Subject Matter Experts' Assessment of Item Statistics

Isaac I. Bejar
Educational Testing Service

This study was conducted to determine the degree to which subject matter experts could predict the difficulty and discrimination of items on the Test of Standard Written English. It was concluded that despite an extended training period the raters did not approach a high level of accuracy, nor were they able to pinpoint the factors that contribute to item difficulty and discrimination. Further research should attempt to uncover those factors by examining the items from a linguistic and psycholinguistic perspective. It is argued that by coupling linguistic features of the items with subject matter ratings it may be possible to attain more accurate predictions of item difficulty and discrimination.

There is some evidence (e.g., Lorge & Kruglov, 1952; Ryan, 1968; Thorndike, 1980; Tinkelman, 1947) that raters are able to accurately rank the difficulty of mathematics test items. There is also some evidence that the statistical characteristics of items can be predicted from the structural characteristics of the items (e.g., Millman, 1978; Searle, Lorton, & Suppes, 1974). The present study, however, is concerned with the question of whether subject matter experts can accurately estimate the difficulty and discrimination of items that measure writing ability. This task would appear to be far more difficult. Unlike mathematics items, where the mathematical operation required to solve the problem largely determines item difficulty, for items that attempt to measure writing skill a much greater variety of factors would seem to determine difficulty. Thus, it is probably not sufficient to determine what error is present in a given item, for the semantic and syntactic context in which that error is presented may influence item statistics significantly.

This suggests that it may be useful to identify which factors contribute to item difficulty. It is clear that subject matter experts are required in this process. This approach was taken by Kirsh and Guthrie (1980), who reported high correlations for the prediction of item difficulty based on subject matter ratings of factors previously identified as affecting the difficulty of literacy items.

The identification of factors that contribute to item difficulty and discrimination could be important both practically and theoretically. From a practical point of view, once these factors have been identified, they can be transmitted to other subject matter experts. Although the creation of items is likely to remain largely a creative endeavor, knowledge of what factors affect difficulty and discrimination may provide better control over the statistical characteristics of items produced and, in principle, may obviate the need to pretest items.

From a theoretical point of view, the identification of facets that account for the variability in difficulty and discrimination across items would seem to be an important step in the construct validation of multiple-choice tests of writing ability.

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 7, No. 3, Summer 1983, pp. 303–310

0146-6216/83/030303-08$1.65

For example, other things being equal, the syntactic context in which a subject-verb agreement error is presented may be related to difficulty in a psychometric sense. Such a relationship might, upon further research, be explained in terms of information-processing constructs. Identification of those facets will, in the end, contribute to the construct validity of the test.

Method

Overview of the Study

The study was based on the Test of Standard Written English (TSWE). The TSWE is a 30-minute multiple-choice test introduced in 1974 as a companion test to the Scholastic Aptitude Test (SAT) with which it is administered. Its purpose is to help colleges place students in appropriate English Composition courses. It is not recommended as an admission instrument. The test consists of 50 items of two types. Items 1–25 and 41–50 are called Usage items and Items 26 to 40 are called Sentence Correction items. The testee is expected to recognize writing that does not follow conventional and standard written English.

The subject matter experts participating in the study were test development staff from the College Board division of Educational Testing Service. Although they were highly skilled and experienced (their experience with the TSWE and other tests of entering skills ranged from 3 to 20 years), several sessions were held to "train" them in the rating task and simultaneously to elicit from them a series of principles that could account for the difficulty and discrimination of items. Once this training phase was completed, the subject matter experts rated the difficulty and discrimination of two sets of 50 items each.

The Training Procedure

Instructions to raters. The raters were assembled as a group and, for a given item, were instructed to examine it and to write down their estimates of its difficulty and discrimination; then, they revealed their ratings and discussed among themselves the rationale behind them. After this discussion, the experts rated the item a second time. At that point estimated difficulty, discrimination, and other statistical information were revealed. The "other" statistical information included the distribution of responses across alternatives for total score quintiles as well as the mean criterion score of students choosing each of the incorrect alternatives.

The experts rated difficulty on the delta metric, since they used delta statistics in their everyday work. The delta difficulty index is a nonlinear transformation of the proportion-correct statistic given by $\Delta = \Phi^{-1}(1 - p)$, where Φ^{-1} is the inverse normal function and p is the proportion of students answering the items correctly. The proportion correct is estimated on those students who attempt the item. The values used in this study were equated deltas. The equating process is performed to account for the fact that the testing populations at different testing times differ in ability and insures that the measure of difficulty for different forms is on the same metric.

