ED 377 229 TM 022 404

AUTHOR	Bergstrom, Betty A.; Stahl, John A.
TITLE	Assessing Existing Item Bank Depth for Computer Adaptive Testing.
PUB DATE	Apr 92
NOTE	17p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (San Francisco, CA, April 21-23, 1992).
PUB TYPE	Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	Ability; *Adaptive Testing; *Computer Assisted Testing; *Evaluation Methods; *Item Banks; Item Response Theory; *Test Construction; Test Items
IDENTIFIERS	*Accuracy; *Information Function (Tests)

ABSTRACT

This paper reports a method for assessing the adequacy of existing item banks for computer adaptive testing. The method takes into account content specifications, test length, and stopping rules, and can be used to determine if an existing item bank is adequate to administer a computer adaptive test efficiently across differing levels of examinee ability. An example is presented that shows that the adequacy of the bank can depend on the stopping rule implemented. The example is from an item bank with 183 items from 4 content subtests, with 50% of items from subtest 1, 24% from subtest 2, 20% from subtest 3, and 6% from subtest 4. The use of information functions for both subtests and the total test gives a picture of the adequacy of the item bank across content areas. The procedure can be modified for use with other item response theory models as long as the item parameters are known. Ten figures illustrate the discussion. (Contains 13 references.) (Author/SLD)

ASSESSING EXISTING ITEM BANK DEPTH

FOR COMPUTER ADAPTIVE TESTING

BETTY A. BERGSTROM

JOHN A. STAHL

American Society of Clinical Pathologists

Paper to be presented at the annual meeting
National Council on Measurement in Education, San Francisco, California, 1992.

ASSESSING EXISTING ITEM BANK DEPTH FOR COMPUTER ADAPTIVE TESTING

Abstract

This paper reports on a method for assessing the adequacy of existing item banks for computer adaptive testing. The method takes into account content specifications, test length and stopping rules and can be used to determine if an existing item bank is adequate to administer a computer adaptive test efficiently across differing levels of examinee ability. The paper contains an example which shows that the adequacy of the bank can depend upon the stopping rule implemented.

ASSESSING EXISTING ITEM BANK DEPTH FOR COMPUTER ADAPTIVE TESTING

Computer adaptive testing (CAT), a method of administering tests via computer and targeting the difficulty of the items presented to the ability of the examinee, may be a useful method of administration, especially for test agencies currently giving long written tests. Computer adaptive testing has been shown to reduce test length without compromising measurement precision (Weiss, 1983, 1985; Weiss and Kingsbury, 1984; McKinley and Reckase, 1980, 1984; Olsen, Maynes, Slawson and Ho, 1986; Lunz and Bergstrom, 1991). As testing agencies contemplate the suitability of CAT as a replacement for existing paper and pencil tests they will need to determine the adequacy of existing item banks.

Since the estimate of ability and the choice of the next item administered requires knowledge of the item parameters, one of the main requirements to implement computer adaptive testing is a calibrated item bank. One suggestion for assessing item bank adequacy has been to compute the information function for the entire bank. This is accomplished by summing the information functions for all items in the bank across the range of ability of the population to be tested (Green, Bock, Humphreys, Linn and Reckase, 1984). However, any particular computer adaptive test will be relatively short, so the adequacy of the item bank must be assessed with regard to the portion of the item bank that any particular examinee will see. It is improbable that examinees will be administered all of the items in the bank.

Another issue in assessing item bank adequacy is the need to fulfill designated content specifications. Content balancing mechanisms of the type described by Kingsbury and Zara (1991) insure that paper and pencil test "blueprints" can be replicated on a CAT. If content balancing is included in the CAT algorithm, any attempt to assess the adequacy of the bank must also take content specifications into account.

A third issue is the requirements placed on an item bank due to the stopping rule. A stopping rule based on a fixed length or a specified precision of measure will impose different requirerments on an item bank than one that is based on a specified level of confidence in a pass/fail decision.

The objective of this paper is to report on a method for assessing the adequacy of existing item banks for computer adaptive testing (CAT).

Method

As part of a national pilot test for ccmputer adaptive examinations, the following method was developed to insure that existing item banks contained an adequate number and distribution of items, within content specifications. The existing item banks were previously calibrated using the Rasch model (Wright and Stone, 1979; Wright, Congdon and Schultz, 1987; Bergstrom and Lunz, In Press) however, the method described is applicable for item banks calibrated with other IRT models as long as the item parameters are known.

