DOCUMENT RESUME

ED 333 024 TM 016 504

AUTHOR Cizek, Gregory J.
TITLE The Effect of Altering the Position of Options in a Multiple-Choice Examination.
PUB DATE Apr 91
NOTE 20p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Chicago, IL, April 4-6, 1991).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Certification; Comparative Testing; *Difficulty Level; Higher Education; *Licensing Examinations (Professions); Medicine; *Multiple Choice Tests; *Physicians; Specialists; Test Construction; *Test Format; Test Items
IDENTIFIERS *Item Position (Tests)

ABSTRACT

A commonly accepted rule for developing equated examinations using the common-items non-equivalent groups (CINEG) design is that items common to the two examinations being equated should be identical. The CINEG design calls for two groups of examinees to respond to a set of common items that is included in two examinations. In practice, this rule has been extended to include the order in which options appear in the two examinations. The performance of a common set of items in which the order of options for one test form was experimentally manipulated was examined to determine if reordering multiple-choice item options resulted in any significant effect on item difficulty. Data from 759 subjects (graduates of medical specialty residency training programs) were gathered as part of the annual administration of a certification examination in a medical specialty area. Each subject responded to 20 multiple-choice items with a projected visual as the stimulus for each item. Examinees had to select from about 30 choices the option that correctly identified the projected visual. Two response booklet forms, differing only in that the position of the 20 options was scrambled, were used. A total of 380 examinees responded to Booklet A, and 379 examinees responded to Booklet B. One examinee was randomly excluded from the analyses for Booklet A. It was found that reordering items often has significant but unpredictable effects on item performance. A linkage is made to previous research on the "response set" construction, and cautions are suggested regarding the effect of reordering options. (SLD)

The Effect of Altering the Position of Options in a Multiple-Choice Examination

Gregory J. Cizek

American College Testing

Paper presented at the annual meeting of the

National Council on Measurement in Education, April, 1991

ABSTRACT

A commonly accepted rule for developing equated examinations using the common-items nonequivalent groups (CINEG) design is that items common to the two examinations being equated should be identical. In practice, this rule has been extended to include even the order in which options appear in the two examinations.

The present study examined the performance of a common set of items in which the order of options for one test form was experimentally manipulated. The study sought to determine if reordering multiple-choice item options results in any significant effect on item difficulty. It was found that reordering options often has significant but unpredictable effects on item performance. A linkage is made to previous research on the "response set" construct and cautions are suggested regarding the effects of reordering options.

The Effect of Altering the Position of Options in a Multiple-Choice Examination

One frequently used design for equated examinations is the Common Items Non-Equivalent Groups (CINEG) design. This design calls for two groups of examinees to respond to a set of common items that is included in two examinations. The performance of the two groups of examinees on the common items is used to calibrate or "equate" the two sets of scores. Full descriptions of the CINEG design for equating have been ably presented elsewhere (Kolen & Brennan, 1987; Braun & Holland, 1982; Petersen, Kolen, & Hoover, 1989; Kolen, 1989; Dorans, 1990).

One frequently encountered test construction requisite for using the CINEG equating design is that the equating item set should represent a miniature form of the total test. That is, the common items should be assembled to match the total test specifications in terms of distribution of content coverage, difficulty, etc. The rationale for this procedure has also been explained and documented elsewhere (Klein & Jarjoura, 1985).

A second frequently encountered test construction rule for equated examinations using common items is that the common items should be identical in the two examinations. Such an admonition is found in Angoff (1971 p. 578) who stressed that the equating item set should "represent the same psychological task to both [examinee] groups. . ." However, in some cases it may be impractical (or impossible) for the test constructor to heed that admonition. For example, in licensure and certification testing programs serving rapidly changing professions, the test constructor can be faced with the dilemma of whether to a) use an item that has been changed in some--hopefully minor--way, b) utilize fewer equating items to make up to common item set or, c) compromise the desired representativeness of the common item

set. None of these options seems desirable. However, each of the options should be investigated to determine the extent to which their implementation might affect the integrity of the equating process.

