Multiple Regression Procedures for the Prediction of Item Difficulties

Ronald J. Nungester
The American College

F J King
Florida State University

Good test construction is at best an arduous and complicated task. Item development requires sophisticated knowledge of both the content area and measurement procedures. To construct a test with optimal measurement characteristics requires that the items selected for inclusion in the test be of appropriate difficulty for the group being measured. The purpose of this paper is to examine a procedure for estimating item difficulties or p values (defined here as the percentage of correct responses to the item) from the structural characteristics of the items through the use of multiple regression techniques.

The early attempts at predicting item difficulty (Thorndike, Bregman, and Cobb, 1924; Tinklemen, 1947) were through the use of judges to rate the item difficulty. These studies indicated that although fairly good relative estimates of item difficulty could be obtained, large numbers of judges were required. Multiple regression techniques were used by Suppes, Loftus, and Jerman (1969) and Jerman and Rees (1972) to analyze the primarily arithmetic structural characteristics of verbal arithmetic items. Structural analysis of items was used by Smith and Shaw (1969) as an aid to instructional design. One hundred elementary addition items were generated from 10 structural variables. Difficulty estimates were then obtained from the administration of the items to 97 fifth grade students. Using stepwise multiple regression procedures, six of the structural characteristics were found to correlate .96 with the item p values. They concluded that if most of the variance in item difficulty could be accounted for by structural variables, then regression analysis could be used to optimally organize and sequence instruction.

Regression analysis has also been used to estimate text readability. King (1974) used the structural characteristics of reading passages to successfully predict the percentage of cloze test restoration scores for selected reading passages. Kane (1973) discussed problems associated with predicting the difficulty of mathematical reading passages. He observed that one of the problems associated with estimating the readability of mathematical passages was the complex combination of mathematical symbols and English text. As an aid to analyzing mathematical text, he distinguished between the language of mathematics (LM) and ordinary English (OE), where LM was defined as composed of OE and formal symbol systems such as Hindu-Arabic numeration. Within LM the terms "word token" and "math token" were defined (Kane, Hater, and Byrne, 1971), where word tokens were words having special meaning in LM which could be in addition to other meanings in OE, such as "plus". Math tokens were the symbols associated with mathematics, such as "-" which might not have a direct phoneme-grapheme relationship.

This paper attempts to integrate these findings and procedures and apply them to the prediction of item difficulties obtained from widely used standardized achievement tests.

METHOD

Materials

One hundred forty word problems were selected for analysis, 100 from the Comprehensive Tests of Basic Skills (CTBS), Level 2, Forms Q and R, and 40 from the California Achievement Tests (CAT), Form A, Level 3. Both the CTBS and the CAT are standardized achievement tests normed on national samples of fourth, fifth, and sixth grade students.

Procedure

Each of 140 items was analyzed for 20 structural variables. In this paper, the definition of a structural variable suggested by Smith and Shaw (1969) was used. This definition stated that a structural variable was "any characteristic, format or content, which distinguishes a problem from other problems." These 20 variables were a combination of predictors found in the Jerman and Rees (1972) study, the readability variables used by King (1974), and other variables added for this study.

The following variables were taken from Jerman and Rees:

NOMC2 (X_1) - a count of 1 was assigned each time a regrouping occurred in each multiplication exercise in the problem.

QUOT (X_2) - a count of 1 was given for each digit in the quotient if division was required, and 0 otherwise.

COLC2 (X_3) - for this variable a count of 1 was given for each column and a count of 1 was given for each regrouping in addition and substraction exercises. This count applied to only the largest exercise in the problem.

DIST (X_4) - this variable was defined as 1 count for each verbal cue which was not a cue for an operation but a distractor. For example, if the word "average" was used but multiplication rather than division was the required operation.

UCONV (X_5) - this factor was present if a conversion (e.g., feet to yards) was required and the equivalent units were not given in the problem (a 0, 1-variable).

LENGTH (X_6) - this variable was defined as the number of words in the problem. To reduce confusion, this variable was renamed NWORD for this paper.

STEP (X_7) - the minimum number of binary operations, steps, needed to reach a solution.

Six readability variables from King (1974) were used. These are:

NDIFW (X_8) - number of different words in the item.

NDWNL (X_9) - number of different words in the item not on the Dale list of words

AVSNLG (X_{10}) - average sentence length for the item.