For discrimination the raters were instructed to rate the biserial correlation of the item. The biserial correlation was computed as follows:

$$r = \{(M_R - M_W)/S_T\}\{[p(1 - p)]/y\} \quad [1]$$

where

M_R is the mean criterion score for students choosing the correct answer;
M_W is the mean criterion score for students not choosing the correct alternative;
S_T is the standard deviation of criterion scores for all students; and
y is the ordinate of the normal density functions corresponding to p or $(1 - p)$, whichever is smaller.

After the estimated statistics were revealed, the raters were encouraged to explain any discrepancies among their own ratings and with the estimated statistics. As part of this process, they formulated hypotheses to account for these discrepancies and were able to test them with items presented subsequently.

The training material. The pool of items from which the items were drawn consisted of five early forms of the test. The items from each form were

separated into two groups corresponding to the two item types. Three sets of 20 Usage items and three sets of 10 Sentence Correction items were taken at random from the pool and assembled into booklets. The 20 Usage items were placed first in the booklet followed by the 10 Sentence Correction items. Since the booklets were formed by randomly choosing from the pool, there was no control over the number of items testing the different types of errors.

For purposes of the study the "true" item statistics were considered to be the equated delta and biserial correlation based on the first national administration of the form. The estimates were based on a random sample of close to 2,000 students from the population of students taking the test. Three 3-hour training sessions were held.

Training results. When the Usage and Sentence Correction items were considered together, the interrater reliability was fairly low and did not seem to increase as a function of experience. Although interrater reliability increased after discussion, to some extent this could be due to correlated errors introduced in the discussion process. More importantly, interrater reliability before discussion did not increase across sessions. The correlations of the composite ratings with the item statistics did not even reach .50 when both Usage and Sentence Correction items were considered simultaneously. Analysis of the residuals suggested that the raters were not equally successful with the two types of items. In fact, the raters were more accurate in predicting the difficulty of Sentence Correction items.

Additional training. It was hypothesized that the raters' performance could be improved by additional training. For this purpose items from five additional TSWE forms were collected and the items sorted into the major error categories to which each item had been previously assigned by Test Development staff.

There were 24 possible error categories, including a no error category. However, only 19 of them were represented in the five forms which were used in the study. Figures 1 and 2 show, respectively, the mean delta and mean discrimination plus and minus two standard errors for the 19 error categories. The error categories in these figures are as follows: (1) no error, (2) subject-verb agreement, (3) tense, (4) verb form, (5) connective, (6) logical agreement, (7) logical comparison, (8) modifier, (9) pronoun, (10) diction, (11) idiom, (12) parallelism, (13) sentence fragment, (14) comma splice, (15) improper subordination, (16) improper coordination, (17) dangling modifier, (18) redundancy/economy/constancy, and (19) vague pronoun reference.

It is evident from Figure 1 that for most categories the mean difficulty is within a narrow interval. However, some categories—most notably, Categories 6 (logical agreement) and 7 (logical comparison)—appear to be more difficult, whereas other categories—Categories 4 (verb form) and 13 (sentence fragment)—appear to be easier. For discrimination, most categories have a mean discrimination clustered about .50 except for Category 1, no error, which has a mean of .40, and Category 13, which has a mean close to .60.

The additional training was conducted in four 3-hour sessions. During each session discussions were focused on specific error categories. This gave the raters an opportunity to examine items which tested a single error yet varied in their difficulty and discrimination. The information on which Figures 1 and 2 are based was made available to the raters as they discussed items within a given category. The raters found it useful to discuss the items among themselves and then to "guess" at the statistics of the items. From time to time they postulated hypotheses that could account for difficulty. Although the raters did not seem able to develop a theory to account for the variation in the difficulty and discrimination of items measuring a single error, they found the exercise professionally useful but frustrating.

Evaluation. Once this training was completed, the raters were gathered for a final session in which two sets of 50 items each were rated under conditions approximating a realistic implementation of the procedure. All four subject matter experts participated in this phase of the study. The rating material consisted of two sets, A and B, of 50 items each. Each set contained 35 Usage items followed by 15 Sentence Correction items.