Background

As noted above, the consequences of violating common item representativeness of the total test have been documented (Klein & Jarjoura, 1985) and the consequences can be serious. Also, many researchers have examined the number of equating items that should be used to comprise the common item set. Research has also generally supported the rule of thumb offered in Angoff (1971, p. 578) that the common item set should consist of at least 20 items or 20 percent of the total test, whichever is greater. However, research is still needed that examines the effect of changes to items on the items' performance characteristics.

Relevant research into the effect of changes on multiple-choice item performance has been focused in two areas: textual changes and format differences. Investigations into effects of textual changes are abundant. Cassels & Johnstone (1984) manipulated the complexity of language in matched pairs of chemistry items and observed small effects on item difficulty for the use of active or passive voice. Larger effects were observed for changes in key words in the item stems, changes in terms involving quantities, changes in overall complexity of the item stems, and for the use of positive versus negatively worded items.

Green (1984) reported that a review of the effect of variation in multiple-choice item "phrasing" yielded differing results. Investigations involving variations in phrasing using samples of children and young adults (Bensen & Crocker, 1979; Bolden & Stoddard, 1980; Lofton & Suppes, 1972) found significant effects on test performance. Investigations using high school and

college subjects (Bornstein & Chamberlain, 1970; Jerman & Mirman, 1971; Millman, 1978) yielded mixed conclusions about the effect of variations on performance. Green's own research revealed no effect of altering the difficulty of the language used in "general information" test items, but a significant effect was observed for option convergence (i.e., the extent to which the response choices were similar).

O'Neill (1986) also investigated the effect on item performance of certain textual changes for multiple-choice items. O'Neill examined changes in abbreviations, symbols, and drug names on a licensure examination and found "no evidence that these types of stylistic manipulations affect examinee performance on individual items" (p. 7).

The possible effects of changes in item performance resulting from alterations in item format is a less well-researched topic. For example, Harris (1990) investigated effects of minor changes in the position of items and passages on the ACT assessment. Harris found that "up to 50% of the examinees [administered one of the scrambled forms] would receive different scale scores if the base from [equating] conversions were used instead of the scrambled Form C conversions" (p. 11).

At the item level, Ace & Dawis (1973) reported on an early study by McNamara and Weitzman (1963) that found placement of the correct response in a multiple-choice item to be a factor influencing item difficulty. For five-option items, those with the correct response in the fourth position were more difficult; those with the correct response in the second or third position were easiest. Ace and Dawis' own work, which concluded that "correct response position is probably a significant determinant of item difficulty" (1973, p. 147), generally supported the earlier work of McNamara and Weitzman.

On the other hand, an experiment conducted by Marcus (1963) found no effect of correct response position on item difficulty. Wilbur's (1966) investigation similarly found no effect of correct response placement on multiple-choice item difficulty.

Purpose

It appears from a review of relevant literature that the question of whether correct response placement does affect multiple-choice item characteristics has not yet been fully resolved. The present study examines the possibly inconsequential--though common--change to multiple-choice items: reordering of response options.

Currently, it is generally believed--and practiced--that equating (common) items should have options that are textually identical and that those options should appear in the same order in the new (current) and old (anchor) forms. There are reasons why, however, that it may be desirable to reorder the options of a multiple-choice item. First, although there is some research to the contrary (cf. Jessell & Sullins, 1975) the practice of key balancing (equalizing the number of times each option is the correct response) is common. To achieve a balanced key, options on equating items may need to be reordered. Second, because of concerns about test anxiety it is sometimes desirable to avoid the situation where one response is the correct answer for several contiguous items. Finally, it is sometimes the case that a stylistic change is desired to place the options in some logically or aesthetically appealing manner. For example, it might be desirable to order the text of options from longest to shortest length, or to order numerical choices to reflect increasing or decreasing magnitude.

The present study attempted to determine if reordering multiple-choice item options results in any significant effect on the items' difficulty.

Practically, if reordering options does not have a significant effect on item performance characteristics, the practice need not be necessarily avoided. Conversely, if option reordering results in changes to item performance statistics, test makers should be cautioned about such practice, especially in the context of equated examinations.

Subjects

Data for the study were gathered as part the annual administration of a certification examination in a medical specialty area. Examinees (n = 759) were all graduates of medical specialty residency training programs who were taking the medical specialty board qualifying examination.

