DOCUMENT RESUME

ED 400 281 TM 025 549

AUTHOR Bergstrom, Betty A.; Lunz, Mary E.
TITLE Equivalence of Rasch Item Calibrations and Ability Estimates across Modes of Administration.
PUB DATE Apr 91
NOTE 21p.; Paper presented at the International Objective Measurement Workshop (Chicago, IL, April 1991).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Ability; *Adaptive Testing; Algorithms; Computer Assisted Testing; *Estimation (Mathematics); *Item Response Theory; *Medical Technologists; Test Format
IDENTIFIERS *Calibration; Item Selection; Paper and Pencil Tests; *Rasch Model

ABSTRACT

The equivalence of pencil and paper Rasch item calibrations when used in a computer adaptive test administration was explored in this study. Items (n=726) were precalibarted with the pencil and paper test administrations. A computer adaptive test was administered to 321 medical technology students using the pencil and paper precalibrations in the item selection algorithms and in the computation of examinee ability estimates. The response data from the computer adaptive test administration were analyzed yielding recalibrated item difficulties and examinee ability estimates. Item precalibrations were compared with item recalibrations. Examinee ability estimates obtained using the item precalibrations on the computer adaptive administration were compared with the examinee ability estimates obtained from using the item recalibrations on the computer adaptive administration. The correlation for examinee ability estimates was 0.99 and for item correlations it was 0.90. Some item calibrations shifted but most remained consistent within the limits of error. Item shift, however, did not affect the ordering of examinee ability estimates. (Contains 1 table, 3 figures, and 23 references.) (Author/SLD)

Equivalence of Rasch Item Calibrations and Ability Estimates Across Modes Of Administration

Betty A. Bergstrom
Mary E. Lunz
American Society of Clinical Pathologists
Board of Registry

Paper presented at the International Objective Measurement Workshop, Chicago, IL, April 1991.

Equivalence of Rasch Item Calibrations and Ability Estimates Across Modes of Administration

Abstract

The purpose of this paper is to explore the equivalence of pencil and paper Rasch item calibrations when used in a computer adaptive test administration. Items were precalibrated with pencil and paper test administrations. A computer adaptive test was administered using the pencil and paper precalibrations in the item selection algorithm and in the computation of examinee ability estimates. The response data from the computer adaptive test administration were analyzed yielding recalibrated item difficulties and examinee ability estimates. Item precalibrations were compared with item recalibrations. Examinee ability estimates obtained using the item precalibrations on the computer adaptive administration were compared with the examinee ability estimates obtained using the item recalibrations on the computer adaptive administration. The correlation for examinee ability estimates was .99 and for item calibrations was .90. Some item calibrations shifted but most remained consistent within the limits of error. Item shift, however, did not affect the ordering of examinee ability estimates.

Key words: CAT, Rasch model, item calibration,

Equivalence of Rasch Item Calibrations and Ability Estimates

Across Modes of Administration

When a computer test is adaptive, a participant is administered items based on a current ability estimate. When an item is answered correctly, an ability estimate is calculated and a more difficult item is presented. When an item is answered incorrectly, a lower ability estimate is calculated and an easier item is presented. The most informative and hence the most useful items are presented to each examinee so that responses to fewer items are required to achieve the same level of precision. In order for an item to be used efficiently with the computer adaptive algorithm, it must be precalibrated using a latent trait model such as the Rasch model which orders items from easy to difficult. A pencil and paper administration or a previous computer adaptive administration can be used for item precalibration.

Many organizations have item pools calibrated from previous pencil and paper administrations. However, the use of these calibrations for a computer adaptive test needs careful consideration. Since the mode of administration is different, there is a possibility that items are somehow "different" when presented on a computer instead of on a piece of paper. If items are "different", pencil and paper calibrations may not be appropriate for a computer adaptive test. In a computer adaptive test each examinee takes an individualized test. Items are presented to different examinees in different contexts and at different points during the test administration. Thus context effects and location effects are potentially different for each examinee. In a paper and pencil test item location and context do not fluctuate. If the location and/or context affect the item calibration, the paper and pencil calibration may not be appropriate for a computer adaptive test.