For this study raters were instructed to rate the delta and discrimination of items after discussing

Figure 1
Mean Delta by Major Error Categories, with the Mean
Plus and Minus Two Standard Errors for Each of 19 Error Categories

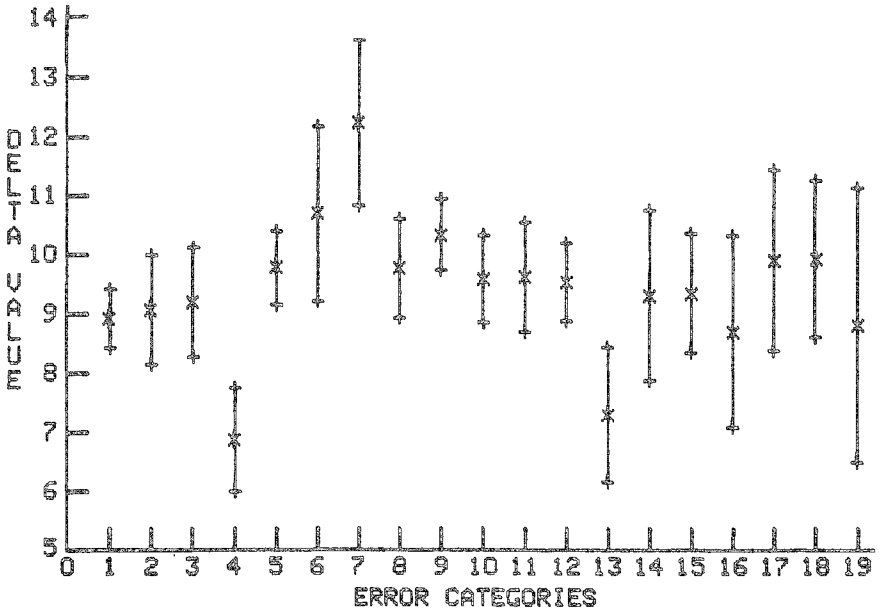
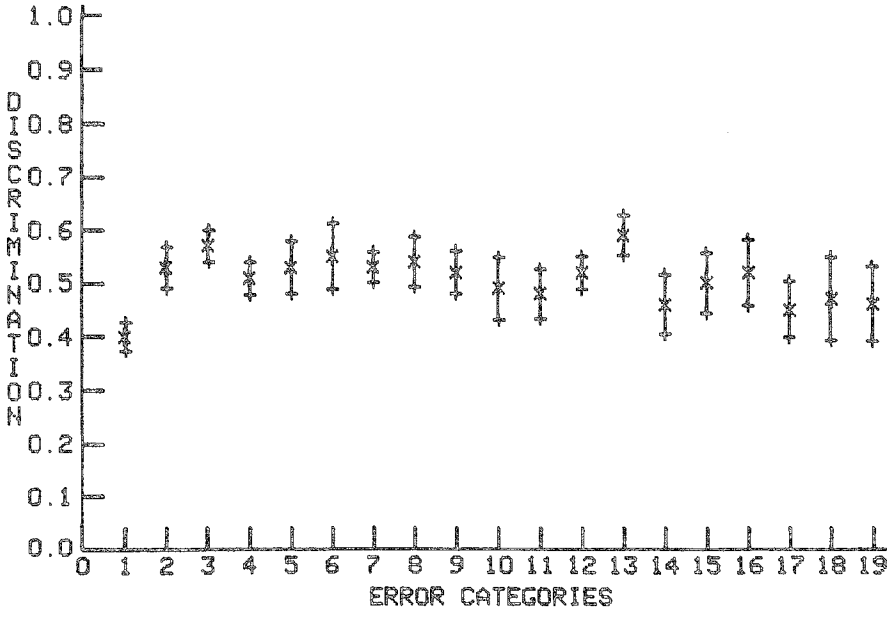


Figure 2
Mean Discrimination by Major Error Categories, with the Mean
Plus and Minus Two Standard Errors for Each of the 19 Error Categories



an item among themselves. However, no feedback was given to the raters after each item. The information in Figures 1 and 2 was made available to the raters, in tabular form, to guide their ratings. In short, the situation was designed to simulate the conditions that would be likely to prevail in a practical situation in which item statistics were estimated by raters. The two sets of 50 items were rated in a 3-hour session.

Results

The interrater reliability for difficulty ratings for Sessions A and B were .95 and .91, respectively. The interrater reliability for Usage and Sentence Correction items, across both sessions, was .94 and .91, respectively. For discrimination ratings, the interrater reliability for Sessions A and B was .88 and .86, respectively. The reliability for Usage and Sentence Correction items across sessions was .88 and .81, respectively.

These reliabilities are substantially higher than those found during the training phase. However, the raters were encouraged to use the available information on item statistics from Figures 1 and 2. Therefore, it is likely that the increase in their agreement is due in part to their use of this information.

Table 1 shows the correlation of each rater with the other raters. The correlations are reported separately for Usage and Sentence Correction items and for both item types combined. The rater-total correlations give an indication of how well a given rater agrees with the other three raters. As can be seen, Rater 1 is the most representative rater for difficulty and discrimination. (This was also true during the training phase).

