DOCUMENT RESUME

ED 106 342 TM 004 457

AUTHOR Reckase, Mark D.

TITLE The Effect of Item Choice on Ability Estimation When Using a Simple Logis'ic Tailored Testing Model.

PUB DATE [Apr 75]

NOTE 14p.; Paper presented at the Annual Meeting of the American Educational Research Association (Washington, D. C., March 30-April 3, 1975)

EDRS PRICE MF-$0.76 HC-$1.58 PLUS POSTAGE

DESCRIPTORS *Ability; Achievement Tests; *Individual Differences; *Item Banks; Multiple Choice Tests; Probability; Response Style (Tests); Simulation; Student Evaluation; Test Bias; *Test Construction; *Testing

IDENTIFIERS Rasch Model; *Tailored Testing

ABSTRACT

This paper explores the effects of item choice on ability estimation when using a tailored testing procedure based on the Rasch simple logistic model. Most studies of the simple logistic model imply that ability estimates are totally independent of the items used, regardless of the testing procedure. This paper shows that the ability estimate is affected by item choice and gives theoretical and empirical information as to the direction and magnitude of the effect. (Author/RC)

The Effect Of Item Choice On Ability Estimation When Using A Simple Logistic Tailored Testing Model

by

Mark D. Reckase

University of Missouri - Columbia

I. Introduction

The simple logistic model has attracted considerable attention since its presentation by Rasch in 1960. One of the reasons for this interest has been the property of the model labeled "objectivity." Rasch (1966) has defined objectivity as follows:

> Hence, the parameter of the subjects in the subgroups [equal score subgroups] may be evaluated without regard to the parameter of the other subjects; and, of course, it has already been shown that these will all be independent of the item parameters. A similar statement holds for the latter. Comparisons capable of being carried out in this way I have called 'specifically objective.'

This statement by Rasch has been interpreted to mean that any set of subjects can be used to calibrate items, and that any set of items can be used to determine the ability parameters of individuals and that the values so determined can very easily be placed on the same scale. These properties of the simple logistic model have been verified theoretically by Rasch (1960) and empirically by Wright (1968).

Tailored testing involves selecting a unique set of items for each individual that is in some way "best" for evaluating his ability, and on the basis of this set determining an ability estimate. Since each individual receives a different set of items, it is almost required that an estimation procedure based on an "objective" model be used. Hence, a tailored testing procedure based on the simple logistic model has been developed and applied to the estimation of achievement (Reckase, 1973; 1974). This application of the model to achievement testing gave the first hints that the workings of an "objective" procedure may not lead to the commonly expected result. The purpose of this paper is to formalize the study of the effects of item selection on ability estimates, and to show that, at least for tailored testing, ability estimates are not totally independent of the items used.

Paper presented at the Annual Meeting of the American Educational Research Association, Washington, D.C., 1975

For the purposes of this paper, the relation between item selection and ability estimation will be shown in two ways. First, the relation between ability estimates and item parameters will be determined theoretically using maximum likelihood estimation for two and three item cases. Second, the relations will be determined empirically again using maximum likelihood estimation for more typical test lengths using simulations and tailored administration of items in a typical achievement testing situation.

II. Theoretical Relationship Between Ability Estimates and Item Parameters

Two Item Case

Suppose an individual s is administered an item i. Assuming the simple logistic model describes the individual's behavior, the probability of a correct response is given by

$$(1) \qquad P\{X_{si} = 1\} = \frac{A_s E_i}{1 + A_s E_i}$$

where X_i is the item score, A_s is the individual's ability parameter, and E_i is the items easiness parameter. If a second item, item j, is administered the probability of an incorrect response to that item is given by

$$(2) \qquad P\{X_{sj} = 0\} = \frac{1}{1 + A_s E_j}$$

where E_j is the easiness parameter of item j. If the items are independent of each other the probability of the 1,0 response pattern for individual s is given by

$$(3) \qquad P\{X_{si} = 1\}P\{X_{sj} = 0\} = \frac{A_s E_i}{1 + A_s E_i} \cdot \frac{1}{1 + A_s E_j}$$

Equation 3 also gives the likelihood of the 1,0 response string for items i and j for a person with ability A_s and hence the maximum likelihood estimate of A_s can be obtained by taking the derivative with respect to A_s and solving for zero.

