ED 388 696 TM 023 943

AUTHOR Lunz, Mary E.; Bergstrom, Betty A.
TITLE Equating Computerized Adaptive Certification Examinations: The Board of Registry Series of Studies.
PUB DATE Apr 95
NOTE 28p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (San Francisco, CA, April 19-21, 1995).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Adaptive Testing; *Computer Assisted Testing; Decision Making; *Equated Scores; *Licensing Examinations (Professions); Medical Laboratory Assistants; *Medical Technologists; *Test Format
IDENTIFIERS Paper and Pencil Tests

ABSTRACT

The Board of Registry (BOR) certifies medical technologists and other laboratory personnel. The BOR has studied adaptive testing for over 6 years and now administers all 17 BOR certification examinations using computerized adaptive testing (CAT). This paper presents an overview of the major research efforts from 1989 to the present related to test equating. The comparability of both candidate ability estimates and pass/fail decisions on fixed-length paper and pencil (P&P) and CAT tests was initially confirmed in a study in which 645 prospective candidates took both PAP and CAT tests. Two additional studies were then completed using actual certification data. The first study divided the test population into two randomly equivalent groups, 1,669 taking a PAP and 1,699 a CAT examination. Items for both modes were equated to the benchmark scale on which the pass point had been established. Mean ability estimates, standard deviations, and pass rates were comparable. Finally, item recalibrations from CAT were studied with samples of 30, 50, and 100 examinees, and ability estimates correlated at 0.99. These studies confirm that equating with PAP or CAT item calibrations produces comparable candidate ability estimates. (Contains 2 tables, 3 graphs, and 35 references.) (SLD)

EQUATING COMPUTERIZED ADAPTIVE CERTIFICATION EXAMINATIONS:

THE BOARD OF REGISTRY SERIES OF STUDIES

Mary E. Lunz

American Society of Clinical Pathologists

Betty A. Bergstrom

Computer Adaptive Technologies, Inc.

2100 West Harrison Street

Chicago, Illinois 60612

Paper presented at the annual meeting of the National Council of Measurement in Education, San Francisco, CA, April, 1995.

EQUATIḶG COMPUTERIZED ADAPTIVE CERTIFICATION EXAMINATIONS:

THE BOARD OF REGISTRY SERIES OF STUDIES

Abstract

The Board of Registry (BOR) has studied adaptive testing for over 6 years, and now administers all 17 BOR certification examinations using computerized adaptive testing (CAT). This paper presents an overview of the major research efforts from 1989 to the present related to test equating.

The comparability of both candidate ability estimates and pass/fail decisions on fixed-length paper and pencil (PAP) and computerized adaptive tests (CAT) was initially confirmed in a study in which prospective candidates took both a computer adaptive and a paper and pencil test. Mean ability estimates, standard deviations, and pass rates were comparable across modes of administration using an item pool calibrated from PAP data.

Two additional studies were then completed using actual certification data. The first study divided the test population into two randomly equivalent groups. One group took their certification examination as a CAT, the other group as a PAP examination. Items for both modes of administration were equated to the benchmark scale on which the pass point had been established. The mean ability estimates, standard deviations, and pass rates were comparable across modes of administration.

Finally item recalibrations from CAT data were studied. A sampling design recalibrated items from samples of 30, 50 and 100 candidates and compared the results to baseline candidate ability estimates. Ability estimates correlated at .99. These studies confirm that equating with PAP or CAT item calibrations produces comparable candidate abili'y estimates.

EQUATING COMPUTERIZED ADAPTIVE CERTIFICATION EXAMINATIONS: THE BOARD OF REGISTRY SERIES OF STUDIES

Introduction

The Board of Registry (BOR) certifies medical technologists and other laboratory personnel. The goal is to make stable and accurate pass/fail decisions using multiple choice certification examinations. To this end, task analysis studies were completed and test specifications developed (see Lunz, Stahl, and James, 1989).

As part of the criterion referenced standard setting process, benchmark scales were constructed and calibrated using the Rasch model. Criterion standards were established on the benchmark scales. At that time, the examinations were administered using fixed length paper and pencil (PAP) examinations. It was the desire of the BOR to maintain the same criterion standards and benchmark scales when the mode of administration changed to computerized adaptive testing (CAT). A series of studies were initiated to determine the effect of using item pools calibrated from PAP examinations to produce equated computerized adaptive examinations.

Since 1985 the Rasch model (Rasch, 1960/1980) has been used to calibrate items and estimate candidate ability. The Rasch model was selected because the sample sizes for some of the tests are not large enough to meet the requirements of 2 and 3 parameter models (Lord, 1983). To insure that all candidates are measured against the same standard, the BOR used Rasch common item equating techniques. Thus, even before adaptive testing was considered, the BOR maintained calibrated benchmark scales on which criterion standards were established, to which examinations were equated.

Test Equating

The purpose of test equating is to place alternate test forms on the same scale so that equivalent performance standards can be implemented. Horizontal equating assumes that alternate test forms are similar with regard to the trait being measured and are intended to measure comparable ability levels. The equating process controls for item sampling differences and puts the items and candidates on the same scale (Baker, 1984). Candidate ability estimates are placed on the same scale after equating so that comparisons of candidate performances are possible, even though they took different test forms (Angoff, 1971).

Test equating techniques for fixed length PAP multiple choice examinations have been studied in some detail (Skaggs and Lissitz, 1981; Lord and Wingersky, 1984; Wright, 1977). Rasch model equating techniques have been found to be practical (Rentz and Bashaw, 1977; Wright and Bell, 1984) and robust even when all assumptions of the model are not met (Forsyth, Sarsangjan, and Gilmer, 1981).