The possibility that item precalibrations might change due to the mode of administration, namely, paper and pencil vs. computer adaptive, has been discussed by several researchers (Kingsbury and Houser, 1989, and Wise, et al., 1989). Green, Bock, Humphreys, Linn and Reckase (1984) suggest several possible problems which might arise when items for a computer adaptive test are precalibrated using data from a paper and pencil test. An overall shift might occur, such that all items become easier or harder, or an "item by mode interaction" might occur where some, but not all, item parameters change. They postulate that items with diagrams or many lines of text may be most vulnerable to an item by mode interaction.

Context effects have been addressed by Kingston and Dorans (1984). They note that the appropriateness of IRT equating based on precalibration requires that changes in position of items in a test between the preoperational calibration and operational administrations of the test have no effect on item parameter estimates. They found some types of complex items, especially those which require extensive instructions, to be particularly sensitive to location effects and thus possibly unsuitable for computer adaptive administration. Yen (1980) also found item characteristics to be affected by the sequence in which items were administered.

One of the consequences of targeting items to the ability level of the examinee is that examinees of different ability levels may be presented with items in different difficulty order. Folk (1990) points out that a high ability examinee will generally answer the initial items of a computer adaptive test correctly and then will receive more difficult items. This results in his test being structured from easy to hard. A low ability examinee will not answer the initial items correctly which results in his test being structured from hard to easy. However, Folk found that the administration of items in different orders did not affect substantially the performance of low or high ability examinees.

Other potential problems in precalibrating items with a pencil and paper test for computer adaptive administration have been addressed by Wainer and Keily (1987). One of these is the differential effect of cross information encountered in computer adaptive testing. If a paper and pencil item provides a cue for another item, all examinees receive the same cue. With a computer adaptive test, examinees are administered different items and items are ordered differently. If an item calibration is influenced by a cueing effect in a pencil and paper administration it may be invalid for the computer adaptive administration. They also point out that one of the virtues of computer adaptive testing--short test length-- may become problematic if item calibrations are unstable. Since the shorter test lacks the redundancy of a conventional test it will be more vulnerable to idiosyncrasies of item performance.

If items have not been precalibrated, an initial pencil and paper administration may be practical for those considering computer adaptive testing. In this case, the size and composition of the sample needed for precalibration of items must be considered. It has been suggested that the sample include a minimum of 1,000 respondents and be comparable to the target population (Rudner, 1989, Green et al., 1984). However, it may be difficult to amass a comparable sample population this large.

Computer adaptive testing has been shown to reduce test length without loss of precision (Weiss, 1983, 1985; Weiss and Kingsbury, 1984; McKinley and Reckase, 1980, 1984; Olsen, et. al., 1986). When the items presented are targeted or tailored to the ability of the examinee, fewer items are required to estimate ability or reach a pass/fail decision. This assumes that the item difficulty calibrations are equivalent regardless of the mode of administration under which the calibrations were obtained.

Purpose

The purpose of this paper is to explore two related issues to determine whether item precalibrations from pencil and paper tests are appropriate for use in computer adaptive testing. The first issue is the equivalence of item calibrations from paper and pencil and computer adaptive administrations. The second issue is the equivalence of examinee ability estimates when item precalibrations from paper and pencil tests versus item calibrations from computer adaptive tests are used for the tailoring algorithm.

Method

Precalibration

Three hundred and twenty-one medical technology students, from 57 educational (training) programs across the country provided data for the precalibration of 726 items. To participate, students had to be eligible to take the first semi-annual administration of the related certification examination.

Each student took one of four different forms of a 200 item conventional pencil and paper test. Each form included a subset of common items for equating so that all forms could be placed on the same scale. Form 1 was taken by 73 students, Form 2 by 86 students, Form 3 by 71 students, and Form 4 by 91 students. Each of the four forms was calibrated by the Rasch model program MSCALE (Wright, Congdon and Schultz, 1987). The analysis of the infit[1], a statistic designed to assess the suitability of the data for constructing ability estimates, revealed that low ability individuals did not answer items correctly more than statistically expected. This confirmed that the precalibrations could be used in the CAT algorithm. The forms were equated using common item equating (Wright and Stone, 1979) and item precalibrations for the 726 items were established.

