DOCUMENT RESUME

ED 121 824	TM 005 255
AUTHOR	Sherman, Susan W.
TITLE	Multiple Choice Test Bias Uncovered by Use of an "I Don't Know" Alternative.
PUB DATE	[Apr 76]
NOTE	29p.; Paper presented at the Annual Meeting of the American Educational Research Association (60th, San Francisco, California, April 19-23, 1976)
EDRS PRICE	MF-$0.83 HC-$2.06 Plus Postage
DESCRIPTORS	Age Differences; *Comparative Analysis; Groups; Guessing (Tests); *Individual Differences; Mathematical Models; *Multiple Choice Tests; National Surveys; Parental Background; Racial Differences; *Response Style (Tests); Rural Urban Differences; Sex Differences; *Test Bias
IDENTIFIERS	*I Dont Know Response Option (Tests); National Assessment of Educational Progress

ABSTRACT

The multiple-choice science exercises used by the National Assessment of Educational Progress include an "I Don't Know" (IDK) alternative to estimate more accurately knowledge of groups of respondents. Group percentages of IDK responses were examined and compared with correct responses to see if the IDK introduces bias. Variance common to IDK and correct responses was reasoned to reflect bias related in part to personality variables. Residuals from a regression analysis predicting correct response from IDK response percentages were analyzed as modified correct response effects. The modification generally reduces differences among groups and smoothes trends across the four age levels. (Author)

MULTIPLE CHOICE TEST BIAS UNCOVERED BY USE OF AN "I DON'T KNOW" ALTERNATIVE[1]

Susan W. Sherman[2]

National Academy of Sciences

The National Assessment of Educational Progress (NAEP) has conducted two surveys of the educational attainments in science of nationwide samples of 9-, 13-, and 17-year-olds, and young adults ages 26 to 35. Most of the exercises, in a multiple-choice format, include an "I don't know" alternative in order to estimate more accurately knowledge of groups of respondents. This paper investigates the possibility that analysis of the "I don't know" alternative data uncovers a form of bias among groups of respondents. Groups' usage of the "I don't know" or uncertainty alternative and age-to-age comparisons are presented. Then a response model relating knowledge and personality variables to responses is explained. Finally, regression analysis is used to adjust correct response percentages for group differences in usage of the "I don't know" alternative.

"I don't know" alternatives have not been included in many widely used multiple-choice tests. A few small scale studies have investigated the effect of including this option. Knapp (1968) compared results of an open-ended mathematics examination with those from two multiple-choice examinations on the same material, one with and one without an "I don't know" alternative. Percent correct on the open-ended exercises was approximated more closely by the multiple-choice form including "I don't know," while the other form over-estimated percent correct for about 75 percent of the exercises. Inclusion

[1] Based on doctoral dissertation, "Group Differences in Responding 'I Don't Know' as an Alternative in Multiple-Choice Exercises," University of North Carolina, 1974.

[2] Formerly at the National Assessment of Educational Progress.

of "I don't know" was judged, therefore, to have reduced the amount of guessing by respondents uncertain of the correct answer. NAEP wanted to avoid using corrections for guessing because the authors of NAEP reports would not defend them. This decision is consistent with an editor's note in Educational Measurement by Robert L. Thorndike: "It has been suggested that the happiest solution to the guessing problem lies not in correcting for guessing but in preventing it." (p. 61) Even though National Assessment may have avoided bias due to respondents' guessing, this paper demonstrates another form (and probable source) of bias that remains in the results and should be recognized.

In the beginning of each NAEP administration, respondents were instructed how to answer the exercises and were shown a sample multiple-choice exercise. A tape recording was played during each administration. The following instructions concerning the uncertainty or "I don't know" alternative were read: "If you don't know the answer to an exercise, just fill in the oval next to I don't know." After each multiple-choice exercise was read to the respondents, the announcer added, "If you do not know the answer, please mark the 'I don't know' response."

