DOCUMENT RESUME

ED 395 014 TM 025 028

AUTHOR Enright, Mary K.; Bejar, Isaac I.
TITLE An Analysis of Test Writers' Expertise: Modeling Analogy Item Difficulty.
INSTITUTION Educational Testing Service, Princeton, N.J.
SPONS AGENCY Graduate Record Examinations Board, Princeton, N.J.
REPORT NO ETS-RR-89-35
PUB DATE Jul 89
NOTE 39p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Construct Validity; *Difficulty Level; Models; Prediction; *Researchers; *Statistical Analysis; Test Construction; *Test Items
IDENTIFIERS *Analogies; *Expertise

ABSTRACT

In this study, the ability of test development staff to predict the difficulty of analogy items was explored. The nature of the item attributes that contributed to test writers' predictions of difficulty as well as actual item difficulty was also investigated. The two expert test writers studied were quite good at predicting item difficulty. Item attributes such as vocabulary difficulty and rationale difficulty contributed to item difficulty. However, a statistical model of item difficulty did not capture all the information that test writers used to judge item difficulty. This research contributes to the construct validation of tests in two ways. First, identification of some item attributes that are associated with item difficulty clarifies what skills and processes are likely to be involved in solving analogies. Second, the expertise of test writers, a crucial ingredient in ensuring the validity of the test, is demonstrated. (Contains 2 figures, 6 tables, and 13 references.) (Author/SLD)

RESEARCH REPORT

# AN ANALYSIS OF TEST WRITERS' EXPERTISE: MODELING ANALOGY ITEM DIFFICULTY

Mary K. Enright
Isaac I. Bejar



Educational Testing Service
Princeton, New Jersey
July 1989

# An Analysis of Test Writers' Expertise: Modeling Analogy Item Difficulty

Mary K. Enright

and

Isaac I. Bejar

Educational Testing Service

This research was supported by the Research Committee of the Graduate Record Examinations Board.

## Abstract

In this study, the ability of test development staff to predict the difficulty of analogy items was explored. The nature of the item attributes that contributed to test writers' predictions of difficulty as well as actual item difficulty was also investigated. The expert test writers studied were quite good at predicting item difficulty. Item attributes such as vocabulary difficulty and rationale difficulty contributed to item difficulty. However, a statistical model of item difficulty did not capture all the information that test writers used to judge item difficulty. This research contributes to the construct validation of tests in two ways. First, identification of some item attributes that are associated with item difficulty clarifies what skills and processes are likely to be involved in solving analogies. Secondly, the expertise of test writers, a crucial ingredient in ensuring the validity of the test, is demonstrated.

An Analysis of Test Writers' Expertise:

Modeling Analogy Item Difficulty

The items currently in use on most aptitude and intelligence tests have been developed from a empirical rather than a theoretical perspective. Although these items are known to be sensitive to individual differences and to predict performance on other tasks such as classroom tests, little is known about which item attributes are associated with variability in the psychometric characteristics of items. Research on this issue has both practical and theoretical implications. From a practical viewpoint, specifying what item attributes are associated with important psychometric characteristics will make the writing of items more efficient and economical. From a theoretical viewpoint, specifying what item attributes are important in differentiating among individuals will contribute to theory development and construct validation by clarifying the meaning of a test.

For example, analogies have long been used as measures of inductive reasoning on aptitude tests (see summary by Pellegrino & Glaser, 1982). However, examinees who take the Graduate Record Examinations often complain that they could solve analogy items if they only knew the words. Indeed, Carroll (1979), has suggested that vocabulary is a major factor on some analogy tests. Carroll examined SAT analogy items, which on the surface resemble GRE analogy items, in an attempt to understand what controls their difficulty. He suggested that two factors, vocabulary difficulty and the complexity of the relationship

between the words in the analogy, control item difficulty. This analysis, however, was based on only 10 analogy items.

In the context of identifying what item attributes are associated with psychometric characteristics, the expertise of the individuals who write test items is, in itself, a scientifically interesting object of study and a crucial ingredient in ensuring the validity of scores derived from that test. Therefore, it is surprising that such expertise is seldom documented. One manifestation of test development expertise is judgments of the psychometric characteristics of items. There is some research that indicates that judges can accurately rank mathematics test items as to difficulty ( Lorge & Kruglov, 1952; Ryan, 1968; Thorndike, 1980; Tinkelman, 1947) and that the statistical characteristics of items can be predicted from their structural characteristics (Millman, 1978; Searle, Lorton, & Suppes, 1974). In the verbal domain, however, Bejar (1983) has reported that even after extended training, expert judges were not able to predict the difficulty of items for the Test of Standard Written English, nor could they identify factors that contributed to item difficulty.