Table 1 also shows the correlation of each rater with the delta and *r*-biserial. The value shown in parenthesis was computed as follows: Each item was assigned the mean value of the difficulty and discrimination for the error category to which that item belonged. This value will be referred to as the "empirical" rating. The correlation of the empirical rating with the estimated statistic is the value shown in parentheses. As can be seen, for difficulty the raters outperformed the empirical rating for Usage items but not for Sentence Correction items. When both item types are combined, each rater outperforms the empirical rating. A similar result is observed for discrimination.

To see this in more detail, Table 2 shows the correlations of the combined raters and the empirical rating with the item statistics. The multiple correlation of the combined raters and the empirical rating with the item statistics is also shown. As can be seen for Usage items, the raters outperform the empirical rating for both difficulty and discrimination. For Sentence Correction items the empirical rating slightly outperforms the raters for difficulty, but for discrimination the raters do slightly better. When both item types are considered together, Table 2 shows that the raters do better on both difficulty and discrimination. Examination of the multiple correlation shows that they are very close to the larger of the two single-order correlations. That is, it seems that the ratings from the subject matter experts and the empirical rating measure the same variable. In the case of Usage items, the subject matter experts measure that variable with more accuracy.

Discussion

The results from this study suggest that even after an extended period of practice and training, the accuracy in estimating item statistics of four subject matter experts does not approach the level that would be required to substitute ratings of item statistics for pretesting. In order for ratings to become practical substitutes for pretesting, their correlation with empirical estimates should approach .80, which is approximately the correlation between deltas on two occasions. The results of the study indicate that the attainment of this goal in a cost-effective manner would not be possible at this level of rater performance.

In principle, the needed level of correlation can be achieved by adding more raters. However, since the correlation of the ratings with the estimated statistics is low in relation to the reliability of the ratings, it is likely that a fairly large number of raters will be required to achieve a high level of correlation. That is, it may be expensive to attain

Table 1
Correlation of Each Individual Rater with the Mean Rating and Delta and Discrimination for Sets A and B

	Difficulty		Discrimination	
Item Type	Rater-Total	Rater-Delta	Rater-Total	Rater r-biserial
Usage (N = 70)				
1	.93	.44	.84	.45
2	.84	.48	.67	.39
3	.78	.28	.76	.44
4	.92	.49	.70	.44
"Empirical Rating"		(.16)		(.40)
Sentence Correction (N = 29)				
1	.92	.20	.72	.10
2	.72	.15	.59	- .05
3	.76	.18	.58	.27
4	.87	.26	.66	.12
"Empirical Rating"		(.30)		(.07)
Both (N = 99)				
1	.93	.38	.82	.39
2	.80	.37	.66	.34
3	.77	.26	.73	.42
4	.90	.43	.69	.30
"Empirical Rating"		(.22)		(.31)

Note: The approximate critical value of the correlation for α = .05 and N of 70, 29 and 99 are .23, .37, and .20 respectively.

a high validity. For example, at a rate of 100 items per 3-hour rating session, the ratings may be estimated to cost approximately $2 per item per rater or $100 per 50-item set per rater. Thus, if 20 raters were to be required to achieve a validity of, say, .90, the ratings cost would be $2,000 (or $100 × 20). Furthermore, this assumes that it is possible to identify such a large number of qualified raters. The cost of processing and assembling a 50-item TSWE pretest in 1981 was about $4,000.

In short, the findings of this study suggest that ratings cannot be substituted for actual pretesting of TSWE items. This is discouraging from a practical point of view, since the savings and the minimization of item exposure that would result from such substitution was one of the motivating factors behind the study.

The second motivating factor behind the study, however, was theoretical and consisted of an intent to uncover certain principles that could account for the variation in difficulty and discrimination among items. Such principles could, in turn, be used to train additional experts or to have more control over the item generation process by alerting item writers to the factors that contribute to difficulty. Although the raters attempted to uncover those principles, by their own account they were not successful. Nevertheless, based on the second study there is

Table 2
Simple and Multiple Correlations of Ratings with Difficulty (Diff) and Discrimination (Disc) for Usage, Sentence Correction, and Both Item Types Combined

Correlation	Usage (N = 70)		Sentence Correction (N = 29)		Both (N = 99)	
	Diff	Disc	Diff	Disc	Diff	Disc
with Rater	.46	.50	.21	.13	.39	.45
w/"Empirical Rating"	.16	.40	.30	.07	.22	.31
Multiple	.48	.51	.30	.13	.39	.45

Note: The approximate critical value of the single order correlation for $\alpha = .05$ and N of 70, 29, and 99 are .23, .37, and .20 respectively.

reason to believe that the raters individually and collectively are able to predict estimated item statistics better than the "empirical rating" (that is, the mean difficulty and discrimination computed on items sorted by error categories). This suggests that subject matter experts can add a unique valid component of their own to the prediction of difficulty and discrimination.