CAT Specifications

The test administration parameters included the following. The minimum number of items was set at 50; the maximum number of items varied based on the performance of the examinee, but 100 was the established maximum. The stopping rule required the ability estimate to be 1.65 times the standard error of measurement (Wright and Masters, 1982) above or below the pass point (.99, for the example test) before testing stopped (a one-tailed 95 percent level of confidence). If, after 50 items, an examinee's estimated measure was far enough above or below the pass point to have 95% confidence in the decision, testing stopped. If an examinee's measure was not sufficiently above or below the pass point to achieve 95% confidence after 50 items, testing continued until the requirement for confidence in the decision was achieved or until the examinee answered 100 items. Test lengths varied to meet this stopping rule.

Content distribution was based on specifications for the certification examination. The example shown is from an item bank which contained 183 items from 4 content subtests. Content specifications for the example test are: Subtest 1 = 50%, Subtest 2 = 24%, Subtest 3 = 20% and

Subtest 4 = 6%. Thus for a 50 item test, 25 items would be selected in subtest 1, 12 items in subtest 2, 10 items in subtest 3 and 3 items in subtest 4 for a given examinee.

Fixed Length 50 Item Test

Since all examinees would be taking at least a 50 item test, the first step was to determine the test information function for a fixed length test of 50 items (minimum test length) across the range of ability measures for the prospective test population. It was hypothesized that the test population ranged in ability from -3.5 logits to 3.5 logits. Given the existing item bank, 50 item tests were simulated for examinees with abilities ranging from -3.5 to 3.5 logits at .10 logit intervals. A computer program was written that chose the specified number of items from the item bank, within each subtest, that were closest in difficulty to the specified examinee ability. The probability that an individual would get each item correct was calculated using the Rasch formula:

$$\mathrm{Ln}(P/1-P) = B - D \qquad (1)$$

where P is the probability of getting the item correct

B is the ability of the examinee

D is the difficulty of the selected item

The formula for computing the information function of an item with the Rasch model is:

$$I = P(1-P). \qquad (2)$$

The subtest information function (ΣI) is the sum of the information contributed by the items chosen that make up that particular subtest. For example, if the ability measure was -2.00 logits, the algorithm selected the 25 items in subtest 1 from the bank closest in difficulty to -2.00 logits to calculate subtest information.

The information for the first subtest is designated ΣI_1 , the second sub-test ΣI_2, and so forth. In order to compute the information function for the total test, the information functions for the four subtests were summed:

$$I_T = \Sigma I_1 + \Sigma I_2 + \Sigma I_3 + \Sigma I_4 \qquad (3)$$

The maximum information obtainable, given perfect targeting, was also calculated.

Perfect targeting assumes that the item difficulty equals the person ability. According to the Rasch model, the probability of getting the items correct is 50% and thus the information function for a perfectly targeted item is $[I=P(1-P)] = (.50 * .50) = .25$. The maximum information obtainable would be .25 times the number of items in the subtest or the total test.

Next, the standard error of measurement (SEM) for each point on the ability measure continuum was calculated using the formula:

$$\mathrm{SEM} = (^1/I_T)^{\frac{1}{2}} \qquad (4)$$

where I_T is the information function for the total test at that particular ability point.

Finally, in order to assess the adequacy of the item bank when the 95% confidence stopping rule is used, the points on the ability continuum where the test would stop at 50 items were determined. A one-tailed confidence interval for each ability measure was estimated by calculating the ratio of the distance of the examinee ability estimate (B) from the pass/fail point (PF) divided by the standard error of measure (SEM) and comparing this

$$(B\text{-}PF/SEM) \qquad (5)$$

ratio to the normal probability distribution table. This provided a level of confidence in the pass/fail decision for each point on the ability continuum for the 50 item test. The test would stop at 50 items if the absolute value of the ratio was 1.65 or greater indicating that the 95% confidence interval excluded the pass/fail point.

Fixed Length 100 Item Test

The entire procedure was repeated for a maximum length test of 100 items but only for examinee measures near the pass/fail point. The procedure was not repeated for examinees whose ability is high enough to pass, or low enough to fail, with 95% confidence in the decision at 50 items. Subtest information functions, the total test information function, the SEM and the confidence levels were obtained.

Results

Fixed Length 50 Item Test

Figures 1, 2 and 3 show that the item bank is adequate in subtests 1, 2,and 3 to provide examinations that yield maximum information for examinees with estimated abilities between -1.5 and 1.5 logits. Figure 4 shows that subtest 4 has enough items to provide a 3 item subtest that yields maximum information to examinees with estimated abilities of -2.0 logits to 2.0 logits.

Figure 5 shows the total test information function, obtained by summing the results from the subtest information functions. The maximum information obtainable is 12.5 for a 50 item test. With this item bank, examinees with ability measures greater than 1.5 or less than -1.5 will be administered some items that do not yield maximum information. High ability (>1.5) examinees will be administered a relatively easy test, so they will correctly answer more than 50% of the items presented. Low ability examinees (<-1.5) will be administered a relatively difficult test, so they will correctly answer less than 50% of the items presented.