Similar findings have been reported for clinical predictions where there is a substantial body of research that indicates that expert judges are often inferior to statistical models (see summary by Wiggins, 1973). None the less, the observations and judgments of clinicians are ar important source of data for statistical predictions. That is, while they may not be very

accurate at predicting outcomes, their data are essential in the construction of a prediction system. For example, if the prediction system is based on regression methodology, the expert or clinician is often the source of the judgmental or observational data which enters into the regression and/or is in the best position to identify the variables that should make up the predictor set.

The present study explored the ability of Educational Testing Service (ETS) test development staff to predict the psychometric characteristics of analogy items. In addition, test writers were asked to make judgments about various item attributes. Thus we were able to explore how well test writers could estimate item difficulty, what item attributes predicted the test writers' estimates of difficulty as well as actual item difficulty, and whether test writers' estimates of item difficulty contributed to the prediction of actual item difficulty information beyond that provided by measures of item attributes.

## Method

Two members of the ETS test development staff, whose duties included the writing of analogy items, were asked to rate 179 GRE analogy items on a number of different attributes and to provide an estimate of item difficulty.

Analogies

The 179 analogy items were drawn from a set of 10 disclosed GRE forms first administered during the years 1981 through 1983. There were 18 analogy items on each form but one item was not included in this study because it was far too easy and difficulty could not be determined precisely. The basic statistical criteria for judging items are the difficulty and discrimination. At ETS it is customary to express difficulty in terms of a transformed metric, called delta. Delta is based on the proportion of the examinees that answer the item correctly and has a mean of 13 and a standard deviation of 4. Furthermore, "raw" deltas obtained from an administration are equated to enhance the metric comparability of items on different test forms administered at different times and to different samples of examinees. In this study the metric of interest is the equated delta which had a mean of 12.39 and a standard deviation of 2.74 for this set of items. The second statistical criterion is discrimination. Basically, this is an indicator of how well the item can differentiate those who know the correct answer from those who do not and is based on a biserial correlation. The mean R-biserial for this set of items was .50 with a standard deviation of .13. A complete description of the items used can be found in Bejar, Chaffin, and Embretson (in press).

Procedure

The analogy items were presented one at a time on an input screen on a IBM/PC XT. The test writers were first asked to solve the item and to type in the item rationale. The item rationale is a statement of the relationship between words in a pair that is shared by the stem word pair and the key word pair (correct answer). They then answered the item and rated attributes of the item. The selection of these item attributes was based on an extensive review of research on analogies as well as discussions with test development staff (Bejar, Embretson, & Chaffin, in press). The five item attributes selected and the procedures for rating each one are described below.

1. Rationale difficulty - a rating of how difficult it was for the test writer to discover the item rationale on a scale of 1 (easy) to 5 (difficult).

2. Rationale complexity - an estimate of the complexity of the test writer's own statement of the item rationale based on the number of significant elements or concepts in the statement of the rationale.

Examples:

| Rationales | Number of Elements |
|---|---|
| A is a member of B | 1 element - membership |
| A is a verbal expression of B. | 2 elements - expression, verbal |
| A is a device through which | 3 elements -device, |

the flow of B is regulated. flow, regulation

3. Syntactic order--a judgment of whether or not the two words in the item pair were in the same order as that in which they would occur in a natural statement of their relationship. For example, BIRD:CAGE would be rated 1, since either the statement "a bird is kept in a cage" or the statement "a cage is where you would keep a bird" is a natural statement of the relationship between the words. HOLES:RIDDLE would be rated 2, because in a natural statement of the relationship between the words, they would occur in the opposite order (to riddle is to mark with holes).

4. Stem-option similarity--ratings of the similarity of the stem relationship to the option relationship for each. For GRE General Test analogy items, there is usually some similarity between the stem and each of the options. The correct answer is the option that is most similar to the stem. The test writers were asked to examine the whole item before rating each option and to take into account what they knew about the similarity of relationships in other items. Options were rated on a scale of 1 (very similar) to 5 (very dissimilar).

5. Judged delta--a rating of the difficulty of the item based on the test writers' past experience with analogy items and using the delta scale.

