DOCUMENT RESUME

ED 243 930 TM 840 241

AUTHOR Rubin, Lois S.; Mott, David E. W.
TITLE The Effect of the Position of an Item within a Test on the Item Difficulty Value.
PUB DATE Apr 84
NOTE 19p.; Paper presented at the Annual Meeting of the American Educational Research Association (68th, New Orleans, LA, April 23-27, 1984).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Difficulty Level; *Item Analysis; Minimum Competency Testing; Reading Tests; Secondary Education; *Test Format; Testing Problems; *Test Items
IDENTIFIERS Item Parameters; *Item Position (Tests); Rasch Model; Virginia

ABSTRACT

An investigation of the effect on the difficulty value of an item due to position placement within a test was made. Using a 60-item operational test comprised of 5 subtests, 60 items were placed as experimental items on a number of spiralled test forms in three different positions (first, middle, last) within the subtest composed of like items. Item data used resulted from Rasch one-parameter item response calibrations. Variations among the mean Rasch difficulties lay well within one standard deviation. Except for a few outliers, the item difficulty values graph within the 95 percent confidence limits for evaluating overall stability of the estimates. Thus, the consistency of these estimates support the notion that Rasch item parameters are not importantly affected by the position of an item. (Author)

The Effect of the Position of an Item Within a Test on the Item Difficulty Value

Lois S. Rubin

David E. W. Mott

Virginia Department of Education

Presented at the annual meeting of the American Educational Research Association in New Orleans, April 1984.

ABSTRACT

The Effect of the Position of an Item Within a Test on the Item Difficulty Value

An investigation of the effect on the difficulty value of an item due to position placement within a test was made. Using a sixty-item operational test comprised of five subtests, sixty items were placed as experimental items on a number of spiralled test forms in three different positions (first, middle, last) within the subtest composed of like items. Item data used resulted from Rasch one-parameter item response calibrations. Variations among the mean Rasch difficulties lay well within one standard deviation. Except for a few outliers, the item difficulty values graph within the 95% confidence limits for evaluating overall stability of the estimates. Thus, the consistency of these estimates support the notion that Rasch item parameters are not importantly affected by the position of an item.

The Effect of the Position of an Item Within a Test on the Item Difficulty Value

A deficiency present in classical test theory approaches is that both item and test characteristics are dependent on the specific attributes of the examinees on which the statistics are gathered. For example, characteristics of items such as difficulty and discrimination vary across groups of examinees with different distributions of ability. Test indices relating to reliability and validity also are affected by the abilities of the examinees taking the test. In contrast, one of the most important attributes of item response theory is the supposed invariance of item parameters across groups (Lord, 1980). That is, the characteristics of each item can be described by one set of values. This quality should allow test developers to gather item statistics on one occasion and use the information subsequently to compile tests having predetermined characteristics.

In item response theory a major assumption made is that the difficulty of individual items is not altered by the test context in which the items are placed. In the classical test theory approach, the validity of this assumption is not crucial since data are collected on the test as a single entity (Whitely & Dawis, 1976). However, since item response theory requires collection of data on items, context effects occurring as the general result of the sequencing of items or as the result of specific characteristics of the other items in the test could have important influences. In practice, statistics are gathered on items either by means of special field test procedures or by placing experimental items on tests administered operationally to examinees. After test form specific statistics are calculated for these new items, the items are linked to the common scale of an item pool or item bank to await use on future tests. Since items are chosen for new tests on the basis of the statistics previously gathered, the accuracy of these statistics is very important to the integrity of any future test. "These item parameter values also influence the trait values that are obtained subsequently for any given examinee's pass/fail responses to the items, and the parameter values influence the standard error of the trait value provided by the latent trait model" (Yen, 1980, p. 297).