CAT Administration

The computer adaptive test administration used 1187 students from 238 educational programs across the country. To participate, students had to be eligible to take the second semi-annual administration of the related

[1] infit statistic is the information weighted mean-square residual that is sensitive to an accumulation of central or inlying deviations. The expected value for the mean squares is one (1.0) and their asymptotic standard errors are approximately the square root of (2/df) where df is the number of independent replications on which the corresponding estimate is based.

certification examination. Programs who participated in the precalibration were not eligible to participate in the computer adaptive test. The two samples were considered comparable.

The computer adaptive testing model used in this study has the following characteristics. It is designed as a mastery model (Weiss and Kingsbury, 1984) to determine whether a person's estimated ability level is above or below a pre-established criterion expressed in the metric (logits) of the calibrated item pool scale. Kingsbury and Houser (1990) have shown that an adaptive testing procedure which provides maximum information about the examinee's ability will provide a more clear indication that the examinee is above or below the pass/fail point than a test which peaks the information at the pass/fail point.

The test was stopped when the examinee ability estimate was either 1.3 times the error of measure above the pass/fail point (a clear pass--one tailed 90% confidence level) or 1.3 times the error of measure below the pass/fail point (a clear fail) *or* when a maximum test length of 240[2] items was reached. Minimum test length was 50 items and the pass/fail point was set at .15 logits.

The Rasch model (Rasch, 1960/1980) was used to calibrate items and estimate person ability. The Rasch model was selected because the sample sizes were not large enough to meet the requirements of 2 and 3 parameter models (Lord, 1983) and there is evidence that the examinee abilities estimated with the Rasch and the 2 and 3 parameter models correlate highly (.90) when tests are administered under a computer adaptive algorithm

[2] In this study 240 items was set as the maximum because it is comparable to the current paper and pencil test. The intent was to verify that fewer items were necessary, but this required that a relatively high number of items be allowed.

(Olsen, et al, 1986). Since the goal is accurate examinee ability estimates, the Rasch model was able to provide sufficient and complete information. The PROX version of the maximum likelihood method of item selection (Wright and Stone, 1979) was used in the adaptive algorithm.

The calibrated item pool contained 726 multiple choice (4-choice) items from six subsets. Table 1 provides a summary of the item pool distribution. Content coverage was designed to be comparable to the test specifications for the conventional paper and pencil certification examination and a content balancing mechanism of the type described by Kingsbury and Zara (1989) was included in the item selection algorithm. In the first 50 items, blocks of ten items were administered from subsets 1-4 and blocks of 5 items were administered from subsets 5 and 6. After 50 items, blocks of 4 items (subsets 1-4) and blocks of 2 items (subsets 5 and 6) were administered. Subset order was selected randomly by the computer algorithm. Maurelli and Weiss (1981) found subtest order to have no effect on the psychometric properties of an existent achievement test battery.

INSERT TABLE 1 ABOUT HERE

The computer adaptive test administrator program (Gershon, 1989) implemented the adaptive algorithm and the content balancing requirements. Items were chosen at random from unused items within .10 logits of the targeted item difficulty within the specified content area. While the examinee considered the item presented, the computer selected two items, one which would yield maximum information should the current item be answered incorrectly and another which would yield maximum information should the current item be answered correctly. This insured that there was no lag time before the next item was presented. Examinees were allowed 4 hours to complete the test.

Recalibration from CAT Administration

To determine the equivalence of item calibrations and to determine whether shifts in item calibration affect examinee ability estimates, the response data from the computer adaptive test administration were recalibrated. Each computer adaptive test yielded an examinee response string. While the entire item pool consisted of 726 items, each examinee response string contained responses from between 50 items (minimum test length) to 240 items (maximum test length). Each item had a unique identifying number. Response strings from all examinees were appended, resulting in a file containing an 1187 (examinee) by 726 (item) matrix with missing data for all items not presented to a particular examinee.

