DOCUMENT RESUME

ED 124 589 TM 005 347

AUTHOR Burton, Nancy W.; And Others
TITLE The Effect of Position and Format on the Difficulty of Assessment Exercises.
PUB DATE [Apr 76]
NOTE 14p.; 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-$1.67 Plus Postage.
DESCRIPTORS Age Differences; *Complexity Level; Educational Assessment; Guessing (Tests); *Multiple Choice Tests; National Surveys; *Response Mode; *Response Style (Tests); Statistical Analysis; Test Bias; Testing Problems
IDENTIFIERS *I Dont Know Response Option (Tests); National Assessment of Educational Progress; Test Format

ABSTRACT

Assessment exercises (items) in three different formats--multiple-choice with an "I don't know" (IDK) option, multiple-choice without the IDK, and open-ended--were placed at the beginning, middle and end of 45-minute assessment packages (instruments). A balanced incomplete blocks analysis of variance was computed to determine the biasing effects of position or format on the national percent correct. Format was found to create a bias, but position did not, except for 9-year-old respondents. (Author)

THE EFFECT OF POSITION AND FORMAT ON THE DIFFICULTY OF ASSESSMENT EXERCISES

by

Nancy W. Burton, Robert C. Larson and Alex M. Pearson

National Assessment of Educational Progress

Paper presented at annual convention of
American Educational Research Association,
San Francisco, April 1976.

The Effect of Position and Format on the Difficulty of Assessment Exercises

Nancy W. Burton, Robert C. Larson and Alex M. Pearson

National Assessment of Educational Progress

Perspective.

The National Assessment of Educational Progress has the charge of gathering and reporting educational achievement data that are

- an accurate representation of absolute performance now: e.g., 82% of the nation's nine-year-olds can multiply 3 x 0 (NAEP, January 1975)
- a precise representation of performance now relative to performance three to seven years ago: e.g., in 1973, 47% of 17-year-olds knew the purpose of an electrical transformer. This is a 13% decline since 1969. (NAEP, May 1975)

The baseline measure of absolute performance must be reliable, though not necessarily in the accepted psychometric sense. It may be better to say that the measures must be "accurate." A measure may be reliable without being accurate: a scale that consistently adds ten pounds to one's weight may be perfectly reliable.

For the relative change measures, reliable biases are unimportant, since the difference between two biased measures is the same as the difference between two unbiased measures, if the bias is simply a constant that cancels out. However, the bias in a measure may also change over time. Suppose, for example, one saw an increase of 2%[1] in 13-year-olds' performance on a certain reading task. Suppose, however, that there was also a 6% decline in the non-response rate: suppose, that is, that 6% more respondents were guessing rather than leaving the item blank. Even if children could guess no better than chance, one would expect 1.5% more children, on a four-alternative item, to get the correct answer because of that change in response patterns. This would change a statistically stable 2% improvement to a non-significant .5% improvement.

National Assessment has always used an "I don't know" foil on all cognitive multiple-choice items to discourage respondents from guessing. Guessing not only inflates the estimation at one point in time of the percent of respondents who can do the task, but also, a change in guessing behavior (as illustrated above) can affect the interpretation of a change in percent of success over time. Unfortunately, the pattern of response to an "I don't know" (IDK) foil

[1] A 2% increase, since there are 3.6 million 13-year-olds in the nation, means that 72,000 more children can perform the reading task.

can also differ for different groups. Sherman (1974) found that some southeasterners, females, blacks, and rural persons use the IDK poorly. Thus the "I don't know" can contribute to bias in a measure at one point in time or change measures over time as much as a differential tendency to omit items.

Other potential sources of bias over time include changes in procedural matters, such as training or experience of test administrators, or school cooperation, or type of print used in packages (test booklets), or the voice which reads the exercises on tape. One of the most serious potential sources of bias arises because National Assessment releases some exercises for publication and does not reuse them. The remaining unreleased exercises are then repackaged and reassessed for change. Since they have been repackaged, they are presented for the second time in different orders, different contexts, and different positions. Any of these variables may affect performance and thus either mask or exaggerate the change in performance over time.