$$(4) \qquad \frac{dP\{X_{si} = 1 \text{ and } X_{sj} = 0\}}{dA_s} = \frac{d\dfrac{A_s E_i}{(1 + A_s E_i)(1 + A_s E_j)}}{dA_s} = 0$$

$$= E_i - A_s^2 E_i^2 E_j = 0.$$

(5) $$A_s = \frac{1}{\sqrt{E_i E_j}}$$

or, in terms of log ability

(6) $$\ln A_s = -\tfrac{1}{2}(\ln E_i + \ln E_j).$$

Equation 5 shows that instead of the ability estimate being independent of the easiness parameters, it is in fact a function of them. However, this result must be kept in perspective. If item i had been responded to incorrectly and item j correctly, and we solved for the maximum likelihood estimate for the ability parameter we would obtain an expression that is exactly the same as is shown in Equation 5. In other words, when using the Rasch model, as long as the same set of items has been administered, it does not matter which items of the set are answered correctly or incorrectly. Only the number correct is important; the number correct being a sufficient statistic for the ability estimate. However, if different individuals are administered different sets of items, they are likely to obtain different ability estimates even if they have the same response pattern.

Figure 1 shows the relationship between ability estimates and easiness parameters for various values when the response pattern is 1,0 or 0,1. E_i refers to the easiness parameter of the first item responded to, and E_j to the second item. Note that if a total score of one is obtained when items with low easiness values are administered, a high ability estimate is the result; if high easiness items are used, a low ability estimate is the result.

This is a conceptually satisfying result since the ability estimates conform to beliefs of what should occur. However, it must be remembered that the 0,1 or 1,0 response pattern will have different probabilities of occurrence depending on the ability of the examinee and the easiness parameters. Even though the response pattern would yield a high ability estimate for a given set of easiness parameters, the actual occurrence of the response pattern may have a very low probability.

Three Item Case

Suppose that three items i, j, and k have been administered to person s and that performance on these items can be described as in Equations 1 and 2 above. The probability of an incorrect response and two correct responses is then given by

$$(7) \qquad P\{0,1,1\} = \frac{1}{1 + A_sE_i} \cdot \frac{A_sE_j}{1 + A_sE_j} \cdot \frac{A_sE_k}{1 + A_sE_k}$$

$$= \frac{A_s^2 E_jE_k}{(1 + A_sE_i)(1 + A_sE_j)(1 + A_sE_k)}.$$

Differentiating with respect to A_s and solving for zero, the following equation is obtained.

$$(8) \qquad \frac{dP\{0,1,1\}}{dA_s} = 2 + A_s(E_i + E_j + E_k) - A_s^3 E_iE_jE_k = 0$$

Solving for A_s yields the maximum likelihood estimate of the ability parameter.

$$(9) \quad A_s = \left(\frac{1}{E_iE_jE_k}\right)^{1/3}\left(\left(1+\left(1+\frac{(-E_i-E_j-E_k)^3}{27E_iE_jE_k}\right)^{1/2}\right)^{1/3} + \left(1-\left(1+\frac{(-E_i-E_j-E_k)^3}{27E_iE_jE_k}\right)^{1/2}\right)^{1/3}\right)$$

Again, as in the two item case, we can easily see that instead of ability estimates being independent of item parameters, they are in fact functions of them. However, the qualification must be made that all three item response patterns yielding a score of two will result in the same ability estimate regardless of which items are responded to correctly as long as the easiness parameters are the same.

Although Equation 9 is rather cumbersome, we can gain some understanding of the relationship expressed by graphing some typical values. Figure 2 shows the results of graphing the equation. Two sets of data have been presented in the following way. Suppose two items have been administered and one correct response is obtained. The ability estimate obtained if the easiness parameters for the first two items were .1 and .2, or .2 and .5 are represented by the upper and lower dashed line respectively. In other words, the ability estimate when easiness values of .1 and .2 are used is about 6.98, and if .2 and .5 are used it is 3.21. The fact that the easier items yield a lower ability estimate is consistent with the results of the previous sections.

Suppose that a third item is now administered and a correct response is obtained. The solid lines on the graph represent the ability estimates for various levels of easiness of the third item for the two situations described above. Note that as the easiness values increases for the third item the ability estimate decreases, but more importantly, the estimate is always above the previous two item estimate. Thus, if an extremely easy item is administered that will surely be responded to correctly, an increase in ability estimate will be obtained though it may be small. A formal proof of this fact is now being attempted.

The question now arises, can the ability estimate be increased to any desired amount simply by adding enough easy items that the examinee can respond to correctly? More generally, can the ability estimate be manipulated by selecting items properly. The exact mathematical specification of this problem is too complex to be used, however, some information concerning this question can be obtained from the simulation data contained in the next section.

