USING THE NON-FINISHERS' RESPONSES TO EXAMINE ITEM BIAS

Jennie P. Yeh
Joan L. Herman

CSE Report No. 152

November, 1980

Center for the Study of Evaluation
Graduate School of Education, UCLA
Los Angeles, California 90024

The research reported herein was supported in whole or in part by a grant to the Center for the Study of Evaluation from the National Institute of Education, U.S. Department of Education. However, the opinions and findings expressed here do not necessarily reflect the position or policy of NIE and no ifficial NIE endorsement should be inferred.

Table of Contents

Introduction

Within the classical theory of measurement, the quality of a test is typically characterized by its reliability. A more important quality of a good test, however, is that it provides valid information about an individual. An achievement test, for example, should provide as complete a picture as possible of the student's mastery of the subject or skill. Stated differently, if a student's knowledge of a given subject is more complete or greater than can be inferred from his performance on the test, then the test is faulty.

Guidelines to help test writers avoid possible pitfalls and assure that items will function as intended are readily available in the literature, e.g., preventing unintended clues to the answer in the wording of the questions, eliminating irrelevant barriers by avoiding ambiguous statements or vocabulary. The issue of format is also of concern. Correct answers, for example, should not be of greater length than incorrect answers----a situation which can easily occur given the need for precision in the correct answer. The quality of an item is routinely evaluated by its clarity, difficulty level, and ability to discriminate between people who know and do not know the subject. This evaluation is usually based on the information provided by those who answered an item, whether correctly or incorrectly. The problem of guessing, that is, responses made by those who don't know the subject but answer a question correctly, has been studied extensively, but little is known about those who make no attempt to answer an item. Can we safely assume that no attempt made indicates no knowledge? Perhaps we can, if a students skips only a few items unsystematically. But what of situations where students are given

enough time to finish the measure (power test), but quit in the middle? What assumptions can be make about their mastery of the subject and what assumptions can we make about the difficulty of the item that students did not attempt to answer? Does a student quit because he does not know the answer or for other reasons? Very little has been done to answer these questions, yet it is common knowledge that stopping in the middle of a test frequently occurs, especially among low-achieving or disadvantaged students.

The results from the Oral Reading Test developed for the Early Childhood Education Program (ECE) evaluation (Baker, 1977) provided the opportunity to study some of the factors that might influence students' test-taking behavior. This test was administered during the ECE study, and similar versions were also used in two small subsequent studies in local schools. Because the procedures for administration permitted students to terminate the test at any point, the results of the three studies allowed an investigation of the item characteristics that might predict students' behavior, and also provided some insight into other factors that might influence students' decisions not to complete a test.

Procedures and Results

The Oral Reading Test required individual students to demonstrate reading skills by reading aloud sentences typical of various grade level texts. The initial plan in constructing the test was to develop linguistic rules in order to select sentences of progressively more difficult syntax. Unfortunately, however, research in this area is equivocal. Additionally, a review of text materials revealed that sentences drawn from different grade levels did not systematically increase in complexity. Therefore, rather than employing a sentence sampling plan based on prior stratification rules, the

test was constructed by a simple random selection procedure. Six sentences were drawn from each of the four most commonly ordered basic reading series in California, according to the following scheme:

1 sentence from the last half of the first grade text.

1 sentence from the first half of the second grade text.

1 sentence from the last half of the second grade text.

1 sentence from the first half of the third grade text.

1 sentence from the last half of the third grade text.

1 sentence from the first half of the fourth grade text.

The resulting twenty-four sentences were randomly assigned, by grade level, to one of four parallel test forms. For each of the test forms, there was both a student version, printed in primary type, and corresponding administrator coding sheets, so that errors could be noted precisely as they occurred.

In order to minimize a child's anxiety and frustration, it was decided that testing would terminate at a child's request. Further, if a child made more than three mispronunciation, substitution, or significant hesitation errors, he would be asked whether he wanted to stop the test. The child's decision would determine whether the test terminated at that point.

In the ECE study, the oral reading test was administered to 8-12 randomly selected students in each of four classrooms in our 72 school sample. Of the 1,380 second-graders who were given the test, 179 or 12.9% did not finish the entire test. Table 1 presents a breakdown of the points on the test at which these non-finishers decided to stop. On Form 1, 18.8% of the non-finishers quit after they read sentence 1, and 15.2% quit after reading sentence 2. Likewise 5.4% of the non-finishers stopped after reading sentence 1 on Form 2, and 13.4% stopped before they attempted to read the last sentence.