Thus, National Assessment measures of change as well as baseline performance must be extremely accurate. From the first, National Assessment has devoted great resources to precise sampling design. In the last few years, it has also begun to devote resources to locating sources of non-sampling error. Many of these non-sampling errors have been dismissed as unimportant on conventional tests. Conventional tests are collections of items, the sum of which is taken to measure a rather globally-defined trait--such as "intelligence" or "arithmetic achievement"--and then only relative to some norm group. Inaccuracies due to individual items and the examinees' response to them can, to some extent, be supposed to average out in the total score. Because of NAEP's item-by-item reporting, these errors once again become important.

Objectives.

The purpose of the present study was to determine the effect of two sources of non-sampling error: position in package (beginning, middle or end of the assessment instrument) and exercise format (multiple-choice with an "I don't know" alternative, multiple choice without IDK, and open-ended). Obviously, position in package is a source of error that one cannot eliminate, since some exercise or other always must be first, middle or last in a package. It is a source of error that can be held constant over time, however, if it is found to be important. Further study of the IDK foil may show that it should be dropped (in future item development: it cannot be dropped from change exercises even though bias is strongly suspected); replaced with corrections for guessing; retained; retained but supplemented with corrections for guessing. The present study can provide some data to answer these questions; however, it must be emphasized that the present data were all collected at one time and so questions about change analyses cannot be fully answered.

Methods and Data Source.

The present study was included in the 1973-74 assessment of Writing and Career and Occupational Development. It is, therefore, based on national probability samples of 2,500 9-year-olds, 13-year-olds or 17-year-olds for each item. At each age, nine different packages were involved, and thus nine different samples of 2,500 respondents. Each package was a block in the 3^3 balanced incomplete blocks design used at each age. The three factors in the design were

- exercise content - three different science questions were developed, such that exactly the same stem was used for both multiple-choice and open-ended formats;
- format - multiple choice with IDK, multiple without IDK, and open-ended;
- position in package - beginning, middle, end.

Each of the nine packages contained three exercises which represented each content, each format and each position. For example, package #1 at age 9 contained

Beginning	Middle	End
Exercise about blood circulation, Multiple-choice without IDK	Exercise about largest living animal, Open-ended	Exercise about lightning and thunder, Multiple-choice with IDK

See attachments 1, 2, and 3 for the wording of the exercises in the multiple-choice and IDK format. Attachments 1, 2, and 3 also give the national percents for each foil (including IDK and no response) for each exercise, format and position.

Results.

The design was set up so that the analysis of variance estimates for the main effects were unconfounded with blocks, but all interactions were partially confounded.[2] To get some independent estimate of these block effects, a marker exercise was placed at the end of each of the nine packages. This marker exercise allowed an empirical estimate of the sampling variability; it also contained variation due to the accumulated effect of differing contexts of presentation, since the nine packages all contained different Writing and COD exercises. This marker exercise contained five parts (five different questions about reading a map). The variance component due to parts within blocks--that is, the natural variation in the difficulty of the five questions--was at least 50 times greater than the component due to the block effect. Thus the block effect, though

[2] Components of the interactions were calculated by the modular arithmetic method described in Winer (1971, p. 606ff).

in some cases statistically significant because of large sample sizes, was very small compared to the normal variation among exercises. There is a second reason for disregarding possible block effects. Inspection of the analysis of variance tables (attachments 4, 5 and 6) shows that the mean squares for confounded interaction parts were about the same size as the mean squares for the unconfounded interaction parts. Both of these pieces of evidence indicate that the main within blocks analysis can be interpreted straightforwardly.

At all ages, there was a large main effect for exercise contents--which is simply to say that some questions were harder than others. There was also a main effect for format. Only at age 9 was there a significant position effect. It did not appear to be a fatigue effect, which might be expected with these young children, but rather a disadvantage in performance to the beginning-of-the-package exercises. It should be noted that these beginning exercises were never <u>first</u> in package, but simply occurred within the first five minutes of testing. Again at all ages there were content by format interactions, which can basically be interpreted as proving that some tasks are more difficult than others in the open-ended format.