Figure 6 shows the standard error of measure (SEM) across the ability continuum obtained for the 50 item test. The minimum possible SEM at 50 items is .28. When examinees do not challenge items that yield maximum information, the error of measurement (SEM) increases.

Levels of confidence in the pass/fail decision are shown in Figure 7. Examinees with ability estimates of 1.5 or greater will pass the test at 50 items with ≥ 95% confidence in the pass decision while examinees with ability estimates < .5 logits will fail the test with ≥ 95% confidence in the fail decision. Examinees with ability estimates between .5 logits and 1.5 logits are so close to the pass/fail point (.99 logits) that making a pass/fail decision with 95% confidence will not be possible and they will take longer tests.

Fixed Length 100 Item Test

Figures 8 and 9 show the total test information function and the SEM, respectively, for the maximum length test of 100 items. The total test information function was obtained by summing the subtest information functions. This item bank is not sufficient to provide a 100 item test that yields maximum information for examinees at any ability level.

Figure 10 shows that for examinees with ability estimates between .50 and .65 greater than 95% confidence in the fail decision is achieved. The variable length CAT will stop before the maximum number of 100 items is reached even though their test is not perfectly targeted. For examinees with ability estimates between 1.35 and 1.5, greater than 95% confidence in the pass decision is also achieved and their test will stop before the maximum test length of 100 items is reached. Examinees with ability measures estimated between .65 logits and 1.35 logits are so close to the pass/fail point (.99 logits) that they will take the maximum number of 100 items. Since a pass/fail decision must be made for these examinees with less than 95% confidence in the decision, it is especially important that their measure be estimated with maximum precision (100 items). Figure 9 indicates that there are insufficient items in the bank near the pass/fail point (.99) to achieve the minimum SEM.

Discussion

Fixed Length Tests

If the test were a 50 item fixed length test, the subtest information functions indicate that the bank would need additional easy (<-1.5) and hard (>1.5) items in all subtests. Increasing the bank at the ends of the continuum would insure that all examinees, including low and high ability examinees, would be tested with comparable levels of precision.

Testing to a Specified Level of Precision

When a specified standard error of measurement (SEM) is used as the stopping rule, test length is variable, since all examinees are tested to the same level of precision regardless of the number of items required to reach the specified SEM. Figure 6 shows that if the stopping rule required reaching a SEM of .28, examinees with estimated ability measures less than .5 logits or greater than 1.5 would have to take longer tests. In order for low and high ability examinees to reach a specified SEM of .28 in 50 items additional easy and hard items would have to be added to the bank.

Confidence Level Stopping Rule

When a specified level of confidence is used as the stopping rule both the distance from the pass/fail point and the SEM influence test length. A confidence level stopping rule results in a more precise estimate of ability for minimally competent examinees (near the pass/fail point), because they take longer tests. Examinees of very high ability and very low ability are estimated less precisely, because they take shorter tests. However, a high level of confidence in their pass/fail decision is obtained because their estimated measures are significantly above or below the pass point.

In order to improve this item bank for use with a confidence level stopping rule, each subtest information function for a 100 item test must be examined. Since low and high ability examinees will pass the test with equal to or greater than 95% confidence in the pass decision, even though their SEM is greater than the minimum SEM, the item bank will not be improved by adding more easy or difficulty items. Content experts will need to write additional items near the pass/fail point, so that items yielding maximum information about the marginal examinee are available. Better targeted items, improve the possibility of achieving 95% confidence in the decision, even for marginal examinees.

Limitations

This procedure assumes that real examinees will respond to items in accordance with the item characteristic curve of the IRT model used. In actual CAT testing, examinees may not respond strictly in accordance with the model. Also, there is always uncertainty in the estimated ability measure at the beginning of the CAT test, resulting in less than optimal items being chosen early in the test with regard to the examinee's final estimated ability measure. An actual observed response pattern is expected be an underestimate of the theoretical information curve (Bejar, Weiss, Gialluca, 1977). Therefore this procedure provides an optimistic look at the information function across the ability continuum.

Conclusion

This procedure gives test developers a method to assess the adequacy of existing item banks which takes into account "blueprint" test specifications and stopping rules. The use of the information function for both subtests and the total test gives a picture of the adequacy of the item bank across content areas.

This procedure can be modified for use with other IRT models as long as item parameters are known. In addition to the factors mentioned above, bank assessment may involve more detailed content specifications, choice of a starting item or items, and/or constraints on use of particular items because of content overlap.

