DOCUMENT RESUME

ED 110 494 TM 004 767

AUTHOR Phillips, Donald L.; And Others
TITLE Stability of Nominal Categories Over Readers, Over Time.
INSTITUTION Education Commission of the States, Denver, Colo. National Assessment of Educational Progress.
PUB DATE [Mar 75]
NOTE 19p.; Paper presented at the Annual Meeting of the American Educational Research Association (Washington, D.C., March 30-April 3, 1975)

EDRS PRICE MF-$0.76 HC-$1.58 PLUS POSTAGE
DESCRIPTORS Analysis of Variance; Elementary Secondary Education; *Essay Tests; *Reliability; *Scoring; Test Bias; *Testing Problems; Test Results; Time

ABSTRACT

The consistency across time and readers of the scoring of National Assessment of Educational Progress (NAEP) open-ended exercises was examined. The procedure studied is a nominal categorical scoring. Ten readers independently read 28 sample responses to each of 12 open-ended exercises at three different times. All ten readers agreed on their assignment on about 75 percent of the sample responses. About 89 percent of the time a reader agreed on the category assignment from one reading to another. (Author)

STABILITY OF NOMINAL CATEGORIES OVER READERS, OVER TIME

Donald L. Phillips

Nancy W. Burton

and

Alex M. Pearson

March, 1975

National Assessment of Educational Progress
The Education Commission of the States

Paper Presented at the American Educational
Research Association 1975 Annual Meeting.
Session no. 6.09

Stability of Nominal Categories over Readers, over Time

Donald L. Phillips

Nancy W. Burton

and

Alex M. Pearson

INTRODUCTION

National Assessment of Educational Progress (NAEP) is a census-like assessment project which collects data on a national probability sample. NAEP has collected data in ten different curriculum areas from four different age classes (9-year-olds, 13-year-olds, 17-year-olds, and young adults 26-35 years old). The major purpose of NAEP is to measure changes across time in performance on objectives-referenced exercises. Many of the exercises NAEP uses in its assessment process are open-ended and must be hand scored. NAEP's hand scoring does not generally consist of assigning responses to points on an ordinal scale. Instead responses are almost always assigned to nominal (descriptive) categories. These nominal categories are, however, classifiable as acceptable or unacceptable.

Because NAEP's objective is to measure changes over time in performance, it is important that hand scoring not depend heavily on the scorer or the time of the scoring. It is known that when essays are scored for quality on ordinal scales, the scores vary with the context in which the papers are read (Coffman, 1971). If these findings are true also for NAEP's data, measurement of change would require that all responses from all points in time be read in the same context and time. If NAEP can show that its semi-professional scoring is consistent across time and scorers, then perhaps change can be measured on these exercises without all responses being re-read each time a change measure is made.

METHODS

The study was designed to answer several questions:

1. To what extent does the score a response receives depend upon the scorer who scores it?

Thanks are due to Janet Bailey for her careful computations.

2. To what extent does the score depend on the time (within the two-three month scoring session) when the response is scored?

For this study sample responses were selected from the actual response data from the Writing and Career and Occupational Development (COD) assessments done in 1973-1974. Three exercises from COD and two exercises from Writing were selected at ages 9, 13, and 17. Three exercises from COD were selected for adults. (See Table 1.) For each exercise one sample response was selected arbitrarily from each of 28 administration units spread throughout the country. Each sample response was assigned a number (1-28), and photo copies of the samples were made.

Table 1.

Age	Content Area	Exercise Number	(NAEP Number)	Number of Parts Analyzed
9	Writing	1	(0-201002)	5
		2	(0-201012)	1
	COD	3	(2-301034)	2
		4	(2-302015)	6
		5	(2-402002)	3
13	Writing	6	(0-201018)	3
		7	(0-202007)	1
	COD	8	(2-102025)	3
		9	(2-302015)	6
		10	(2-306012)	3
17	Writing	11	(0-201018)	3
		12	(0-301008)	4
	COD	13	(2-102025)	3
		14	(2-306006)	5
		15	(2-306012)	3
Adult	COD	16	(2-302005)	10
		17	(2-306009)	12
		18	(2-306012)	3

The readers were members of the professional scoring department at Measurement Research Center in Iowa City, IA. All of these scorers had at least bachelors degrees; most had teaching experience. The experimental papers were scored as a part of the normal COD and Writing scoring, which involved approximately 100,000 student responses per age. All scorers were trained together on the use of up to 40 different scoring guides before the scoring for each age began. The scoring guides consist of a descriptive title for each category, illustrated by up to twenty sample responses.