Whitely and Dawis (1976) investigated context effects on classical (p-value) and Rasch (one-parameter logistic model) item difficulties by using a verbal analogies test. A core of fifteen items were placed on seven different tests. Each test consisted of sixty items, the fifteen core and forty-five unique items. The tests, administered in sixty-minute sessions, were distributed randomly in each of seven sessions. Of the fifteen items, six had statistically significant differences in Rasch difficulties and nine had statistically significant differences in classical difficulties across the seven tests. Yen (1980) compared differences in context effects for mathematics and reading items in both the three- and one-parameter logistic models. It was found that item parameters estimated from the same context are more highly related than item parameters estimated from different contexts. Also, context effects appeared for both the three-parameter and the Rasch item difficulties, weaker for the three-parameter than for the Rasch on the reading items, and the reverse on the mathematics items. In addition, although context effects were found to influence the shape of the obtained item and test characteristic curves, these influences were less for the Rasch model than for the three-parameter model.

As Kingston and Dorans (1982) point out, there are only two alternatives when considering the use of precalibrated items on a test. Either the item must be placed in the same position on the new test as on the test used for item parameter calibration, or the position of an item must make no appreciable difference on the item difficulty. Because the first alternative is usually not feasible on account of the administrative complexities involved, a systematic investigation was made on the effect on obtained item difficulties when the item's position varied. Experimental items were placed at the beginning, middle, and end of subtests composed of like items.

METHOD

The data for this study came from the March 1983 administration of the Virginia Minimum Competency Reading Test given to approximately 80,000 students. The Rasch one-parameter logistic model had been chosen as the basis for test development and longitudinal equating for this program. Thus, items used are selected to fit this model. The regular editions of the reading test are comprised of sixty operational items divided into five competencies or subtests (twenty items in the first competency and ten items on each of the other four competencies). The test is similar in format and content to the IOX Basic Skill Tests: Secondary Level, Reading (IOX, 1978). When experimental items are placed on the forms, the total number of items per form is usually raised to eighty; however, for this investigation the total number of items per form was eighty-four. The test is administered with no time limit. For this study, a total of sixty experimental items were placed on different forms in each of three positions (first, middle, last) within their respective competency.

Eighteen forms of the test, containing the same operational items but different experimental items, were administered in a spiralled fashion (i.e., packaging the forms in sequential order, with packages beginning with as many different form numbers as forms being administered). This type of administration resulted in randomly parallel groups taking each form. After all student answer sheets were scanned, a random sample of 10,000 students was drawn for the purpose of calibrating the items using the BICAL III computer program (Wright, Mead, & Bell, 1979). Thus, the items in each of the forms were calibrated with a sample of approximately 550 students. When the item difficulties in a form are calibrated by BICAL III, the mean of the item difficulties is set to zero. Item difficulty calibrations are anchored relative to the other items in the test form. The sixty operational items of the March administration constituted the core of items used to link the eighteen forms together and to the common scale of the existing item bank (Wright & Stone, 1979). No experimental items were used in the linking process. The item difficulty parameters reported are those adjusted to the scale of the existing bank. The p-values are the actual obtained values.

RESULTS

The means and standard deviations of all the item difficulty estimates in each of the three positions (first, middle, last) are presented in Table 1. The greatest difference in mean difficulty values is between the first and middle positions and that is .144. Between first and last position the difference in means is .049 and between middle and last the difference is .095. The means of the difficulty estimates for the items within each subtest is displayed graphically in Figure 1. The greatest variation in means is in the fifth subtest and the least is in the fourth subtest.

In Table 2 the means and the standard deviations of the p-values of all the items in their respective positions are presented. These mean p-values differed by .003 to .012, with the greatest difference between the first and middle positions. The mean p-values for the items within each of the subtests as shown in the graph in Figure 2.

The mean ability estimates for each of the forms wherein the items under discussion were placed are presented in Table 3. These range from a high of 2.99 to a low of 2.71. Table 3 also contains the person separability indices (PSI) for each experimental form. This index calculated during the BICAL III calibrations of the item difficulty estimates is similar to the index of subject separability (ISS) described by Gustafsson (1977).

Figures 3-5 display graphs of the difficulty estimates of each item in one position plotted against the difficulty estimate of the same item in another position. The correlation coefficients relating to the graphs are also presented. All three correlations are 0.95 or higher.