In addition to the above ratings for each item, a measure of

the average vocabulary difficulty for each item was calculated and the items were categorized according to a higher-order classification based on semantic relations. The vocabulary difficulty measure was based on the word frequency data obtained from Kucera & Francis (1967). All words in an item were assigned a frequency value. If the word was not listed, it was assigned a zero value. For each of the five word pairs in an item, the smaller of the two values was used as the frequency for that pair, and these minimum frequencies were averaged to produce a mean word frequency measure for the item. The analogies were classified as intensional or extensional as part of an attempt to develop a taxonomy based on semantic relations (Bejar et al., in press). The intensional classification is based on an overlap of attributes or properties between two concepts; intensional relations include class inclusion, similarity, contrast, attribute, and nonattribute. The second classification is based on extensional relations including: use, cause/purpose, and space/time relations.

The analogy items were presented in the same random order to the two test writers. The items were presented in blocks of 10, and at the end of each block, the test writers were given the correct answers and the actual equated delta value for each item. The test writers worked individually and the duration and scheduling of sessions were arranged to suit their convenience.

## Results

The database for the analysis included the test writers'

ratings of the item attributes and their judgments of delta, the equated deltas, the logs of the word frequency index, and the higher-order taxonomic classification for each item. An overall stem-option similarity rating was computed for each item by taking the mean of the rating of the five stem-option similarity pairings. Data were grouped into three blocks of 60 trials.

The data analyses proceeded in five phases. First, the effect of practice on each test writer's ability to predict item difficulty and interrater differences in prediction were assessed. Second, the effect of practice on the model test writers used to estimate difficulty was evaluated by examining the contribution of word frequency and the test writers' judgment of item attributes to their judgments of delta. Third, the relationship between the test writers' model of item difficulty and equated delta was investigated. Fourth, the issue of whether the test writers' judgment of delta contributed to the prediction of equated delta information beyond that contained in measures of item attributes was explored. Finally, the question of whether the test writers' model of item difficulty differed for the two classes of analogies was investigated.

Prediction of Item Difficulty

The test writers' ability to predict item difficulty was assessed by calculating the regression of delta on the test writers' judged delta and the standardized mean-squared difference separately for each test writer and block. The mean-squared difference was standardized using the standard

deviations of equated deltas within a block. These results are presented in Figures 1 and 2. When judged delta was used to

---

Insert Figure 1 here

---

predict equated delta, the $R^2$ for the two test writers combined increased from .24 in Block 1 to .46 in Block 3. In Figure 1 the $R^2$ data over the three blocks is presented separately for the two test writers. Both test writers improved with practice. The test writers also demonstrated improvement in the accuracy of their predictions with practice as illustrated in Figure 2 which presents the standardized mean-squared differences for the test writers individually. The standardized mean-squared differences were computed as

$$SMD = \sum_i f_i(d_i-d)^2/ns^2_t + d^2/s^2_t$$

where $d_i$ is the difference between the judged and equated delta and $s_t$ is the standard deviation of the equated deltas.

The test writers' errors in prediction were of very similar magnitude and decreased with practice. Overall, the differences between the two test writers were not large; therefore, in subsequent analyses the data were combined for the two test writers by taking the mean of their judgments.

Test Writers' Model of Item Difficulty

The model the test writers used to judge delta was explored

by regressing judged delta on word frequency and on the test writers' mean judgment of other item attributes using a forced-entry method. Changes in this model, with practice, were evaluated by calculating the regression separately for each block. The results of these regressions are presented in Table 1.

---

Insert Table 1 here

---

As can be seen in the table, the proportion of variance accounted for by the model increased from about 35% to 41%. Initially, the test writers' judgment of delta was influenced primarily by word frequency and rationale difficulty. With practice, the importance of word frequency decreased; rationale complexity becomes a significant factor in Blocks 2 and 3 and stem-option similarity approached significance in Block 3. Thus, the test writers' model of item difficulty appeared to become more complex with practice. Note, however, that is difficult interpret what meaning the judgments of rationale complexity had because the direction of its contribution was opposite from what might have been expected.

Prediction of Equated Delta

In the next phase of the analysis, the relationship between the test writers' model of item difficulty and equated delta was explored by regressing judged delta on word frequency and on the test writers' judgment of other item attributes using a

forced-entry method. Subsequently, the test writers' mean judged delta was entered into the model to determine if their estimate of delta contained information not already incorporated into the model.