Reference

Bejar, I.I., Weiss, D.J. and Gialluca, K.A. (1977). An information comparison of conventional and adaptive tests in the measurement of classroom achievement (Research Report 77-7). Psychometric Methods Program Department of Psychology University of Minnesota, Minneapolis, MN.

Bergstrom, B.A. and Lunz, M.E. (In Press). Equivalence of Rasch item calibrations across modes of administration. Objective Theory into Practice, M. Wilson (Ed.). Norwood, N.J.: Ablex Publishing Corporation.

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 Zara, A.R. (April, 1991). A comparison of procedures for content-sensitive item selection in computerized adaptive tests. Paper presented to the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Lunz, M.E. and Bergstrom, B.A. (1991). Comparability of decision for computer adaptive and written examinations. Journal of Allied Health, 20, 1, 15-23.

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.

Olsen, J.B., Maynes, D.D., Slawson, D. and Ho, K. (1986, April). 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.

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

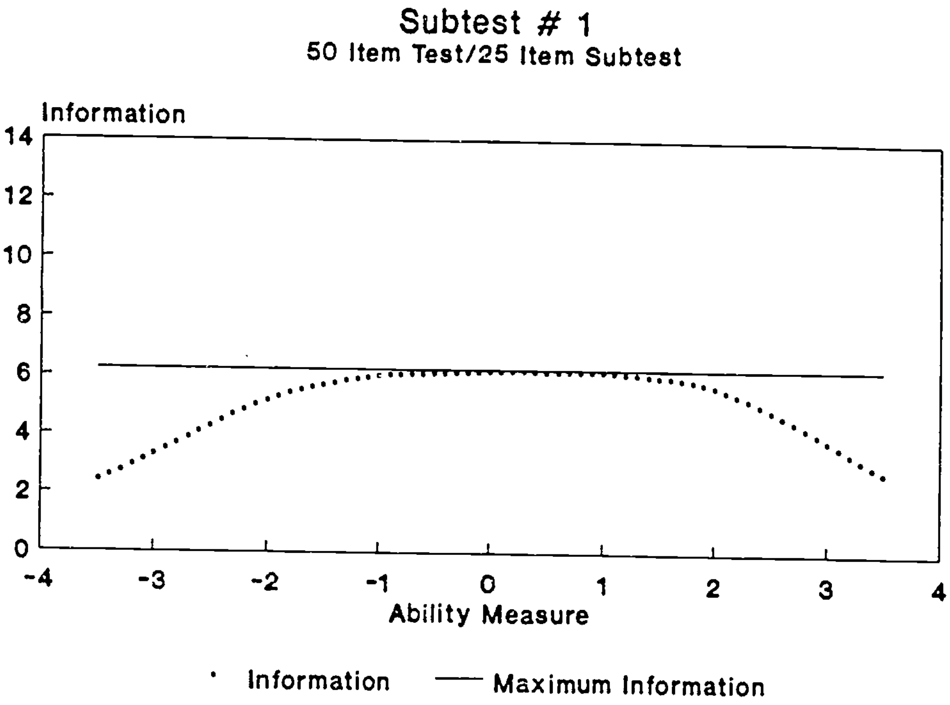
Weiss, D.J. and Kingsbury, G.G. (1984). Application of computerized adaptive testing to educational problems. Journal of Educational Measurement, 21, 4, 361-375.

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

Wright, B.D., Congdon, R. and Schultz, M. (1987). MSCALE (Computer Program). Chicago: University of Chicago.

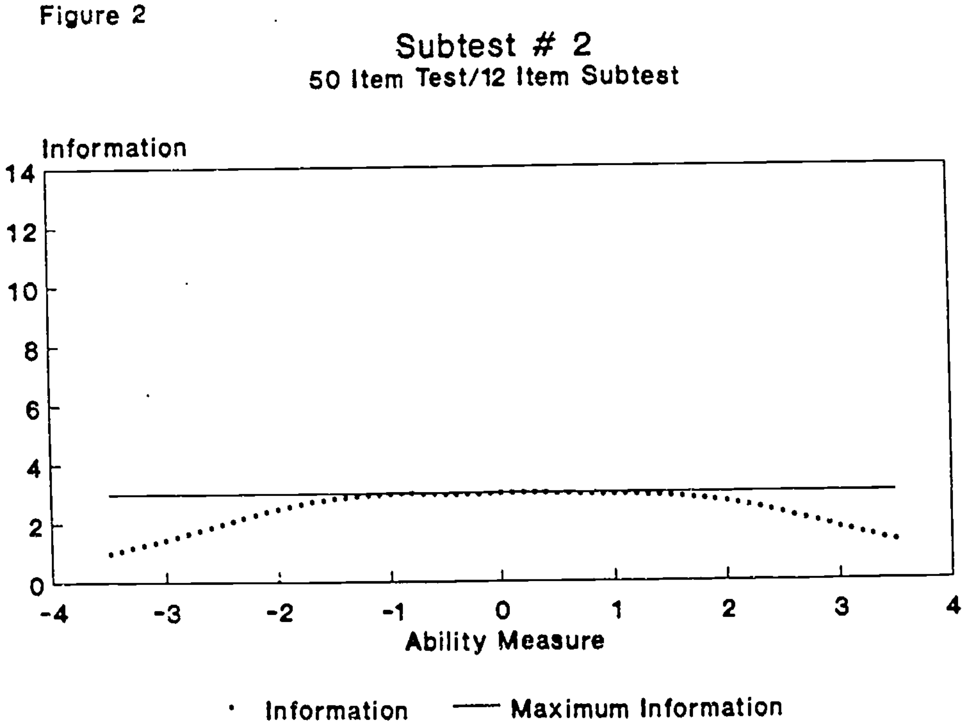
Wright, B. D. and Masters, G.N. (1982). Rating scale analysis. Chicago: MESA Press.

