality of the provisional calibration suggests a modification of the adaptive calibration approach that would take advantage of later item positions as the provisional calibration became more precise. As an example, the first 100 responses to an item might come from a field-test position early in the operational test. This would provide substantial variability in the trait levels of the examinees chosen. The next 100 responses could be obtained from a later field-test position, resulting in a smaller variance of examinee trait levels with the more precise provisional difficulty estimate. This process could proceed, with the item being tested in later positions, until the desired level of accuracy was obtained.

The adaptive calibration approach has been shown to have potential utility in parameter estimation in the context of an adaptive test. Some situations would benefit from this type of procedure more directly. In many certification and licensure situations, the number of candidates tested within a year is fairly small. This has been a stumbling block in moving these programs to adaptive testing because it is quite difficult to create an item pool of the necessary size, and it is difficult to maintain a fresh item pool by adding enough new items. the adaptive calibration approach might reduce some of these barriers to adaptive testing by allowing smaller sample sizes to provide more accurate item parameter estimates. While this is only an initial research effort, the results are promising enough to warrant further study.

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. (1983). Subject matter experts' assessment of item statistics. *Applied Psychological Measurement, 7,* 303-310.

Bejar, I. I., Lawless, R. R., Morley, M. E., Wagner, M. E., Bennett, R. E., & Revuelta, J. (2003). A feasibility study of on-the-fly item generation in adaptive testing. *Journal of Technology, Learning, and Assessment, 2(3).* Available from http://www.jtla.org

Buyske, S. G. (1998). Optimal design for item calibration in computerized adaptive testing: The 2PL case. *New Developments and Applications in Experimental Design, 34,* 115-125.

Burdick, D., Stenner, A. J., & Kyngdon, A. (2010). From model to measurement with dichotomous items. *Journal of Applied Measurement, 11,* 112-121.

Holman, R. & Berger, M. P. F. (2001). Optimal calibration designs for tests of polytomously scored items described by item response theory models. *Journal of Educational and Behavioral Statistics*, 26, 361-380.

Impara, J. C. & Plake, B. S. (1998). Teachers' ability to estimate item difficulty: A test of the assumptions in the Angoff standard setting method. *Journal of Educational Measurement, 35,* 69–81.

Kingsbury, G. G. & Houser, R. L. (1999). Developing computerized adaptive tests for school children. In Drasgow, F. and Olson-Buchanan, J. B. (Eds.) *Innovations in computerized assessment.* Mahwah, NJ: Lawrence Erlbaum Associates.

Kingsbury, G. G. & Zara, A. (1991). A comparison of procedures for content-sensitive item selection in computerized adaptive tests. *Applied Measurement in Education*, 4, 241-261.

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.

Mislevy, R. J. (1988). Exploiting auxiliary information about items in the estimation of Rasch item difficulty parameters. *Applied Psychological Measurement, 12,* 281-296.

NWEA (2003). *Technical manual for the NWEA Measures of Academic Progress (MAP) and Achievement Level Tests (ALT).* Portland, OR: Author.

Sands, W. A., Waters, B. K., & McBride, J. R. (1998). *Computerized adaptive testing: From inquiry to operation.* Washington, DC: American Psychological Association.

Swaminathan, H. & Gifford J. A. (1980). 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.

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.