AVWDLG (X_{11})- average word length for the item.

PDIFW (X_{12}) - percentage of different words in the item.

PDWNL (X_{13}) - percentage of different words in the item not on the Dale list.

In that the content of the items analyzed in this study differed from previous studies, the following structural variables were defined by the present authors.

FRAC (X_{14}) - this factor was present if the problem involved fractions (a 0, 1-variable).

DECI (X_{15}) - this factor was present if the problem involved decimals (a 0, 1-variable).

ALGE (X_{16}) - this factor was present if the problem was presented in algebraic form (a 0, 1-variable).

TIME (X_{17}) - this factor was present if the problem involved operations with time (a 0, 1-variable).

GEOM (X_{18}) - this factor was present if the problem involved geometry (a 0, 1-variable).

MONY (X_{19}) - this factor was present if the problem involved the use of money (a 0, 1-variable).

ALTERN (X_{20}) - this factor indicated the nature of the response alternatives (0 - numerical, 1 - verbal, 2 - mixed).

The use of King's readability variables with mathematical text required that protocols for the analysis of math tokens be created. For numerals, each numeral was counted as one familiar word with a word length equal to the number of digits in the numeral. Thus, 1215 would be tallied as one familiar word with a length of four letters. All other mathematical symbols were expressed in word form for analysis.

Analysis

A total of 140 items were selected from the three standardized tests. Each of these items was analyzed for each of the 20 structural variables. Separate item p values were available from the respective test manuals for all three grade levels (4th, 5th 6th), thus giving a total sample size of 420 (3 x 140). These 20 structural variables, two dummy variables, GRADE4 (X_{21}) and GRADE5 (X_{22}), plus the interactions between grade level and other variables were used as predictors of item p values (expressed as percentages for analysis) in a stepwise regression analysis. In addition, an a priori decision was made that only predictors with a nominal significance of $\underline{p} < .01$ would be included in the equation.

RESULTS

With the criterion of a nominal significance level of $\underline{p} < .01$, ten variables entered the regression equation (Table 1), yielding an $\underline{R}$ of .68.

Table 1

Summary of Stepwise Regression Analysis of 140 Verbal Arithmetic Items

STEP	VARIABLE	F TO ENTER OR REMOVE	PROBABILITY	MULTIPLE R
1	INDWNLG4 (X_9X_{21})	68.535	.000	.375
2	GEOM (X_{18})	38.393	.000	.461
3	IFRG4 ($X_{14}X_{21}$)	35.039	.000	.524
4	IFRG5 ($X_{14}X_{22}$)	22.214	.000	.588
5	PDIFW (X_{13})	24.996	.000	.592
6	QUOT (X_2)	16.997	.000	.613
7	COLC2 (X_3)	16.460	.000	.633
8	AVSNLG (X_{10})	23.195	.000	.657
9	INDWNLG5 (X_9X_{22})	9.333	.002	.667
10	GRADE4 (X_{21})	8.856	.003	.676

Of these variables, four were interactions. INDWNLG4 (X_9X_{21}) and INDWNLG5 (X_9X_{23}) were interactions of the number of different words not on the Dale list and grades 4 and 5. IFRG4 ($X_{14}X_{21}$) and IRRG5 ($X_{14}X_{22}$) were interactions of fractions and grades 4 and 5. The resulting regression equation was:

$$(1) \quad Y_i = 74.9 - 29.68X_{18} + .15X_{13} - 4.72X_2 - 2.31X_3 - .65X_{10} - 8.86X_{21} - 1.93X_9X_{21} - 1.42X_9X_{22} - 23.6X_{14}X_{21} - 19.16X_{14}X_{22}.$$

Regression coefficients, standard errors of regression coefficients, computed F values, level of significance, β weights and elasticity are given in Table 2. In order to compare the relative importance of each weight, the regression coefficients were normalized to β weights.

