ED 231 854 TM 830 383

AUTHOR Green, Kathy E.

TITLE Multiple-Choice Item Difficulty: The Effects of Language and Distracter Set Similarity.

PUB DATE Apr 83

NOTE 17p.; Paper presented at the Annual Meeting of the American Educational Research Association (67th, Montreal, Quebec, April 11-15, 1983).

PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.

DESCRIPTORS *Difficulty Level; Foreign Countries; Higher Education; Item Analysis; *Language; *Multiple Choice Tests; *Response Style (Tests); Semantics; Syntax; Test Construction; Testing Problems; *Test Items

IDENTIFIERS *Distractors (Tests); New Zealand

ABSTRACT

The purpose of this study was to determine whether item difficulty is significantly affected by language difficulty and response set convergence. Language difficulty was varied by increasing sentence (stem) length, increasing syntactic complexity, and substituting uncommon words for more familiar terms in the item stem. Item wording ranged from very simple sentences to sentences involving excess verbiage and syntactic transformations. The semantic similarity of item response options was also varied, creating three levels of response set convergence. Seventeen judges confirmed the orderings of language difficulty and response set convergence for the test items, and 990 undergraduate students at the University of Washington responded to the items. Analyses of item difficulty calculated for each level of each of the two factors included an additive conjoint method, repeated measures analysis of variance, and a test of proportions. Response set convergence was an effective manipulation in the ranges studies. Language difficulty had a non-significant effect on difficulty level. (CM)

MULTIPLE-CHOICE ITEM DIFFICULTY: THE EFFECTS OF LANGUAGE AND DISTRACTER SET SIMILARITY

Kathy E. Green, Ph.D.

Victoria University of Wellington

Wellington, New Zealand

PRINTED IN NEW ZEALAND.

Numerous investigations have been conducted exploring the effects of item manipulations on responses to multiple-choice test items (e.g., Campbell, 1961; Dudycha & Carpenter, 1973). One purpose of these studies has been to identify which characteristics of items affect item difficulty, test reliability, and esp ecially test validity. If a test is designed to measure knowledge of a specific content area, variation in item scores with minor item changes may call into question the test's content validity. For example, if syntactic differences in item phrasing produced items with significantly different difficulty levels, subject area knowledge clearly would not be the only process involved in item solution. If knowledge is to be measured at some predetermined level, as in mastery testing, the effects of item characteristics become even more important. Are two items with different difficulty levels written from the same content measuring the same skills? If it cannot be assumed that they are, it would be informative to understand how the processes involved in responding to the two items differ. In conjunction with this, identification of item characteristics which contribute to differences in difficulty is important, in particular identification of those characteristics which are not content-related. Analysis of item difficulty alone cannot answer the question of whether and to what extent a test is valid. Items measuring the same underlying attributes may, and generally do, differ in difficulty level. But a clearer understanding of item difficulty may suggest a framework for characterizing items. Carroll (1976) suggested that items presented in conventional tests are highly similar to the tasks studied in experimental cognitive psychology and that a logical difficulty analysis method could be applied. By comparing items differing in difficulty, inferences may be made about the processing and knowledge demands of items.

Variation in item phrasing has been an item characteristic examined by several researchers. A significant effect on test performance due to the level of item language difficulty has been found with samples of children and young adults (Benson & Crocker, 1979; Linville, 1970; Loftus & Suppes, 1972); however, studies employing high school and college samples have not consistently found a significant effect (Bornstein & Chamberlain, 1970; Jerman & Mirman, 1971; Millman, 1978). Two possible explanations of the inconsistencies found are that language difficulty has no effect on item difficulty once individuals have reached some criterion level of verbal proficiency , or that the item manipulations employed were too weak to produce detectable effects. If item phrasing were allowed to vary across a wide range of language difficulty, the more plausible explanation might be isolated. However, in the interest of practical concerns, it would be of marginal value to use unrealistic multiple-choice items as stimuli. The range of usable items is assumed to be narrower than the range of all possible items.. It would also be of interest to examine whether the selection of distracters varies with different item phrasings.

Another characteristic of multiple-choice items is the degree of similarity (convergence) among the response options. Guttman and Schlesinger (1967) argued that difficulty could readily be manipulated by altering the similarity of distracters; Maier (1970) suggested that items with highly similar alternatives would be difficult in spite of an individual having sufficient mental capacity and the essential knowledge needed to answer a question. While convergence of response options is assumed by introductory measurement texts (e.g., Chase, 1978) to affect multiple-choice item difficulty, empirical evidence substantiating this and ot'er common assumptions is lacking.