Instruments

Each subject responded to 20 multiple-choice items as part of the qualifying examination. The stimulus for each of the 20 items was a projected visual. Examinees were required to select, from a list of approximately 30 choices, the option that correctly identified the projected visual. Two forms of an examinee response booklet were utilized. The two response booklet forms differed only in that the position of the 30 options was scrambled. Thus, for example, the correct response to item 1 might appear a choice "C" (the third option) on Form A, but would be in position "BB" (the 28th option) on Form B.

The unusual design of the response booklets allowed for the direct comparison of nearly identical items attempted by randomly equivalent groups of examinees where the only difference in the items was the position of the correct response. One particularly intriguing characteristic of the two response booklets was that the possibilities for distance between options was increased over that possible in a typical four- or five-option multiple-choice item. That is, the options for a five-option item can only be reordered such

that the maximum difference in the position of the correct answer would be four places (e.g., changed from A to E). In the present study, a much wider range of possibilities existed; specifically, a maximum distance difference in correct option position of 29 places was observed. This characteristic allowed the research to be especially sensitive to how differences in the distance of correct response placements might affect item performance. Table 1 shows the position of correct responses in the two forms and the absolute form-to-form difference in positions.

INSERT TABLE 1 ABOUT HERE

Procedures

The total group of 759 examinees was randomly assigned to two groups; one group received Form A of the examination response booklet and the other group received Form B. In group 1, 380 examinees responded according the order of options in Booklet A; Group 2 had 379 examinees responding according to the order options in Booklet B. One examinee was randomly excluded from Group 1 for the rest of the analyses. A check on the equivalence of ability in the two examinee groups was performed based on their total score on the full (200-item) examination; a t-test revealed no significant difference ($t = .54$, $p = .587$) in group mean scores. An F-test using the ratio of the variances (Group 1 = 22.08^2; Group 2 = 22.66^2) showed that the group scores were also likely to be of equally variability ($F_{378,378} = 1.05$; ns).

Results

Also shown in Table 1 are the item difficulty and discrimination values for the 20 items in the two forms. For 14 of the 20 items, the Form B version

was more difficult than the Form A version. In four cases, the Form B version was easier than the Form A version. For two items, Form A and Form B p-values were the same. Mean difficulties for the Form A and Form B items were .643 and .621, respectively. Mean discriminations were .372 for Form A and .385 for Form B.

Figure 1 is a plot of the Form A and Form B p-value pairs. The plot reveals a nearly perfect linear relationship between the pairs. The product-moment correlation between the pairs was +.992.

INSERT FIGURE 1 ABOUT HERE

Despite the high correlation between item p-values on the two forms, some fairly large differences in form-to-form performance were noted. Although two items showed identical performance (in terms of difficulty) for the two forms, the remaining items displayed changes in difficulty. To examine the extent of change, the difference between Form A and Form p-values was taken (DIFFA-DIFFB). These differences ranged from -.02 to +1.0.

To examine whether any of the changes in difficulty were statistically significant, t-tests between the pairs of p-values were conducted. Of the 18 pairs tested, four showed statically significant differences. Referring to Table 1, item pairs 5 and 18 and pairs 7 and 16 exhibited statistically significant differences at alpha levels .05 and .01, respectively.

To further assess the possibility of changes in difficulty being attributable to changes in position of the correct response, differences in correct response position were plotted against differences in item difficulty (see Figure 2).

INSERT FIGURE 2 ABOUT HERE

First, two variables were created. DIFFA-DIFFB represents the differences in p-values for an item pair, obtained by subtracting the item's Form B difficulty value from its Form A difficulty value. The second variable (POSA-POSB) represents the simple difference in correct response position for an item pair, obtained by subtracting the numerical position of the correct response in Form B from the position of the correct response in Form A. A correlation was calculated between the two variables and found to be -.0423.

At first glance, the near absence of any linear association between the two variables is confirmed by examination of the plot shown in Figure 2. However, further examination of the relationships shown in Figure 2 suggested some interesting observations. However, to more easily interpret the relationships shown in Figure 2, a second version of the plot was generated.