The results for the regression of equated delta as a function of word frequency and test writers' judgment of other item attributes are presented in Table 2 for the three blocks.

---

Insert Table 2 here

---

The results parallel those for the prediction of judged delta in that there was improvement with practice. In comparing Tables 1 and 2, we can see that initially the model predicted judged delta ($R^2$ = .35) better than equated delta ($R^2$ = .18). However, by the end of training, prediction of equated delta ($R^2$ = .43) was slightly better than that of judged delta ($R^2$ = .41). Again, the test writers' model of item difficulty appeared to become more complex with training. Initially, the item attributes that contributed most to the prediction of equated delta were word frequency and rationale difficulty. However, by Block 3, all the item attributes appeared to contribute to the prediction of equated delta although the role of rationale complexity is difficult to interpret.

Next, judged delta was added to the regression for predicting delta from item attributes with the purpose of

estimating the contribution of expert judgment. The results for these regressions are presented in Table 3. The change in $R^2$

---

Insert Table 3 here

---

when the test writers' estimated delta was added to the model was significant. This suggests either that some important item attributes, which test writers use to predict equated delta, were not included in our predictor set or that the test writers' integration of item attribute information is different from our statistical integration of the information. In particular, a linear mathematical model may not be an appropriate description of the test writers' model which may take into account complex interactions between variables. Some evidence for this point of view is presented next.

Test writers' performance and type of analogy

The analysis just presented focused on the ability of test writers to benefit from practice on the estimation of item difficulty and the nature of their model of analogy difficulty. As we have seen, the test writers were relatively good at predicting difficulty to begin with and were then able to improve as a result of practice. It is also valuable to ask whether the their performance and their model of difficulty are different for the two types of analogies, intensional and extensional.

Table 4 presents the regression of judged delta on the selected item attributes. The results indicate that for both

Insert Tables 4 & 5 here

types of items, rationale difficulty is an important component of the test writers' judgment. For the intensional items, word frequency and stem-option similarity also appear to be important components of the judgment. For the extensional items, rationale complexity appears to be the only component, other than rationale difficulty, that contributes to expert judgement. This suggests that the set of item attributes we chose to study is not sufficiently comprehensive, particularly for extensional items as $R^2$ is considerably smaller for these items. This view is supported by Table 5, which presents the results for the regression of equated delta on these item attributes. The $R^2$ is somewhat higher for intensional items.

Furthermore, when the item attributes and the judged deltas are placed together in a regression equation, the $R^2$ is still somewhat higher for intensional items, as is evident in Table 6.

Insert Table 6 here

This suggests that extensional items are inherently more difficult to predict. On the whole, however, the test writers

appear to be quite adept at predicting difficulty and seem to contribute to the prediction information that is not captured by the our set of item attributes. The last column of Table 6 reports the probability of the increase in $R^2$ from Table 5, to Table 6 in which judged delta is added to the predictor set. As can be seen, that probability is very small. Thus, judged delta appears to include information that goes beyond the item attributes identified. It is reasonable to call that information expertise. Moreover, as can also be seen in Table 6, the intercept, at least for the intensional items, is not significant. This suggests that the test writers not only are able to accurately rate the difficulty of the items, but also to do so in the same metric as the delta statistic is expressed.

## Discussion

In contrast with the literature on clinical judgment (Wiggins, 1983) and work on other verbal item types (Bejar, 1983), the test writers in this study were quite good at predicting item difficulty. Furthermore, the test writers showed some important gains in the prediction of item difficulty with training. The effects of this training appear to be due to an increase in the complexity of the test writers' model of item difficulty. However, it appears that test writers' judgments of difficulty either contained more information than was captured by the item attributes measured in this experiment or that their synthesis of this information was better than the statistical synthesis. We were only partially successful at documenting what

item attributes contributed to their judgments. Initially, test writers' judgments of delta were predicted primarily by word frequency and rationale difficulty. By the end of training, however, the other item attributes they were asked to rate also contributed to their judgement of difficulty. Nevertheless, it was difficult to interpret the role of one of these item attributes. The contribution of the rationale complexity measure was not in the expected direction. The fact that similar results were found in a separate attempt to scale this set of analogies in terms of rationale complexity (Bejar et al., in press) suggests that this measure is inappropriate. Finally, the fact that the test writers had different models of item difficulty for different classes of analogies suggests that complex interactions among item attributes need to be taken into account in predicting item difficulty.