The purpose of the present empirical work was to determine whether item difficulty was significantly affected by language difficulty and response set convergence. Language difficulty was varied by increasing sentence (stem) length, increasing syntactic complexity, and substituting uncommon words for more familiar terms in the item stem. An attempt was made to create items whose wording ranged from very simple sentences to sentences involving excess verbiage and at least one syntactic transformation. This study also varied the semantic similarity of item response options to create three levels of response set convergence. Demonstration of a significant main effect of either variable bears implications for test constructors concerned with content validity and with the abilities assessed by multiple-choice test items.

METHODS

Subjects

Two groups of subjects participated in this study: seventeen judges provided confirmation of the orderings of language difficulty and response set convergence for the test items used; 990 students in 19 separate undergraduate classes at the University of Washington responded to the test items. Students were requested to complete the quiz only once; their responses were voluntary and anonymous.

Materials

Twenty-one general information items were constructed or drawn from an existing source (Nelson & Narens, 1980). Nine variations of each item were written, crossing three levels of language difficulty with three levels of response set convergence. Appendix A presents two test items with all variations. For each of the 21 items, judges were asked to order the item variations on perceived language difficulty and on perceived response set convergence. The method of comparison by triples was used (Rounds, Miller, & Dawis, 1978). Agreement was assessed by calculation of Kendall's Coefficient of Concordance (W, Siegel, 1956). In six cases, the levels of response set convergence assigned by the judges disagreed with those intended by the experimenter. In these instances, the modal ordering of the judges was used to assign levels of response set convergence. For all other items, Kendall's W was significant at $p < .05$. Judges' rankings of items on level of language difficulty agreed in all cases with that intended by the experimenter. Kendall's W was significant at $p < .01$ for all items for language complexity.

One of the nine variations of each item was assigned to each of nine test forms. Assignment was random within the following constraint: At least two items from each level of language difficulty by response set convergence (9 cells) were assigned to each test form. This was done to prevent any form from being composed solely of items from one level of language difficulty or response set convergence. It was felt that using this method of assigning items to forms would produce materials with less possibility of overall bias due to respondent achievement differences. Each test form contained 21 items, randomly ordered, and was three pages in length.

Procedure

Students were randomly assigned to test forms. Of the 990 students receiving a form, 75 responses were discarded due to missing data. The remaining 915 responses were complete or had fewer than four skipped items. Students were given 10-15 minutes at the beginning or at the end of class to complete the test. The number of students responding to each variation of an item ranged from 96 to 105. Students were naive to the specific purposes of the experiment.

Item difficulties were calculated for each level of each of the two factors. One level each of two items were discarded from the analysis due to errors in wording on test forms, leaving 19 complete items. The data analysis was carried out in three stages: (1) Initially, an additive conjoint method was applied to the data to determine the likelihood of language difficulty and response set convergence additively affecting item difficulty (Van der Ven, 1980). Additive conjoint models require assumptions to be met which are weaker than those required when performing parametric tests. (2) The data were then analyzed using a repeated measures analysis of variance to determine whether language difficulty/response set convergence had an overall effect on item difficulty. (3) Provided a significant overall effect was obtained, the effects of variables were assessed for each item using a test of proportions (Guilford & Fruchter 1973). Differences between extreme levels of the treatment variables were tested. Finally, the pattern of distracter selection across levels of language difficulty was examined.

Neither item discrimination indices nor the internal consistency reliability of each test form were calculated. Since each form was a conglomerate of general information questions plus experimental variations in items, the total score was considered to be of marginal use.

RESULTS

No evidence for the existence of an interaction was found using an additive conjoint model. An analysis of variance was then performed (Table 1). The design used was a fully crossed ANOVA with a repeated measure on items. Item content had a marked effect on item difficulty as did response set convergence. The effect of language difficulty was not significant nor was there any evidence of a two-way interaction. Planned comparisons showed levels 1 and 3 and levels 2 and 3 of response set convergence to differ at $p < .05$.

(Table 1 here)

A significant difference (p .05) between levels 1 and 3 of response set convergence was found for 15 of the 19 items (Table 2). In 14 of the 15 cases, the direction of effect was that hypothesized. In the last case (item 5), it was clear that instructions to judges had resulted in an inappropriate ordering of levels of convergence. Judges' instructions were to select as most difficult the response set in which alternatives were the closest. This resulted in the alternative set with numbers more nearly equal in value being chosen as more difficult than the alternative set, in which each option was a possible outcome, given an error in some step in the numerical calculation. For three items, level 2 difficulty did not fall intermediate to levels 1 and 3.