Table 2

Regression Coefficients, Standard Errors of Regression Coefficients Computed F Values, and β Weights

VARIABLE	β	STANDARD ERROR OF b	F	PROBA-BILITY	β WEIGHT	ELA TICI
INDWNLG4 (X_9X_{21})	-1.926	0.665	8.386	.004	-.200	-.03
GEOM (X_{18})	-29.686	2.878	106.389	.000	-.415	-.03
IFRG4 ($X_{14}X_{21}$)	-23.602	3.552	44.142	.000	-.267	-.01
IFRG5 ($X_{14}X_{22}$)	-19.158	3.405	31.658	.000	-.217	-.01
PDIFW (X_{13})	0.148	0.056	6.962	.009	.114	.15
QUOT (X_2)	-4.715	0.860	30.093	.000	-.209	-.03
COLC2 (X_3)	2.314	0.480	23.221	.000	-.186	-.02
AVSNLG (X_{10})	-0.647	0.140	21.434	.000	-.187	-.11
INDWNLG5 (X_9X_{22})	-1.424	0.394	13.030	.000	-.148	-.02
GRADE4 (X_{21})	-8.860	2.977	8.856	.003	-.217	-.04
(CONSTANT)	74.588	4.606	262.135	.000		

This transformation to normalized form was necessary since each variable had been expressed in different units. The normalized equation was:

(2) $Z_i = -.41X_{18} + .11X_{13} - .21X_2 - .19X_3 - .19X_{10} -$
$.22X_{21} - .20X_9X_{21} - .15X_9X_{22} - .27X_{14}X_{21} -$
$.22X_{14}X_{22}$

Once the equation is in normalized form, examination of the β weights provides a sensible method for comparison of the relative contribution of a variable. In equation 2, a one standard deviation change in GEOM (X_{18}) will be accompanied by the largest change in Z_i. Additional feeling for the importance of a variable may be obtained by examining the percentage of total variance accounted for by the variable (Guilford, 1965). This value is obtained by finding the product of the variable's β weight and the variable's zero order correlation with the dependent variable. The obtained value for the percentage of variance accounted for by each variable respectively is expressed as follows:

$46 = 12_{(18)} + 0_{(13)} + 4_{(2)} + 2_{(3)} + 2_{(10)} + 8_{(21)} + 8_{(9,\ 21)} + 0_{(9,22)} +$
$8_{(14,\ 21)} + 3_{(14,\ 22)};$

where the subscripts indicate the variable's number. It should be kept in mind that these values are appropriate only within this 10-predictor model. There are no assurances that these values would remain constant if variables were added or deleted. Table 3 shows for each variable the order of entry, order of importance within the regression equation, variance added by each step, and variance from regression.

Table 3

Order of Importance of Predictors
Within the 10-Predictor Model

VARIABLE	ORDER OF ENTRY	ORDER OF IMPORTANCE	VARIANCE ADDED STEP BY STEP	VARIANC FROM REGRESSI
GEOM (X_{18})	2	1	.072	.12
IFRG4 ($X_{14}X_{21}$)	3	2	.061	.08
GRADE4 (X_{21})	10	3	.011	.08
IFRG5 ($X_{14}X_{22}$)	4	4	.037	.03
QUOT (X_2)	6	5	.026	.07
IDWNLG4 (X_9X_{21})	1	6	.141	.08
AVSNLG (X_{10})	8	7	.032	.02
COLC2 (X_3)	7	8	.024	.02
IDWNLG5 (X_9X_{22})	9	9	.013	.00
PDIFW (X_{13})	5	10	.039	.00

Examination of Tables 2 and 3 shows the inappropriateness of using raw regression coefficients, order of entry into stepwise regression, or the increase in variance accounted for by each step as a criterion for importance of a predictor.

Since equation 1 was computed from estimated item difficulties for grade levels four, five and six, it was possible to write a separate equation for each grade. This was done by replacing the dummy variables GRADE4 (X_{22}) and

GRADE5 (X_{23}) with the appropriate values for the grade level in question. The following three equations were determined:

(3) $Y_i = 74.9 - 29.68X_{18} + 15X_{13} - 4.72X_2 - 2.31X_3 - .65X_{10}$ (Sixth Grade)

(4) $Y_i = 74.9 - 29.68X_{18} + .15X_{13} - 4.72X_2 - 2.31X_3 - .65X_{10} - 1.42X_9 - 19.16X_{14}$ (Fifth Grade)

(5) $Y_i = (74.9 - 8.86) - 29.68X_{18} + .15X_{13} - 4.72X_2 - 2.31X_3 - .65X_{10} - 1.93X_9 - 23.6X_{14}$ (Fourth Grade)

If all predictors were equally important for each grade level, it would be expected that the regression lines for each grade level would be parallel or coincidental. For this to occur, the regression equations could differ only in intercepts. The same predictors and regression coefficients would be found in each equation. Since grade level was represented by two dummy variables, parallel lines would be forced unless interactions occurred between structural and grade level variables. Examination of equation 1 reveals that interactions did occur.