Figure 1 shows the frequency of item use on the CAT compared with the difficulty of the item precalibration. The mean number of examinees to whom an item in the CAT was administered was 161.62 with a standard deviation of 88.43. The minimum number of examinees was 13 and the maximum number of examinees was 382. Items with calibrations between -1 and 1 logits were administered most frequently. Thus the number of examinees used to recalibrate each item after the CAT administration varied considerably.

INSERT FIGURE 1 ABOUT HERE

The 1187 by 726 response matrix was analyzed with BIGSCALE (Wright, Linacre, and Schultz 1989) an update of the MSCALE program which processes large data sets that have missing data. Examinee ability estimates were not held constant but recalculated based on the item recalibrations. This procedure produced a new set of item calibrations and a new set of examinee ability estimates based upon responses from the CAT administration only. The fit of the item recalibrations to the model was reviewed, and the items were found to be suitable for constructing ability estimates. The infit statistic for all items indicated that low ability examinees did not get items correct more frequently than statistically expected.

Comparison of Item Calibrations.

The item precalibrations, obtained from the pencil and paper test, were compared with the item recalibrations from the CAT administration. Item precalibrations were based on sample sizes of 71 to 91 examinees. Item recalibrations were based on sample sizes of 13 to 382 examinees. Summary statistics, and correlations were calculated. Then the recalibrated item distribution was adjusted for the difference in means and standard deviations and the logit differences were calculated.

Comparison of Ability Estimates

During the CAT administration the computer adaptive algorithm calculated examinee ability estimates based on the pencil and paper precalibrated items. In the recalibration, examinee ability estimates and item recalibrations were calculated simultaneously from the response data collected during the CAT administration. The examinee ability estimates obtained in the CAT administration using the paper and pencil item precalibrations were compared with the examinee ability estimates obtained after the item recalibration. Summary statistics, correlations and logit differences were calculated.

Results

Item Calibrations

The correlation for pencil and paper item precalibrations and the computer adaptive test recalibrations was .90 indicating that some shift did occur for some of the items (See Figure 2). Precalibrations ranged from -3.61 to 3.84 logits, with a mean of -.02 and a standard deviation of 1.00.

INSERT FIGURE 2 ABOUT HERE

Recalibrations ranged from -3.84 to 3.60 with a mean of 0.00 and a standard deviation of 1.22.

Two types of shift occurred. The first is an overall shift, indicated by the difference in the mean and standard deviation of the item precalibrations compared to the mean and standard deviation of the item recalibrations. The spread of the item recalibrations (S.D. =1.22) is wider than the spread of the item precalibrations (S.D. =1.00). The effect of the difference in standard deviations is that the hard items appear harder when recalibrated and the easy items appear easier.

The second type of shift occurred with specific items. After the distribution of the recalibrated items is adjusted for differences in the two means and standard deviations, some item calibrations still shift and the order of item difficulty is altered. A few items recalibrate as more difficult than they did originally on the precalibration and a few items recalibrate as less difficult.

The shifts in item calibrations from precalibration (small sample, pencil and paper administration) to recalibration (varying sample per item, computer adaptive administration) may be due to the mode of administration or to item bias (a difference in the intent or preparation between the precalibration sample population and the recalibration sample population). For example, of the 7 items with the largest shift in the direction of easier on the recalibration, 5 were from the same content area, indicating possible differential preparation between the two sample populations.

Examinee Ability Estimates

The mean ability estimate calculated with precalibrated items was .21 with a standard deviation of .51. The mean ability estimate calculated with recalibrated items was .24 with a standard deviation of .58. The mean logit difference between ability estimates was -.03 with a standard deviation of .10.