(Table 2 here)

In three cases , level 2 language difficulty values did not fall intermediate to levels 1 and 3.

(Table 3 here)

Holding response set convergence fixed, the selection pattern of distracters for each level of language difficulty was examined for each item. This was done by sorting selection patterns into the following five categories. Of the 61 inspections possible (21 items x 3 levels, minus 2 for items which were incomplete), the numbers in each category were:

11 Too few incorrect responses to classify (i.e., $p > .85$).

28 Rank ordering of distracters by selection frequency identifcal across three levels of language difficulty.

13 Allowing ranks to be equated when the number of persons choosing two distracters differed by 4 or less, "ranks" identical across levels of language difficulty.

4 Allowing ranks to be equated as above, "ranks" identical but selection patterns disproportionate.

5 "Ranks" differ.

Of the 50 inspections made, 41 (82%) showed distracter selection patterns to be similar across levels of language difficulty. The remaining 9 (18%) cases were cast into two-way contingency tables (3 x 3) and χ^2-values calculated to see if there was a significant lack of independence between level of language difficulty and distracter selection for any item. Values of χ^2 significant at $p < .05$ were obtained for three items.

DISCUSSION

One purpose of this research was to contribute to the identification of item characteristics that are trivial and those that have significant effects on how accurately people respond to items. Response set convergence was found to be an effective manipulation, in the ranges studied, while language difficulty was not. The variations in phrasing in this study were not powerful enough to produce significant effects for most items. While making the differences in phrasing even more marked might well have increased the power of the test, the actual phrasing variations employed in this study were designed to produce questions that were potentially usable in the classroom. Results of this study indicate that language difficulty has a small, nonsignificant effect on difficulty level. This suggests that minor variations in phrasing and vocabulary may be considered trivial and permissible in generation of parallel test items.

In the individual case, however, it seems that the effects of language difficulty on item difficulty are not necessarily predictable. This point is underscored by finding items for which level 2 difficulties were significantly lower than level 1 or significantly higher than level 3. In some cases, increasing the language difficulty may provide subjects with additional information or cues to a correct response, making the item easier. Contrarily, the additional information may cue subjects to an incorrect response. One item which has an inappropriate level 2 difficulty was:

Item 19: L1--Preventing unlicensed individuals from charging for medical care will tend to ---
L2--Laws that prevent people from offering medical services for pay unless they have been licensed will tend to ---

In item 19 , L1 produced more difficult items than L2 despite judges' agreement upon L1 as the less difficult wording. A post hoc explanation of this might involve viewing "preventing unlicensed" as a confusing double negative.

Items whose distracter selection patterns differed across levels of language difficulty were also examined. Post hoc explanations of selection patterns for the three items with irregular distracter patterns (Items 13, 17, and 19) would be tenuous but seemed to involve cueing. Specific words in the stem were associated with certain distracters.

When pooled across a large number of test items, the next effects of phrasing on an individual's score are not likely to be large. But there are effects for some individual items. The implications of this are relevant in three areas: First, methods of item construction (e.g., item forms, facet theory, linguistic approaches) which mandate fixed syntax in generation of parallel items may be overly cautious. Second, consistent prediction of differences in item difficulty linked to language complexity will require a much finer analysis and specification of linguistic variables than was the case in this study. Third, models of the difficulty of individual items will need to include a component(s) related to language difficulty if precise predictions about individual items are required. The need for this component would be even more striking if one's purpose were to model the examinee's effort in responding to an item. Anecdotal reports from students suggested that language complexity affected how difficult they perceived the item to be; judges were clearly able to discern and order differences in items by language difficulty. Yet there was no reliable effect of language complexity on the accuracy of responses. The amount of information to be processed increased as level of language difficulty increased, but adults' ability to understand language rapidly must be sufficient to make increases in nonessential information part of encoding. And, given sufficient time for encoding, item phrasing may thereafter play little part in further response. Presenting items with minimal time allowed for response would allow a test of this hypothesis. Also, obtaining measures of subjects' reading ability would allow analysis of the effects of language difficulty and response set convergence for subjects of different ability levels. Examinees' test anxiety would also be a useful measure to obtain in further research of the effects of item variations.

The effects of response set convergence on item difficulty were significant both overall and for the majority of individual items. These results are in agreement with introductory texts' suggestions for manipulation of item difficulty through homogeneity of options.