Test equating issues with PAP and CAT are simultaneously the same and different, in that the purpose is the same, but the implementation of the technique is different. Both modes of administration within the certification environment use horizontal equating and rely on common items to assess item sampling differences. However, for PAP tests, equating usually occurs after the test administration while for CAT, equating occurs during test administration. For PAP examinations, alternate test forms are anchored to the benchmark scale on which the criterion standard is established using a common group of items. After the examination administration, the stability of the item calibrations is verified and items may be deleted from the anchor or from the scoring process. It is not necessary for all items to be calibrated to the benchmark scale prior to examination administration. All candidates take the same items, so the sample of candidate responses is consistent across items, regardless of item difficulty.

For CAT, all items presented to the candidate must be calibrated to the benchmark scale prior to administration of the examination. Because the examination is scored using the adaptive algorithm during the test administration, there is no post administration opportunity to verify the stability of the item calibrations before candidate scores are reported. All calibrated items link the adaptive test to the benchmark scale and all tests administered from the calibrated pool are automatically equated. This means that the item calibrations must have been shown to be stable and reliable prior to CAT administrations.

The possibility that item calibrations might change due to the mode of administration, PAP or CAT, has been discussed by several researchers (Kingbury and Houser, 1989; and Wise, Barnes, Harvey and Plake, 1989). Green, Bock, Humphreys, Linn and Reckase (1984) suggest several possible problems which might arise when items for a CAT are calibrated using data from a PAP test. Since, the Board of Registry wished to begin adaptive testing using the existing calibrated item pools, several studies were initiated to gain empirical evidence as to the stability of item calibrations from PAP examinations when the items are used on CATs.

General Description

Three studies are reported in this paper. The stability of candidate ability estimates is the focus of the studies, since the goal of certification is to make stable and accurate pass/fail decisions. All candidates are measured against the same criterion standard, on the benchmark scale to which examinations are equated.

The computerized adaptive testing model used in these studies is a mastery model (Weiss and Kingsbury, 1984) designed to determine whether a candidate's estimated ability is above or below a criterion standard or pass point on the calibrated benchmark scale. As previously mentioned, all items were calibrated with PAP certification examinations using the Rasch model and unconditional (joint) maximum likelihood estimation (UCON). The PROX formula (Wright and Stone, 1979) was incorporated into the item selection algorithm for on-line candidate ability estimation ($B_n = \bar{D}_i + \log(R/L-R)$, when B_n is the estimated ability of the candidate, $\bar{D}_i$ is the mean difficulty of the items presented to a candidate, R is the number of correct responses, and L is the length of the test. It was decided to use the computationally simpler PROX formula for on-line ability estimates because prior research confirmed the comparability of PROX and UCON candidate ability estimates (Bergstrom, 1990). Several versions of the CAT Administrator program (Gershon, 1990), each slightly more refined with regard to administrative capabilities, were used to administer CATs in these studies.

The test specifications were the same for all three studies reported in this paper. Content coverage was designed to be comparable to the test specifications for the PAP certification examination, and a content balancing mechanism of the type described by Kingsbury and Zara (1989) was included in the item selection algorithm. There were variations in test administration conditions for some studies (see Lunz and Bergstrom, 1994), but the examination equating techniques were essentially the same for all studies.

Study One: The Pilot Study

Study One was a pilot study accomplished with the cooperation of medical technology programs across the country. A Rasch calibrated item pool containing 726 multiple choice items from the six related subtests of this examination was developed using item calibrations from PAP examinations. Participants (N=645) agreed to take a PAP test and a CAT. Both tests were developed from the same calibrated item pool. After the test administration period, the item pool was recalibrated using data from the CAT administration. Thus, three ability estimates were collected for each participant, 1) fixed length PAP test (L=109 items); 2) variable length CAT from paper and pencil item calibrations (L=50-100 items); 3) variable length CAT from CAT item calibrations.

Results from this initial study were very encouraging and persuaded the Board of Registry to further pursue adaptive testing. When ability estimates were compared across modes of administration and source of item calibrations, no significant differences were found. Mean ability estimates on the fixed length PAP was .22 (SD = .48); mean ability estimates on the CAT from PAP calibrations was .23 (SD = .56) and on the CAT from CAT calibrations was .23 (SD = .53) (see Bergstrom and Lunz, 1994 and Bergstrom, 1992 for more details). The correlation of participant ability estimates PAP to CAT was .84 (corrected for error) and comparable pass/fail decisions were made for 77% of the participants (Lunz and Bergstrom, 1991). On the CAT, the correlation of ability estimates obtained from the PAP calibrations and the CAT calibrations was .99. Thus the candidate ability estimates were ordered comparably whether the item calibrations were obtained from PAP or CAT data.

The following additional observations were noted from this initial study. First, a difference in the standard deviation of ability estimates from fixed length PAP examinations and variable length CAT was noted with the CAT administration having the larger standard deviation (.48 vs .56). Second, more spread in the item calibrations was noted when items were recalibrated with CAT data (PAP SD=1.00; CAT SD=1.22). Third, due to the adaptive algorithm, item exposure varied greatly. This was most obvious at the ends of the scale. Few participants were administered the very easy and the very hard items. Thus in the future when only CAT data is available, items will be calibrated from unequal sample sizes.

The results of Study One, indicate that candidates were measured comparably across modes of administration when items were drawn from the same calibrated pool and persuaded the Board of Registry to make a definitive decision to proceed to computerized adaptive testing.

Study Two: Parallel Groups Study

During the transition from PAP tests to CAT, the opportunity to compare the performance of parallel groups of candidates taking CAT and PAP tests was made available. The medical technologist candidate population was divided into two groups for the 1993 administration of the certification examination. Candidates were assigned to take the PAP test or the CAT. This was a real certification examination, so the decisions counted. A group of 1,669 candidates took the PAP test and 1,699 took the CAT exam. As always, the PAP test was given to all candidates on the same day at 45 test sites. The CAT was administered during a 3 month period at 35 CAT sites across the country. Both the PAP and CAT items were calibrated to the same scale.