Figure 1



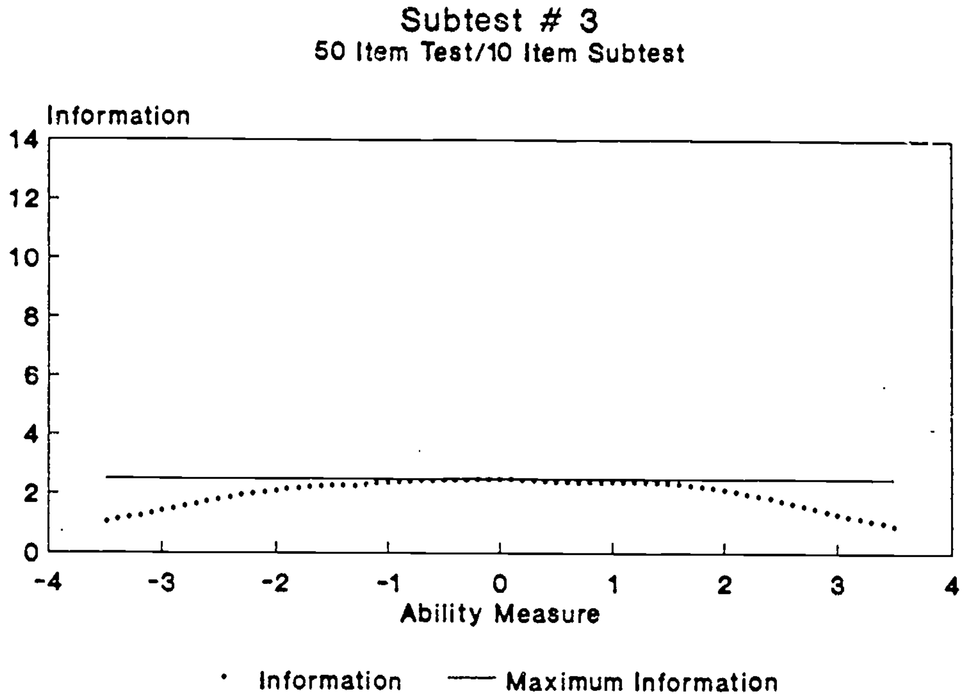
Maximum Information • 6.25

Figure 2



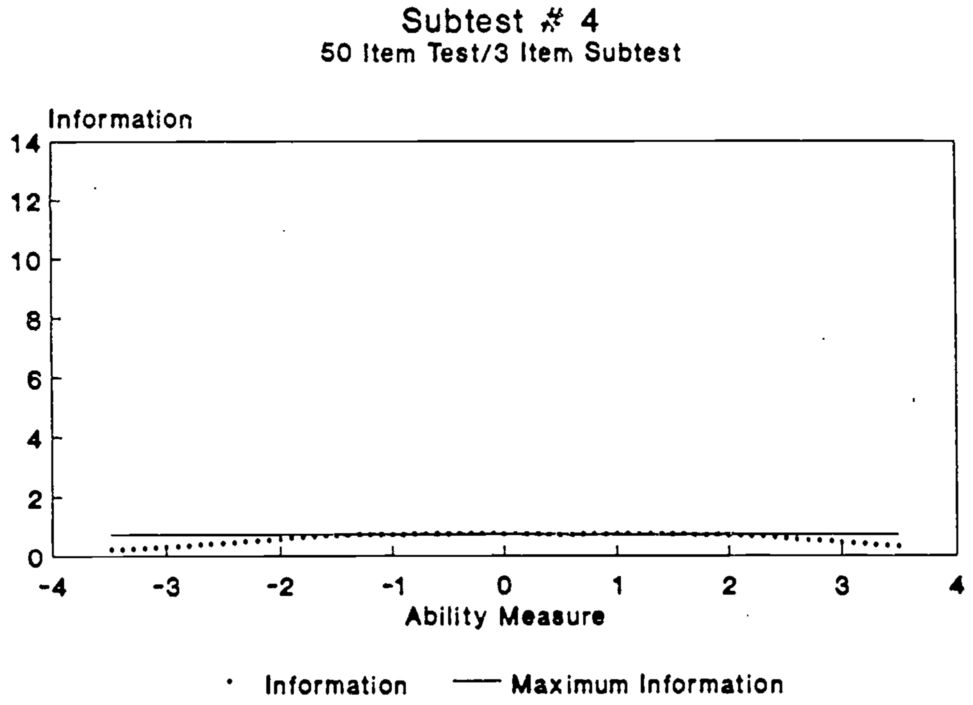
Maximum Information • 3

Figure 3



