ED 034 308 EA 002 631

AUTHOR	Vincent, William S.; Casey, John J.
TITLE	Statistical Report on Indicators of Quality.
INSTITUTION	Columbia Univ., New York, N.Y. Inst. of Administrative Research.
PUB DATE	May 68
NOTE	4p.
JOURNAL CIT	IAR-Research Bulletin; v8 n3 p1-3 May 68
EDRS PRICE	EDRS Price MF-$0.25 HC-$0.30
DESCRIPTORS	*Educational Quality, Elementary Grades, *Measurement Instruments, *Rating Scales, School Districts, Secondary Grades, *Statistical Analysis, Statistical Studies

ABSTRACT

As a measure of school quality, the instrument Indicators of Quality has met the demands of reliability and item discrimination in a statistical study based on in-class observations in all the school districts in metropolitan New York. In this initial application, the 51 observation items were scored positive, zero, or negative and four aggregate scores were devised. The scores were categorized into elementary, secondary, and district levels. The high and low scores were separated by roughly four standard deviations in the district-wide calculation of each aggregate score. In addition, an intercorrelation of the four scores was run showing no dependence. In general, each item had at least one component, either positive or negative, that discriminated well, and most items had high difficulty levels and high discrimination indices. The reliability of the total instrument was estimated by means of the Spearman Brown formula, which provided a reliability coefficient (r = .91). Related documents are EA 002 619, EA 002 620, EA 002 621, and EA 002 622. (Author/LN)

Research Bulletin

Volume 8, No. 3 May, 1968

Statistical Report on Indicators Of Quality

William S. Vincent • John J. Casey

A year ago a new instrument for obtaining a quantitative measure of school quality was described in these pages.[1] At that time it was stated that the manual and observation guide were not being offered for public distribution until a substantial statistical examination could determine its general applicability for the intended task. The first steps of statistical analysis have now been completed and it is evident from this that *Indicators of Quality* meets well the demands of reliability and item discrimination that are needed in an instrument of 51 polarized items (scorable on a 103-item scale) applied to a representative sampling of all class meetings that comprise a school's instructional setting. The application procedure is observation of a standard time span in each location of the sample.

What also is clear is that the instrument is not a device for amateurs. That is to say, the training and screening of observers are fundamental to obtaining a reliable quantitative appraisal of school. It appears that a three-day period of familiarization and trial application is minimal, and that during the first six days of an observer's work a sampled "cross-check" against other observers is advisable in order to identify highly variable or grossly divergent observers. When these are identified, even after the training session, their work must be discarded and their schedule redone.

The problem is of course the typical one of human variability. While it is the intent of the training session to bring all observers to the state of applying the instrument with like mind, so that any observer will see in a situation what any other observer would have seen and score it the same, it is apparent that human beings vary greatly in their capacity to follow the stringent rules that enhance reliability. It also appears advisable, where some time has elapsed since an observer's last experience in applying *Indicators,* that a brief period of retraining be undertaken. Essentially, training, and retraining consist of attaining memorized familiarity with the items, attending to the rules which govern the application of items to observed behavior, and clarifying borderline events through questioning and discussion. *Indicators* is not the first behavioral measurement device that is subject to such limitations, a famous example being the *Stanford Binet.*

Thus it is that the decision has been made not to offer the materials used in the application of *Indicators of Quality* for general distribution at all. What is available, then, is not copies of the instrument permitting people to apply it themselves to their own situation, but rather a scoring and report service. *Indicators of Quality* is available to be applied by trained observers whose reliability has been or will be computed. They will use observation schedules constructed by a staff coordinator, since consistency in the sampling of class meetings is a requisite. Scores are reported in relation to norms being developed and *Quality Control Charts* (based on standard score scales) facilitate interpretation.

[1] "Indicators of Quality," *IAR Research Bulletin*, Volume 4, No. 3, May, 1967.

Preliminary Statistical Analysis

What follows are some details from the preliminary analysis of results obtained in the application of *Indicators* to forty-seven school districts in the metropolitan area of New York. A total of 4287 observations were obtained in the forty-seven districts. This first application was limited to the 3rd, 4th, 5th, and 6th grades in the elementary schools and the 10th, 11th, and 12th grades in the high schools. The observations constitute a stratified sampling of all credit class meetings occurring at the indicated grade levels.

Each of the fifty-one items is "polarized." This means that either the positive case or its opposite may be observed. The total number of positive events provides a measure of favorableness with respect to individualization, interpersonal regard, creativity, and group activity, the four areas of educational process upon which the instrument is validated. The total number of negative events provides a measure of unfavorableness. The net score is the number of negatives subtracted from the number of positives.[2]

Timed observations obtained from the forty-seven school districts were scored by four different procedures. The first, *mean positive score,* is simply the arithmetic average of the number of positive signs scored during the sample of observations in each district. Second, the *mean negative* score is an arithmetic average of the number of negative signs scored. Third, a *mean difference* score was calculated whereby a difference was derived from single observations by subtracting the number of negative signs scored from the number of positive signs scored.[3] As a fourth method of scoring the *percentage* of the *difference scores* which were in the *positive* range in each district was calculated. Subsequently, a categorization of scores was developed for the three levels: elementary, secondary, and district total. In all, twelve differing techniques were utilized and profile scales were developed to graphically display the results. Included in the profile were scales showing the means of district means, high and low scores, and the standard deviations of the scores.