III. Empirical Studies into the Relationship Between Ability Estimates and Item Parameters

Simulation Studies

The first question for which an answer was sought using simulation data, was the one presented in the previous section. Can ability estimates be increased to any desired amount simply by administering enough easy items? To obtain an answer to this question, a very simple simulation was run. Suppose an individual has an ability parameter of 7.00. We can get a rough idea of the ability estimate this individual would obtain from a tailored testing procedure, by assuming that he will answer correctly those items with a probability of correct response of greater than .5, and incorrectly those items with probaoility of correct response less than .5. The result of this procedure is the most likely response pattern.

Once the correct and incorrect response patterns have been specified, an ability estimate can be found using an iterative maximum likelihood procedure. The iterative procedure for finding the mode of the likelihood distributions is not as accurate as the algebraic procedures of the previous sections, but it can be used for much more complex cases.

As a simple example of this procedure, suppose that the first item administered has an easiness parameter of .1. Using Equation 1, the probability of a correct response is found to be 7 x .1/(1+7x.1) = .41. Since this value is below .5, the assumption will be that a wrong response is obtained. If a second, easier item is administered with easiness parameter .2, the probability of a correct response is .57, so it is assumed that a correct response will be made. Once a correct and incorrect response has been made, ability can be estimated yielding 6.98 using the iterative technique and 7.07 using the exact procedures.

Suppose a very easy item with a parameter value of 10.00 is now administered. The probability of correct response to this item is .986. Assuming a correct response, the new ability estimate is 7.13. If another item with easiness 10.0 is administered and responded to correctly, the subsequent ability estimate is 7.23. The upper line in Figure 3 shows the relationship between the ability estimates and the number of items with easiness 10.00 administered.

Notice that the plot yields a straight line that increases at about .07 with each correct response to an item with easiness parameter 10.0. As long as the individual can continue to respond correctly to the items, his ability estimate will continue to climb and, based on the 0.986 probability of response, he has a better than .50 probability of getting 32 items with easiness 10.0 correct.

The lower line of the graph shows a similar result for an individual first getting an item with easiness 1.0 correct, then an item with easiness 0.5 incorrect, followed by a string of correct responses to a set of items with easiness 10.0. Again, a straight line function of the number of items is obtained, this time increasing by increments of 0.06.

A similar result is obtained when items of great difficulty are administered except, of course, that the slope of the line is negative. In practice, however, the result with difficult multiple choice items is distorted because of guessing effects. The tailored testing program used for the administration of items avoids the complex guessing problem by always administering items equal to or easier than the reciprocal of the ability estimate. That is, errors in item selection are always in the direction of easier items.

The implication of this simple simulation is that a definite bias can be induced in the ability estimation procedure by administering items that are extreme in easiness parameter - either very hard or very easy. The size of the bias is determined by the number of extreme items administered and how deviant the items are from those optimal for the individual. How this problem affects the result of the tailored administration of items, both for simulation and empirical data will now be discussed.

The tailored testing procedure used to administer items is based on the premise that items with traditional difficulty of 0.50 are optimal for evaluating an individual. In terms of the Rasch model, items with easiness parameter equal to the reciprocal of the ability parameter of an individual will have a difficulty of 0.50 for that individual. Thus, once an ability estimate has been obtained, the tailored testing program searches the item pool for an item with easiness value equal to the reciprocal of the estimate. If an item with exactly the required value is not found, the next easier item is used. It is the fact that easier items are always selected that relates the tailoring procedure to the previous material.

Once the tailored testing procedure defined above administers all of the items of optimal easiness, progressively easier and easier items are administered. If the preceeding section is a correct model of what occurs when very easy items are administered, the results of the procedure should be to overestimate ability. To verify the conjecture for the tailored testing procedure, the same type of simulation as described above was used. If the probability of correct response to the administered item was greater than 0.50, a correct answer was assumed; if below 0.50, an incorrect response was assumed.

Figure 4 shows the ability estimate after a set of items have been administered to an individual with ability 1.00, the first item administered had an easiness 1.00 and was responded to correctly, the second item had easiness 0.50 and was responded to incorrectly. Subsequent items were picked to have easiness equal to or greater than the reciprocal of the ability estimates. All items were selected from a 57 item pool with log easiness values equally spaced from -3.00 to +3.00.

The figure shows fairly clearly that after eight items have been administered, the procedure stops converging and begins to yield increasing estimates of ability. At that point, all of the appropriate items have been used and only easy items are selected from the pool, yielding a bias in the ability estimates.

A second simulation on a 225 item pool yields a similar result. After eight items have been administered the procedure over estimates, corrects itself, but again overestimates and starts an upward climb again after ten items have been administered. The large pool has more appropriate items, but eventually they are depleted and a bias in estimation is the result. Thus, the simple simulations using the tailored testing procedure yields the same result as the theoretical data shown earlier. The only question now is as to whether the same effect will be present using human subjects.