A one-way analyses of variance indicated that there was no significant difference between the means of the Rasch difficulty estimates of the items placed in each of the three positions, $\underline{F}$ (2,118) = 2.57, $\underline{p}$ .05.

DISCUSSION

Some variation can be seen among the mean difficulty value estimates for the items in the different positions (first, middle, last) within their respective subtests; however, the differences between these means lay well within one standard deviation. The mean p-values show less variation among the different positions.

Since the examinees taking each form are randomly assigned from the same population, the ability estimates were expected to be similar and they were. All forms contained at least two experimental items other than those used for this study, so the effect of the different positions of the items on the ability estimates can not be determined conclusively. However, the information presented shows that the ability estimates of the examinees taking each form are very similar.

The person separability index (PSI) is almost identical for all forms. This index serves as a counterpart to the coefficient of reliability when direct estimates of the variance of the errors of measurement can be obtained (Gustafsson, 1977). The PSI is sample specific and is lower when the ability level of the examinees is not measured precisely. This seems to be the case in this study. The ability estimates on the forms vary from 2.99 to 2.71, when the items have a mean difficulty value of zero. Because the data were derived from an administration of a high school minimum competency test, it might be expected that the mean of the difficulty estimates of the items would be much lower than the mean of the examinee ability. In such circumstances, values of the PSI are expected to be in the 0.80 to 0.85 range because the test is not precise (Wright & Stone, 1979).

Viewing the graphs of the comparative item difficulty values, the invariance property of the Rasch model becomes evident. Except for a few outliers, all points lie within the 95% confidence limits for evaluating the overall stability of the difficulty estimates for the same items as described by Wright and Stone (1979). The statistic suggested by Wright and Stone is $t_{i12} = (d_{i1} - d_{i2})/(s_{i1}^2 + s_{i2}^2)^{1/2}$, with an approximate normal distribution having a mean of 0 and a standard deviation of 1. $(s_{i1}^2 + s_{i2}^2)^{1/2}$ is an estimate of the expected standard error of the difference between two difficulty estimates d_{i1} and d_{i2}, independently calibrated, for one parameter δ_i. The reasons for the items producing inconsistent difficulty parameters are not apparent. No item appears as an outlier on all three graphs.

The consistency of the difficulty estimates of the items placed in different positions seems to support the notion that Rasch item parameters are not importantly affected by the position of an item. Context effects such as those produced by the individual characteristics of adjacent items, as opposed to general position effects, may play a part in causing the few items to be outliers. However, this study concentrated only on the general position effects. Other investigations are planned to look at specific context effects within the subtests.