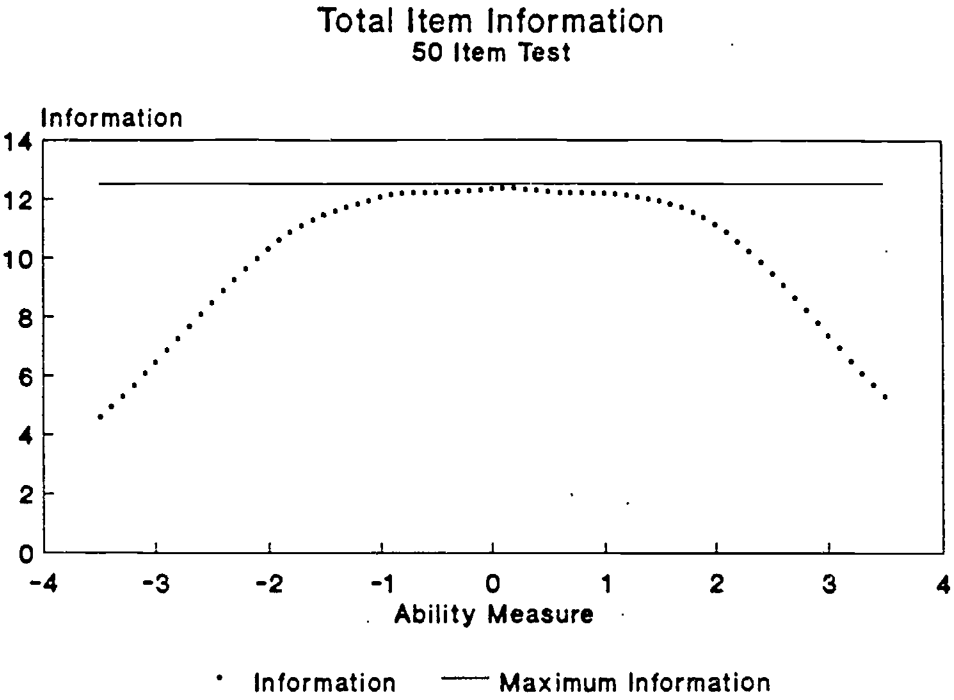
Maximum Information • 2.5

Figure 4



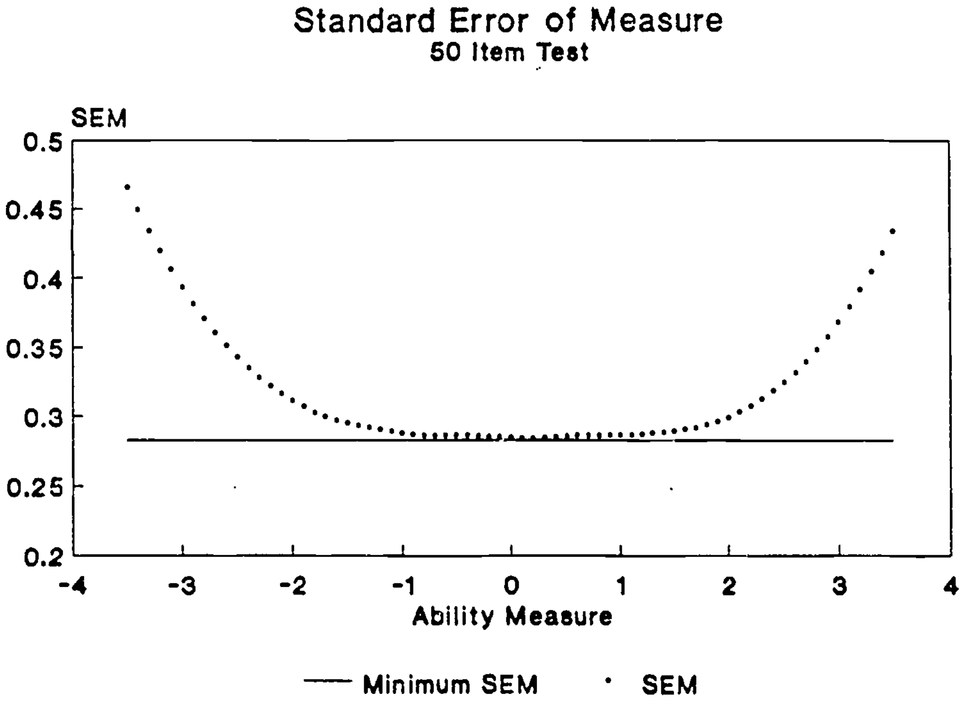
Maximum Information • .75

Figure 5



Maximum Information • 12.5