Real Data Study

Seventeen graduate and undergraduate students were administered a statistics and measurement test using the tailored testing procedure described above. A sixty item pool of multiple choice items was stored in the computer for use by the program. The results for two individuals are shown in Figure 5. Note that the same effect is present as in the simulation. After eight to ten items have been administered, the ability estimates begin to increase regularly as easier items are administered. This result is typical of fourteen of the seventeen cases. In the three cases that did not show the same result, responses seemed either almost random, or, in one case, convergence was very quick and stable. Thus, tailored administration of items to college students confirms the theoretical and simulation results.

V. Summary and Discussion

The purpose of this paper has been to show that the ability estimates obtained using a tailored testing procedure based on the Rasch model are dependent on the item pool. This result is in opposition to the normally assumed relationship between item parameters and ability estimates. In general, the results of this research has shown that biases can be induced in the estimation procedure by using either very easy or very difficult items. Overestimates are obtained with easy items and underestimates are obtained with hard items.

The relationship between item parameters and ability estimates has been studied in three ways; algebraically for simple cases, simulation studies for more complex cases, and finally the tailored administration of items to college students for a more realistic test. Each of the analysis techniques has yielded the same result: bias was induced in the estimation procedure by the administering of easy items. Unfortunately, the amount of bias cannot yet be specified exactly. However a few general rules can be stated.

First, the amount of bias in the raw ability estimate seems to be linearly related to the number of extreme items administered. As more items are administered, the amount of bias in the estimate increases. However, the extreme items must all be of the same type, either very hard or very easy, for the biasing effect to be present.

Second, the more extreme the item the less bias there seems to be in the ability estimate. Thus less increase in ability estimate will result from administering an item with easiness parameter 100 then with parameter 10. However, there is always some change induced in the ability estimate, and the change does accumulate as more items are administered.

Third, if a stopping rule for the tailored administration of items can be determined so that extreme items are not administered, good estimates of a persons ability can be obtained with relatively few items. The simulation studies show convergence to the true abilities in eight to ten items. In practice more items would be required, the increase being dependent on the amount of measurement error.

Several implications can be drawn from the results of this paper. First, one must be cautious in generalizing the mathematical properties of a model to practical testing situations. This author, for one, held several misconceptions concerning the nature of "specific objectivity" that have been clarified by this study.

Second, the Rasch model yields a useful technique for application to tailored testing, but one that must be applied carefully to avoid inducing bias in estimation. Bias in estimation can be avoided by building in checks for items deviating from those optimal for an individual.

Third, a relatively large item pool is required to make enough optimal items available to a tailored testing procedure to yield quick accurate estimates. Simulations seem to indicate that between 100 and 200 equally spaced items are sufficient, but fewer items may be adequate for homogeneous groups of individuals.

In general the Rasch model yields a viable technique for tailored testing and the results of this paper in no way negates that fact. Several cautions concerning the technique have been presented and it is hoped that users of the model will evaluate the implications for other testing situations.

References

Rasch, G. *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research, 1960.

Rasch, G. An item analysis which takes individual differences into account. *British Journal of Mathematical and Statistical Psychology*, 1966, 49-57.

Reckase, Mark D. An application of the Rasch simple logistic model to tailored testing. Paper presented at the 1974 Annual Meeting of the American Educational Research Association, April, 1974. (ERIC Document Reproduction Service No. ED 092 585).

Reckase, Mark D. An interactive computer program for tailored testing based on the one-parameter logistic model. *Behavior Research Methods and Instrumentation*, 1974, 6(2), 208-212.

Wright, B.D. Sample-free test calibration and person measurement. *Proceedings of the 1967 Invitational Conference on Testing Problems*. Princeton: Educational Testing Service, 1968, 85-101.

FIGURE 1
RELATIONSHIP BETWEEN ABILITY ESTIMATES AND ITEM PARAMETERS FOR TWO ITEM CASES.