TABLE 1

NON-FINISHER'S DROP-OUT PATTERN ON ECE ORAL READING TEST

(PERCENTAGE OF STUDENTS WHO QUIT)

Form	Sentence				
	After 1	After 2	After 3	After 4	After 5
1	18.8%	15.2%	23.6%	18.8%	23.6%
2	5.4%	8.0%	39.3%	33.9%	13.4%
3	12.1%	18.5%	42.7%	14.5%	12.1%
4	13.3%	18.1%	20.0%	13.3%	35.5%

One factor that may influence a student's decision to quit in the middle of a test is that the next item is too difficult. Another possible factor in this decision might be that the student had just finished a difficult item and was reacting to feelings of frustration. The difficulty level of each sentence, as indicated by the ratio of the average number of errors made to the maximum possible errors, is presented in Table 2. A strong relationship between the difficulty level of the previous or following sentence and the percentage of non-finishers who quit was not found.

TABLE 2

Sentence Difficulty for ECE Oral Reading Test

(Mean Percentage of Errors)

Form	Sentence					
	1	2	3	4	5	6
1	11.1	11.1	11.6	7.6	14.5	6.2
2	10.2	7.8	14.5	15.6	14.1	13.4
3	6.2	5.5	12.0	17.0	6.4	14.2
4	9.7	7.3	11.5	15.0	7.4	14.8

Table 3 presents the number of words in each of the 24 sentences (six sentences in each of the 4 forms). Compare Table 1 and 3, in Form 1. The length of sentence 3 is 8 words and of sentence 4, 19 words. 15.2% of the non-finishers quit before attempting to read sentence 3, and 23.6% quit before sentence 4. Also, in Form 4, there are 7 words in sentence 5 and 17 words in sentence 6. 13.3% quit before attempting to read sentence 5 and 35.5% quit before sentence 6. A scatter plot of percentage of students quitting and the length of the following sentence is presented in Figure 1. A strong linear relationship was observed. The Pearson Product Moment Correlation Coefficient was 0.69 with df=18. This was significant at α=.0003.

TABLE 3

Number of Words per Sentence for Each Form

Form	Sentence 1	2	3	4	5	6	Form Total
1	9	17	8	19	16	22	91
2	6	9	4	18	17	10	64
3	5	15	16	21	12	14	83
4	5	12	14	14	7	17	69

In two other follow-up studies, the results from the Oral Reading Test provided further information on the problem of non-finishers. One study was conducted at a school district that was known to have low SES and low student achievement. The Oral Reading Test was re-organized into two forms (Form L and J) each composed of 8 sentences and given to 159 students. In comparing the Oral Reading Test results from the ECE study and that of the low SES district study (Table 4), we found that, as expected, the latter had a greater percentage of students who chose not to finish the test (31.45% vs. 12.97%). Those who finished the test, when compared with non-finishers on the same sentence, made more errors on each of the 16 sentences. On the average, students from the low SES district made 1.22 errors on sentence 2 on Form 1. The students in the ECE study made only 0.72 errors. However, a comparison of the non-finishers from the two studies tell a different story. The non-finishers from the low SES district actually made significantly fewer errors on 10 out of 16 sentences than the ECE sample. The fact that the non-finishers in the lower SES district made fewer errors than the non-finishers from the cross-sectional ECE sample indicated that there is a difference between the two groups in the relationship between knowledge (number of errors made) and tendency to quit in the middle of a test. Given the same ability level as those in the statewide sample, students from the lower SES district were more likely to give up trying.

TABLE 4

COMPARISON OF ECE AND LOW SES DISTRICT ORAL READING RESULTS

AVERAGE NUMBER OF ERRORS

Form L/Sentence	1	2	3	4	5	6	7	8
ECE Finishers	.49	.72	.88	.82	.97	2.83	1.61	1.75
ECE Drop-Outs	2.17	2.33	4.05	5.07	3.80	8.20	6.12	NA
Low SES District Finishers (N=55)	.84	1.22	1.89	1.69	1.80	4.69	3.16	3.42
Low SES District Drop-Outs (N=23)	2.00	3.39	4.84	4.31	4.08	6.44	2.00	NA

Form J/Sentence	1	2	3	4	5	6	7	8
ECE Finishers	.88	.82	1.83	2.05	3.22	2.91	2.83	2.91
ECE Drop-Outs	4.05	5.07	5.81	5.03	7.70	7.64	8.20	NA
Low SES District Finishers (N=54)	1.18	.96	1.87	2.17	3.43	2.87	2.78	2.81
Low SES District Drop-Outs (N=27)	4.85	4.60	3.81	5.33	6.20	2.37	1.00	NA

An adapted version of the ECE Oral Reading Test was also used in a study conducted in a local school serving a higher SES community. Four forms of the test were administered, each containing seven sentences. Two forms of the test (Forms 1 and 2) contained sentences randomly selected from those used in the ECE study, while the other two forms contained newly sampled sentences. Procedures for test administration were the same as those used in the ECE study. The distribution of finishers and non-finishers for the 185 students who took the test is shown in Table 5.