Figure 3 shows the second version of Figure 2, with some modifications. First, "cross-hairs" have been added that divide the plot into four

---

INSERT FIGURE 3 ABOUT HERE

---

quadrants. Clockwise, from the upper-left, the quadrants isolate items that were: 1) easier in Form B and whose correct response appeared earlier in Form B then in Form A; 2) more difficult in Form B and whose correct response appeared earlier in Form B than in Form A; 3) more difficult in Form B and whose correct response appeared later in Form B than in Form A; and 4) easier in Form B and whose correct response appeared later in Form B than in Form A. Also, rough (and arbitrarily placed) concentric circles have been added to

the plot in Figure 3. Finally, several item pairs have been highlighted: the four pairs with statistically significant differences in difficulty (5, 7, 16, 18) have been circled and three "outlier" item pairs (8, 9, 16) have been identified with a box.

Several interesting relationships are revealed in Figure 3. First, the innermost concentric circle shows that most item pairs fell into an area of weak association. That is, 13 of the 20 item pairs possessed both the characteristic of a small positional change for the correct response and a small change in p-value. Generally, small changes in correct response position seemed to have little effect on item difficulty. However, it is somewhat disconcerting to note that the slight effect on difficulty appears to be nearly unpredictable in terms of direction; i.e., it cannot be concluded from this data that placing the correct option later in the response list is associated with an increase or decrease in item difficulty. Second, for each of the four circled pairs (i.e., those that showed statistically significant mean differences), the Form A version was easier than the Form B version and the correct response appeared later in the Form B list of response options. However, although these pairs showed statistically significant form-to-form differences in p-values, there was not a consistent pattern with respect to change in correct response position. For example, note that items 5 and 16 differed in correct response placement by only 2 positions, whereas item 18 differed by 8 positions and item 7 differed by 23 positions in placement of the correct response.

It is also interesting to note what might be called "outliers." Three item pairs (8, 9, and 16) have been identified with a box. These items seem to escape any explanation for their performance. For example, pair 16 exhibited only slight form-to-form difference in correct response position,

but displayed the second highest magnitude for change in p-value. On the other hand, pairs 8 and 9 showed little and no difference, respectively, in difficulty although these pairs showed extreme positional changes.

Discussion

Much of the earlier research on the effects of correct response option placement in multiple-choice testing focused on a construct called "response set." The present research examined possible effects of correct response placement changes on item performance without response to that construct. However, in placing this research into that theoretical framework, it appears that any effect a response set construct might exert may be strongly mitigated by increasing the number of response options. This point might be obvious: as the number of possible choices in a multiple-choice item increases, the item effectively approaches similarity to an open-ended type item, an item type in which any effects of response set would disappear. Previous research on the existence of a response set construct has yielded mixed results. The present research suggests that even if the construct exists, it exists in a very limited sense, possibly only in multiple-choice format test items with a limited number of response choices. Thus, as an explanatory tool, response set may have limited usefulness.

The primary purpose of this research was to examine what effect alteration in correct response position might have on item performance, an especially important issue in the context of equated examinations. Results showed that even with a fairly small number of item pairs (20) and a relatively small number of examinees per group (378), several items displayed significant changes in overall performance. The majority of the item pairs however, differed little in terms of form-to-form changes in difficulty.

Overall, these findings suggest that altering the position of the correct

response in equated examinations is dangerous practice. Because equating conversion equations depend so heavily upon inferences regarding the cause of changes in item performance, it is recommended that test constructors exercise extreme caution when altering common items. This study suggests that item performance can be effected very little or to a great degree by changes in correct option placement. Most disconcertingly, the direction and magnitude of any changes seem almost unpredictable. This result should be cause for great concern whenever altered items are considered for use as equating items; in fact, the practice should apparently be proscribed in most usual circumstances. It might also be advisble to avoid reordering options on any examination where target levels of difficulty are sought. Because of the unknown direction and magnitude of p-value changes observed in this study, it seems apparent that in order to have the greatest confidence in individual item statistics (to be aggregated for total test estimations), the position of options should not be changed.

In summary, this research made some progress toward answering a practical question: whether changes in correct response position affect item performance. The research made little progress toward addressing the more theoretical issue: how or why item performance changes occur. The answer to the practical question should assist test makers and lends supports to current guidelines and "rules-of-thumb." The answer to the theoretical question--maybe the more interesting question--should be addressed further in future research efforts.

References

Ace, M. C., & Dawis, R. V. (1973). Item structure as a determinant of item difficulty in verbal analogies. Educational and Psychological Measurement, 33, 143-149.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), Education Measurement, 2nd Ed., pp. 508-600. Washington, DC: American Council on Education.