These findings have implications for the validity of tests of analogical reasoning and for theories of performance on such tasks. This report is a validity study in two senses. First, by modeling and documenting the expertise of test writers, we have validated the test development process which is critical in maintaining the quality of the test across test forms and administrations. Secondly, this study aimed to be "a scientific inquiry into score meaning" (Messick, 1988) and as such has theoretical implications. As a consequence of exploring the attributes that contribute to analogy item difficulty, we are better able to answer questions about what GRE analogy items

measure. While vocabulary difficulty plays a role in determining item difficulty, other item attributes such as rationale difficulty, stem-option similarity, and syntactic-order of word pairs are also important. Thus the description of analogy tests as tests of reasoning rather than vocabulary tests receives support. However, exploring the role of item attributes in the context of expert judgment also allowed us to see that there are factors contributing to item difficulty that we were unable to specify.

## References


Bejar, I. I. (1983). Subject matters experts' assessment of item statistics. Applied Psychological Measurement, 7, 303-310.

Bejar, I. I., Chaffin, R., & Embretson, S. (in press). Cognitive and psychometric analysis of analogical problem solving. New York: Springer-Verlag.

Carroll, J. B. (1979, October). Measurement of abilities constructs. In A. P. Maslow & R. H McKillip (Eds.), Construct validity in psychological measurement: Proceedings of colloquium on theory and application in education and employment (pp. 23-41). [Jointly sponsored by US Office of Personnel Management and Educational Testing Service]. Princeton, NJ: Educational Testing Service.

Kucera, H., & Francis, W. N. (1967). A computational analysis of present day American English. Providence, RI: Brown University Press.

Lorge, I., & Kruglov, L. A. (1952). A suggested technique for the improvement of difficulty prediction of test items. Educational and Psychological Measurement, 12, 554-561.

Messick, S. (1988). Validity. In R. L. Linn (Ed.), Educational Measurement (3rd. Ed.). New York: MacMillan.

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

Pellegrino, J. W., & Glaser, R. (1982). Analyzing aptitudes for learning: Inductive reasoning. In R. Glaser (Ed.), Advances in instructional psychology, Vol 2 (pp. 269-245). Hillsdale, NJ: Erlbaum.

Ryan, J. J. (1986). Teacher judgments of test item properties. Journal of Educational Measurement, 5, 301-306.

Searle, B. W., Lorton, P., & Suppes, P. (1974). Structural variables affecting CAI performance on arithmetic word problems of disadvantaged and deaf students. Educational Studies in Mathematics, 5, 301-306.

Thorndike, R. L. (1982). Item and score conversion by pooled judgement. In P. W. Holland & D. B. Rubin (Eds.), Test equating. New York: Academic Press.

Tinkelman, S. (1947) Difficulty prediction of test items. (Teachers College Contributions to Education Report No. 941). New York: Columbia University Press.

Wiggins, J. S. (1973). Personality and prediction: Principles of personality assessment. Reading, MA: Addison-Wesley.

Table 1

Regression Analysis for Predicting Judged Delta from Other Item Attributes

| Block | Intercept | Log Word Frequency | Rationale Difficulty | Rationale Complexity | Syntactic Order | Stem-Option Similarity | $R^2$ |
|---|---|---|---|---|---|---|---|
| 1 | 13.97*** | -.712*** | 1.026*** | -.174 | - .748 | - .564 | .345 |
| 2 | 11.12*** | -.547** | 1.341*** | -.566* | .954 | - .246 | .326 |
| 3 | 13.74*** | -.427* | 1.283** | -.651* | 1.358 | -1.300+ | .409 |

+p<.10 *p<.05 **p<.01 ***p<.001

Table 2

Regression Analysis for Predicting Delta from Item Attributes

| Block | Intercept | Log Word Frequency | Rationale Difficulty | Rationale Complexity | Syntactic Order | Stem-Option Similarity | $R^2$ |
|---|---|---|---|---|---|---|---|
| 1 | 16.55*** | -.842* | .801* | - .786 | - .698 | - .669 | .180 |
| 2 | 15.08*** | -.712* | 1.201+ | - .403 | -2.23 | - .229 | .153 |
| 3 | 13.99*** | -.789* | 1.684** | -1.109** | 3.25* | -1.785+ | .426 |