After scoring normal assessment responses for two to three weeks, each scorer was given in random order sets of photocopies of each sample response. Each scorer independently read each response and recorded the score on a separate sheet. The sets of sample responses were then collected and given new random orders. The sets were presented again to scorers for rescoring when about one half of all of the data for an age class had been scored and again immediately after the scoring for the age class was completed. The scoring for each age took from two to three months to complete.

ANALYSES

Introduction

The major analysis problem was that most exercises had nominal scoring categories. Many conventional summary statistics, however, require ordinal data. We finally defined several percent-of-agreement summaries that made sense to us and were based on the raw data.

The first answers the question:

(1) What is the probability that all of the scorers or all scorers but one will agree on a random paper?

A second type of percent of agreement focuses not on the agreement among scorers, but on the agreement within scorers over the three scoring times. It answers the question:

(2) What is the probability that a random scorer will assign the same category at least twice, or all three times?*

*For age 9 the sample responses were only scored twice, since the entire scoring session took only six weeks.

Further analysis required transformation of the data to an ordinal scale. Since the scorers are believed to be competent judges, the score assigned by most scorers to a given response was assumed to be the true score for that response. The most common score variable was defined as presence or absence of this true score. (The most common score variable is denoted by MCS.)

For the MCS (most common score) variable, one further percentage was computed. It was a more general percent of agreement than the percent of agreement on responses (1) or the percent of agreement over time (2), defined above. This overall percent of agreement answers the question:

(3) What is the probability that a random scorer, on a random response, at any one of three times, will assign the true category?

All three percents of agreement are discussed below and presented in Attachments 1-4.

The MCS variable was also used to compute a repeated-measures analysis of variance, with responses and scorers as random factors, and time and (where applicable) exercise part* as fixed factors. These Responses x Scorers x Times x Parts analyses were meant to answer the questions:

Do different scorers vary in their ability to assign true scores?

Does time affect scorers' ability to assign true scores?

Does the exercise part* affect that ability?

Does the specific response affect that ability?

A second ordinal variable was created by collapsing the data into acceptable and unacceptable categories. The A/U (acceptable/unacceptable) variable was also used in the Respondents x Scorers x Times x Parts analysis of variance

---

*Exercises have parts for several reasons. Parts may be two aspects of the same task (such as scores for spelling and punctuation) or they may be two attempts at the same task (as when respondents are asked to give two reasons for something).

design, but the questions to be answered differed. One would expect both respondent and part scores to vary in acceptability--to vary, that is, in difficulty. The questions to be answered by this analysis are:

> Do scorers differ in their assignment of acceptable scores?
>
> Does time of scoring affect the assignment of acceptable scores?
>
> Or, in other words, do either scorers or time affect the difficulty of an open-ended measure?

The analysis of variance for both MCS and A/U are presented in Attachments 5-8 below.

One final analysis of the A/U variable was made, based on analysis of variance data. This analysis is related to the intra-class correlation or Cronbach's alpha. It differs, however, in that it is based on a multi-factor design. It involves estimating the generalizability of the scoring from a ratio of relevant components of variance.* (For a general discussion of the technique, see Stanley, 1971).

Specifically, the among-respondents component of variance is taken as an estimate of variance in the population of the ability to perform, or not perform, the exercise. That is, it is taken as an estimate of the variance of the population true score. That variance component is divided by the sum of the variance components judged to be relevant to the actual (as opposed to the experimental) scoring situation. The resulting ratio can be interpreted as a reliability estimate: a ratio of true variance to total variance.

Those components of variance determined to be relevant were:

(1) among persons, the estimate of true variance.

(2) among times: since a normal hand scoring takes two-three months to complete.**

---

*) Expected Mean Squares were constructed by the BMD 08V analysis of variance program (Dixon, 1973).

**See Glass and Hakstian (1968) for a critical discussion of including fixed effects in an analysis of this type. We felt justified, since our selection of times would result in a maximum variance due to time, and thus create a conservative estimate of reliability.

(3) among scorers: since NAEP data are based on sums of items scored by different scorers.

(4) all interaction of the above factors.