The PAP, a 200 item fixed length test, was equated after the examination using common item equating methods. The CAT was equated during the administration because all items presented to candidates were anchored to the benchmark scale. The CAT, was a fixed length 90 item adaptive test, tailored to the estimated ability of the candidate. Thus, each CAT presented a unique set of items, selected to provide a precise test for the candidate. The percentage of common items among candidate examinations varied (see Stahl and Lunz, 1993).

It was assumed in the study design that the PAP and CAT groups were equivalent because they had met the qualifications to enter the certification process and had been randomly assigned to groups. The results are presented in Table 1 and show that the two modes of administration yielded almost identical candidate results.

After the examination, the items were recalibrated using CAT data only, and a second ability estimate for each CAT candidate was calculated. This yielded two ability estimates for each candidate who took the CAT (N=1,699). The first ability estimate was equated to the benchmark scale (constructed from PAP data), the second ability estimate was not equated to the benchmark scale (CAT data collected in 1993 only). A comparison of the candidate ability estimates from the PAP and CAT item calibrations yielded a correlation of .99 indicating that the candidates were rank ordered comparably using item calibrations from PAP or CAT data.

The CAT item recalibrations were not equated to the benchmark scale (original PAP calibrations), so the pass/fail decisions were not comparable until the appropriate linear transformation, using an equating constant, placed the CAT item recalibrations onto the benchmark scale. After the adjustment, pass/fail decisions were comparable, 77% pass and 23% fail. However, 2.2% of the candidates, namely those in the error of measurement, altered status from pass to fail (1.1%) or fail to pass (1.1%). The SD for the ability estimates was noted to be larger when the CAT item recalibrations were used (PAP = .66 and CAT = .93). It was felt that the increased SD was due to the recalibration of items using candidate samples of different sizes and ability levels (due to test tailoring). The CAT data matrix contains a lot of missing data, since candidates were presented with selected groups of items based on the adaptive algorithm. Some items were presented to many candidates while other items were presented to very few candidates. This issue is addressed in the next study.

Study Three: Stability of Item Calibrations

The stability of Rasch candidate ability estimates, when tests are administered using CAT, depends in part, upon the stability of the item calibrations. Concerns about item calibration stability relate to the mode of test administration (PAP vs. CAT), the restricted range of candidate ability due to test tailoring, and the variability in the size of the calibrating sample. Previous research on the stability of item calibrations has been mixed. The research shown from Studies One and Two in this paper indicate that in actual testing, when items are calibrated from PAP data, candidate ability estimates remain stable when items are recalibrated with CAT data. Ito and Sykes (1994), however, found that when using simulated data, item difficulties were not well replicated when difficult items were calibrated using responses from able candidates and easy items were recalibrated using responses from less able candidates.

Before it becomes necessary for the BOR to recalibrate the benchmark scale and update the criterion standard, when only CAT data is available, it seems advisable to understand the impact of using CAT data only to calibrate an item pool. The hypothesis of this study is that candidate estimated ability will not be significantly affected when items are recalibrated on candidate samples using CAT data with its restricted range and varying sample sizes. Since each candidate sees a unique test, the impact of item drift may differentially impact candidates depending upon the particular set of items that are administered. However, it is expected that differences in item recalibrations will minimally affect candidate ability estimates due to the nature of the adaptive algorithm.

When the PROX formula is used to calculate candidate ability estimates from known item calibrations, as is the case with this computerized adaptive test, the effect of item calibration drift can be projected. The PROX formula estimates candidate ability (B_n) in CAT as:

$$B_n = \bar{D}_i + \log(R/L-R)$$

where B_n is the ability of the candidate, and $\bar{D}_i$ is the mean difficulty of the test items presented to a candidate; R is the number correct and L is the length of the test. Using the Rasch model, a drift in one or more item calibrations (seen in recalibration) alters the mean difficulty of the test presented to a candidate. For example, if 10% of the items on a particular test are recalibrated as more difficult by .10 logits, the potential change in a candidate's ability estimate can be calculated as:

10% x .10 logits = .01 increase in candidate ability estimate

In practice, some of the items in an item pool recalibrate as more difficult, while others recalibrate as easier. Any change in the candidate ability estimate relates to the change in the <u>mean</u> item difficulty of the it ms presented to the candidate.

Even radical changes, if they occur in only a few items will have little impact on candidate ability estimates as long as sufficient test length is maintained. For example, a 50 item CAT, may include 2 or 3 percent of the items that drift by as much as 1.00 logit. This will result in a minimal change in the estimate of candidate ability.

2% x 1.00 logit = .02 logit change in candidate ability estimate

3% x 1.00 logit = .03 logit change in candidate ability estimate

This change is less than the standard error measurement for a candidate taking a CAT of 50 items that is tailored to the current estimated candidate ability (SEM for 50 items = $(L/R*W)^{½}$ or $(50/25*25)^{½}$ = .28). Thus, from a theoretical perspective, item recalibration using CAT data should have a minimal impact on candidate ability estimates, even though candidates take different tests, some of which may have a higher or lower percentage of items that drifted.

Data from the CAT administration in 1993 were analyzed. Obviously, there was a lot of missing data in the data matrix since each candidate was presented with a unique set of items tailored to his/her ability. Items near the center of the scale were presented more frequently than items that were relatively easier or more difficult.