From a theoretical perspective, however, it is troubling that the raters—subject matter specialists and skilled test developers—were not able to articulate "theories" that could account for the variations among items in difficulty and discrimination. Clearly, a fairly elaborate theory must be required to account for that variation. In considering how to proceed in formulating an initial theory, an appropriate place to look for inspiration is linguistic and psycholinguistic research. One of the most significant concepts in modern linguistic theory is the dichotomy between the deep and surface structure of sentences. The deep structure of a sentence is a representation of the meaning of that sentence (e.g., Langendoen, 1969). The surface structure is a manifestation of that meaning, that is, the sentence as we read it.

It would appear that further research on the problems of predicting or anticipating the statistical characteristics of items could benefit from a linguistic analysis of test items. The initial hypothesis would be that the degree of difficulty in judging the grammaticalness of a sentence is related to the linguistic features of the stem. We may, for example, examine the deep structure of the stems of the easiest and the most difficult items testing a specific grammatical error. Contrasting the resulting deep structures may give some clues as to why the difficult items are difficult and the easy ones easy. Alternatively, the key to the difficulty of items may not lie in their deep structure as such, but in the transformations that are applied to the deep structure to produce the surface structure. Consider the following two sentences taken from Langendoen (1969, chap. 8).

1. The rumor that that the report which the advisory committee submitted was suppressed is true is preposterous.
2. The rumor is preposterous that it is true that the report which the advisory committee submitted was suppressed.

These two sentences have the same deep structure; that is, they mean the same thing and yet have surface realizations which differ substantially in their comprehensibility. If TSWE items were to be based on these sentences, it is likely that an item

derived from Sentence 1 would be more difficult. This hypothesis, proposed by Miller (1962), is known in psycholinguistic circles as the "derivational theory of complexity." Although the theory has run into difficulties, Valian (1979) has noted "that no experiment 'disproved' DTC [derivational theory of complexity], and as a first approximation DTC may be correct" (Valian, 1979, p. 5).

It is to be expected that a syntactic analysis alone will not account fully for the statistical characteristics of items. After all, the TSWE intends to measure correct usage, not just correct grammar, and usage refers to ". . . the attitudes speakers of a language have toward different aspects of their language . . ." (Postman & Weingartner, 1966, p. 80). The problem is compounded by the fact that those attitudes as well as other nonsyntactic factors probably are not invariant across subpopulations. Nevertheless, so long as the mixture of subpopulations taking the test is fairly constant, it may be feasible to use subject matter experts as a means of tapping the nonsyntactic determinants of item difficulty and discrimination. The integration of syntactic and nonsyntactic factors is likely to improve the predictability of item statistics.

References

Kirsch, I. S., & Guthrie, J. T. Construct validity of functional reading tests. *Journal of Educational Measurement,* 1980, *17,* 81–93.

Langendoen, D. T. *The study of syntax: The generative-transformational approach to American English*. New York: Holt, Rinehart, & Winston, 1969.

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

Lorge, I., & Kruglov, L. The improvement of estimates of test difficulty. *Educational and Psychological Measurement,* 1953, *13,* 34–36.

Miller, G. A. Some psychological studies of grammar. *American Psychologist,* 1962, *17,* 748–762.

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

Postman N., & Weingartner, C. *Linguistics: A reevaluation in teaching*. New York: Dell, 1966.

Ryan, J. J. Teacher judgment of test item properties. *Journal of Educational Measurement,* 1968, *5,* 301–306.

Searle, B. W., Lorton, P., & Suppes, P. Structural variables affecting CAI performance on arithmetic word problems of disadvantaged and deaf students. *Educational Studies in Mathematics,* 1974, *5,* 371–384.

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

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

Valian, V. The wherefores and therefores of the competence-performance distinction. In W. H. Cooper & E. C. T. Walker (Eds.), *Sentence processing: Psycholinguistic studies presented to Merrill Garrett*. Hillsdale NJ: Erlbaum, 1979.

Acknowledgments

I am grateful to the Test Development staff of the College Board division at ETS for serving as raters in this study, as well as for contributing important insights to the design of the study, and to Thomas Donlon for useful suggestions.

Author's Address

Send requests for reprints or further information to Isaac I. Bejar, Educational Testing Service, 07-R, Princeton NJ 08541, U.S.A.