+p<.10 *p<.05 **p<.01 ***p<.001

Table 3

Effect of Adding Judged Delta as a Predictor to the Regression of Delta on Item Attributes

| Block | Intercept | Log Word Frequency | Rationale Difficulty | Rationale Complexity | Syntactic Order | Stem-Option Similarity | Judged Delta | $R^2$ | Probability of Change in $R^2$ |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 8.08* | -.410 | .179 | -.681 | - .245 | -.327 | .606** | .300 | <.004 |
| 2 | 3.33 | -.135 | -.216 | .194 | -3.235+ | .031 | 1.057*** | .434 | <.0000 |
| 3 | 4.256 | -.486 | .774 | -.648 | 2.292 | -.865 | .709*** | .545 | <.0005 |

+p<.10 *p<.05 **p<.01 ***p<.001

Table 4

Regression of Judged Delta on Item Attributes for Intensional and Extensional Analysis

| Class | Intercept | Log Word Frequency | Rationale Difficulty | Rationale Complexity | Syntactic Order | Stem-Option Similarity | $R^2$ |
|---|---|---|---|---|---|---|---|
| Extensional | 12.09*** | -.245 | 1.099*** | -.484* | .474 | -.5711 | .23 |
| Intensional | 13.23*** | -.937*** | 1.095*** | -.017 | .101 | -.659* | .45 |

*p<.05 **p<.01 ***p<.001

Table 5

Regression of Delta on Item Attributes for Intensional and Extensional Analysis

| Class | Intercept | Log Word Frequency | Rationale Difficulty | Rationale Complexity | Syntactic Order | Stem-Option Similarity | Multiple $R^2$ |
|---|---|---|---|---|---|---|---|
| Extensional | 15.85*** | - .547* | 1.116** | -.953** | .041 | -1.112+ | .17 |
| Intensional | 13.676*** | -1.026*** | .953* | -.170 | .160 | - .388 | .21 |

+p<.10 *p<.05 **p<.01 ***p<.001

Table 6

**Effect of Adding Judged Delta as a Predictor to the Regression of Delta on Item Attributes for Intensional and Extensional Items**

| Class | Intercept | Log Word Frequency | Rationale Difficulty | Rationale Complexity | Syntactic Order | Stem-Option Similarity | Judged Delta | $R^2$ | Probability of Change in $R^2$ |
|---|---|---|---|---|---|---|---|---|---|
| Extensional | 6.81* | -.364 | .?94 | -.591+ | -.314 | -.685 | .747*** | .34 | <.0000 |
| Intensional | 2.375 | -.225 | .017 | -.155 | .074 | .174 | .854*** | .39 | <.0000 |