The correlation of the examinee ability estimates calculated using the paper and pencil item precalibrations and the examinee ability estimates calculated using computer adaptive item recalibrations was .99. Figure 3 shows that there is virtually no difference between the examinee ability estimates using item precalibrations (paper and Pencil) or recalibrations (computer).

INSERT FIGURE 3 ABOUT HERE

Discussion

In this study, even though the precalibrations were obtained from a pencil and paper administration with relatively few participants, most of the Rasch item calibrations remained stable when recalibrated from the computer adaptive administration. The results demonstrate that items precalibrated in a pencil and paper administration can be used for computer adaptive tests. The item calibrations were equivalent given varying numbers of examinees, different contexts, and varying modes of administration. The pencil and paper precalibrations used a sample of examinees of varying ability levels so each item was calibrated from a range of examinee abilities. Items on the computer adaptive administration for each item were targeted to the examinee ability so the recalibrations were based on a smaller range of ability levels.

Two types of shifts occurred in the item calibrations. The first type, an overall shift in mean and standard deviation, can be corrected for by using an equating transformation. The second type of shift, a shift in the calibration of certain items is potentially more problematic because examinees take different items. This means that when some items shift, some examinees may be differentially affected depending upon how many of the shifted items are presented to them.

The examinee ability estimates correlation of .99 indicates that even though a small percentage of the item calibrations shift, the examinee ability estimates are not affected. No examinee ability estimates differed beyond the variance expected due to error of measure. However, if shift in item calibration is a concern, the items can be identified and revised or discarded for subsequent CAT administrations. The examinee ability estimates however, can be considered valid even if it is necessary to re-evaluate some items.

This study confirms that organizations desiring to utilize the technology of the computer for administering tests can do so without the expense of recalibrating their item pools. Of course, the item pool must be continually monitored for drift, validity and quality of item content whether tests are administered in a paper and pencil or computer adaptive mode.

REFERENCE

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.

Gershon, R.C. (1989). CAT Administrator (Computer Program) Chicago: Micro Connections.

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.

Kingsbury, G.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.G. and Houser, R.L. (1990). Assessing the utility of item response models: Computerized adaptive testing. Paper presented to the annual meeting of the National Council on Measurement in Education, Boston, MA.

Kingsbury, G.G. and Zara, A.R. (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.

Lord, F.M. (1983). Small N justifies Rasch model. In David J. Weiss (Ed.) New Horizons in Testing. Academic Press, Inc., New York.

McKinley, R.L. and Reckase, M.D. (1980). Computer applications to ability testing. Association for Educational Data Systems Journal, 13, 193-203.

McKinley, R.L. and Reckase, M.D. (1984). Implementing an adaptive testing program in an instructional program environment. Paper presented at the meeting of the American Educational Research Association, New Orleans.

Maurelli, V.A. and Weiss, D.J. (1983). Factors influencing the psychometric characteristics of an adaptive testing strategy for test batteries (Research Report 81-4). Minneapolis: University of Minnesota Department of Psychology, Psychometric Methods Program, Computerized Adaptive Testing laboratory.

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.

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

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. (1983). New horizons in testing: Latent trait test theory and computerized adaptive testing. Academic Press Inc., New York.

Weiss, D.J. (1985). Final Report: Computerized Adaptive Measurement of Achievement and Ability (Project NR150-433, N00014-79-C0172 Minneapolis, MN: University of Minnesota.

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 Ability estimatesment in Education, 2, 235-241.

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

Wright, B.D., Linacre, J.M. and Schultz, M. (1989). BIGSCALE (Computer Program). Chicago: MESA Press.