A baseline group of items was identified. The criteria for including items in the baseline were: 1) a minimum of 100 candidates answered the item, and 2) candidates who answered the item had minimum test lengths of 30 items (Linacre, 1994) from the baseline group of items. From the pool of 792 items, 92 items and 549 candidates met the criteria and were included in the baseline analysis that calibrated items and estimated candidate ability. The baseline sample was selected to provide the most accurate item calibrations and candidate ability estimates possible from the available data, so that comparisons with recalibrations and the original PAP benchmark calibrations would be meaningful. Because a subset of items from the pool was used, candidates had variable length tests depending, upon how many items from their CATs were included in the baseline sample of items. Test lengths ranged from 30-62 items. Item recalibrations were based on different candidate sample sizes and the candidate ability estimates were based on different numbers of items. The items were recalibrated using BIGSCALE, (Wright, Schultz, Linacre, 1993) a computer program for Rasch analysis of items and candidates. The baseline item calibrations were based on 113 - 395 candidate responses.

A major issue in calibrating items from CAT data is the range in the sample size among items. To explore the effect of decreased sample size, recalibrations using random samples of 30, 50, and 100 candidates from the original baseline sample of 549 candidates were selected. For the sample of 30 candidates, items were recalibrated based on the responses of 8 - 24 candidates. For the sample of 50 candidates, items were recalibrated based on the responses of 9 - 40 candidates, and for the sample of 100 candidates, items were recalibrated based on responses from 16 - 73 candidates. Obviously, the number of candidate responses used for the recalibration affects the error of measurement associated with the item calibration, but should have minimal impact on the relative difficulty of the items on the scale.

As a separate analysis, the 92 baseline items were then anchored to the calibrations from the benchmark scale, which were originally from PAP data. It was expected that the item and candidate calibrations would be ordered comparably, but that a linear transformation using an equating constant would be necessary to adjust the CAT recalibrations to the benchmark scale.

Results are reported in Table 2. Candidate summary statistics for the PAP calibrations and each CAT recalibration are presented. The mean candidate abilities and SEMs were comparable, but the standard deviations varied slightly. When ability estimates are recalibrated using the CAT derived item calibrations then compared to CAT ability estimates derived from PAP item calibrations, the correlation is .99. However, a linear transformation equating constant of 1.07 logits was needed to adjust the CAT scale and reproduce pass/fail decisions of the benchmark scale. The stability in item calibrations is, of course, reflected in the candidate ability estimates.

Similar correlations of .99 were found for ability estimates derived from item recalibration with small sample sizes. Zscore analysis identified no significant differences in candidate ability estimates due to item recalibration using less than 30, 50, or 100 candidate responses.

The correlation between item calibrations from the benchmark (PAP) scale and the baseline CAT recalibrations was .96, indicating that the ordering of the item difficulties was stable. However, calibrations differed by the equating constant (1.07).

Item recalibrations tend to drift when the size of the candidate samples on which they are recalibrated gets smaller and the estimates are less precise, but very few items drift by more than one logit. For the sample of 30 candidates, 10 items drifted by more than one logit; for the sample of 50 candidates, 5 items drifted by more than one logit; and for the sample of 100 candidates, only one item drifted by more than one logit. The ordering of candidate ability estimates are minimally affected by these drifts in calibrated item difficulty. As expected the correlations between item recalibrations and the baseline calibration improve as the number of candidates in the calibration sample increases.

Item recalibrations were relatively stable, even though different candidate samples were used for each recalibration. Even though item recalibrations varied slightly, candidate ability estimates were stable compared to the baseline estimates. The correlation of .99 with the baseline ability estimates, indicates that candidates were ordered comparably even after items were recalibrated. Candidate ability estimates did not change significantly, due to item drift, or imprecision of item estimation.

Conclusions from the Three Studies

It is important to remember that a calibrated item pool means that all test forms are drawn from the same pool and "automatically" equated to all other test forms drawn from that pool. For PAP this means linking and anchoring of selected items on tests. For CAT, this means linking and anchoring all test items. Candidates are essentially protected from unexpected changes in item recalibrations because test difficulty is accounted for during the administration of the CAT. During CAT administrations, items cannot be selectively anchored or deleted, all items are anchored to the benchmark scale through their calibrations, whether that scale was constructed from PAP or CAT data. For certification examinations, anchoring to the benchmark insures that candidates are measured against the criterion standard established on that scale.

For PAP tests, equating occurs after the examination. Candidates are protected from poorly performing items because they can be deleted or unanchored before the final scoring. This is important because PAP tests are fixed length and not tailored to candidate ability, so there may be some items that do not perform as expected.

The Rasch item calibrations have been shown to be stable when calibrated from the CAT or PAP data. Thus, for these tests, it is reasonable to use item calibrations from a PAP administration or CATs. When items are recalibrated using CAT data, despite its incomplete data matrix and restricted range, item recalibrations remain relatively stable, although the error of measurement associated with each item difficulty estimate increases as the sample size decreases. Also the passing standard must be established on the new recalibrated scale. Results presented by Ito and Sykes (1994) may have been due to the design of their simulation. In actual CAT administrations, some item overlap across candidate ability levels does occur (see Stahl and Lunz, 1993).

One interesting observation is that candidate ability estimates from CAT item calibrations tend to have larger standard deviations than PAP tests. While there is not sufficient evidence for a definitive explanation, one conjecture is that because item difficulty calibrations spread out at the ends of the scale, so do candidate ability estimates. Since only the most able or least able candidates are presented with the more extreme items, item calibrations at the ends of the scale may be more definitively calibrated due to test tailoring. While the SD increases on CAT for both item and candidate estimates, the ordering of candidates, and the pass/fail decisions are not affected. This phenomenon may be more crucial for CATs that report actual scores rather than pass/fail results only.

Interpretation of test equating requires a different perspective on PAP and CAT. On PAP tests, all candidates answer the same items. When tests are equated with the Rasch model, some candidates may answer items correctly by guessing, making these items inappropriate as anchors. The selection of items used to equate to the benchmark scale must be done carefully on PAP tests. The primary concerns when using CAT data for item calibrations are, accumulating adequate sample sizes for reasonably precise recalibration of items, and assessing the impact of restricted range of candidate ability on calibration. Guessing is less likely due to the tailoring algorithm.