Table 1

Item Difficulty Values

Position	Mean	S.D.
First	.936	1.200
Middle	.792	1.313
Last	.887	1.205

Table 2

p-values

Position	Mean	S.D.
First	.833	.103
Middle	.845	.103
Last	.836	.101

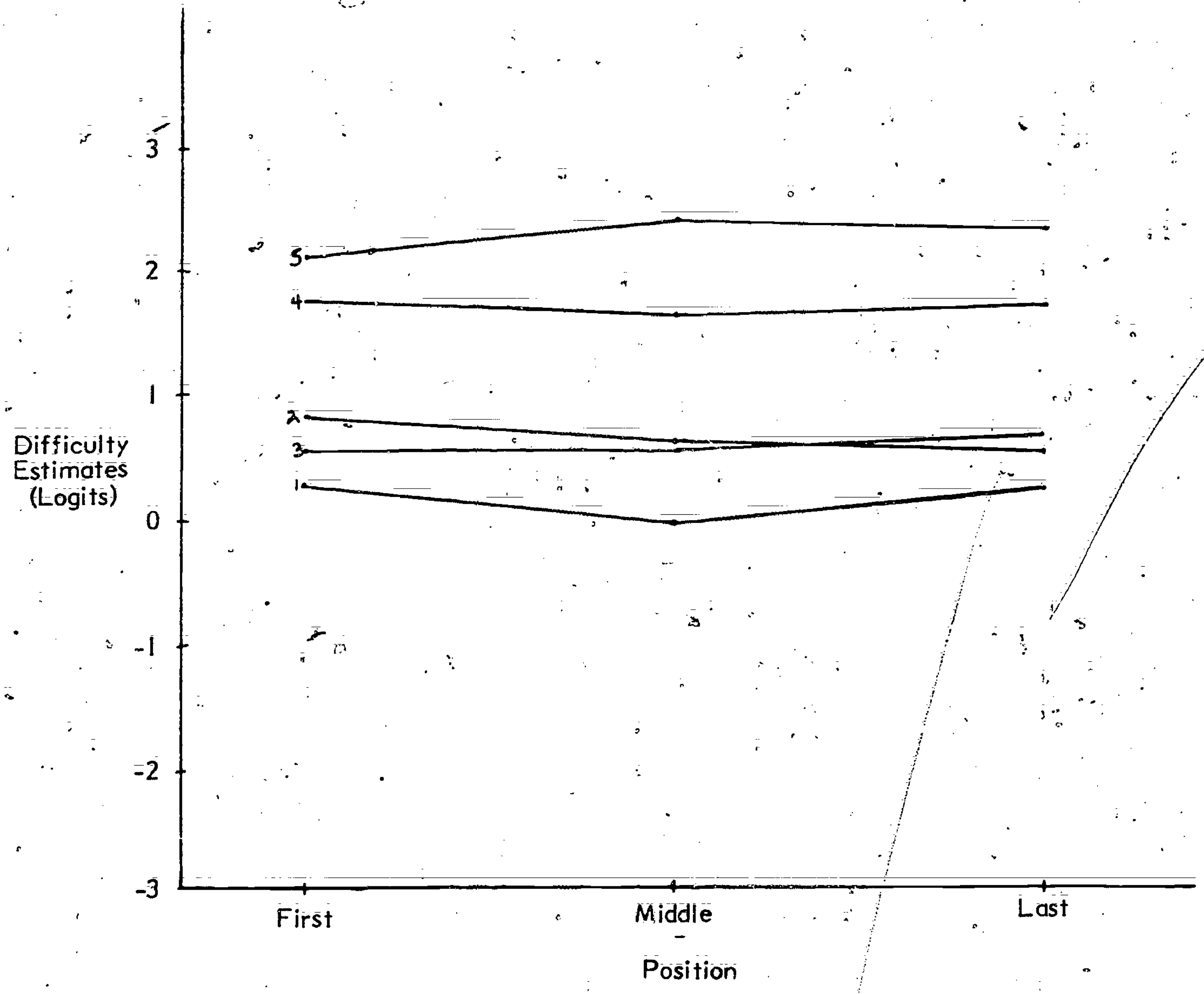

Figure 1. Mean Item Difficulty Estimates

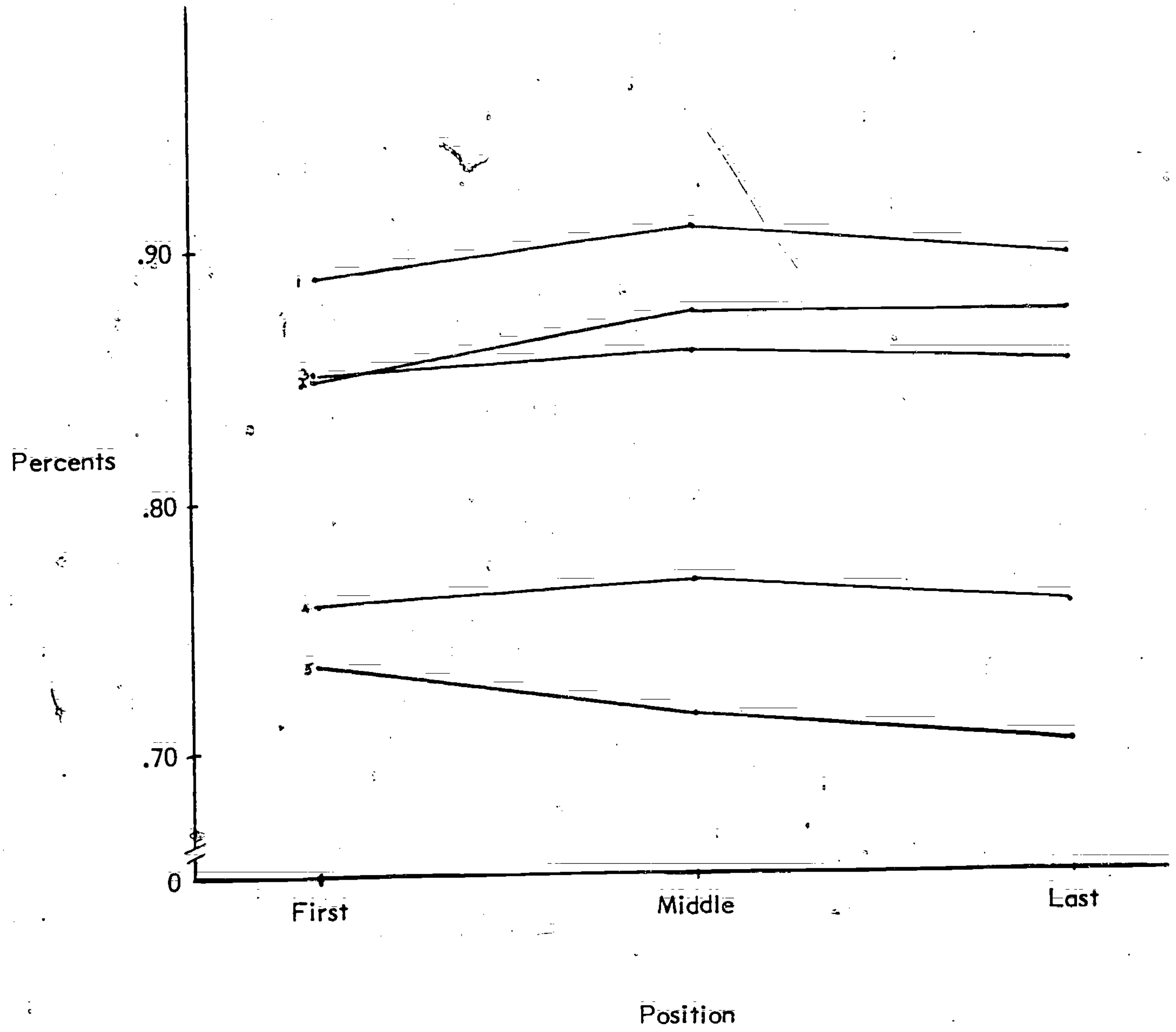

Figure 2. Mean Item p-values

Table 3

Form Statistics

Form	Ability Estimates (Logits) Mean	 S.D.	Person Separability Indices
1	2.73	1.14	.84
2	2.75	1.20	.85
3	2.89	1.22	.83
4	2.74	1.19	.85
5	2.91	1.15	.82
6	2.71	1.19	.83
7	2.89	1.12	.82
8	2.99	1.20	.83
9	2.75	1.12	.84
10	2.94	1.10	.81