TABLE 5

Suburban Higher SES School Oral Reading

(Non-Finishers Drop-Out Pattern
By Form Number and Percentage of Students Quit)

Form	Last Sentence Read							Total Number Of Students
	1	2	3	4	5	6	7	
1	-	6%(3)	-	4%(2)	-	-	90%(46)	51
2	-	-	4%(2)	-	-	-	96%(46)	48
3	2%(1)	-	-	-	-	-	98%(44)	45
4	2%(1)	-	-	-	-	-	98%(40)	41

As Table 5 indicates, only 5% of the students from this higher SES school did not finish the test, compared with 12% non-finishers in the ECE sample. While one plausible explanation for the lower rate of non-finishers in the suburban school might be level of reading achievement, comparison of the results of identical items on a criterion-referenced reading test revealed no significant differences.(In fact, the scores of students of comparable age from the suburban school were slightly lower than those in the ECE sample.)

However, an interesting difference in student attitudes was found between the two groups. In response to the question, "Are you a good reader?" 91% of the suburban student population answered affirmatively, compared with 80% of the ECE sample. While attitude data was not available at the student level, this difference suggests that students' self-concept and self-confidence may be an important factor in test-taking behavior.

Summary and Conclusions

Bias in the Oral Reading Test appears to be a real possibility. Three potential sources of bias are test wiseness, cultural differences in verbal behavior, and examiner expectations.

In the first study, involving the ECE children, sentence length and quitting were found to be related. Given that the difficulty of vocabulary was kept constant, it would appear that children quit on the basis of their expectations of difficulty, rather than the actual difficulty of the sentences. Because the sentences used in the test were not contextualized (and often quite awkward in construction), the reading task was actually more difficult than what would occur in a more natural oral reading situation. If the reading task was difficult to begin with, then the likelihood of quitting on the basis

of sentence length seems to be greater even among students with normal reading ability if their self-confidence in dealing with new situations is low. For example, in the follow-up study with the high-SES school children, the lower rate of quitting (5% as compared to 12.9 from the ECE group) is hypothesized to be related to higher self-confidence as indicated by a higher rate of self-evaluation as a "good reader."

Among the lower SES children, the quitting rate was much higher than among the original ECE group--31% vs. 13%. In addition to the possibility that the lower SES children lacked self-confidence, cultural bias in the testing situation may have taken place; most of the low SES group was either Black or Mexican-American. Ratusnik and Koneigsknecht (1977) cite several studies in which bi-racial testing situations resulted in reduced verbal responsiveness and defensive language behavior among Blacks. Houston (1969) reports that the 9-12 year old sample of Black children he studied, hypercorrected their language response in bi-racial testing situations, indicating anxiety over the disparity between their own dialect and the perceived language norm. If such was the situation among the low SES group studied here, a child may have chosen to quit early rather than suffer further embarrassment. This would be particularly true among the children of lower ability. Where a middle-class child of normal ability may continue to perform in a task which is actually more difficult than his ability level, the same level child from a minority background may tend to give up. Results of research suggest that differences in motivation under conditions of testing vary among SES groups with the higher SES group displaying the greatest motivation.

One other factor which must be considered is that of examiner bias--

especially with respect to the administrative procedure of terminating a testing session if a certain number of mistakes are observed. If expectations are lower, it is more likely that a pause will be interpreted as frustration, rather than meaningful pondering, for example, and that the test will be terminated on that basis. The "Pygmalion effect" is a well-documented phenomena, and there is no reason to believe it is not a factor in testing situations despite training of examiners. In dealing with certain dialects of English, an examiner may be more likely to interpret a certain pronunciation of a word as a mistake. (See, for example, Grill and Bartel's article on Language Bias in the Grammatic Closure subtest of the ITPA (1977). In other words, possibilities of bias extend to test examiners as well as student background characteristics.

References

Baker, E. Evaluation of the California Early Childhood Education Program, Volume I. Los Angeles: Center for the Study of Evaluation, University of California, 1976.