While for most items judges' orderings of response set convergence levels coincided with the empirical orderings, in three cases they did not. For three items, level 2 difficulties were not intermediate to levels 1 and 3. This result is in accord with research suggesting judges to be effective, but less than perfect predictors of item difficulty (e.g., Prestwood & Weiss, 1977; Willoughby, 1980). It also confirms that subjective judgment of item difficulty on factors other than overall item difficulty may prove useful. The correlation (Pearson's *r*) between the significance of L3-L1 differences in item difficulty and Kendall's *W* was .76 ($p < .01$). The degree of agreement among judges on ordering of levels of response set convergence, then, predicted the significance of item difficulty differences quite well.

When more information is available about the effects on item difficulty of well-defined variations in items, a model of response to multiple-choice items may be within reach. One component of this model would be response set convergence. Others may be factors such as item complexity (number of steps to solution), level of generality of item content, and physical format differences (e.g., physical appearance of items, placement of keyed option, etc.). This study's scope was narrow, concentrating solely upon item difficulty and employing a severely limited number of items. The research design used in this study required a large number of subjects to obtain relatively little information. Alternative designs (such as repeated measures on subjects rather than items) may be more efficient. Further studies may profitably include item discrimination, subjective estimates of item difficulty, or response time as dependent measures. The model developed from synthesizing results of these suggested studies should aid in the continued development of a methodology of item generation. Identification of the item variations having significant and nonsignificant effects on item difficulties should provide crude boundary conditions for the production of parallel test items and thereby facilitate test construction. Content validation of tests will also be conceptually clarified when items can be described in terms of the level of thought as well as the information requirements most central to the correct solution.

This study has served to suggest that the content validity of achievement tests for adults may not be greatly affected by the language difficulty of items per se. If the time allowed is sufficient, phrasing complexity may add to the burden of the examinee but was found by this and other work to have little effect upon overall score. Response set convergence, however, had a strong effect upon item difficulty. The implication of this result is that the degree of discrimination required by an item needs to be specified for clear communication of exactly what is being tested.

Table 1

Analysis of Variance Summary Table[a]

Source of Variation	Sum of Squares	d.f.	Mean Square	F
Language Difficulty (LD)	.033	2	.017	1.84
Response Set Convergence (RSC)	1.337	2	.668	72.21[b]
LD x RSC	.032	4	.008	.89
Items (I)	7.634	20	.382	42.44[b]
Items x Treatments[c] (Error)	1.479	160	.009	
Total	10.515	188		

[a] Language difficulty and response set convergence were treated as fixed and items as random.

[b] Significant at $p<.01$.

[c] The I x LD, I x RSC, and I x LD x RSC sums of squares were pooled to form the error term.

Table 2

Effects of Response Set Convergence on Item Difficulty

	Level 1		Level 2		Level 3		
Item	Difficulty	N	Difficulty	N	Difficulty	N	*
1	-	-	.57	302	.46	296	-
2	.58	302	.55	297	.57	301	.85
3	.72	302	.35	300	.62	304	.01
4	.59	303	.59	300	.56	300	.40
5	.22	303	.42	285	.70	293	.00
6	.93	300	.89	299	.83	301	.00
7	.91	300	.94	295	.94	296	.23
8	-	-	.84	305	.80	305	-
9	.93	303	.90	300	.80	303	.00
10	.92	306	.91	298	.81	295	.00
11	.75	303	.64	302	.36	304	.00
12	.83	304	.82	303	.61	301	.00
13	.34	306	.28	298	.13	297	.00
14	.89	301	.78	300	.76	306	.00
15	.32	299	.42	299	.30	298	.56
16	.72	291	.55	298	.48	286	.00
17	.48	291	.48	293	.37	290	.01
18	.96	300	.82	305	.74	303	.00
19	.44	293	.61	303	.19	299	.00
20	.81	302	.82	302	.40	300	.00
21	.58	304	.27	297	.15	299	.00

*Significance level of differences between level 1 difficulty and level 3 difficulty

Table 3

Effects of Language Difficulty on Item Difficulty

	Level 1		Level 2		Level 3	
Item	Difficulty	N	Difficulty	N	Difficulty	N
1	.61	294	–	–	.65	301
2	.61	303	.51	302	.59	295
3	.59	307	.53	298	.58	301
4	.54	299	.61	300	.59	304
5	.50	298	.39	291	.44	292
6	.92	302	.87	304	.86	294
7	.97	297	.94	301	.88	299
8	.94	302	–	–	.82	301
9	.88	302	.87	302	.88	302
10	.88	302	.87	301	.89	296
11	.60	299	.58	306	.56	304
12	.72	302	.78	302	.79	304
13	.26	297	.28	299	.22	305
14	.85	300	.77	305	.80	303
15	.36	300	.33	303	.34	293
16	.60	294	.60	295	.49	286
17	.51	300	.41	296	.42	278
18	.81	298	.76	304	.83	306
19	.38	299	.51	299	.38	297
20	.70	308	.68	300	.66	296
21	.35	305	.32	300	.33	295