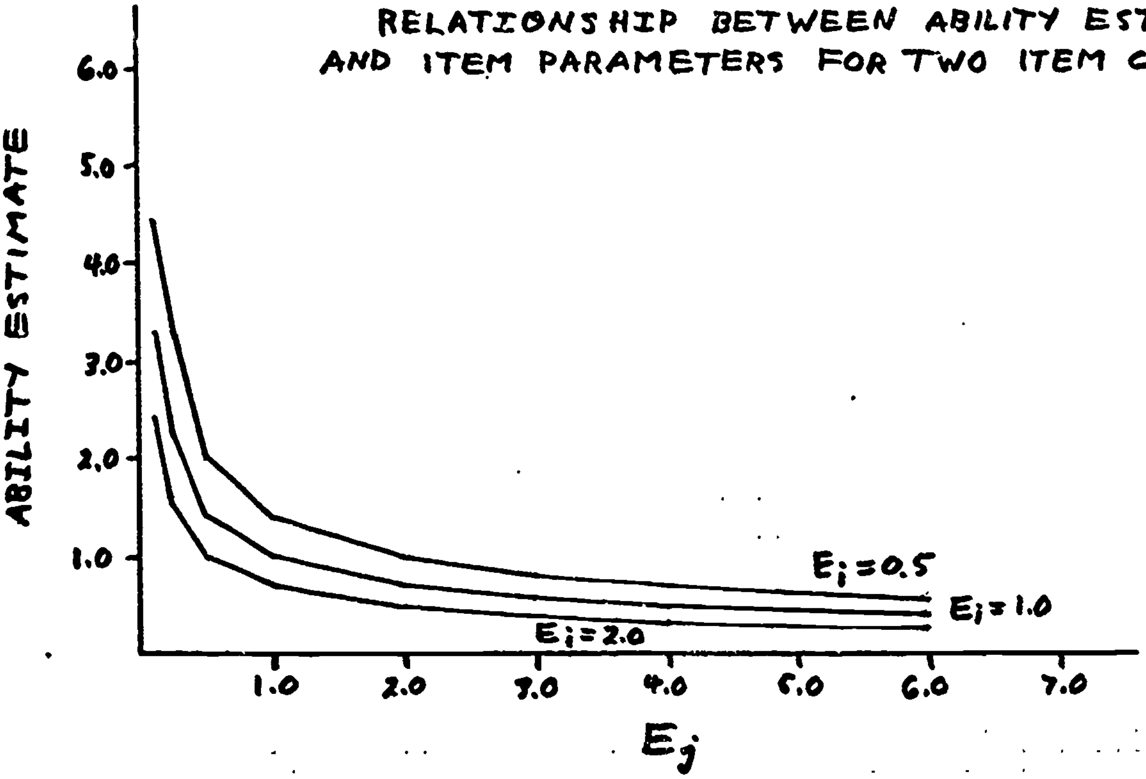
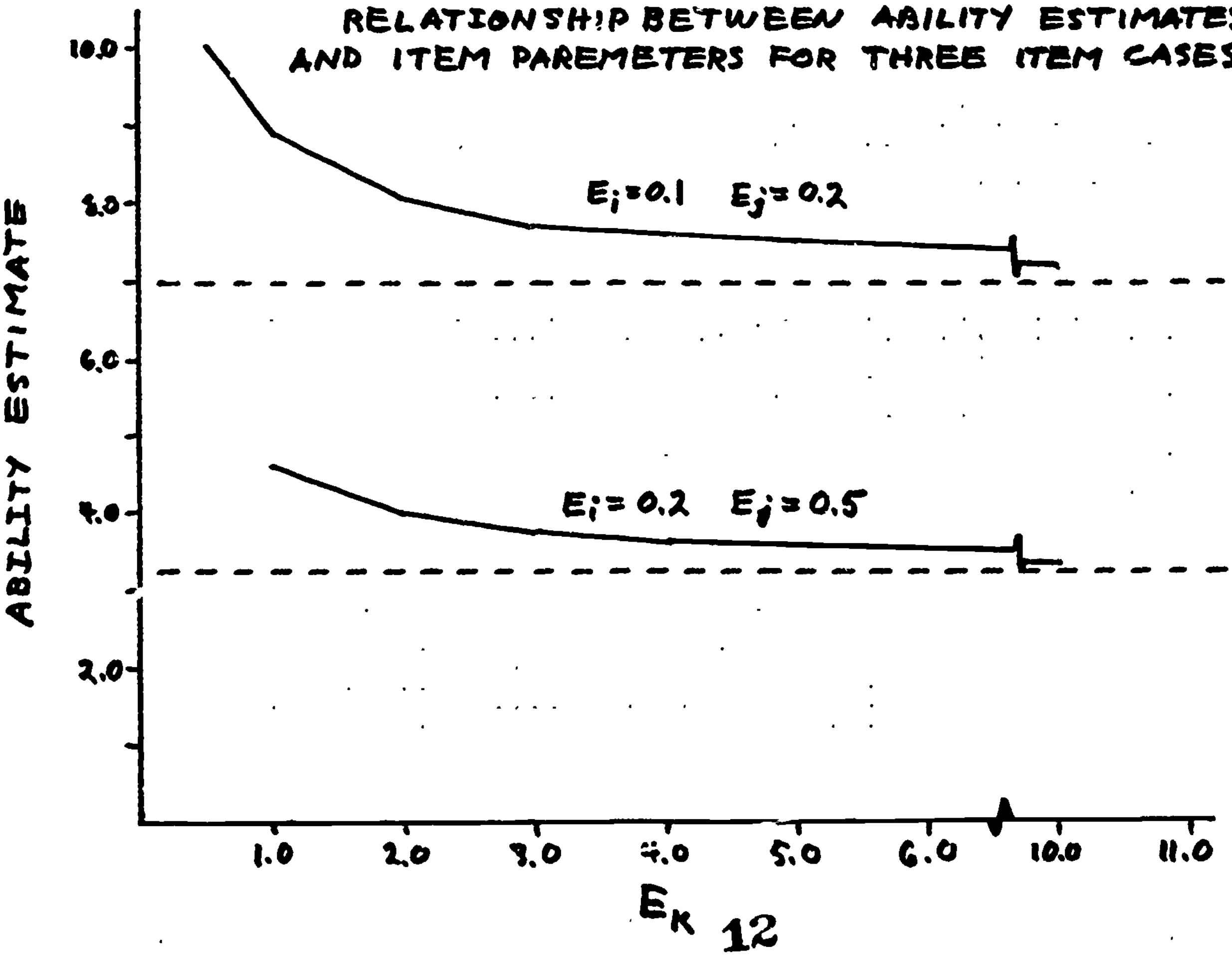