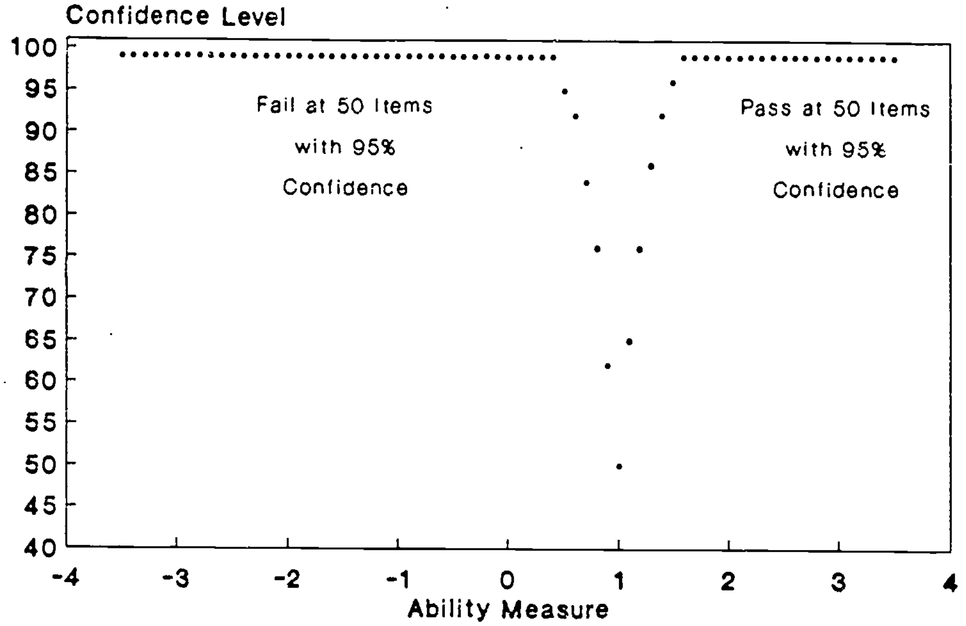
Figure 6



Minimum SEM • .28

Figure 7

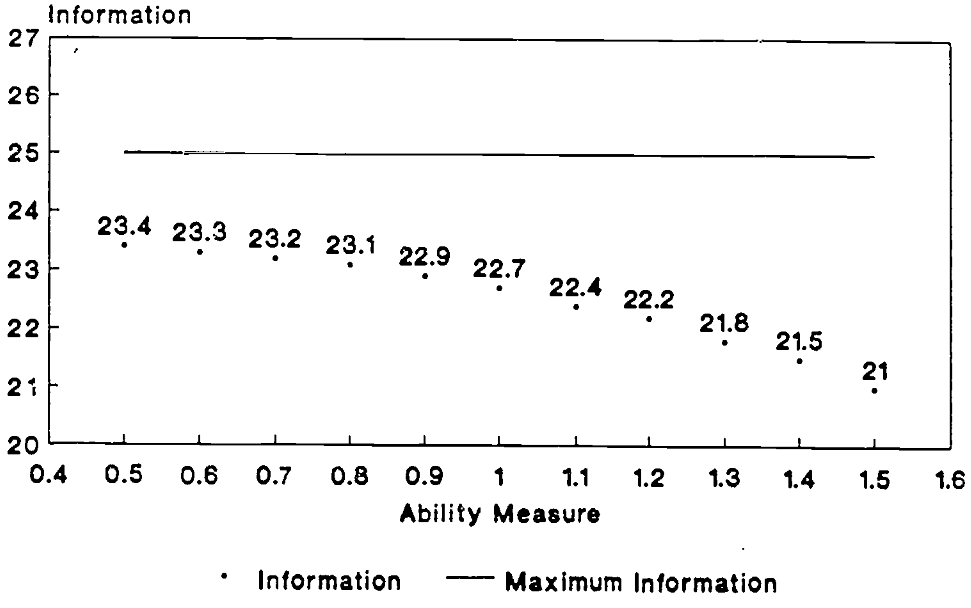
Confidence in the Pass/Fail Decision
50 Item Test