+p<.10 *p<.05 **p<.01 ***p<.001

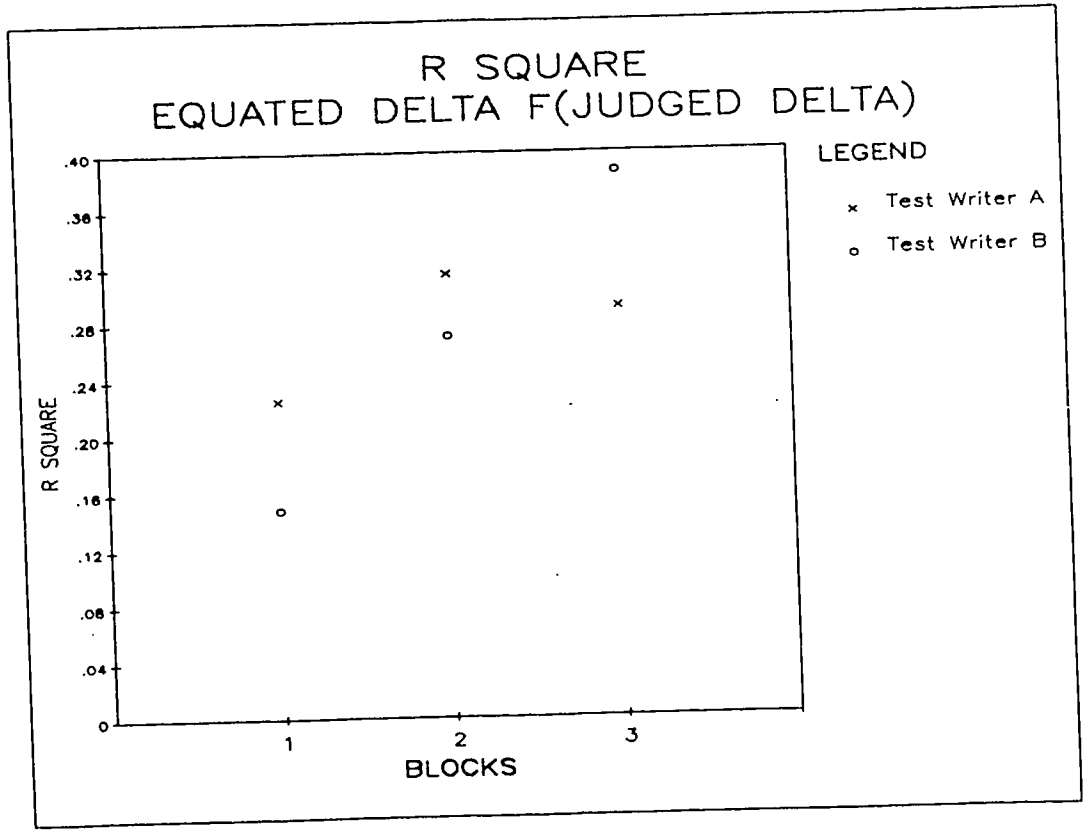


Figure 1. Improvement in $R^2$ for the regression of equated delta on judged delta over blocks for test writers A and B.

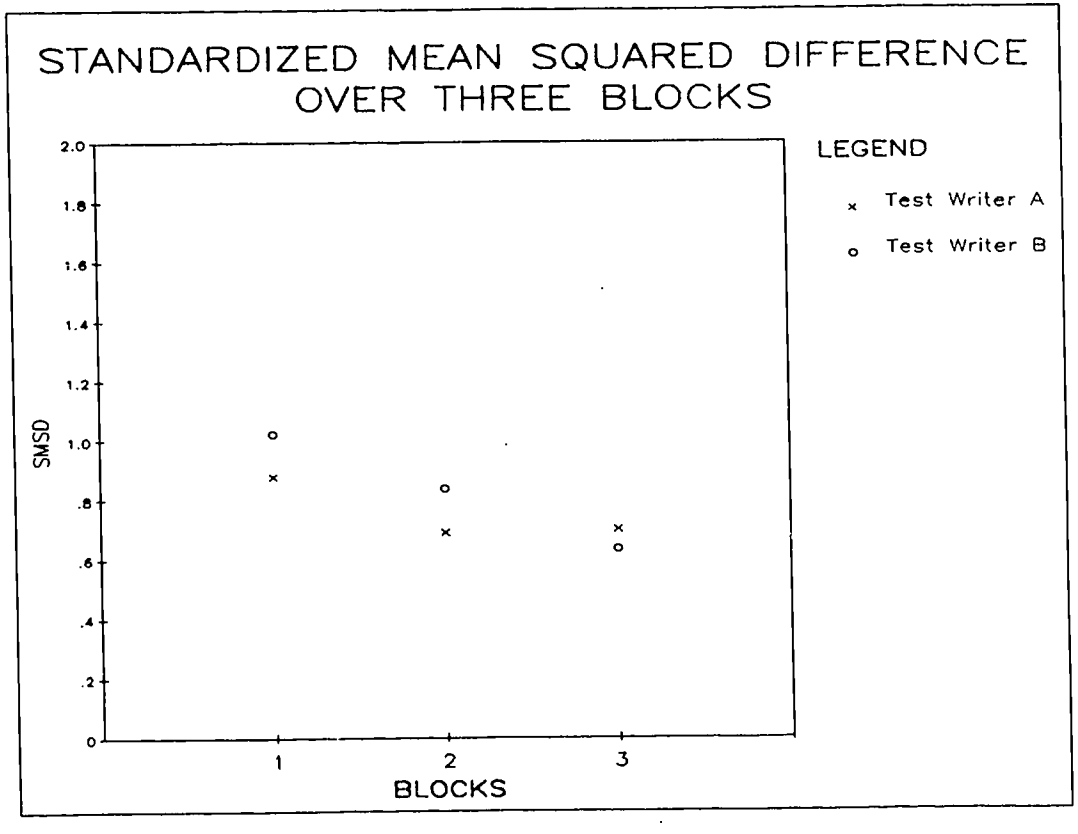


Figure 2. Standardized mean squared difference as a function of practice for test writers A and B.