Benson, J., & Crocker, L. (1979). The effects of item format and reading ability on objective test performance: A question of validity. Educational and Psychological Measurement, 39, 381-387.

Bolden, B. J., & Stoddard, A. (1980, April). The effects of language on test performance of elementary school children. Papers presented at the annual meeting of the American Educational Research Association, Boston, MA.

Bornstein, H., & Chamberlain, K. (1970). An investigation of the effects of "verbal load" in achievement tests. American Educational Research Journal, 7, 597-604.

Braun, H. I., & Holland, P. W. (1982). Observed score test equating: A mathematical analysis of some ETS equating procedures. In P. W. Holland and D. B. Rubin, (Eds.), Test equating, pp. 9-49. New York: Academic.

Cassels, J. R. T., & Johnstone, A. H. (1984). The effect of language on student performance on multiple choice tests in chemistry. Journal of Chemical Education, 61, 613-615.

Dorans, N. J. (1990). Equating methods and sampling designs. Applied Measurement in Education, 3, 3-18.

Green, K. (1934). Effects of item characteristics on multiple-choice item difficulty. Educational and Psychological Measurement, 44, 551-561.

Harris, D. J. (1990, April). Effects of passage and item scrambling on equating relationships. Paper presented at the annual meeting of the American Educational Research Association, Boston, MA.

Jerman, M. E., & Mirman, S. (1971). Linguistic and computational variables in problem solving in elementary mathematics. Educational Studies in Mathematics, 5, 317-362.

Jessell, J. C., & Sullins, W. L. (1975). The effect of keyed response sequencing on multiple-choice items on performance and reliability. Journal of Educational Measurement, 12, 45-48.

Klein, L. W., & Jarjoura, D. (1985). The importance of content representativeness for common-item equating with nonrandom groups. Journal of Educational Measurement, 22, 197-206.

Kolen, M. J. (1989). Traditional equating methodology. Educational Measurement: Issues and Practice, 7(4), 29-36.

Kolen, M. J., & Brennan, R. L. (1987). Linear equating models for the common-item nonequivalent-populations design. Applied Psychological Measurement, 11, 263-277.

Loftus, E. R., & Suppes, P. (1972). Structural variables that determine problem-solving difficulty in computer-assisted instruction. Journal of Educational Psychology, 63, 531-542.

Marcus, A. (1963). The effect of correct response location on the difficulty of multiple-choice questions. Journal of Applied Psychology, 47, 48-51.

McNamara, W. J., & Weitzman, E. (1945). The effect of choice placement on the difficulty of multiple-choice questions. Journal of Educational Psychology, 36, 103-113.

Millman, J. (1978, July). Determinants of item difficulty: A preliminary investigation (CSE Report No. 114). Center for the Study of Evaluation, University of California at Los Angeles.

O'Neill, K. A. (1986, April). The effect of stylistic changes on item performance. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Wilbur, P. H. (1966). Positional response set in the multiple-choice examination. Dissertation Abstracts International, 29/09-A, 2902-3051. (University Microfilms No. AAD67-02131)

Table 1

Correct Response Position (CRP), Difficulty, and Discrimination for Items in Forms A and B

	------Form A------			------Form B------			Absolute Difference	
Item	CRP	Diff	Disc*	CRP	Diff	Disc*	in CRP	DIFFA-DIFFB
1	6	.95	.39	11	.95	.35	5	0.00
2	17	.75	.30	15	.72	.31	2	.03
3	27	.83	.40	22	.82	.45	5	.01
4	13	.71	.46	14	.69	.50	3	.02
5	28	.97	.21	30	.94	.29	2	.03
6	5	.68	.41	9	.65	.53	4	.03
7	3	.64	.32	26	.54	.39	23	.10
8	2	.77	.30	23	.76	.32	21	.01
9	6	.87	.35	.31	.87	.40	25	0.00
10	22	.32	.46	19	.34	.45	3	-.02
11	4	.41	.45	11	.39	.45	7	.02
12	30	.37	.43	1	.31	.48	29	.06
13	18	.19	.31	6	.21	.29	12	-.02
14	14	.95	.28	10	.94	.25	4	.01
15	13	.78	.43	11	.80	.50	2	-.02
16	28	.23	.27	30	.15	.18	2	.08
17	32	.37	.54	31	.35	.43	1	.02
18	4	.71	.42	12	.64	.39	8	.07
19	8	.92	.30	3	.90	.31	5	.02
20	6	.43	.40	8	.44	.42	2	.01
Means	--	.643	.372	--	.621	.385	8.290	.023