This series of studies demonstrates that Rasch item calibrations from PAP or CAT data produce comparably ordered candidate ability estimates and can be used to create a calibrated item pool. In addition, there is evidence that item calibration drift is minimal. It should be noted that the items in this study were carefully constructed and reviewed to meet the test specifications and taxonomy requirements. In addition they were reviewed for quality of content and relevance to the field of practice. Overall, the item pools are very carefully constructed. These studies show that equating works regardless of the mode of administration under which the data are collected, especially when the item pools are carefully constructed.

REFERENCE

Angoff, W.H. (1971). Scales, norms and equivalent scores in R.L. Thorndike (ed.) Educational Measurement; American Council on Education, Washington DC.

Baker, F.B. (1984). Ability metric transformations involved in vertical equating under item response theory. Applied Psychological Measurement, 8, 261-271.

Bergstrom, B.A. (1990). Comparability of UCON and PROX estimation for CAT. ASCP Research Report 10. Chicago, IL: American Society of Clinical Pathologists.

Bergstrom, B.A. (1992). Ability measure equivalence of computer adaptive and pencil and paper tests: A research synthesis. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Bergstrom, B.A. and Lunz, M.E. (1994). The equivalence of Rasch item calibrations and ability estimates across modes of administration. M. Wilson (ed.) Objective Measurement Theory into Practice, 2, 122-128.

Folk, V.G. (1990). Adaptive testing and item difficulty order effects. Paper presented at the annual meeting of The American Educational Research Association, Boston, MA.

Forsyth, R., Sarsangjan, B. and Gilmer, J. (1981). Some empirical results related to the robustness of the Rasch Model. Applied Psychological Measurement, 8, 2, 213-218.

Gershon, R.C. (1990). CAT ADMINISTRATOR (Computer Program). Chicago: Computer Adaptive Technologies.

Green, B.F., Bock, R.D., Humphreys, L.G., Linn, R.L. and Reckase, M.D. (1984). Technical guidelines for assessing computerized adaptive tests. Journal of Educational Measurement, 21, 4, 347-360.

Ito and Sykes (1994). The estimation of item difficulties: Implications for a CAT implementation. Paper presented at the annual meeting of the National Council of Measurement in Education, New Orleans, L.A.

Kingsbury, G. and Houser, R. (1989). Assessing the impact of using item parameter estimates obtained from paper-and-pencil testing for computerized adaptive testing. Paper presented to the annual meeting of the National Council of Measurement in Education, San Francisco, CA.

Kingsbury, G. and Houser, R. (1990). Adapting adaptive testing: Using the MicroCAT testing in a local school district. Educational Measurement: Issues and Practice, 3-29.

Kingsbury, G. and Zara, A. (1989). Procedures for selecting items for computerized adaptive tests. Applied Measurement in Education, 2, 4, 359-375.

Kingston, N.M. and Dorans, J.J. (1984). Item location effects and their implications for IRT equating and adaptive testing. Applied Psychological Measurement, 8, 2, 147-154.

Linacre, J.M. (1994). Sample size and item calibration stability. Rasch Measurement Transactions, 7, 4, 328.

Lord, F.M. (1983). Small N Justifies Rasch Model. In David J. Weiss (ed.) New Horizons in Testing. Academic Press, Inc. New York.

Lord, F.M. and Wingersky, M.S. (1984). Comparison of IRT true-score and equipercentile observed-score "equatings". Applied Psychological Measurement, 8, 453-461.

Lunz, M.E. and Bergstrom, B.A. (1991). Comparability of decisions for computer adaptive and written examinations. Journal of Allied Health, 15-23.

Lunz, M.E. and Bergstrom, B.A. (1994). Computer adaptive testing: A national pilot study In Mark Wilson (Ed). Objective Measurement. New Jersey: Ablex.

Lunz, M.E., Stahl, J.A. and James, K. (1989). Content validity revisited: Transforming job analysis data into test specifications. Evaluation and the Health Professional, 12, 2, 192-206.

Olsen, J.B., Maynes, D.D., Slawson, D. and Ho, K. (1986). Comparison and equating of paper-administered, computer administered and computerized adaptive tests of achievement. Paper presented at the American Educational Research Association Meeting, San Francisco, CA.

Rasch, G. (1960/1980). Probabilistic models for some intelligence and attainment tests. Chicago: University of Chicago Press.

Rentz, L.R. and Bashaw, W.L. (1977). The national reference scale for reading an application of the Rasch model. Journal of Educational Measurement, Vol. 14, 161-178.

Rudner, L.M. (1989). Notes from Eric/TM. Journal of Educational Measurement Issues and Practice, 8, 4, 25-26.

Skaggs, G. and Lissitz, R. (1982). Test equating: relevant issues and a review of recent research. Paper presented at the annual meeting of the American Educational Research Association, Los Angeles, CA.

Stahl, J.A. and Lunz, M.E. (1993). Assessing the extent of overlap of items among computerized adaptive tests. Paper presented at the annual meeting of the National Council of Measurement in Education, Atlanta, GA.

Wainer, H. and Kiely, G. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement*, *24*, 3, 185-201.

Weiss, D.J. and Kingsbury, G.G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement*, *21*, 4, 361-375.

Wise, S.L., Barnes, L.B., Harvey, A.L. & Plake, B.S. (1989). Effects of computer anxiety and computer experience on the computer-based achievement test performance of college students. *Applied Measurement in Education*, *2*, 235-241.

Wright, B.D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement*, *14*, 97-116.

Wright, B.D. and Bell, S.R. (1983). Item banks: What, why and how. *Journal of Educational Measurement*, *21*, 4, 331-345.

Wright, B.D., Congdon, R. and Shultz, M. (1987). *MSCALE* (Computer Program). Chicago: MESA Press.