Response Model

The following model is proposed as a way to look at different types of responses and knowledge:

Responses

Knowledge	Correct	Wrong	Omit	Uncertainty	
Know	(1,1)*	(1,2)	(1,3)	(1,4)	K_{ij}
Don't Know	(2,1)	(2,2)	(2,3)	(2,4)	$100-K_{ij}$
	C_{ij}	$100-C_{ij}-D_{ij}$		D_{ij}	100

*The table cells and marginals represent a classification of responses of group i to exercise j.

Because of NAEP procedures such as reading exercises aloud to respondents and having generous time allotments for every exercise, the percentages of careless errors, cell (1,2), can be assumed to be zero. The percentages of omissions (column 3) are very close to zero for most exercises and can be assumed to be zero for present purposes.

National Assessment currently reports C_{ij} as an estimate of K_{ij}, the percentage of respondents in group i who know the correct answer to exercise j. C_{ij} is a biased estimate of K_{ij} because of respondents who give the correct answer without really knowing it, cell (2,1), and because of respondents who do not give the correct answer even when they know it, cell (1,4). Similarly, D_{ij} is a biased estimate of those who know they do not know the correct answer. Those who respond "I don't know" may have personality traits such as timidity, shyness, fear or dislike of risk-taking, fear of being wrong, and lack of motivation. The "I don't know" response percentages may provide the best measure available in the NAEP data for adjusting correct response percentages in order to better reflect real differences in knowledge about science.

Sheriffs and Boomer (1954) deal with corrections for guessing in a classroom situation. They claim to have uncovered some psychological factors related to guessing. One group of students was instructed to answer all items on a test. A second group was told to circle items of which they were uncertain and then answer those items. All students were also rated on the A-Scale of the Minnesota Multiphasic Personality Inventory (MMPI). High scorers on the A-Scale are "characterized by introversion, rumination, anxiety, low self-esteem and undue concern with the impression they make on others" (Sheriffs and Boomer, p. 84). Students in the second group who had a high score on the A-Scale were more often penalized by the right-minus-wrong correction for guessing because they omitted more items. However, they scored as well as others when the number of correct responses was used as the test score.

Welsh (1968) investigated the relationship between two intelligence tests and measures of anxiety, self-confidence, and impulsiveness. The two intelligence tests used were the D-48, a non-verbal, timed test, and the Terman Concept Mastery Test (CMT), a basically verbal, untimed test. Anxiety, as measured on the A-Scale of the MMPI, was found to be negatively related to performance on the D-48 of bright high school boys (r = -0.17) and girls (r = -0.09). The correlations were more nearly zero for the CMT. This measure of anxiety was not found to be positively related to the number of items omitted by respondents, as Sheriffs and Boomer (1954) might have predicted. Further, Welsh concludes that, although anxiety is negatively related to these intelligence test scores, it is apparently an insignificant variable in accounting for individual differences in the scores. The Self-Confidence scale of Gough's Adjective Check List was found to be positively related (correlation coefficients in the teens) to the number of items attempted on both the D-48 and the CMT for both boys and girls. Students with more self-confidence tend to attempt more test items than do other students. Impulsive students, as determined by the Pd and Ma scales of the MMPI, tend to answer more items wrong and fewer correct on both tests of intelligence.

These results cannot be directly generalized to the National Assessment respondents. Welsh studied only intelligent, talented, high school students. Nevertheless, the results do indicate that personality variables are related to some measures of performance as well as to test-taking strategies. The relationships found by Welsh may be even stronger in a heterogeneous sample like that used by NAEP. One would predict on the basis of Welsh's findings that a high proportion of impulsive respondents would tend to give wrong answers, to be in the second column of the matrix. The self-confident respondents would be less likely to say "I don't know" than would others. From the research of Sheriffs and Boomer, one would expect highly anxious people to fall in cell (1,4), to know the correct answer but say "I don't know."