The twelve scales described above were devised in order to permit an initial assessment of the data and also to provide a suitable vehicle for presentation of the results. It became apparent that the district scores were approximately normally distributed on each of the scales, and that four or more standard deviations separated the high and low scores on nine of the scales. The remaining three scales evidenced 3.9 standard deviations respectively between the high and low mean scores. A further assessment was made by intercorrelating elementary level and secondary level scores from thirty-three comparable districts.[4] The resulting correlation coefficients showed a considerable lack of relationship among the four scoring techniques at these levels. The conclusion to be drawn from this is that elementary school scores are not highly related to secondary school scores, irrespective of the manner of scoring.

A third type of assessment of the data was made by means of a visual scrutiny of the scores obtained by each district on each of the scales. It was observed that while some districts were uniformly high and others were uniformly low on the various scales, the trend, however, was toward some variability of performance on each of the scales and at each level. All of these assessments, collectively and individually, tended to support the supposition that all four methods of scoring might be suitable for an instrument of this type.[5] Fuller analyses of the data and future reports of scores will be developed at each level and for each of the several scoring techniques outlined above.

Item Discrimination and Reliability

The 4287 time samples provided a substantial base of data from which was derived a record of how well each of the instrumental items functioned in the application. A 103-point scale was utilized to rank the time samples from highest to lowest, and this was made possible by choosing the *difference* score as a criterion. Each of the 51 polarized items was analyzed in terms of its positive and negative components, and both difficulty levels and discrimination indices were computed. The difficulty levels of the 51 items ranged from a high of 78.8% to a low of 2.3% with a median difficulty of 27.3%. Flanagan's tables from item analysis[6] were used to estimate the discrimination indices for each of the 102 positive and negative signs. All save one of these 102 signs discriminated in the expected direction, that is, the positive signs were seen more often in the time samples ranked highest

[2] For a complete discussion of the validation categories and the nature of the polarized items see *IAR Research Bulletin*, Vol. 4, No. 3, loc. cit.

[3] For a fuller discussion see Robert Chisholm, "Development of a Procedure for Selecting Indicators of School Quality for Inclusion in an Observer's Guide," Ed.D. Project, New York, Teachers College, Columbia University, 1966, p. 133.

[4] Included were twenty-eight MSSC and five CSS districts which were organized on a K-12 basis.

[5] Chisholm, op. cit., p. 132.

[6] See John Flanagan, "General Considerations in the Selection of Test Items and a Short Method of Estimating the Product Moment Coefficient from the Data at the Tails of the Distributions," *Journal of Educational Psychology*, 30:674-80, 1939.

and the negative signs were seen more often in the time samples ranked lowest. Discrimination indices ranged from a high of +.65 to a low of +.12 for positive signs and from a high of —.61 to a low of .00 for negative signs. The single sign which contradicted expectation in discrimination performance and the few signs with low discrimination indices were uniformly associated with low difficulty items. In general, each item had at least one component, either positive or negative, which discriminated well, and most items had high difficulty levels and high discrimination indices.

Item analysis is essentially an atomistic procedure well suited to quantify the signs of *Indicators of Quality* in terms of how well they functioned on this initial application. It also provides basic information which can be used again in other analyses. The difficulty level of each of the items was instrumental in settig up split-halves for the purpose of determining a reliability coefficient of internal consistency. Items of approximately equal difficulty were selected for inclusion in each half, and the districts were rescored on the basis of signs in each of the halves. Again, the *difference* score was used as a criterion, and on the basis of mean *difference* scores for districts a correlation coefficient ($r = .84$) was obtained. The reliability of the total instrument was estimated by means of the *Spearman Brown* formula, which provided a reliability coefficient ($r=.91$)—a measure of the instrument's reliability.

Research Bulletin

Devoted to research carried on by the Institute of Administrative Research and related groups, review of results, report of work in progress, discussion of theory and design, and implications for educational policies

Published in November, February and May
by the Institute of Administrative Research

●

William S. Vincent, *Director*
Ernst Auerbacher, *Associate*
Charles E. Danowski, *Associate*
Norman J. Walsh, *Associate*

Paul R. Mort Fellow
David L. Donovan

U.S. Office of Education Research Trainees
E. Robert Bagley
John J. Casey
Anthony M. Cresswell
Charles W. Laabs
Thomas P. Wilbur

Research Fellows
John J. Battles
Howard M. Coble
Richard Lerer
Harbison Pool
Brian Simpson
George C. Zima

●

Subscription rate: $1.50 per year

May, 1968