*Point-biserial correlations

Figure 1

Plot of Form A and Form B Item Difficulty Values

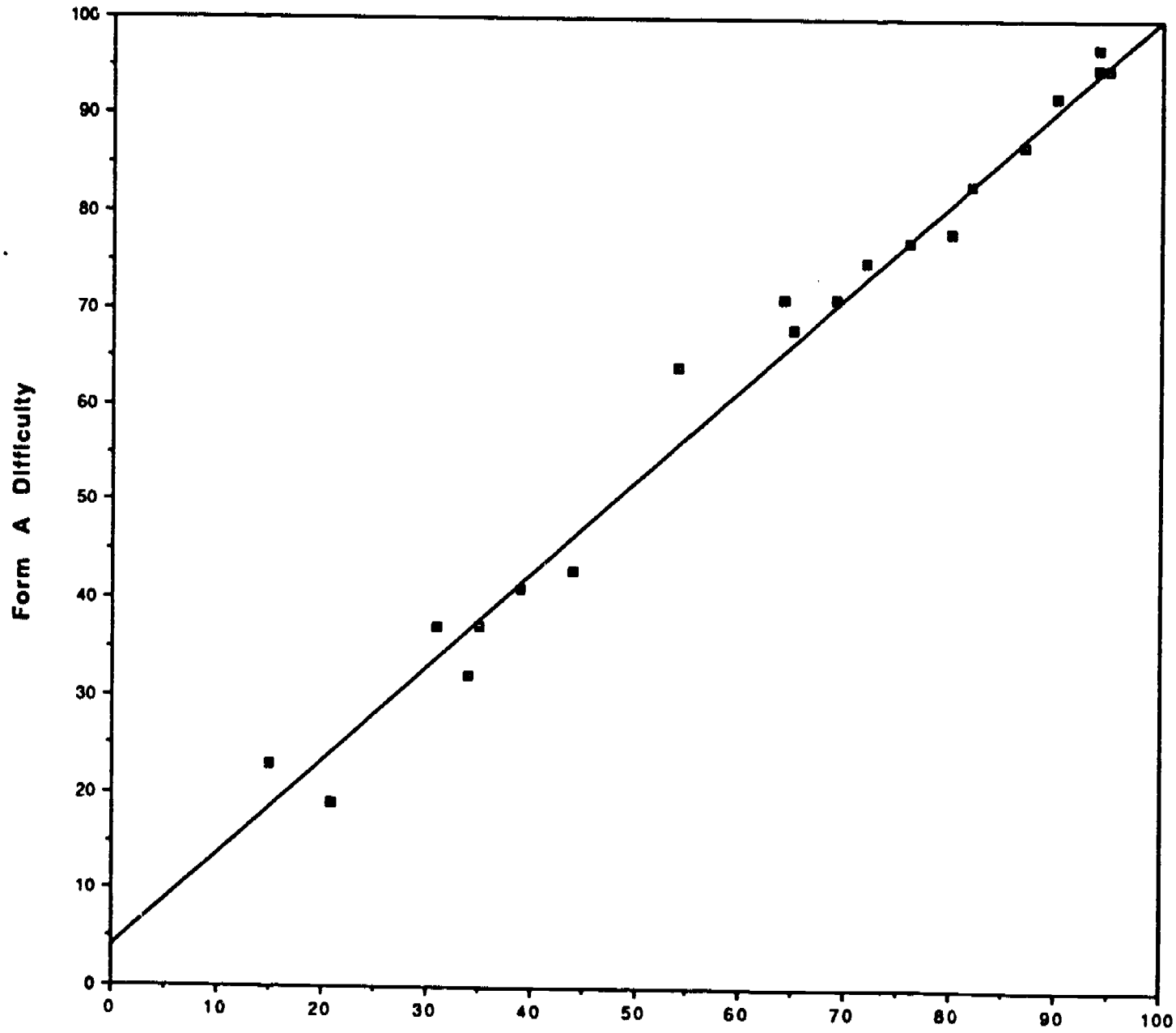

Figure 2