Group Comparisons

National Assessment presents results in terms of five factors classifying respondents at each age level. The five factors or reporting variables in 1969-70 were

Region	Northeast Southeast Central West
Sex	Male Female
Color	Black Nonblack
Parental Education	No high school Some high school Graduated from high school Post high school Unknown
Size and Type of Community	Extreme affluent suburb Extreme rural Extreme inner city Inner city fringe Suburban fringe Small places Medium sized places

The following groups were found to say "I don't know" more than the nation as a whole and to be correct less often in the 1969-70 science assessment: Southeastern adults, age 17 and adult females, black adults, extreme rural adults, and 17s and adults who did not report their parents' education or whose parents had no high school education. Generally, groups that are correct less often than others say "I don't know" more often. However, the correspondence is far from perfect, suggesting that lack of information is only one of several reasons for saying "I don't know." It could be argued that members of all of these groups tend to be more submissive and less self-confident than members of other groups. According to research by Sheriffs and Boomer (1954) and Welsh (1968), such people would be expected to say "I don't know" more than others. Hence, these findings may be viewed as supportive of the idea that personality variables affect the frequency of usage of the uncertainty response.

Age-to-Age Comparisons

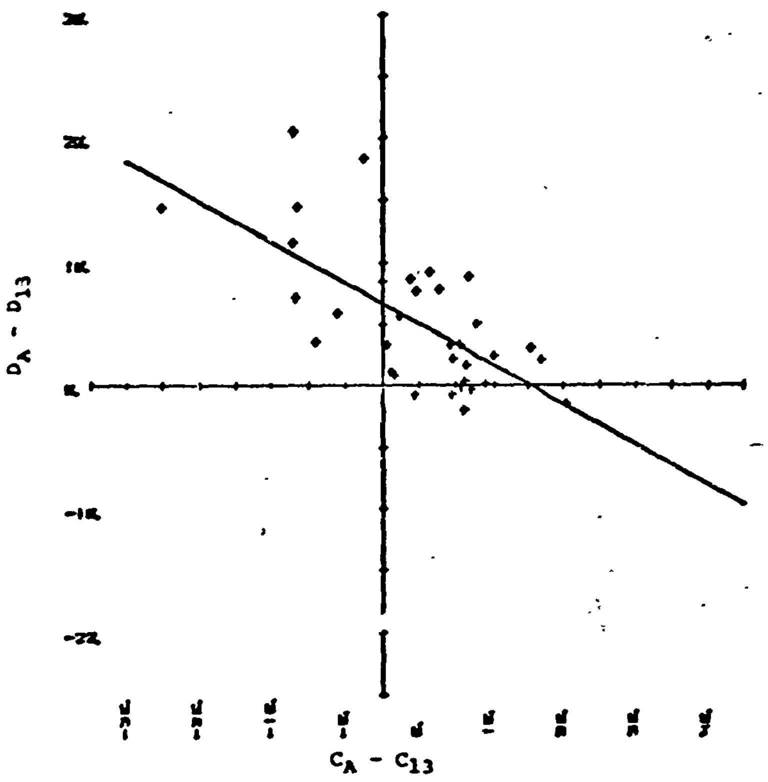
The analysis of overlap exercises (those administered at more than one age level) is important since these exercises form a common ground on which age-to-age comparisons can be made with a degree of confidence lacking when comparisons are based on different sets of exercises administered at different age levels.

The most interesting findings in this part of the study resulted from an analysis of the relationship between age-to-age differences in percentage of "I don't know" responses and differences in percentage of correct responses. Scatterplots for age 13-adult and age 17-adult comparisons are presented in Exhibit 1. On the ordinate in each scatterplot is graphed the difference between two age levels in percentage of "I don't know" responses on the overlap exercises, labeled $D_A - D_{13}$ or $D_A - D_{17}$. Comparable age-to-age differences in percentages of correct responses are plotted on the abscissa, labeled $C_A - C_{13}$ or $C_A - C_{17}$. Pearson product-moment correlation coefficients for all ages compared are presented in Exhibit 2.

The comparisons between both 13- and 17-year-olds and adults show not only moderately strong negative correlations but also fairly large y-intercepts.