The significant format effect deserves further discussion. Exhibit 1 displays the mean percent correct (averaged over the three positions) for each exercise in each format at each age.

Exhibit 1. Means and Standard Deviations (in parentheses) for Three Different Formats of Exercises

		Multiple Choice		Open-Ended
		- IDK	+ IDK	
	Exercise #			
	1	76.9 (1.80)	68.9 (0.76)	52.6 (0.67)
Age 9	2	27.5 (2.07)	26.4 (4.74)	5.5 (0.75)
	3	65.2 (1.61)	62.3 (1.37)	22.7 (2.15)
	average*	56.51 (1.83)*	52.52 (2.88)*	26.91 (1.87)*
	1	47.4 (1.29)	41.0 (1.13)	5.8 (0.61)
Age 13	2	59.1 (1.55)	58.6 (1.53)	32.9 (3.50)
	3	23.8 (0.53)	22.9 (1.66)	10.7 (0.65)
	average*	43.42 (1.20)*	40.86 (1.45)*	16.44 (2.08)*
	1	53.5 (0.67)	44.8 (1.15)	13.5 (1.00)
Age 17	2	72.7 (2.73)	71.3 (0.49)	48.9 (1.25)
	3	17.7 (2.22)	14.6 (1.06)	11.0 (0.74)
	average*	47.96 (2.06)*	43.57 (.94)*	24.49 (1.01)*
	overall average*	49.30 (1.74)*	45.65 (1.94)*	22.61 (1.55)*

*Standard deviations in the "average" column are the square root of pooled within-cell variances.

The overall average (the last line in the table) shows a large difference between open-ended and multiple-choice formats and a smaller--but still statistically significant[3]--difference between the two multiple-choice formats. Having the "I don't know" foil does reduce the overall percent correct. Having the "IDK" may slightly increase the variance, but this experiment was not sensitive enough to detect it.[4]

Importance of the Study.

This is one of a series of studies to locate sources of non-sampling errors in the estimates of performance on assessment tasks. The goal is to increase the accuracy of baseline and change estimates. Considering that, in the first assessment of change in Science, the average decline in 9-year-olds' performance was 1.8%; in 13s', 1.9%; and in 17s', 2.3% (NAEP, February 1975), it is obvious that great precision is required to detect changes.

This study has resulted in several further investigations. Because of the stability of the small block (package) effect, NAEP staff is now looking at methods of adjusting sampling weights to make packages more comparable. Because of some inconsistency in performance of IDK vs no-IDK multiple-choice exercises, staff is continuing to examine the use of item-scoring formulas (versions of correction-for-guessing techniques) as an alternative to the "I don't know" foil. These investigations will hopefully lead to new techniques for increasing the accuracy of assessment results.

[3]

$$t_{36} = \frac{49.3 - 45.65}{\sqrt{\frac{1.74}{18} + \frac{1.94}{18}}} = 8.11$$

[4]

$$\frac{(1.94)^2}{(1.74)^2} = 1.24; \quad F_{18,\ 20\,(\alpha = .75)} = 1.38$$

REFERENCES

NAEP Reports.

Math Fundamentals, Selected Results from the First National Assessment of Mathematics, Report 04-MA-01, 1972-73 assessment. Washington, D.C.: Government Printing Office, January, 1975.

National Assessments of Science, 1969 and 1973, A Capsule Description of Changes in Science Achievement, Report 04-S-00. Washington, D.C.: Government Printing Office, February 1975.

Selected Results from the National Assessments of Science: Energy Questions, Report 04-S-01. Washington, D.C.: Government Printing Office, May, 1975.

Sherman, Susan W. Group Differences in Responding "I don't know" as an Alternative in Multiple-Choice Exercises. Unpublished doctoral dissertation. Chapel Hill, N.C.: 1974.

Winer, B. J. Statistical Principles in Experimental Design, second edition. New York: McGraw-Hill, 1971.