Note that the variance due to parts was omitted. Variance among parts might be construed as part of the true variance. It was nevertheless not included, because parts were a fixed factor (see note on preceding page) and thus might seriously bias the ratio. These variance ratios are presented in Attachments 1-4.

RESULTS AND DISCUSSION

Percent Agreement

The data were initially analyzed by calculating, for each exercise, the percentage of the sample responses on which all scorers agreed upon the category assignments. These percentages showed considerable variation across exercises ranging in value from 58.4% to 93.8% with the mean percentage being 76.5%. The overall mean for agreement of all but one scorer was 86.3%.

Next the categories assigned to the responses by each reader for all readings were compared. The average percent of reader agreements were calculated on each exercise. These percentages varied from 82.6% to 98.7% and a mean of 90.4%.

The overall agreement, based on presence or absence of most common score, ranged from 88.8% to 99.5%, with an average of 94.1%.

All the percentages seem to be high enough to indicate that NAEP hand scoring is not heavily dependent upon the scorer. The percentages are displayed in Attachments 1-4.

There is a slight advantage for COD over Writing exercises on all three indices. Since the average advantage across ages 9, 13 and 17 is no greater than 5% on any index, we conclude that the scoring is essentially equivalent for both subjects. Indeed, since the COD exercises studied cover topics appropriate to mathematics and citizenship as well as career education, we are tempted to generalize to all NAEP scoring (except art, literature, and music!).

Analysis of Variance - MCS

If NAEP's scoring procedure were perfectly generalizable over scorers, times, respondents, and different exercise parts, there would be no variation at all in assignment of the most common score. We would, therefore, prefer to find no significant analysis of variance effects. The worst possible result would be to find large effects for scorers or times. That would mean that the baseline NAEP data would have to be rescored every time NAEP wanted to measure change or a state or local assessment wanted to compare its results with the NAEP baseline. Even if the expense of such re-scoring were tolerable, it would be extremely inelegant for NAEP's baseline, criterion results to change for every different comparison.

Inspection of Attachments 5-8, "Probability Levels Associated with the F-Ratios for the Analysis of Variance" for the MCS results show two strong and consistent effects. These are the effects for responses and for the responses by exercise parts interaction. It is not surprising--though regrettable--that the consistency of the scoring depends on how people respond and what they are responding to.

There do not appear to be any consistent effects for times or for scorers but the scorers by times interactions appear more often than one would like. To evaluate the importance of these effects, components of variance were estimated and, from these, the percentage of total variance was calculated for each effect. This analysis showed that approximately 17% of the variance (over all exercises) could be attributed to responses and parts combined, and less than 1% could be attributed to scorers and times combined. Thus the effects, even for responses and parts, are small, though statistically stable.

Analysis of Variance - A/U

In contrast to the MCS analysis, one would expect large variations among responses and parts for the acceptable/unacceptable variable. In fact, the variance among responses is--as mentioned above--an estimate of the variance among true scores in the sample. However, as with the MCS, variance among scorers or times is a strong blow at the generalizability of the scoring procedure.

The results of the U/A analyses of variance are summarized in Attachments 5-8. Again, the only strong and consistent effects are for responses and the responses by parts interaction. These two effects combined account for over 73% of the variance across all exercises. In contrast, the two effects account

for only 17% of the variance in the MCS variable, above. Thus, for the A/U variable, the effect of responses and parts is both statistically stable and large.

The effects for scorers and times are negligible.

Components of Variance Estimates of Generalizability

Components of variance for the A/U variable were also used to compute an estimate of the ratio of true variance to total variance in the ability to perform the exercise acceptably. The results are displayed in column (4) of Attachments 1-4. Note that this coefficient is affected by the lack of variance among responses on very easy (or very difficult) exercises. In particular, exercises 9 and 18--which were answered correctly by 99% of respondents--have a coefficient of less than .35 simply because there was almost no variation among responses. Exercises 8, 13, and 15 also were answered correctly by more than 90% of the respondents.

The median percentage of variance accounted for was .80. This is a conservative estimate of the generalizability of NAEP scoring, since it is based on single exercises, answered by only 28 respondents; since variations due to scorers and times are included in the error variance; and since 5 of the 18 exercises included were extremely easy. While the question always remains of how reliable is reliable enough, the present investigators were extremely pleased with a median coefficient of .80. We feel that good evidence now exists that NAEP scoring procedures will generalize to other times--for change measures--and other users--for local comparisons.