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

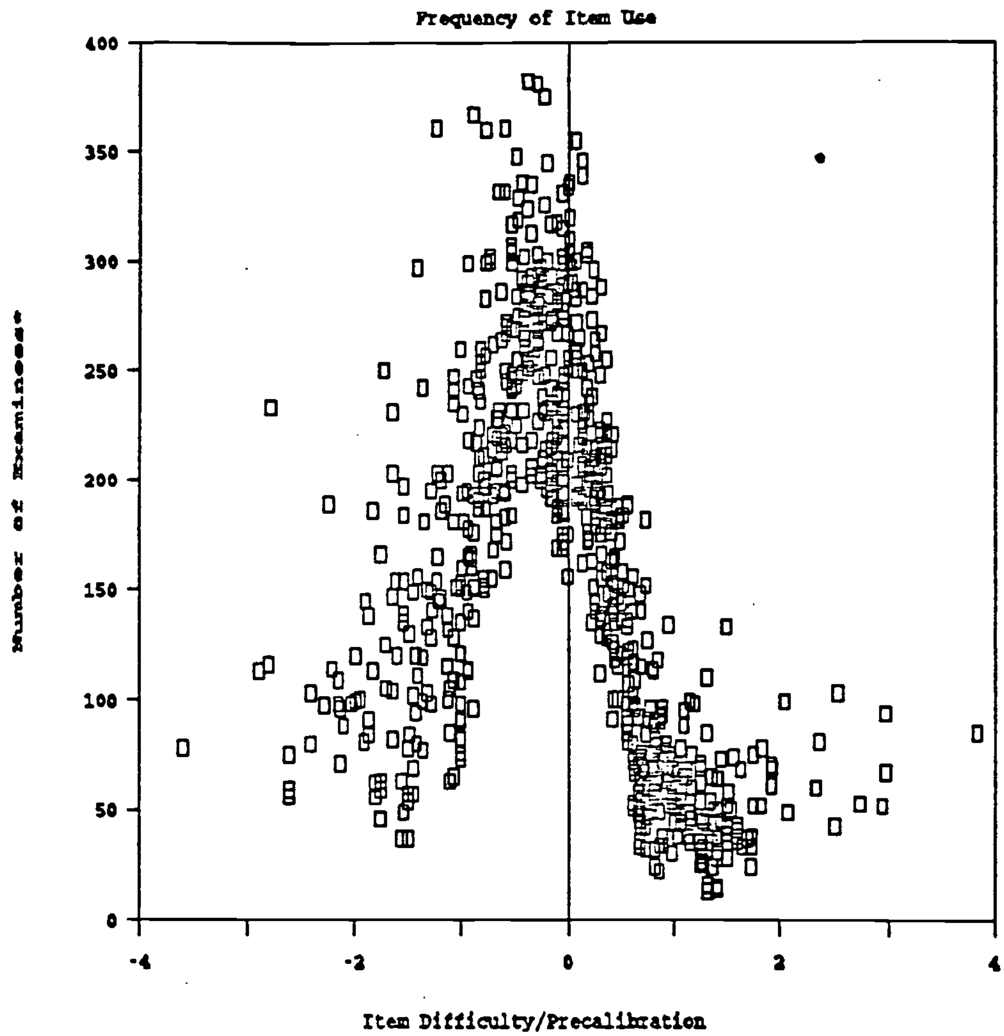
Yen, W.M. (1980). The extent, causes and importance of context effects on item parameters for two latent trait models. Journal of Educational Measurement, 17, 4, 297-311.

Table 1

Item Pool Distribution

Subtest	Content Specifications	Number of Items	Calibrations Mean	Std Dev	Minimum	Maximum
1	20%	147	-.06	.96	-2.89	2.36
2	20%	165	-.07	1.00	-2.21	2.94
3	20%	142	-.07	1.06	-3.61	2.97
4	20%	135	-.05	.97	-2.80	2.97
5	10%	72	.09	.97	-2.24	3.84
6	10%	65	.25	.96	-2.78	2.04
Total	100%	726	-.02	1.00	-3.61	3.84