Exhibit 1

Scatterplots showing relationships between age-to-age comparisons for correct and "I don't know" responses for ages 13 and adult and ages 17 and adult



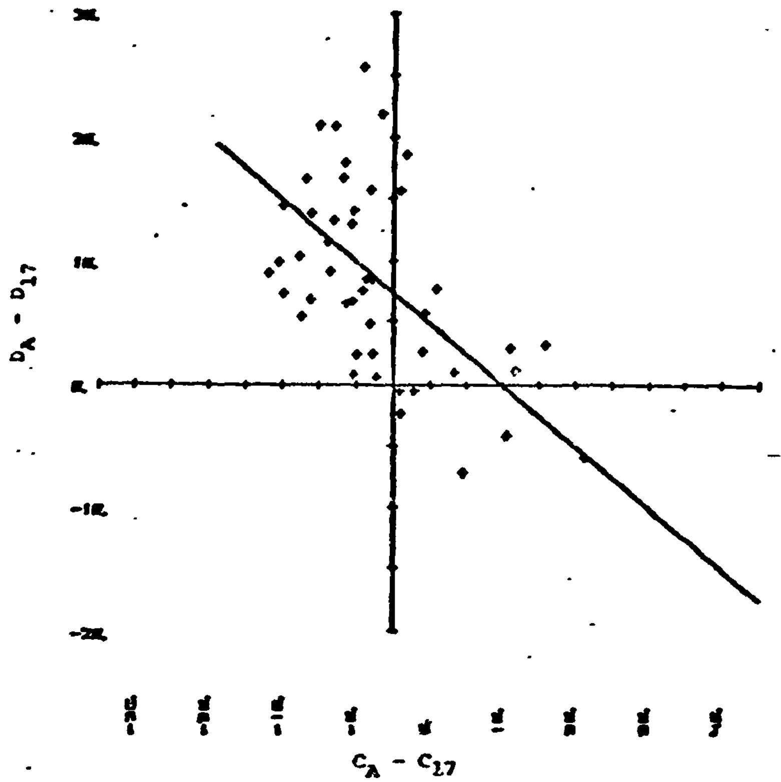


EXHIBIT 2

Ages Compared	# Exercises	Pearson r
9, 13	15	-0.44
9, 17	0	-
9, Adult	14	-0.67
13, 17	16	-0.21
13, Adult	34	-0.67
17, Adult	48	-0.61

The y-intercept corresponds to the difference in uncertainty alternative percentages expected when there is no difference in the percentage of correct responses for the two age levels. Based on the 34 overlap exercises for ages 13 and adult, it appears that adults say "I don't know" about 6.7 percent more often than 13s when they perform about the same in terms of correct responses. These 34 exercises give strong evidence that adults either are less certain that their wrong answers are correct or have a greater proclivity to admit that they do not know the answer to a science exercise than do 13-year-olds.

The scatterplot for 17-year-olds and adults is very similar to the one for the exercises common to 13-year-olds and adults. The third quadrant is completely empty, indicating that 17-year-olds never say "I don't know" more than adults on a given exercise if they respond correctly more often. The y-intercept for these 48 exercises is 7.3 percent, indicating that adults say "I don't know" about 7 percent more than 17s when they give the same percentage of correct responses. The fact that the pattern of use of the "I don't know" response is different between adults and other respondents should be considered when examining the NAEP results.

Regression Analyses

Method

Regression analyses were performed both to further investigate the relationship between correct and "I don't know" responses and to construct an alternate measure of the groups' knowledge about science. The regression model for each exercise as originally conceived was:

$$\hat{C}_i = \bar{C} + b_1 (D_i - \bar{D}), \qquad i = 1,2,\ldots,20$$

or equivalently,

$$\hat{C}_i = b_0 + b_1 D_i$$

where $b_0 = \bar{C} - b_1\bar{D}$,

$\hat{C}_i$ = predicted percentage of correct responses for group i, weighted by size of group i,

$\bar{C}$ = national percentage of correct responses for a given exercise,

$(D_i - \bar{D})$ = observed effect (group minus national percentages in "I don't know" responses) for group i, weighted by size of group i.

The regression model was fitted separately for each exercise, thereby allowing for a different relationship between correct and uncertainty responses for each exercise. C_i, the predicted correct response percentage for each group and for a given exercise, is a linear function of the national percentage of correct responses and the national and group percentage of uncertainty responses.