The equation for sixth grade, equation 3, was composed of a combination of five verbal and quantitative predictors plus the intercept. The equation for the fifth grade, equation 4, contained two predictors, NDWNL (X_9) and FRAC (X_{14}), in addition to the five predictors in equation 3. The addition of these predictors indicated that items would be predicted to be more difficult for fifth graders than for sixth graders, if the items contained fractions or unfamiliar words. For fourth graders, equation 5 was found to contain the same predictors as equation 4; however, a different intercept and different regression coefficients for NDWNL (X_9) and FRAC (X_{14}) were obtained. From equation 5 we see that all items will be predicted to be more difficult for fourth graders than for either fifth or sixth graders. In addition, since the regression coefficients for NDWNL (X_9) and FRAC (X_{14}) differ in magnitude from equation 5 to equation 4, the estimated p values for items containing fractions or unfamiliar words would be predicted to be even more difficult for fourth graders.

In addition to the aforementioned analysis, a logistic transformation (Cox, 1970) of the dependent variable was made. This transformation provided a procedure for handling problems associated with the boundedness of p values. However, no appreciable increase in R was obtained.

DISCUSSION

It has been suggested that structural analysis of items can provide useful information to test developers. This information should be useful on at least two levels. First, the generation of regression equations would enable item writers to have quick, inexpensive estimates of item difficulties. For example, from equation 1, for the average verbal arithmetic item having an average sentence length of five words, no words not on the Dale list, and involving geometry, the predicted percentage of correct responses for sixth grade students would be:

$$74.9 - 29.68(1) + .15(0) - 4.72(0) - 2.31(0) - .65(5) = 41.97.$$

Such estimates would give the test constructor at least a feeling for the suitability of the item.

A second use of the regression technique would be in examining the measurement properties of existing tests for different groups. Equations 4 and 5 could be interpreted as indicated that the items analyzed are measuring different traits for students in different grade levels. For grades four and five, these items are more a measure of ability to read and to do fractions than they are for grade six. This technique could be similarly used to examine differences by sex or race in the mesurement properties of a set of items, thus providing yet another method for examination of test bias. However, a great deal of care must be taken in interpreting the results of such an analysis. It must be kept in mind that this procedure is correlational in nature and that there is no statistical justification for making causal inferences.

In summary, item structural analysis should be useful in the construction and use of standardized achievement tests and selection examinations. The development of these equations would be particularly useful in secure testing situations where no preliminary administration of items can be made. It seems reasonable that regression equations could be developed for many types of items and subject populations. To obtain optimal predictability, items should be grouped into generally homogeneous types. Thus, many equations would need to be developed, catalogued, and maintained.

References

Cox, D.R. Analysis of binary data. London: Metheun & Co., Ltd., 1970.

Guilford, J.P. Fundamental statistics in psychology and education. New York: McGraw-Hill, 1965.

Jerman, M. and Rees, R. Predicting the relative difficulty of verbal arithmetic problems. Educational Studies in Mathematics. 1972, 4, 306-323.

Kane, R.B. Assessing the readability of elementary instructional materials in mathematics. Paper presented at the annual meeting of American Educational Research Association, New Orleans, Louisiana, February, 1973.

Kane, R.B., Hater, M.A., and Byrne, M.A. A readability formula for mathematical English. Paper presented at the annual meeting of the National Council of Teachers of Mathematics, Los Angeles, California, Spring, 1971.

King, F.J. A content referenced interpretive system for standardized reading tests. Florida State University, 1974.

Smith, T.A. and Shaw, C.N. Structural analysis as an aid in designing an instructional system. Journal of Educational Measurement. 1969, 6, 137-143.

Suppes, P., Loftus, E., and Jerman, M. Problem-solving on a computer based teletype. Educational Studies in Mathematics. 1969, 2, 1-15.

Thorndike, E.L., Bregman, E.O., and Cobb, M.V. The selection of tasks of equal difficulty by a consensus of opinion. Journal of Educational Research, 1924, 9, 133-139.

Tinklemen, S. Difficulty prediction of test items, New York: Bureau of Publications, Teachers College, Columbia University, 1947.