Pass/Fail Point • .99

Figure 8

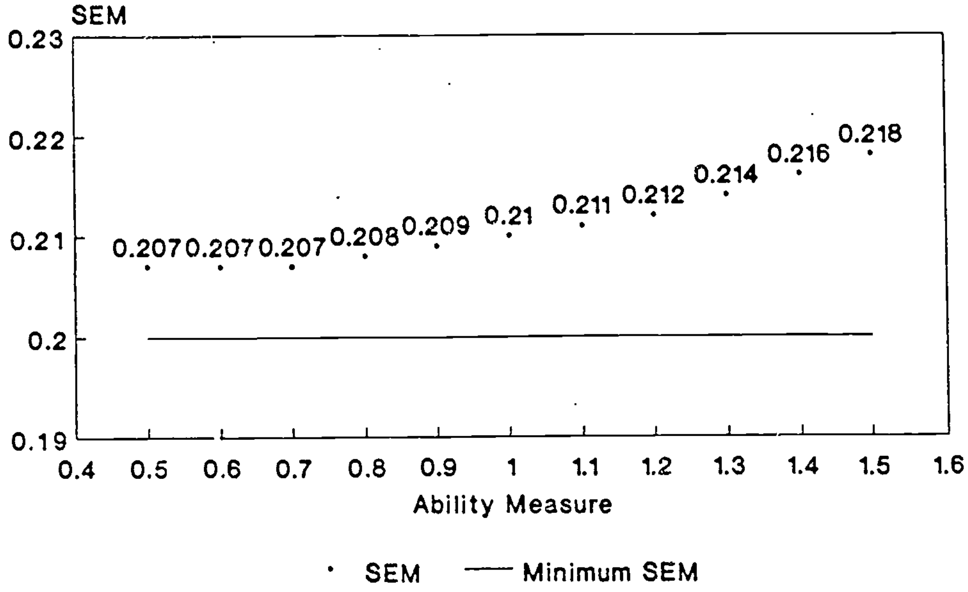
Total Item Information
100 Item Test



Maximum Information • 25

Figure 9

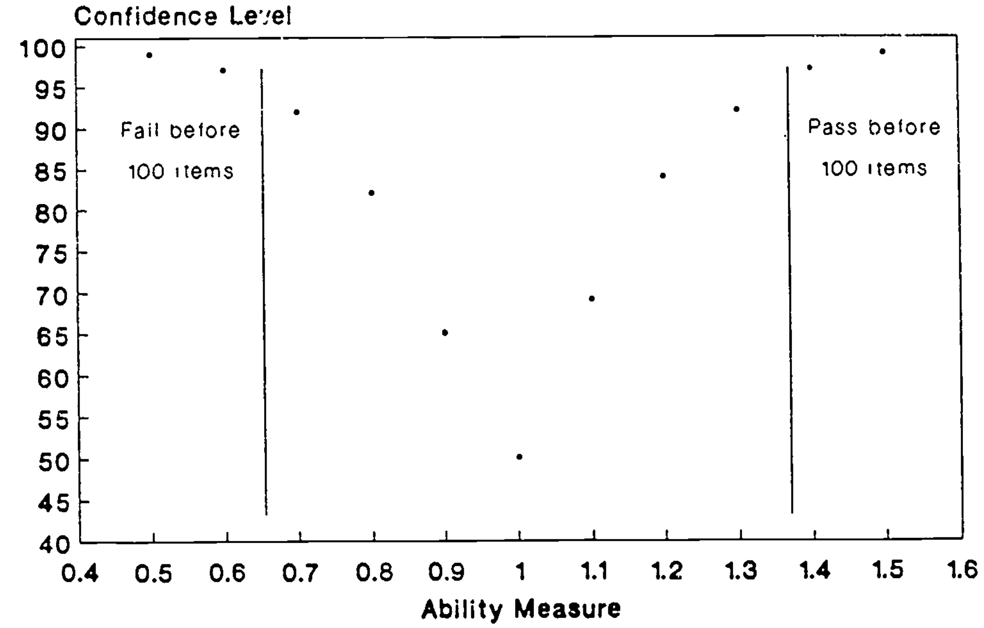
Standard Error of Measure
100 Item Test



Minimum SEM • .20 Pass/Fail Point • .99

Figure 10

Confidence in the Pass/Fail Decision
100 Item Test



Pass/Fail Point•.99