Not all exercises were included in the regression analyses. Those exercises with very small percentages of "I don't know" responses, less than 5%, were excluded since they can show only very small group differences in the usage of that response. Exercises with percentages of correct responses less than 20 or greater than 80 were also excluded. In very easy or very difficult exercises outside that range the relationship between correct and uncertainty responses is not terribly interesting. Further, within the range 20% to 80%, percentages are linear with arcsin $\sqrt{C}$, a transformation commonly used to normalize percentages. By the criteria $D \geq 5\%$ and $20\% \leq C \leq 80\%$, 49 exercises were selected for inclusion at age 9 years, 56 at 13, 65 at 17, and 66 at the adult level.

There is a restriction on the percentages which adversely affects the results of these regression analyses. The correct and "I don't know" percentages for any exercise must be less than or equal to 100 percent. The analyses were recalculated using another criterion variable not so dependent upon the "I don't know" percentages. The ratio of correct to all but "I don't know" percentages, "attempted-correct" percentages,

$$A_i = \frac{C_i}{100-D_i}$$

was substituted as the criterion variable. The resulting regression model is

$$\widehat{\left(\frac{C_i}{100 - D_i}\right)} = \overline{\left(\frac{C_i}{100 - D_i}\right)} + b_1(D_i - \overline{D}),$$

$$i = 1,2,\ldots,20.$$

where all of the variables are the same as those defined before. The new prediction answers the questi "Based upon the percentage of respondents who say 'I don't know' to a given exercise, what percentage of the other respondents are predicted to be correct?"

The primary reason for performing the regression analyses was to adjust or modify correct response percentages for group differences in "I don't know" response percentages. Hence, the focus of the discussion of the results of the regression analyses is upon the residuals, the differences from the regression lines. Each group i has a residual, $A_i - \hat{A}_i$, the observed percentage of attempted-correct responses minus the predicted percentage of attempted-correct responses, for each exercise. A residual for one group represents that part of the attempted-correct response percentage not predictable from the "I don't know" response percentage. The residuals are rescaled by multiplying them by (100 - D_i) in order to make them directly comparable to the correct response effects as reported by NAEP. The resulting residuals, called "modified" correct response effects, are conceptually independent of the "I don't know" responses.

How does this regression model relate to the response model presented earlier? Both correct and "I don't know" responses are given by respondents who know and those who do not know the correct answer to an exercise. Which respondents give which answers is determined in part by knowledge and in part by personality variables.

Personality factors may be most highly related to cell (2,1) (boldness) and cell (1,4) (timidity) of the response model. Knowledge factors probably fall mostly in cells (1,1) and (2,4). If cells (1,1) and/or (2,4) were nearly empty, then predicting A_i from D_i would adjust for mainly personality and response style variability. If, on the other hand, cells (2,1) and (1,4) were empty, then knowledge would be properly reflected by C_i and lack of knowledge by D_i. Any changes made by the regression analysis modification in this case would be overadjustments. Because one cannot actually distinguish between cells (1,1) and (2,1) and between cells (1,4) and (2,4), one cannot know for certain when one or two cells are nearly empty. However, there is a far from perfect negative correlation between C_i and D_i, indicating that there are group differences in the relative usage of the uncertainty response unrelated to group differences in science performance. These differences are proposed to be related to psychological variables. Therefore, one can posit that cells (2,1) and (1,4) are not both empty and that the regression analysis is adjusting for some personality variability. Further, since cells (1,1) and

(2,4) can be assumed to be nonempty, the regression analysis is removing some knowledge variability from the residuals. Without extensive further research it is impossible to know what proportion of the variability removed by the adjustment is associated with knowledge and what with personality and response style.

Results

Regional differences in modified percentages were larger than unmodified NAEP percentages at ages 9 and 13 and smaller at ages 17 and adult. This suggests that there may be larger true differences between younger respondents than represented in NAEP balanced data. The regional differences at ages 17 and adult as reported by National Assessment may be inflated by response styles in usage of the uncertainty alternative.

The modification of correct response percentages has a large impact on sex differences in science performance. Sex differences at the three younger ages are reduced by the regression analysis modification. They are virtually eliminated at the adult level over the 66 exercises analyzed. Sex differences in correct response percentages for many of the exercises at the adult level can be explained almost completely by differences in usage of the "I don't know" alternative. Some exercises continue to show a clear advantage for one sex or the other after the data modification, but there are fewer showing the overwhelming male advantage as depicted in NAEP data.