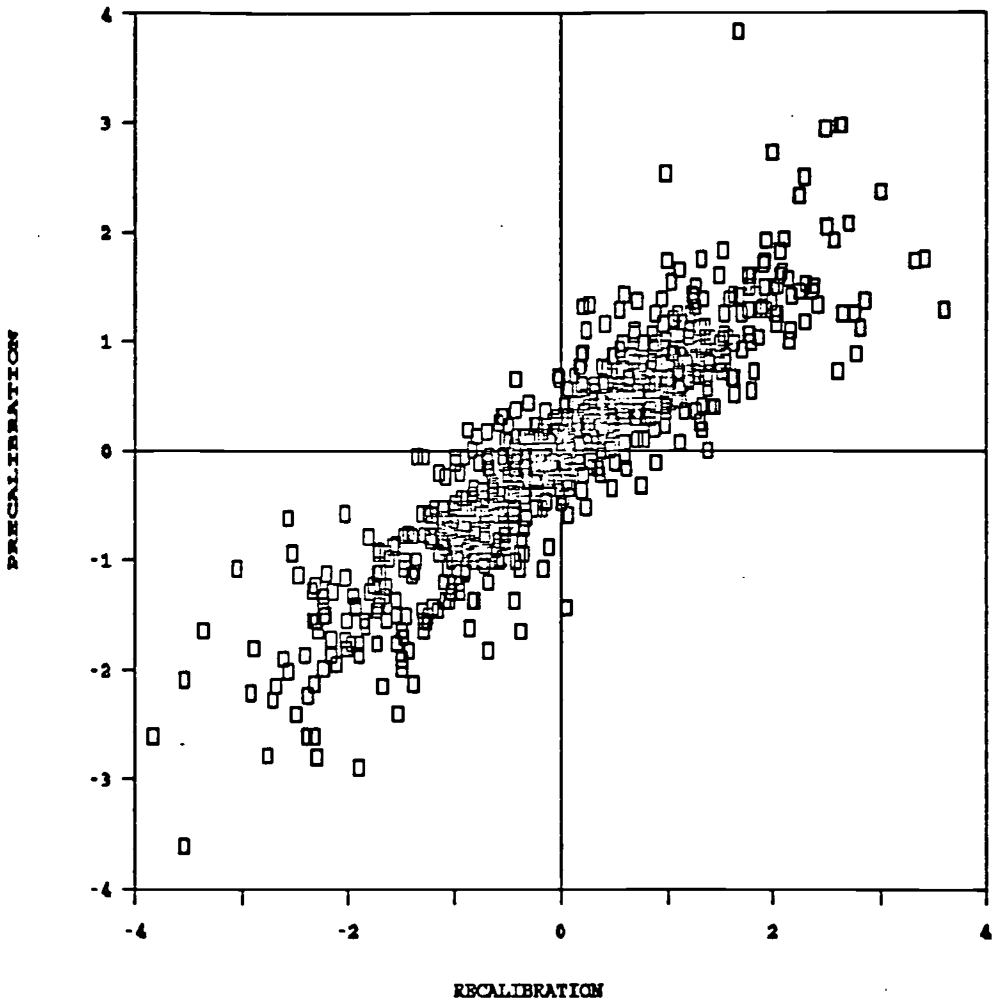
Figure 1
Cat Administration



*Number of Examinees Per Item
Mean=161.63 SD=88.43 Min=13 Max=382

Figure 2

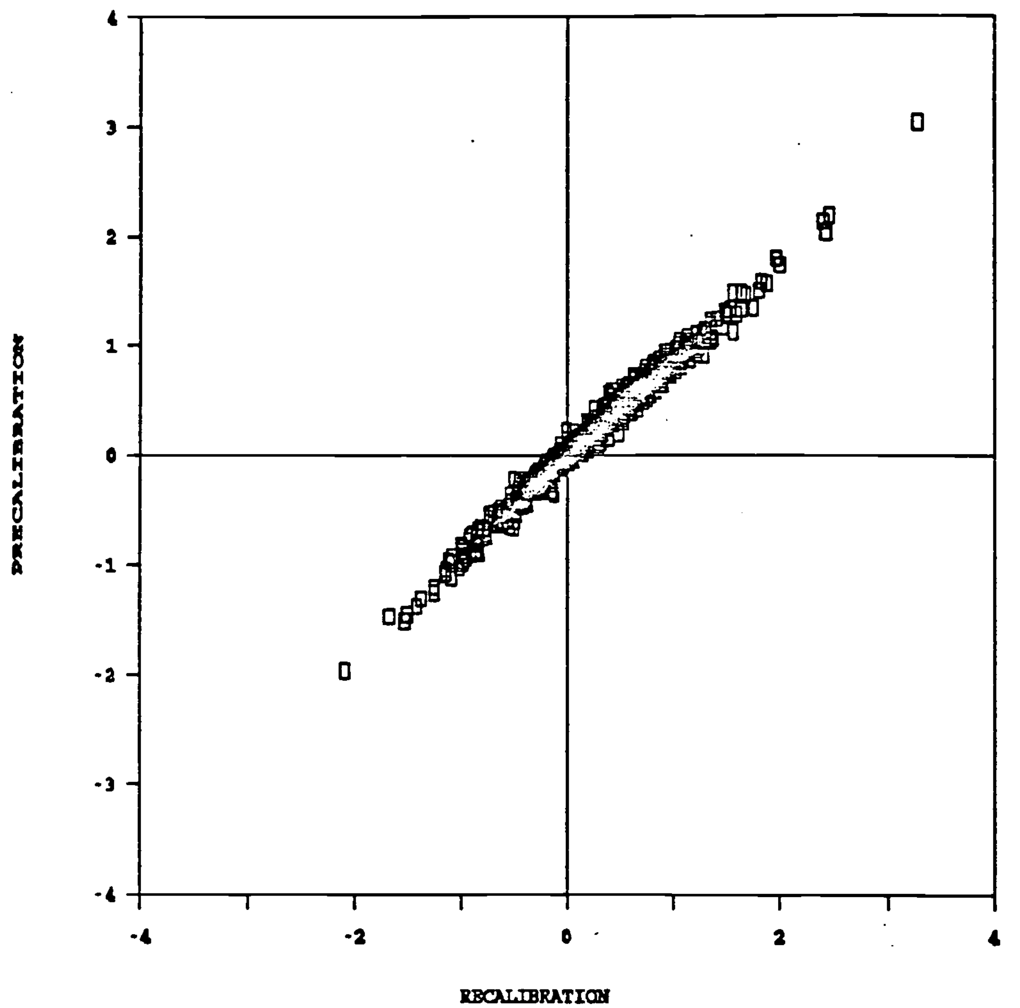
Comparison of Item Calibrations



r=.90

Figure 3

COMPARISON OF ABILITY ESTIMATES



r=.99

TM 025549

NCME Annual Meeting, April 9-11, 1996

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement (OERI)
Educational Resources Information Center (ERIC)

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: Equivalence of Rasch Item Calibrations and Ability Estimates Across Modes of Administration

Author(s): Betty Bergstrom, Mary Lunz

Corporate Source: Computer Adaptive Technologies

Publication Date: April, 1991

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic/optical media, and sold through the ERIC Document Reproduction Service (EDRS) or other ERIC vendors. Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce the identified document, please CHECK ONE of the following options and sign the release below.

	Level 1	Level 2	
[✓] **Check here** Permitting microfiche (4"x 6" film), paper copy, electronic, and optical media reproduction	Sample sticker to be affixed to document: "PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY ____ Sample ____ TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."	Sample sticker to be affixed to document: "PERMISSION TO REPRODUCE THIS MATERIAL IN OTHER THAN PAPER COPY HAS BEEN GRANTED BY ____ Sample ____ TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."	[ ] **or here** Permitting reproduction in other than paper copy.

Sign Here, Please

Documents will be processed as indicated provided reproduction quality permits. If permission to reproduce is granted, but neither box is checked, documents will be processed at Level 1.

"I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce this document as indicated above. Reproduction from the ERIC microfiche or electronic/optical media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries."

Signature: B. Bergstrom	Position: Dir. Psychometric Services
Printed Name: Betty Bergstrom	Organization: Computer Adaptive Technologies
Address: 2609 W Lunt Ave #2E Chicago IL 60645	Telephone Number: (312) 274-3286
	Date: 4/22/96

OVER