Attachment 1: Percents of Agreements

Age 9

Subject	Exercise	(1) Responses: 7 Scorers	(1) Responses: At Least 6 Scorers	(2) Times: 2 Times	(3) Overall	Ratio of Variance Components (Generalizability of Acceptable/Unacceptable Scoring)
Writing	1	65.7	77.2	87.0	88.8	.65
	2	67.9	75.0	92.4	89.3	.82
COD	3	93.8	96.4	98.7	98.4	.99
	4	74.2	86.4	90.1	92.6	.90
	5	58.4	78.0	85.7	89.0	.49

Attachment 2: Percents of Agreements

Age 13

Subject	Exercise	(1) Responses: 10 Scorers	(1) Responses: at least 9 Scorers	(1) Responses: at least 8 Scorers	(2) Times: 3 Times	(2) Times: at least 2 Times	(3) Overall	(4) Ratio of Variance Components (Generalizability of Acceptable/Unacceptable Scoring)
Writing	6	64.3	80.2	85.3	82.9	98.2	91.3	.91
	7	71.4	82.1	86.9	86.8	98.9	92.6	.80
	8	68.6	79.4	85.7	82.6	97.9	92.6	.62
COD	9	91.5	92.6	98.6	98.7	100.0	99.5	.33
	10	72.6	84.5	89.3	86.8	98.9	94.0	.85

Attachment 3: Percents of Agreements

Age 17

Subject	Exercise	(1) Responses: at least 9 Scorers	(1) Responses: at least 8 Scorers	(2) Times: 3 Times	(2) Times: at least 2 Times	(3) Overall	Ratio of Variance Components (Generalizability of Acceptable/Unacceptable Scoring)
Writing	11	88.4	93.8	86.9	99.5	97.3	.83
	12	76.8	90.5	92.9	100.0	95.5	.75
COD	13	72.6	84.1	86.9	99.5	93.1	.61
	14	72.9	84.3	89.4	99.0	93.2	.87
	15	92.1	95.6	96.7	100.0	98.2	.65

Attachment 4: Percents of Agreements

Adults

Subject	Exercise	(1) Responses*			(2) Times		(3) Overall	Ratio of Variance Components (Generalizability of Acceptable/Unacceptable Scoring)
		10 Scorers	at least 9 Scorers	at least 8 Scorers	3 Times	at least 2 Times		
	16		93.0	97.2	97.5	99.9	98.4	.93
COD	17	62.8	82.3	91.7	88.0	99.6	92.9	.80
	18	90.1	94.5	94.5	96.6	99.6	96.9	.29

*For exercise 16 only 9 scorers scored the sample responses.

ATTACHMENT 5

Probability Level Associated with the F-Ratios for the Age 9 Analysis of Variance

Exercise # (See Table 1)	Most Common Score 1	2	3	4	5	Acceptable/Unacceptable 1	2	3	4	5
Effect										
R (Responses)	.01	.01	.01	.01	.01	.01	.01	.01	.01	.01
T (Times)**	*	*	*	.10	*	*	*	*	*	.05
S (Scorers)	*	*	*	*	*	.10	*	*	*	.25
P (Parts)**	.01		.10	*	.25	.01		.01	*	.01
R x T	.01	*	*	.05	*	.25	*	*	.01	.25
R x S	-	-	-	-	-	-	-	-	-	-
T x S	*	.25	*	.05	.05	.10	*	*	*	.05
R x P	.01		.01	.01	.01	.01		.01	.01	.01
T x P**	.25		*	*	.10	*		*	*	.10
S x P	*		*	.05	*	.01		*	.05	.05
R x T x S	-	-	-	-	-	-	-	-	-	-
R x T x P	.01		*	.01	.05	.05		*	.01	.05
R x S x P	-		-	-	-	-		-	-	-
T x S x P	.25		*	.25	*	*		*	*	.05
R x T x S x P	-		-	-	-	-		-	-	-

* Indicates a probability level of greater than .25.
**The F-statistic computed for these effects was F' (see Winer, 1971).