FIGURE 2
RELATIONSHIP BETWEEN ABILITY ESTIMATES AND ITEM PAREMETERS FOR THREE ITEM CASES.



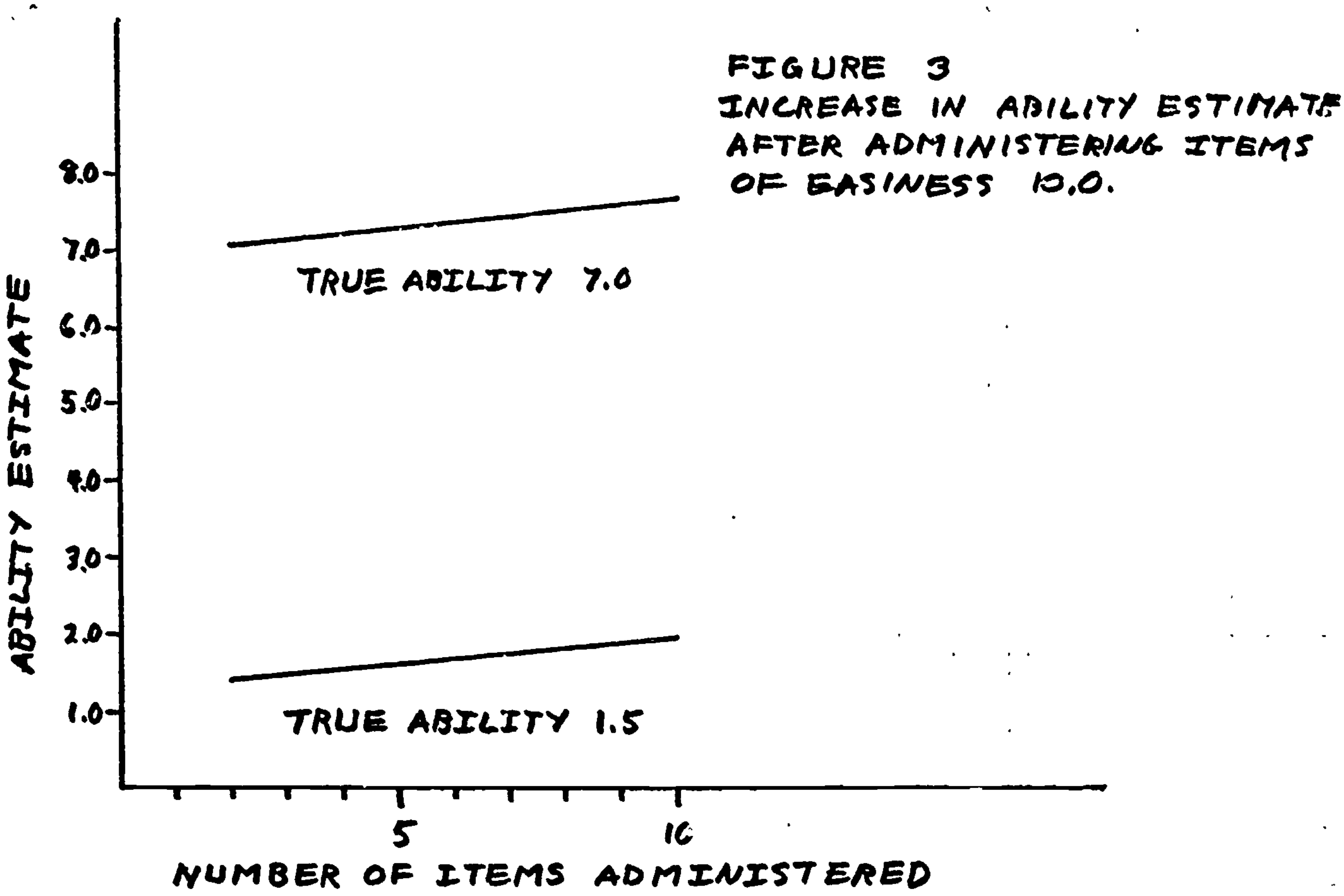


FIGURE 3
INCREASE IN ABILITY ESTIMATE
AFTER ADMINISTERING ITEMS
OF EASINESS 10.0.