Plot of Form-to-Form Difficulty and Positional Differences

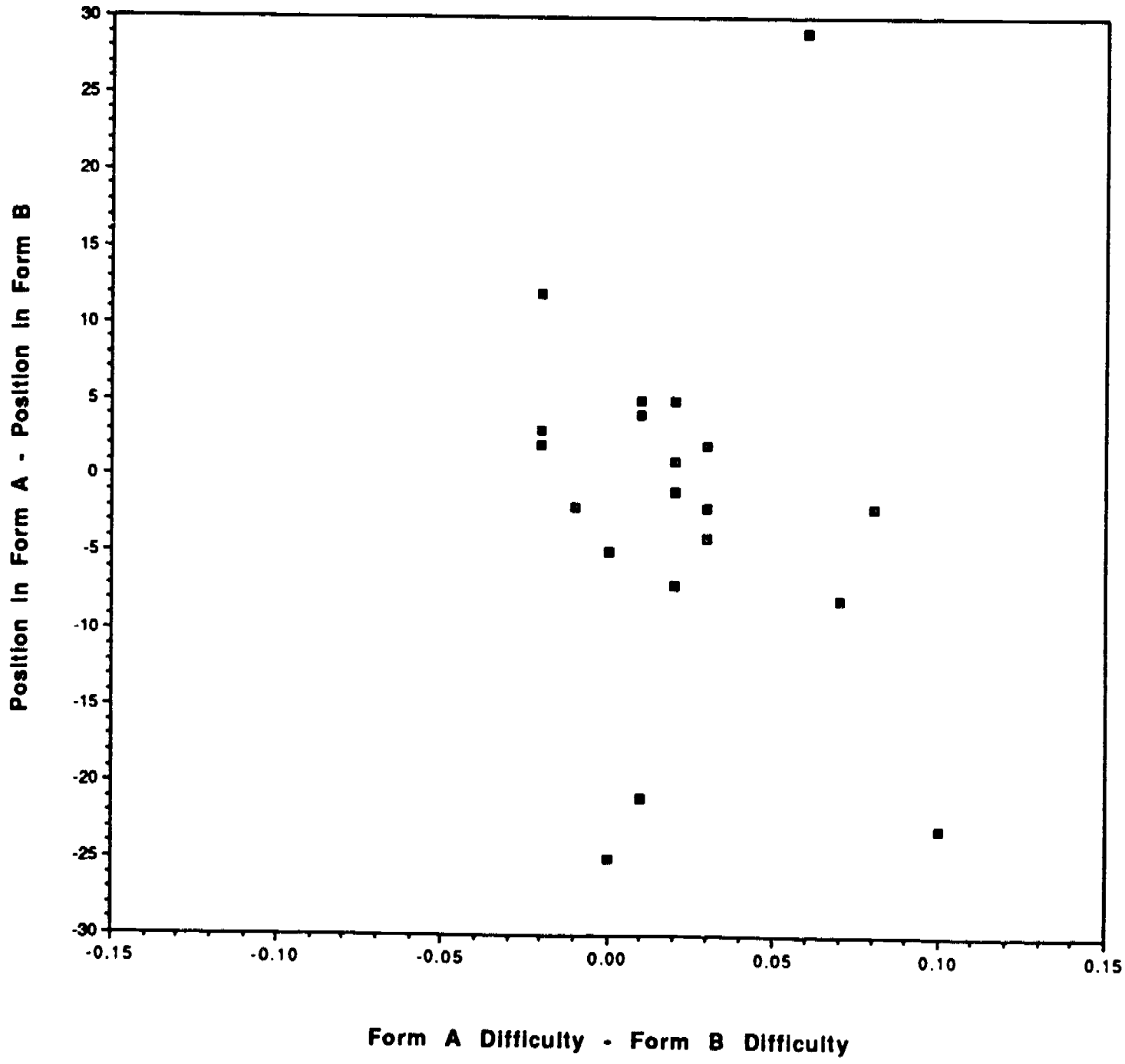

Figure 3

Modified Plot of Form-to-Form Difficulty and Positional Differences