Attachment 1: Exercise Text and Results - Age 9

	Multiple Choice without "IDK"				Multiple Choice with "IDK"					Open Ended			
	Beginning	Middle	End	Average	Beginning	Middle	End	Average		Beginning	Middle	End	Average
The one part of the body which makes blood circulate is the									No Response	3.1	4.0	4.3	3.8
(No Response)	2.1	1.1	1.3	1.5	0.8	0.6	0.3	0.6					
● heart.	75.4	78.9	76.4	76.9	69.6	69.1	68.1	68.9	Acceptable	52.4	53.3	52.0	52.6
○ brain.	5.0	4.0	3.4	4.1	3.3	3.0	4.8	3.7					
○ hippocampus.	5.3	4.7	6.9	5.6	3.9	3.6	4.7	4.1	Unacceptable	22.4	24.9	27.1	24.8
○ lungs.	12.1	11.4	12.0	11.8	8.4	9.6	9.3	9.1					
○ I don't know.					13.9	14.2	12.7	13.6	IDK	22.0	17.7	16.6	18.8
What is the largest animal now living?									No Response	1.9	2.4	4.7	3.0
(No Response)	0.7	0.6	1.2	0.8	0.9	0.9	0.7	0.8					
○ The elephant	20.8	17.8	19.1	19.2	18.6	18.7	16.8	18.0	Acceptable	20.6	22.6	24.9	22.7
○ The giraffe	8.8	9.1	9.7	9.2	10.9	10.0	10.3	10.4					
○ The dinosaur	5.7	5.5	5.4	5.5	6.1	5.3	6.4	5.9	Unacceptable	71.1	69.0	64.3	68.1
● The whale	64.0	67.0	64.5	65.2	60.7	63.2	62.9	62.3					
○ I don't know.					2.9	2.0	2.9	2.6	IDK	6.5	6.0	6.0	6.2
Why do you usually see the lightning before you hear the thunder?													
(No Response)	2.8	1.5	3.7	2.7	1.1	1.1	0.7	1.0	No Response	3.8	4.0	4.3	4.0
○ Eyes respond faster than ears.	7.9	6.0	5.8	6.6	6.2	6.5	5.7	6.1					
○ Sound travels faster than light.	7.2	6.7	6.8	6.9	8.4	5.7	5.3	6.5	Acceptable	6.2	4.7	5.5	5.5
● Light travels faster than sound.	25.1	28.4	28.9	27.5	21.1	27.7	30.3	26.4					
○ Lightning causes thunder.	48.1	48.0	45.6	47.2	34.8	39.4	42.7	39.0	Unacceptable	39.9	51.2	50.1	47.1
○ Lightning and thunder have nothing to do with each other.	8.8	9.4	9.1	9.1	7.8	7.1	5.5	6.8					
○ I don't know.					20.6	12.4	10.1	14.4	IDK	50.1	40.1	40.2	43.5