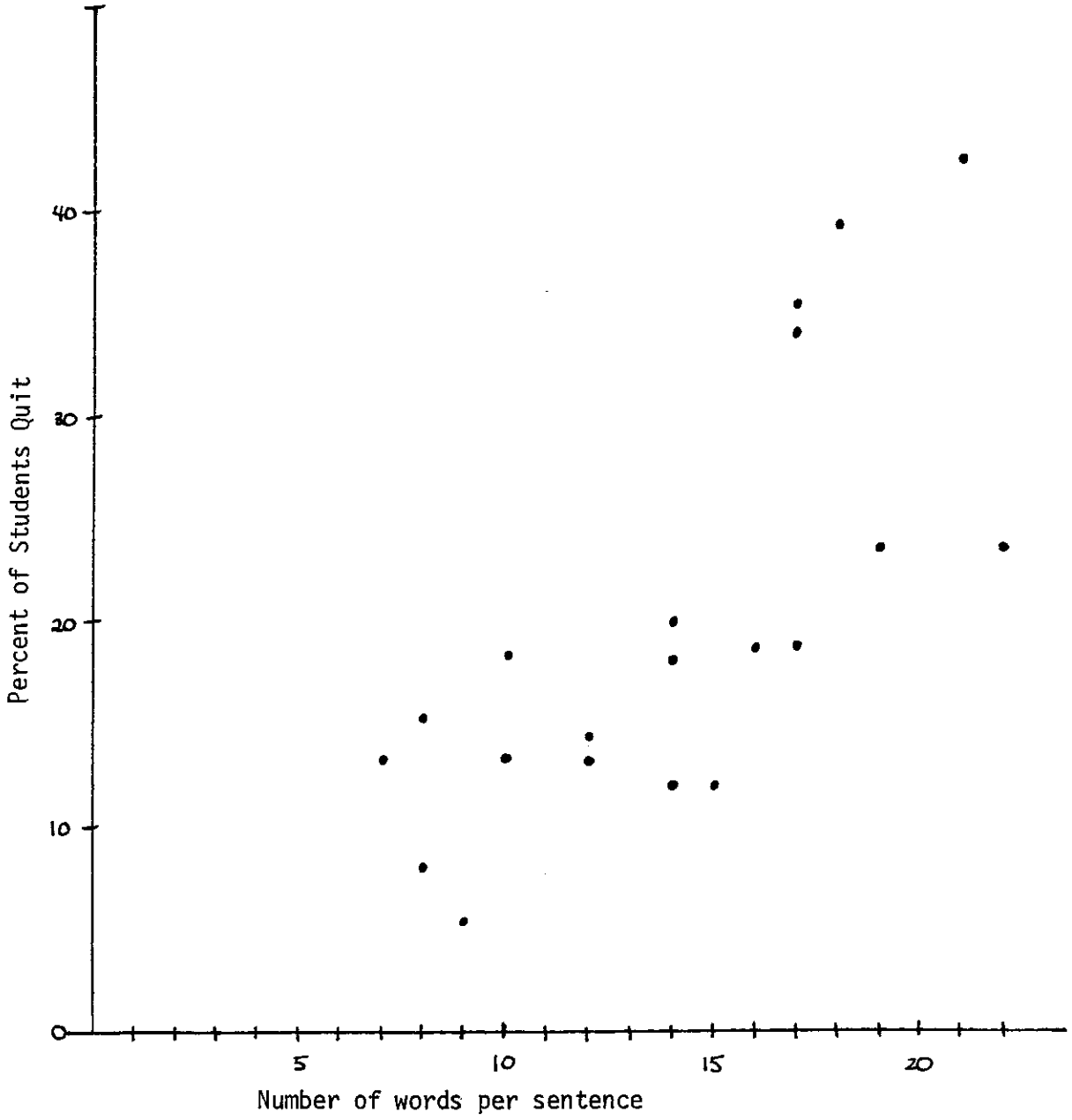
Grill, J. J., & Bartel, N. R. Language bias in tests: ITPA grammatic closure. Journal of Learning Disability, 1977, 10(4), 41-47.

Houston, S. A. A sociolinguistic consideration of Black English of children in northern Florida. Language, 1969, 45, 599-607.

Ratusnik, D. L., & Koneigshnecht, R. A. Bi-racial testing: The question of clinicians' influence on children's test performance. Language, Speech, and Hearing Services in Schools, 1977, 8, 5-14.

FIGURE 1

Scatterplot of Number of Words per Sentence
by Percent of Students Quit



Correlation (R) = 0.692
Significance = .00035
R Squared = 0.480
Std Err of Est = 3.499
Intercept (A) = 7.425
Slope (B) =31.641
Plotted values = 20