Wright, B.D., Linacre, J.M. and Schultz, M. (1990). *BIGSCALE* (Computer Program). Chicago: MESA Press.

Wright, B.D. and Stone, M.H. (1979). *Best Test Design*. Chicago: MESA Press.

Yen, W.M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement*, *5*, 2, 245-262.

TABLE 1

STUDY TWO: SUMMARY CANDIDATE STATISTICS IN LOGITS

	CAT	PAP
N Candidates	1699	1669
N Items	90	186**
Mean	.87	.82
SD	.66	.56
SEM	.22	.16
Separation R*	.89	.92
% Pass	77%	77%
% Fail	23%	23%

* Reliability of person separation: internal reliability statistic generally comparable to the KR-20

** After deletion of items

TABLE 2

STUDY THREE: SUMMARY STATISTICS

		CAT	CAT	CAT	CAT
Recalibration Sample Size	PAP	30	50	100	Baseline
Comparison of Candidates					
N Candidates	547	30	50	100	547
Candidate Mean	1.69***	.65	.63	.62	.63
Candidate SD	.49	.55	.52	.54	.54
Candidate SEM	.32	.32	.32	.32	.32
Candidate Separation R	.56	.65	.61	.65	.63
Zscore Analysis**					
N Candidates > ∓ 2.00	0	0	0	0	___
N Candidates < ∓ 2.00	0	30	50	100	___
Correlation to Baseline Ability Estimates	.99	.99	.99	.99	
Comparison of Items					
N Items	92	92	92	92	92
Item Mean*	1.07***	.00	.00	.00	.00
Item Calibration SD*	.46	.78	.75	.66	.53
Item SEM	.14	.60	.47	.33	.14
Zscore Analysis**					
N Items > ∓ 2.00	4	1	3	1	___
N Items < ∓ 2.00	88	91	89	91	___
Correlation to Baseline Item Calibrations	.96	.57	.74	.89	

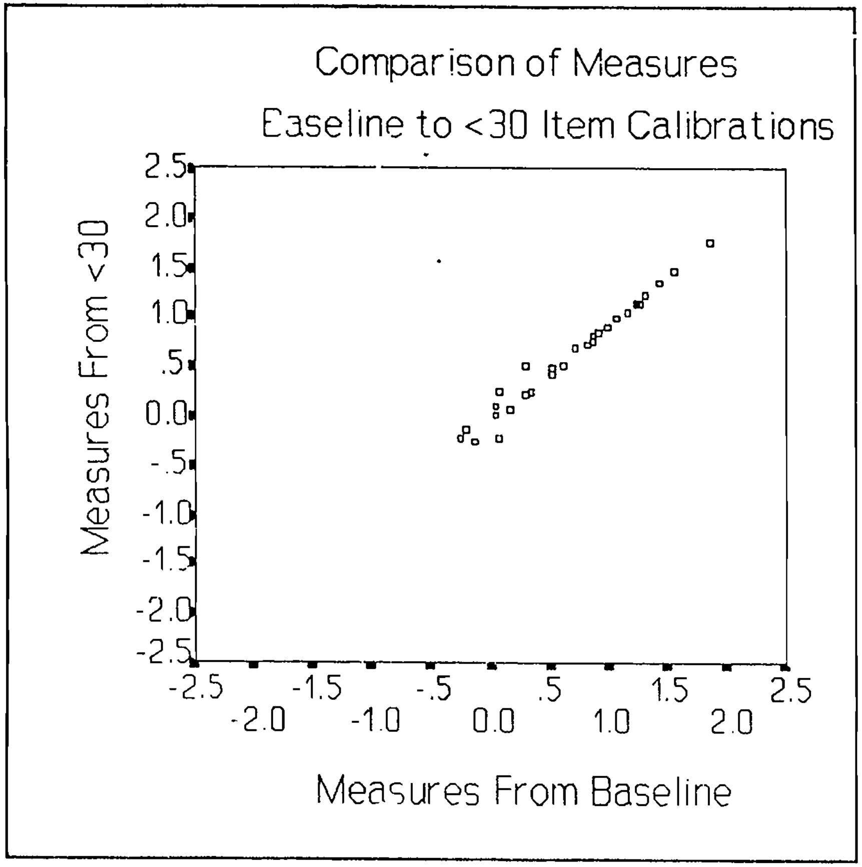
* Each recalibration was completed independently so the mean difficulty of the items was .00 - no equating

** Comparison of sample candidate ability estimates/item recalibrations to baseline candidate ability estimates and item calibrations