Attachment 2. Exercise Text and Results - Age 13

	Multiple Choice without "IDK"				Multiple Choice with "IDK"					Open Ended			
	Beginning	Middle	End	Average	Beginning	Middle	End	Average		Beginning	Middle	End	Average
Why doesn't the filament in a light bulb burn up when it gets hot?									No Response	1.2	5.7	1.5	2.8
(No Response)	0.5	1.2	1.1	0.9	0.1	0.2	0.4	0.2	Acceptable	5.1	5.9	6.3	5.8
○ The filament material never burns.	6.6	5.2	7.9	6.6	6.8	5.1	5.7	5.9					
○ The filament material doesn't get hot enough to burn.	5.7	7.8	6.1	6.5	7.7	5.0	5.5	6.1	Unacceptable	51.1	60.1	68.5	59.9
● There is no oxygen in the bulb.	47.0	46.3	48.8	47.4	40.3	40.4	42.3	41.0					
○ The argon in the bulb prevents burning.	40.3	39.5	36.1	38.6	26.4	23.1	34.7	28.1	IDK	42.6	28.3	23.6	31.5
○ I don't know.					18.8	26.2	11.4	18.8					
What is the name given to all rocks formed by heating and melting?									No Response	5.4	5.0	5.3	5.2
(No Response)	0.5	0.3	0.5	0.4	0.7	0.0	0.3	0.3	Acceptable	10.7	10.0	11.3	10.7
○ Lava	62.4	58.8	60.4	60.5	59.1	54.6	49.0	54.2					
○ Sandstone	1.9	3.3	2.5	2.6	1.9	2.1	1.5	1.8	Unacceptable	45.8	50.8	51.1	49.2
● Igneous	24.4	23.4	23.6	23.8	21.2	24.5	23.1	22.9					
○ Metamorphic	10.9	14.2	13.0	12.7	11.3	11.8	14.4	12.5	IDK	38.1	34.2	32.3	34.9
○ I don't know.					5.8	6.9	11.7	8.1					
Why do you usually see the lightning before you hear the thunder?									No Response	0.7	1.5	1.3	1.2
(No Response)	0.2	1.0	0.7	0.6	0.1	0.1	0.0	0.1					
○ Eyes respond faster than ears.	2.9	3.1	1.9	2.6	3.0	2.0	3.2	2.7	Acceptable	29.8	32.2	36.7	32.9
○ Sound travels faster than light.	4.0	3.0	2.7	3.2	3.2	2.1	2.7	2.7					
● Light travels faster than sound.	60.6	57.5	59.2	59.1	58.3	60.3	57.3	58.6	Unacceptable	43.8	43.2	46.8	44.6
○ Lightning causes thunder.	26.5	29.0	29.0	28.2	26.6	27.3	26.3	26.7					
○ Lightning and thunder have nothing to do with each other.	5.9	6.5	6.5	6.3	5.1	5.0	5.7	5.3	IDK	25.6	23.1	15.2	21.3
○ I don't know.					3.8	3.1	4.6	3.8					

Attachment 3. Exercise Text and Results - Age 17

	Multiple Choice without "IDK"				Multiple Choice with "IDK"					Open Ended			
	Beginning	Middle	End	Average	Beginning	Middle	End	Average		Beginning	Middle	End	Average
Why doesn't the filament in a light bulb burn up when it gets hot?									No Response	1.5	2.2	3.6	2.4
(No Response)	1.8	2.1	2.0	2.0	0.4	0.4	0.4	0.4					
○ The filament material never burns.	7.7	8.5	9.6	8.6	7.4	5.2	6.0	6.2	Acceptable	14.3	12.4	13.9	13.5
○ The filament material doesn't get hot enough to burn.	10.8	9.5	9.4	9.9	7.0	6.8	8.1	7.3					
● There is no oxygen in the bulb.	52.9	54.2	53.3	53.5	43.7	44.7	46.0	44.8	Unacceptable	52.2	60.7	56.0	56.3
○ The argon in the bulb prevents burning.	26.7	25.8	25.7	26.1	19.8	14.0	19.0	17.6					
○ I don't know.					21.7	29.0	20.4	23.7	IDK	32.1	24.7	26.6	27.8
Theoretically, how far does the earth's gravity extend?									No Response	3.7	2.1	4.1	3.3
(No Response)	0.8	0.6	1.4	0.9	0.4	0.3	0.6	0.4					
○ To a point between the earth and the moon	54.4	55.2	46.7	52.1	49.4	47.6	45.9	47.6	Acceptable	11.3	11.6	10.2	11.0
○ To a point between the earth and the sun	17.1	16.7	18.6	17.5	16.0	16.6	15.1	15.9					
○ Throughout the solar system	11.2	11.1	13.1	11.8	10.1	10.3	11.4	10.6	Unacceptable	46.1	52.2	47.8	48.7
● Throughout the universe	16.5	16.4	20.3	17.7	13.6	14.4	15.7	14.6					
○ I don't know.					10.5	10.7	11.3	10.8	IDK	38.9	34.0	37.7	36.9
Why do you usually see the lightning before you hear the thunder?													
(No Response)	0.2	0.6	0.6	0.5	0.2	0.1	0.4	0.2	No Response	1.1	1.0	2.0	1.4
○ Eyes respond faster than ears.	1.6	2.2	2.3	2.0	1.2	1.8	1.8	1.6					
○ Sound travels faster than light.	2.5	1.4	1.7	1.9	2.0	1.3	1.8	1.7	Acceptable	50.2	48.8	47.7	48.9
● Light travels faster than sound.	75.7	71.9	70.4	72.7	71.1	71.0	71.9	71.3					
○ Lightning causes thunder.	15.4	19.1	17.6	17.4	18.0	15.7	15.8	16.5	Unacceptable	34.4	37.3	34.9	35.5
○ Lightning and thunder have nothing to do with each other.	4.7	4.7	7.4	5.6	4.3	6.4	4.9	5.2					
○ I don't know.					3.3	3.7	3.3	3.4	IDK	14.2	13.0	15.6	14.3