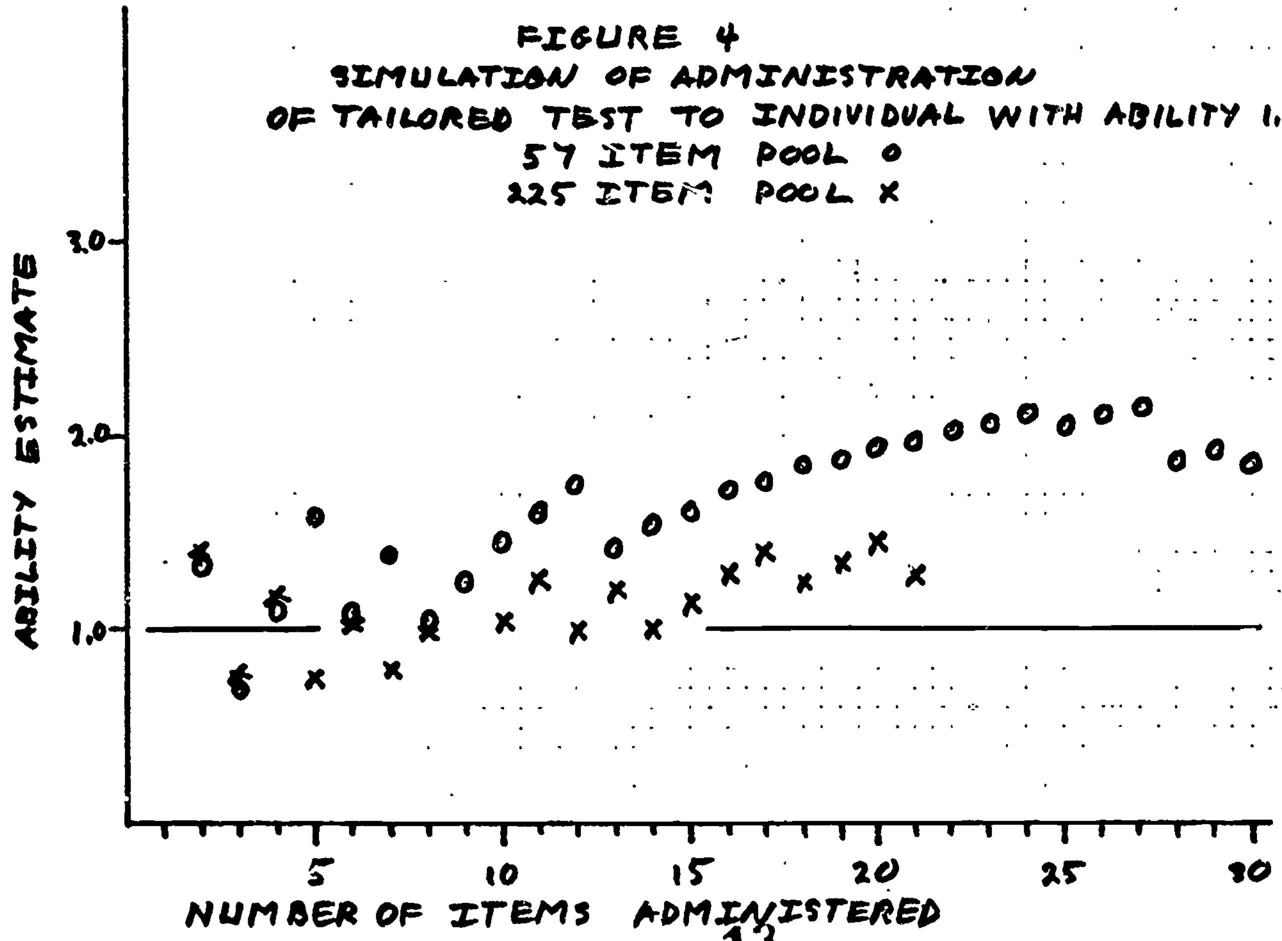
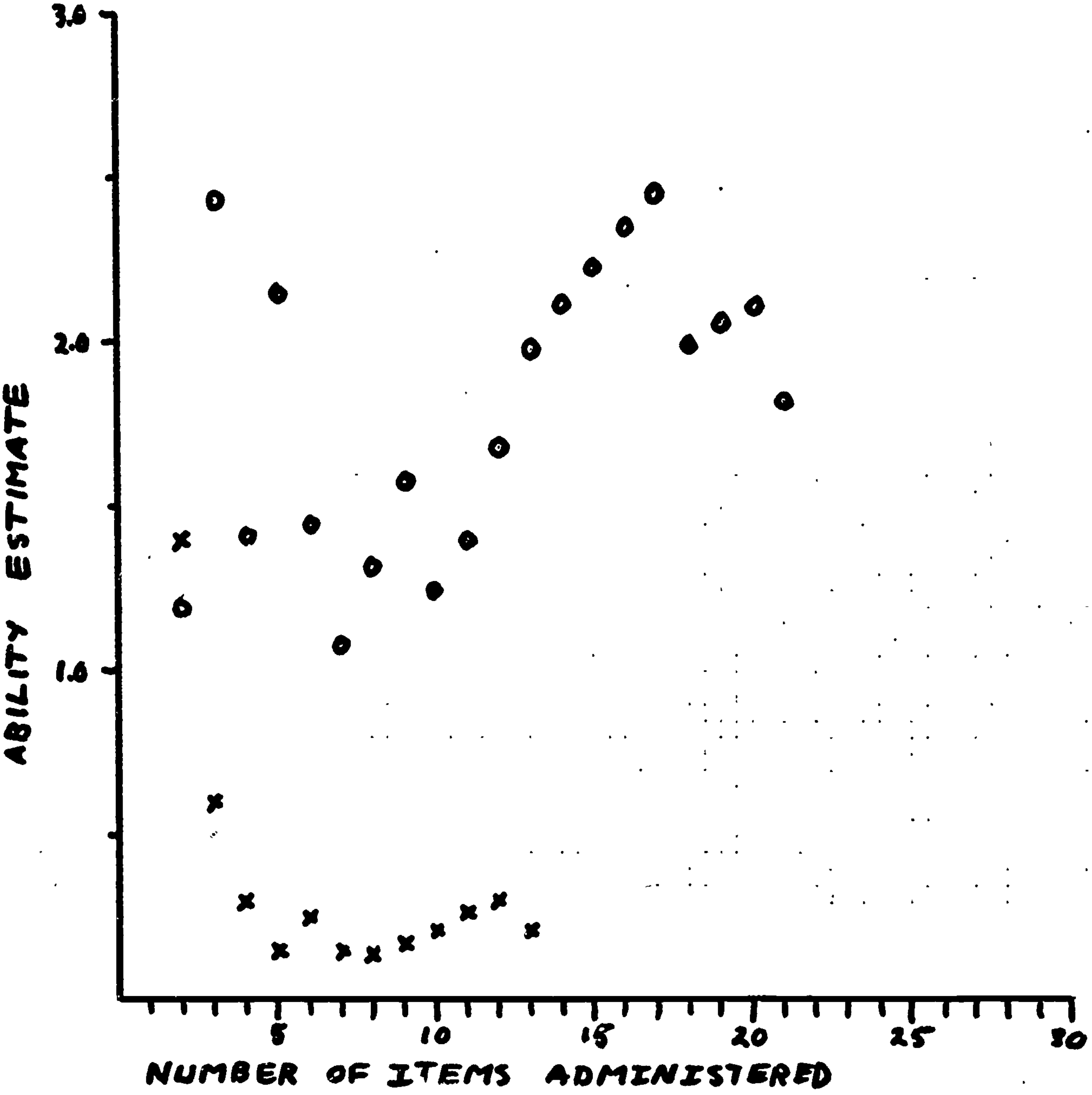


FIGURE 4
SIMULATION OF ADMINISTRATION
OF TAILORED TEST TO INDIVIDUAL WITH ABILITY 1.0

FIGURE 5
TAILORED TEST RESULTS
FOR TWO INDIVIDUALS O, X.


ABILITY ESTIMATE
NUMBER OF ITEMS ADMINISTERED