ATTACHMENT 6

Probability Levels Associated with the F-Ratios for the Age 13 Analysis of Variance

Exercise # (See Table 1)	Most Common Score 6	7	8	9	10	Acceptable/Unacceptable 6	7	8	9	10
Effect										
R (Responses)	.01	.01	.01	.01	.01	.01	.01	.01	.01	.01
T (Times)**	*	.10	*	.25	*	*	*	*	*	.10
S (Scorers)	.01	.10	*	.25	*	.25	*	.05	*	*
P (Parts)**	.25		*	.25	.05	.01		.10	.25	.10
R x T	.10	.05	*	.10	.05	.05	*	*	*	.05
R x S	-	-	-	-	-	-	-	-	-	-
T x S	.01	.10	.01	.25	.01	*	*	.01	.25	*
R x P	.01		.01	.01	.01	.01		.01	.01	.01
T x P**	*		*	.25	*	*		*	*	*
S x P	.01		*	.25	*	*		*	*	*
R x T x S	-	-	-	-	-	-	-	-	-	-
R x T x P	.01		*	.01	.10	.01		*	.01	.25
R x S x P	-		-	-	-	-		-	-	-
T x S x P	.01		.25	.25	.05	.10		*	.25	*
R x T x S x P	-		-	-	-	-		-	-	-

* Indicates a probability level of greater than .25.
** The F-statistic computed for these effects was F' (see Winer, 1971).

ATTACHMENT 7

Probability Levels Associated with the F-Ratios for the Age 17 Analysis of Variance

	Most Common Score					Acceptable/Unacceptable				
Exercise # (See Table 1)	11	12	13	14	15	11	12	13	14	15
Effect										
R (Responses)	.01	.01	.01	.01	.01	.01	.01	.01	.01	.01
T (Times)**	*	.25	.10	.01	*	*	*	*	*	*
S (Scorers)	.25	.25	.05	.05	*	.01	*	*	.01	*
P (Parts)**	*	*	*	.25	*	*	.01	.25	*	*
R x T	*	*	.10	.05	*	*	.25	.01	.25	.01
R x S	-	-	-	-	-	-	-	-	-	-
T x S	.05	*	.05	.25	*	*	.05	*	.05	*
R x P	.01	.05	.01	.01	.01	.01	.01	.01	.01	.01
T x P**	*	*	*	.10	.25	*	*	*	*	*
S x P	.05	.25	*	*	*	*	.01	*	*	.25
R x T x S	-	-	-	-	-	-	-	-	-	-
R x T x P	*	*	.05	.05	*	*	*	.01	*	.01
R x S x P	-	-	-	-	-	-	-	-	-	-
T x S x P	.25	*	.10	*	*	*	.01	*	.05	*
R x T x S x P										

* Indicates a probability level of greater than .25.
**The F-statistic computed for these effects was F' (see Winer, 1971).

ATTACHMENT 8

Probability Level Associated with the F-Ratios for the Age Adult Analysis of Variance

Exercise # (See Table 1)	Most Common Score 16	17	18	Acceptable/Unacceptable 16	17	18
Effect						
R (Responses)	.01	.01	.01	.01	.01	.01
T (Times)**	*	.25	*	*	*	*
S (Scorers)	*	.10	.05	.01	.01	.10
P (Parts)**	.01	.25	.10	.01	.01	.25
R x T	*	*	*	*	*	*
R x S	-	-	-	-	-	-
T x S	*	.01	*	.25	.01	*
R x P	.01	.01	.01	.01	.01	.01
T x P**	*	*	.10	*	.25	.25
S x P	*	.01	.05	.01	.01	*
R x T x S	-	-	-	-	-	-
R x T x P	.10	.25	*	*	.25	*
R x S x P	-	-	-	-	-	-
T x S x P	.10	.01	*	.01	.01	*
R x T x S x P	-	-	-	-	-	-

* Indicates a probability level of greater than .25.
**The F-statistic computed for these effects was F' (see Winer, 1971).

References

Coffman, W.E. "Essay Examinations." In R.L. Thorndike, ed., Educational Measurement (2nd Ed.). Washington, D.C.: American Council on Education, 1971.

Dixon, W.J. Biomedical Computer Programs, University of California Press: Berkeley and Los Angeles California, 1973, p.699.

Glass, G. V and Hakstian, A.R. "Measures of Association in Comparative Experiments: Their Development and Interpretation." Research Paper No. 14. Boulder, Co.: Laboratory of Educational Research, University of Colorado, 1968.

Stanley, J.C. "Reliability." In R.L. Thorndike, ed., Educational Measurement (2nd Ed.). Washington, D.C.: American Council on Education, 1971.

Winer, B.J. Statistical Principles in Experimental Design (2nd Ed.) New York: McGraw Hill, 1971.