REFERENCES

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

Bolden, B.J. & Stoddard, A. The effects of language on test performance of elementary school children. Paper presented at the Annual Meeting of the American Educational Research Association, Boston, 1980.

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

Campbell, A.C. Some determinants of the difficulty of non-verbal classification items. Educational and Psychological Measurement, 1961, 21, 899-913.

Carroll, J.B. Psychometric tests as cognitive tasks: a new "structure of intellect". In L. Resnick (Ed.), The Nature of Intelligence. Hillsdale, N.J.: Erlbaum & Associates, 1976.

Chase, C.I. Measurement for Educational Evaluation (2nd ed.). Reading, Massachusetts: Addison-Wesley, 1978.

Dudycha, A.L. & Carpenter, J.B. Effects of item format on item discrimination and difficulty. Journal of Applied Psychology, 1973, 58, 116-121.

Guilford, J.P. & Fruchter, B. Fundamental Statistics in Psychology and Education (5th ed.). San Francisco: McGraw-Hill, 1973.

Guttman, L. & Schlesinger, I.M. Systematic construction of distracters for ability and achievement test items. Educational and Psychological Measurement, 1967, 27, 569-580.

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

Linville, W.J. The effects of syntax and vocabulary upon the difficulty of verbal arithmetic problems with fourth grade students. Ph.D. dissertation, Indiana University, 1969. (Dissertation Abstracts International, 1970, 30(9-A),4310)

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

Maier, N.R.F. What makes a problem difficult? In Maier, N. (ed.), Problem Solving and Creativity. Belmont, Ca.: Brooks/Cole Pub. Co., 1970.

Millman, J. Determinants of item difficulty: a preliminary investigation. CSE Report No. 114, Center for the Study of Evaluation, UCLA, July 1978. (ERIC ED 163 071)

Nelson, T.O. & Narens, L. Norms of 300 general-information questions: accuracy of recall, latency of recall, and feeling-of-knowing ratings. Journal of Verbal Learning and Verbal Behavior, 1980, 19, 338-368.

Prestwood, J.S. & Weiss, D.J. Accuracy of perceived test-item difficulties. JSAS Catalog of Selected Documents in Psychology, 1977, 7, 100.

Rounds, J.B., Miller, T.W., Dawis, R.V. Comparability of multiple rank order and paired comparison methods. Applied Psychological Measurement, 1978, 2, 415-422.

Siegel, S. Nonparametric Statistics for the Behavioral Sciences. New York: McGraw-Hill, 1956.

Van der Ven, A.H.G.S. Introduction to Scaling. New York: Wiley, 1980.

Willoughby, T.L. Reliability and validity of a priori estimates of item characteristics for an examination of health science information. Educational and Psychological Measurement, 1980, 40, 1141-1146.

APPENDIX A

Examples of Test Items with All Variations

Item 14

Language Difficulty

Level 1: What navigation instrument is used at sea to plot position relative to the magnetic north pole?

Level 2: Which device is used in navigation at sea to determine geographical location by alignment with the magnetic field of the earth?

Level 3: Which device consisting of metal mounted horizontally or suspended freely is used in navigation at sea to determine geographical location using as a referent alignment with the magnetic field of the earth?

Response Set Convergence

Level 1:
a. binoculars
b. compass
c. protracter
d. radar

Level 2:
a. anemometer
b. compass
c. gyroscope
d. sonar

Level 3:
a. astrolabe
b. compass
c. loran
d. sextant

Item 6

Language Difficulty

Level 1: What bird can't fly and is the largest on earth?

Level 2: Which of the following is a member of a genus of birds which are characterized by their large size (the largest on earth) and by being unable to fly?

Level 3: Any of several members of the genus Struthio which are characterized by their large size (the largest on earth) and by being flightless are referred to as a/an

Response Set Convergence

Level 1:
a. crane
b. flamingo
c. ostrich
d. swan

Level 2:
a. auk
b. crane
c. ostrich
d. swan

Level 3:
a. auk
b. emu
c. ostrich
d. penguin