*** Equating value places the candidate ability estimate on the benchmark scale

N of Items = 92 of 726 items

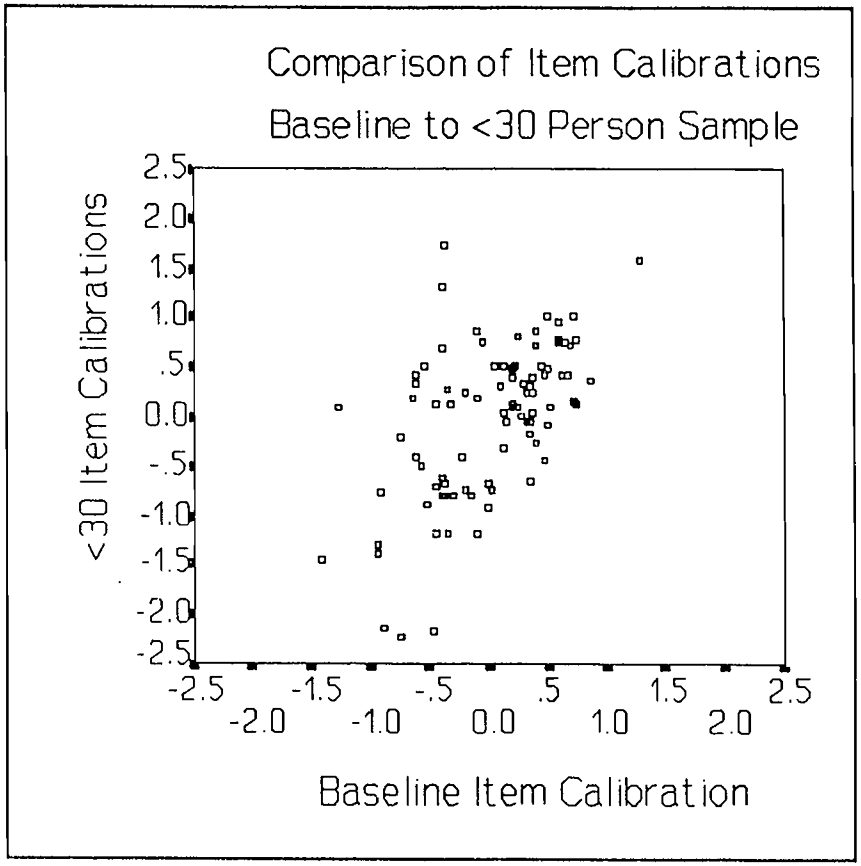
Graph 1



Comparison of Measures

Easeline to <30 Item Calibrations

Correlation = .99

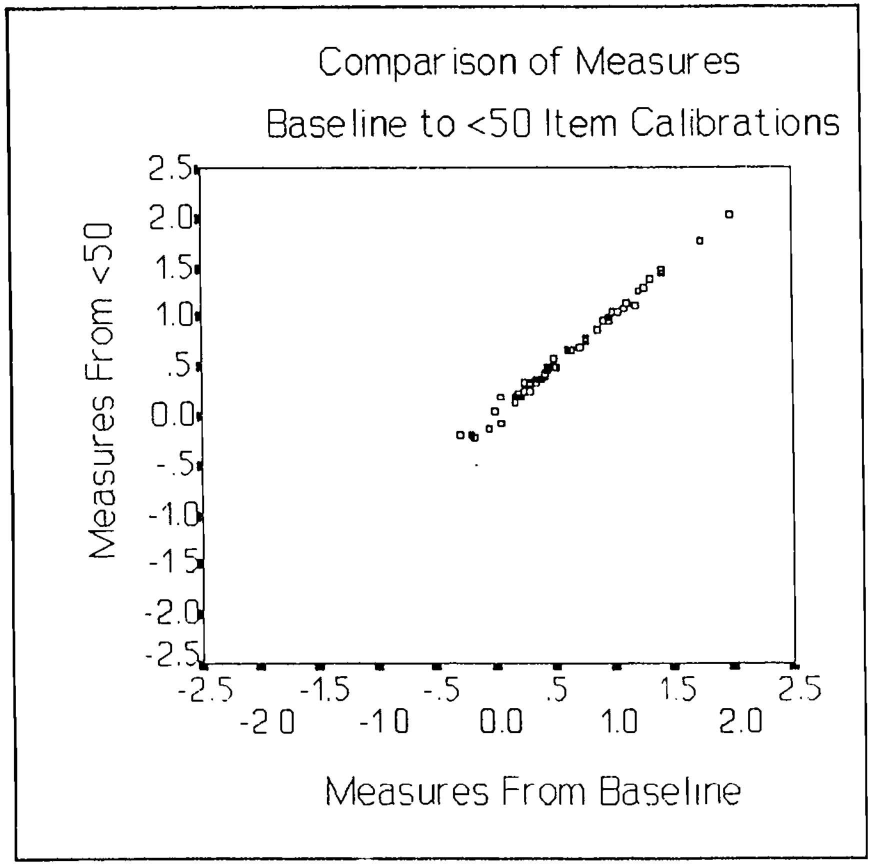


Comparison of Item Calibrations

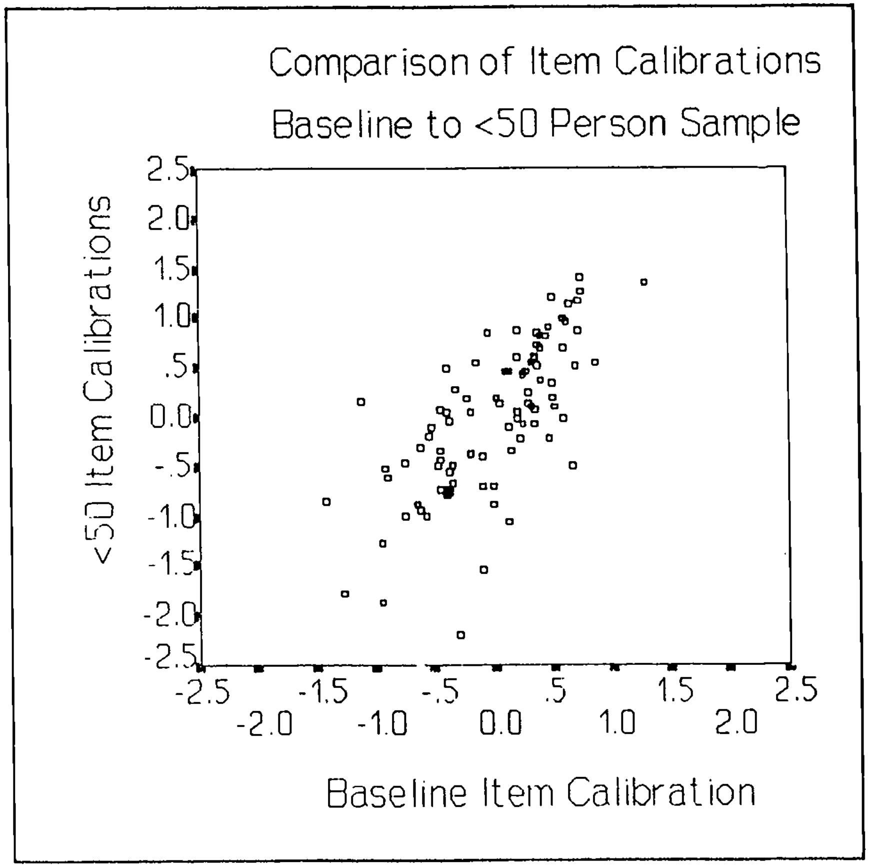
Baseline to <30 Person Sample

Correlation = .57

Graph 2