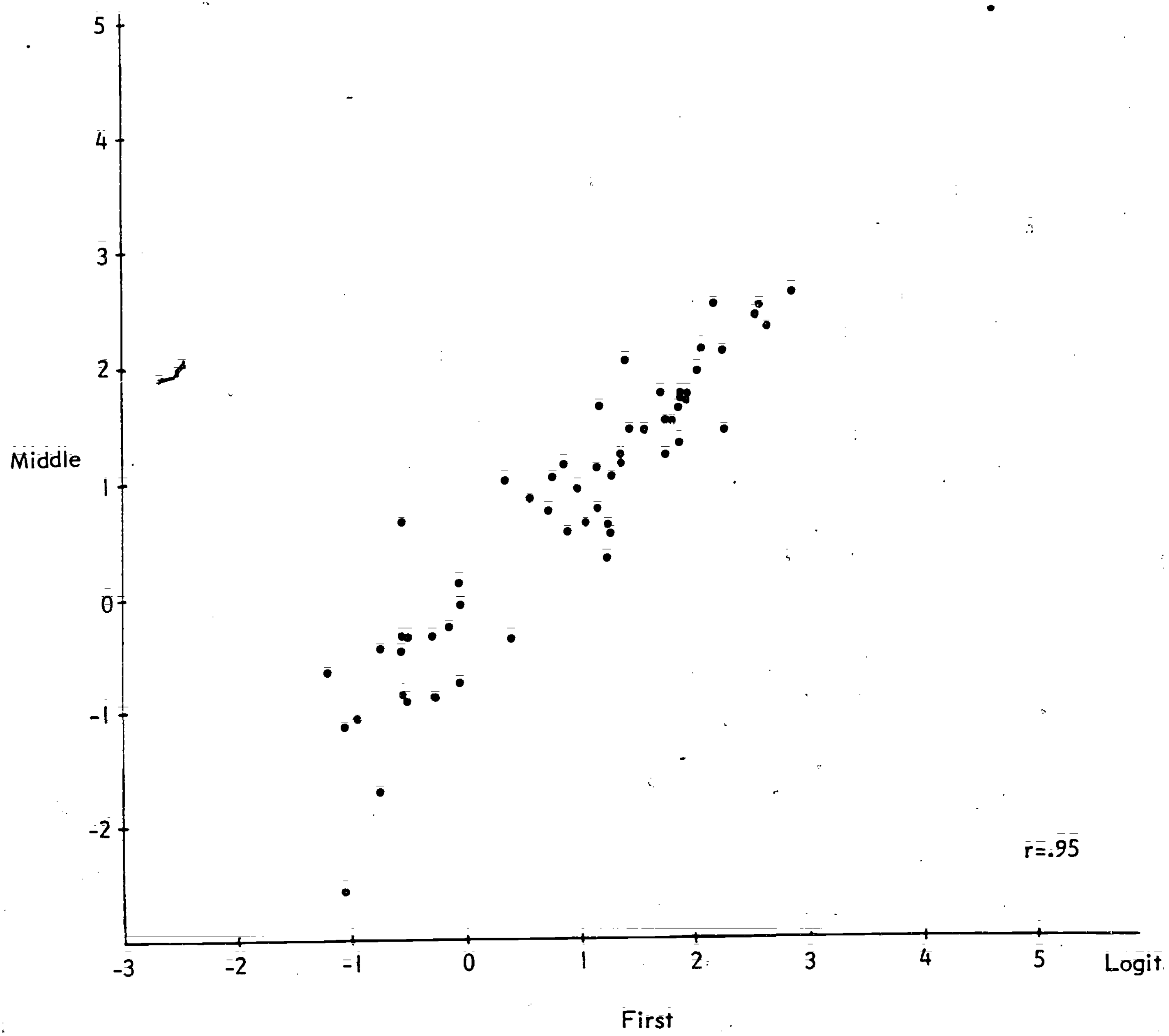

Figure 3. Item Difficulty Estimates (Logits)
First and Middle Positions

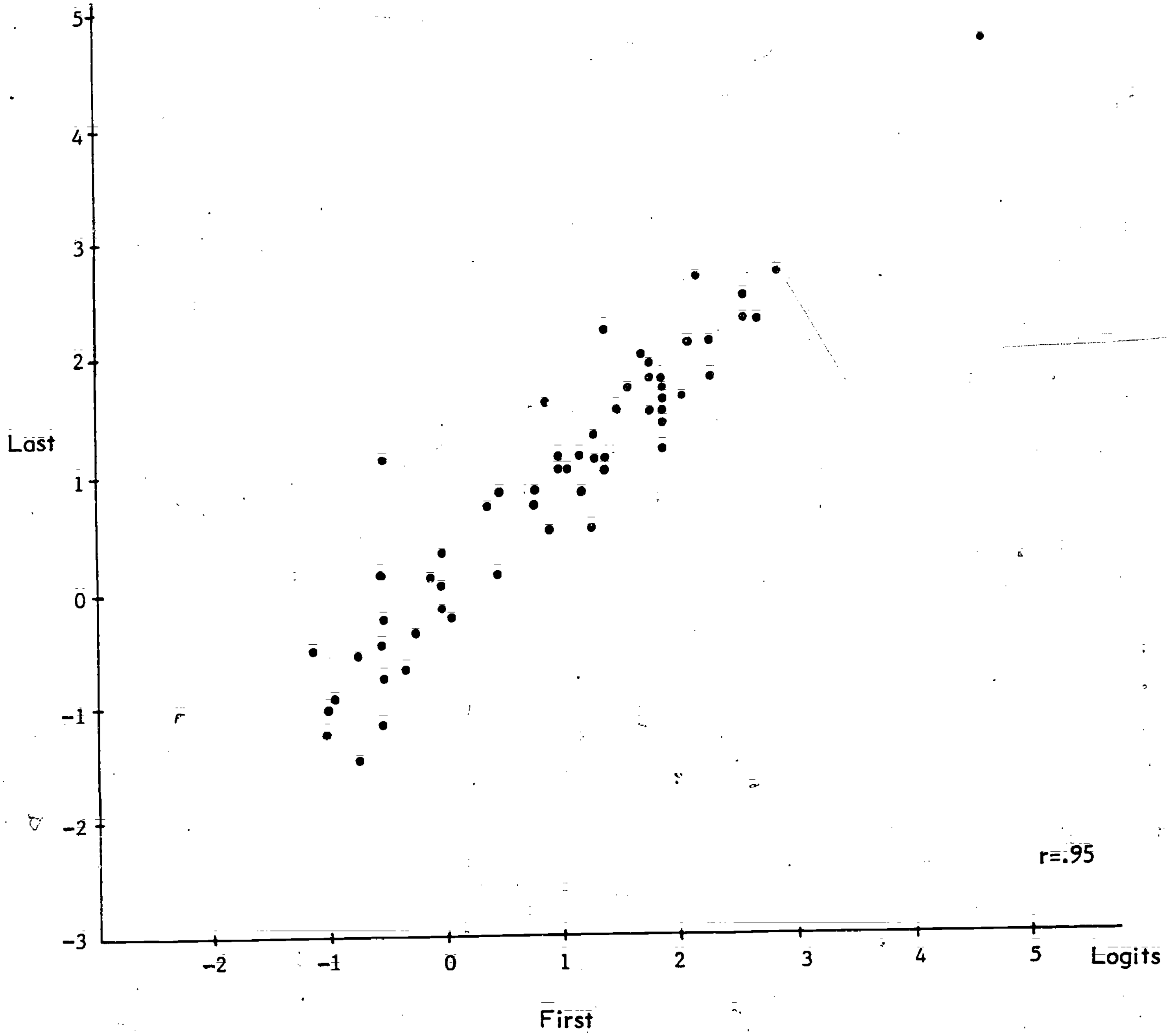

Figure 4. Item Difficulty Estimates (Logits)
First and Last Positions

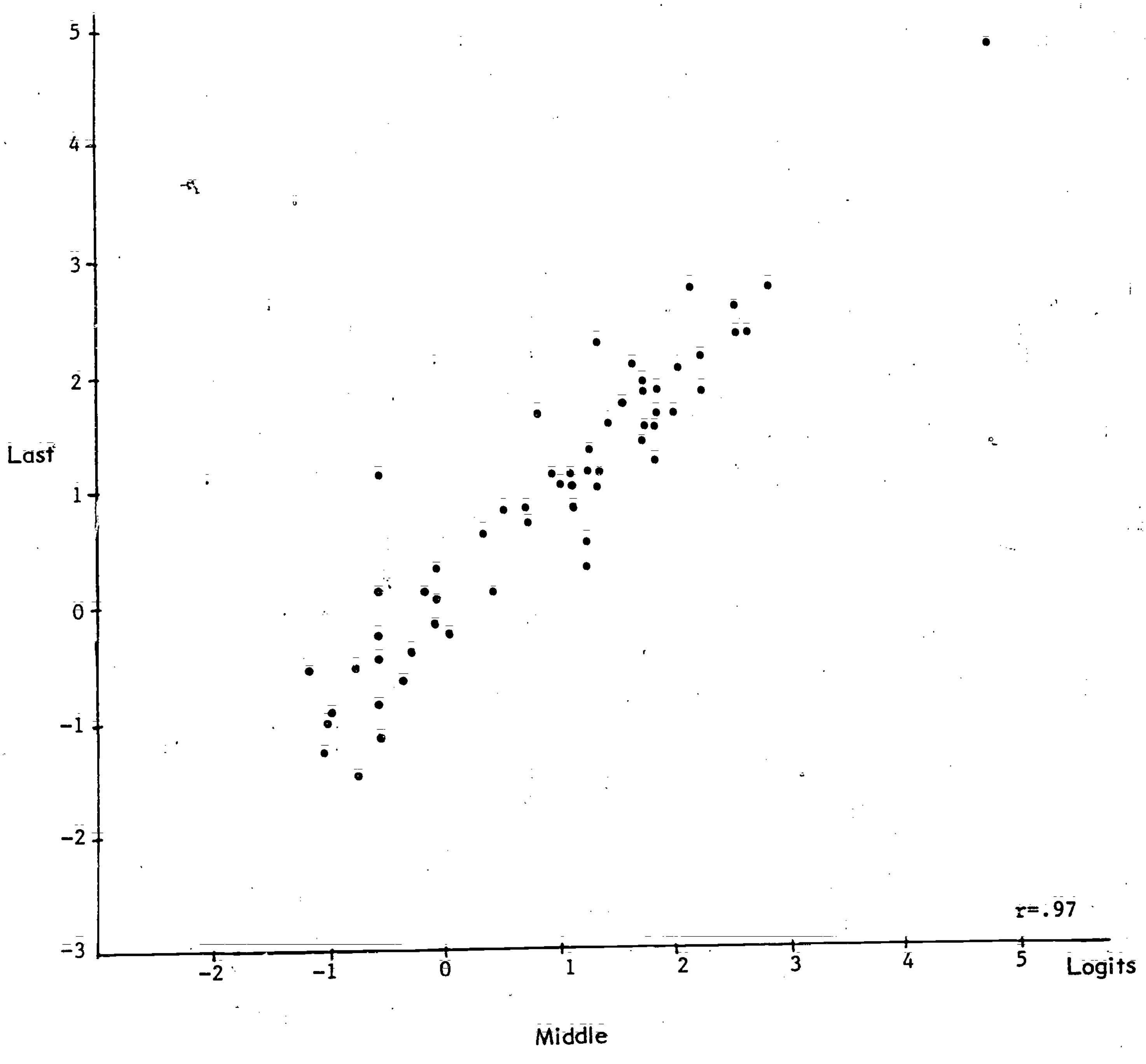

Figure 5. Item Difficulty Estimates (Logits). Middle and Last Positions

References

Burton, N.W., Larson, R.C., & Pearson, A.M. (1976, April). The effect of position and format on the difficulty of assessment exercises. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Gustafsson, J.E. (1977). The Rasch model for dichotomous items: Theory, application, and a computer program. No. 63. Sweden: Gothenburg University, Institute of Education. (ERIC Document Reproduction Service No. ED 154 018).

IOX Basic Skill Tests: Secondary Level, Reading. (1978). Los Angeles: The Instructional Objectives Exchange.

Kingston, N.M. & Dorans, N.J. (1982). The effect of the position of an item within a test on item responding behavior: An analysis based on item response theory. (GRE Board Professional Report No. 79-12bP). Princeton, NJ: Educational Testing Service.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum.

Whitely, S.E. & Dawis, R.V. (1976). The influence of test context on item difficulty. Educational and Psychological Measurement, 36, 329-337.

Wright, B.D., Mead, R.J., & Bell, S.R. (1979). BICAL: Calibrating items with the Rasch model. (Research Memorandum No. 23B). The University of Chicago: Statistical Laboratory, Department of Education.

Wright, B.D. & 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.