Attachment 4. Analysis of Variance Table for Age 9 Design

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square	F
Within Blocks	18			
Exercise: E	2	9976.4	4988.2	1609**
Format: F	2	4644.0	2322.0	749**
Position: P	2	27.1	13.6	4
(EF^2)	(2)	(153.1)		
E x F	2	153.1	76.5	25*
(EP^2)	(2)	(6.3)		
E x P	2	6.3	3.1	1
(EP)	(2)	(0.4)		
F x P	2	0.4	.2	> 1
(EFP)	(2)	(5.5)		
(EFP^2)	(2)	(0.5)		
(EF^2P)	(2)	(12.8)		
E x F x P	6	18.8	3.1	
Between Blocks	(8)			
(EF)	(2)	(425.8)		
(EP)	(2)	(8.9)		
(FP^2)	(2)	(9.5)		
(EF^2P^2)	(2)	(10.2)		

*$\alpha < .05$
**$\alpha < .01$

Attachment 5. Analysis of Variance Table for Age 13 Design

Source of Variation		Degrees of Freedom		Sum of Squares		Mean Square	F
<u>Within Blocks</u>			<u>18</u>				
Exercise: E			2		4411.3	2205.7	689**
Format: F			2		3990.9	1995.4	624**
Position: P			2		7.4	3.7	1.2
	(EF^{2})	(2)		(430.3)			
E x F			2		430.3	215.2	67*
	(EP^{2})	(2)		(.721)			
E x P			2		.721	.36	> 1
	(FP)	(2)		(.134)			
F x P			2		.134	.07	> 1
	(EFP)	(2)		(.867)			
	(EFP^{2})	(2)		(15.836)			
	(EF^{2}P)	(2)		(2.614)			
E x F x P			6		19.3	3.2	
<u>Between Blocks</u>		<u>(8)</u>					
	(EF)	(2)		(263.8)			
	(EP)	(2)		(1.4)			
	(FP^{2})	(2)		(15.4)			
	(EF^{2}P^{2})	(2)		(3.2)			

*α < .05
**α < .01

Attachment 6. Analysis of Variance Table for Age 17 Design

Source of Variation		Degrees of Freedom		Sum of Squares		Mean Square	F
Within Blocks			18				
Exercise: E			2		11211.7	5605.8	2156**
Format: F			2		2801.7	1400.9	539**
Position: P			2		1.2	.6	
	(EF^2)	(2)		(646.3)			
E x F			2		646.3	323.1	124**
	(EP^2)	(2)		(5.5)			
E x P			2		5.5	2.8	1.1
	(FP)	(2)		(4.9)			
F x P			2		4.9	2.4	> 1
	(EFP)	(2)		(7.1)			
	(EFP^2)	(2)		(1.6)			
	(EF^{2}P)	(2)		(6.8)			
E x F x P			6		15.4	2.6	
Between Blocks		(8)					
	(EF)	(2)		(336.7)			
	(EP)	(2)		(7.5)			
	(FP^2)	(2)		(2.7)			
	(EF^{2}P^2)	(2)		(0.2)			

**α < .01