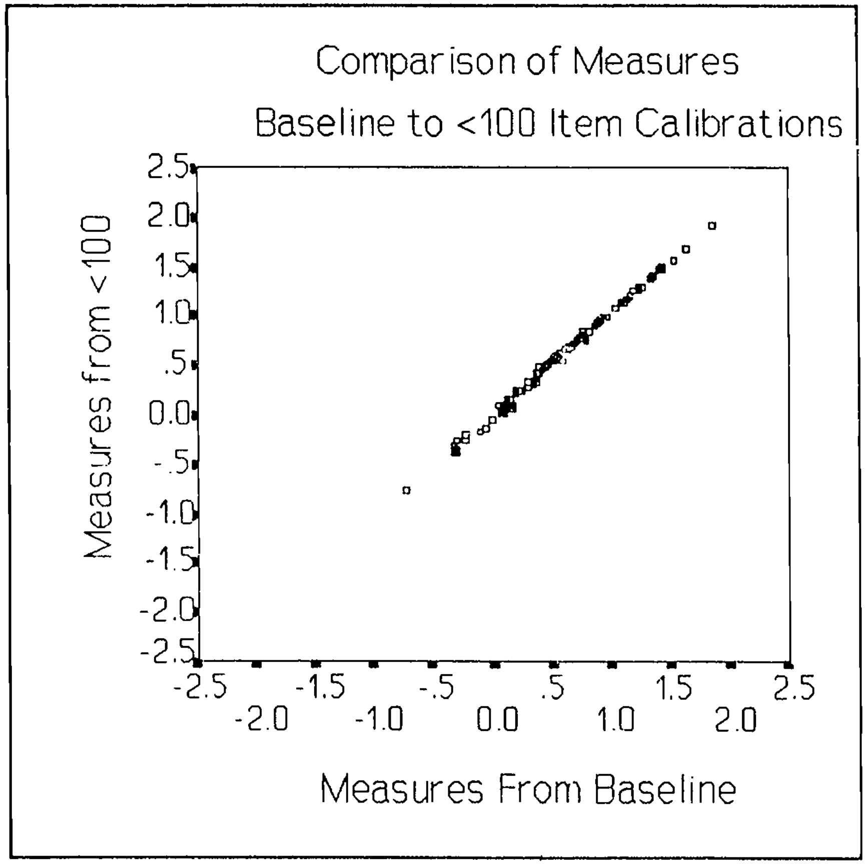

Comparison of Measures
Baseline to <50 Item Calibrations

Correlation = .99


Comparison of Item Calibrations
Baseline to <50 Person Sample

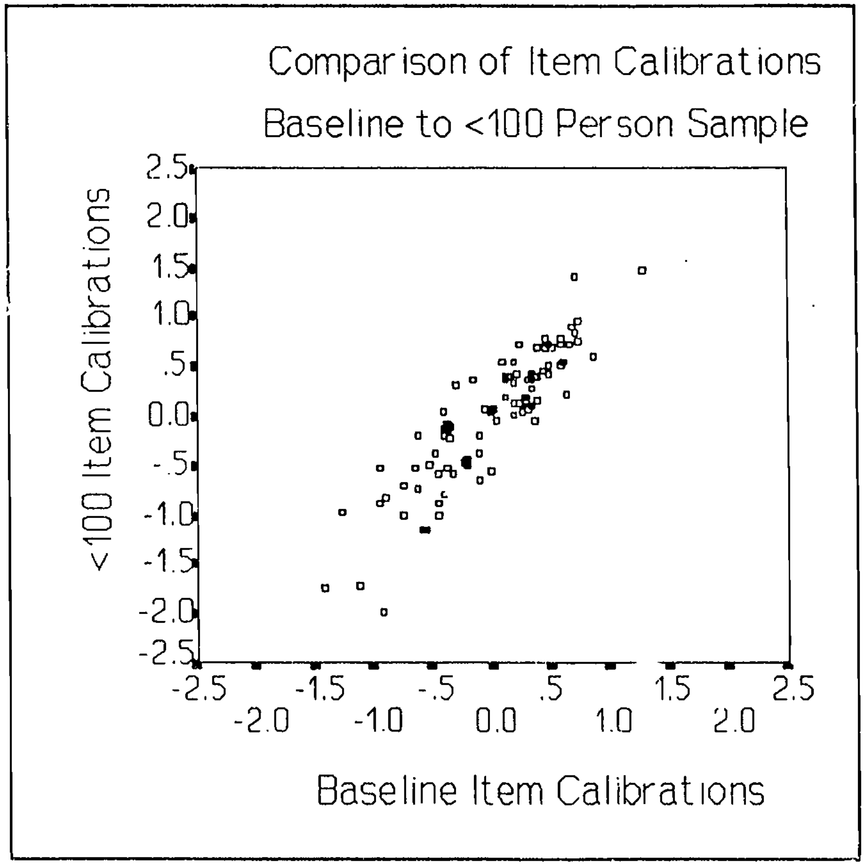
Correlation = .74

Graph 3



Comparison of Measures

Baseline to <100 Item Calibrations



Comparison of Item Calibrations

Baseline to <100 Person Sample

Correlatio

n = .99

Correlation = .89