One can conclude that much of the sex difference in science performance as reported by National Assessment is an artifact of sex differences in test-taking behavior. It would be interesting to perform similar analyses on NAEP data from other subject areas, say reading or literature, where females commonly outperform males.

The deficits in science performance for black respondents at the three upper age levels are substantially reduced but not eliminated by the modification. There are color differences in percentages of correct responses which cannot be predicted from "I don't know" response data. This suggests that there are differences in knowledge over and above differences in response style which distinguish blacks from nonblacks.

The regression analysis modification generally reduces differences for the three extreme size and type of community groups. The effects for rural 13-year-olds and young adults are greatly reduced. Rural residents' deficit in science knowledge relative to other groups is apparently not as severe as is suggested by their correct response percentages as reported by NAEP. The effects for inner city respondents do not change as much with the modification as do those for rural residents. The advantage of affluent suburban respondents is reduced considerably by the modification. The affluent suburban respondents still show performance above the national level after adjustment for differences in response style.

Differences between the parental education groups are also reduced by the modification. Those who did not report parental education and those whose parents had no high school training showed markedly reduced deficits. Changes for 17-year-olds and adults were the largest. The some high school group showed slightly reduced deficits. Those whose parents had post high school training had reduced advantages. It appears that differences between parental education groups in percent of correct responses may be inflated by differences in usage of the uncertainty response.

The fact that a large majority of the group effects is reduced by the modification is significant. The technique itself does not ensure that all effects will be reduced. That many were reduced indicates that there is variability common to the two types of responses for groups and that this variability emphasizes group differences. The reduction and smoothing of group differences is an intuitively appealing outcome. The modifications seem to be reasonable: They smooth out group trends across the four age levels; they provide sensible ordering of groups within variables. In the present report the results themselves provide the strongest justification for the analysis.

The observation that most group differences are reduced by the regression analysis modification does not imply that the "I don't know" alternative should not be used in multiple-choice exercises. If the alternative is not offered, group differences in the attribute of interest could be affected in other ways. Respondents who would otherwise say "I don't know" could guess blindly or omit exercises, even if instructed to leave none blank.

The results of the modification do suggest that percentages of correct responses are impure or biased measures of knowledge. Correct response percentages could well be adjusted by some technique such as that used here to better reflect knowledge differences between groups. The regression analysis modification used in this report is only one of many possible adjustment procedures.

Summary

This paper has presented research investigating group differences in knowledge and factors that influence the measurement of that knowledge. It is believed that measures of knowledge and skills should not contain built-in biases for or against groups of respondents. A form of bias in current National Asses.ment data has been identified. The bias results from groups' differing usage of the "I don't know" alternative, included by NAEP in every multiple-choice exercise. One method, designed to remove the effect of this bias from the existing data, has been developed and used in this paper.

Psychologists, educators, and others who use multiple-choice tests should carefully weigh all available evidence to decide whether or not to include an "I don't know" alternative and whether or not to adjust the resulting data for response style differences. The importance of psychological and background variables, normally assumed to be irrelevant to many attributes measured, should not be overlooked or underemphasized in this context. For the advancement of science and especially in times such as these when the social trend is toward equality and fairness, this source of response bias should be further explored.

REFERENCES

Knapp, T. R. The mathematics study. Unpublished report. Exploratory Committee on Assessing the Progress of Education, 1968.

Sheriffs, A. C. & Boomer, D. S. Who is penalized by the penalty for guessing? Journal of Educational Psychology, 1954, 45, 81-90.

Sherman, S. W. Group Differences in Responding 'I Don't Know' as an Alternative in Multiple-Choice Exercises. (Doctoral dissertation, University of North Carolina) Chapel Hill, North Carolina, 1974.

Thorndike, R. L. Educational Measurement. Washington, D. C.: American Council on Education, 1971.

Welsh, G. S. A study of two intelligence tests (D-48 and Terman Concept Mastery Test) and their relationship to measures of anxiety, impulsiveness and verbal interests in gifted adolescents. Unpublished report. University of North Carolina, Chapel Hill, North Carolina, 1968.