Proceedings of the Iowa Academy of Science

Volume 55 | Annual Issue Article 45

1948

Quantification of the Wechsler-Bellevue Vocabulary

Charles F. Haner
Grinnell College

Wilse B. Webb
Washington University

Quantification of the Wechsler-Bellevue Vocabulary

CHARLES F. HANER AND WILSE B. WEBB

The importance of the vocabulary test in mental testing is unquestioned. Such tests have been evaluated and used as the best single index of general intelligence when only one measure can be used. The inclusion of the vocabulary test in a general intelligence test is almost a foregone conclusion. In addition the vocabulary test has been extensively used in studying mental deterioration. Considering the importance of the vocabulary test it is surprising that so little attention has been given to scoring problems involved. Yacorzynski (1941) has suggested that the vocabulary score used in the deterioration ratio may obscure real deterioration in vocabulary functioning due to the gross scoring method now employed. One of Terman's students, Helen Green (1931) attempted to quantify qualitative differences in definitions by establishing categories of responses and giving differential credit for these categories. She was able to distinguish clearly between CA levels with this method, but it was considered impractical for purely administrative work.

Instructions to test administrators of the Stanford-Binet and Wechsler-Bellevue point out the possibility of qualitative variations of definitions and in some cases have given half credit for definitions not up to a given standard. There seems to be no evidence available as to the reliability of test administrators in dealing with these qualitative differences, or of the range or distribution of qualitative differences met with.

This paper concerns itself with the quality of definitions encountered in administering the Wechsler-Bellevue and it attempts preliminary quantification of them. The paper is in the nature of a status report on what it is hoped will be a series of research projects on vocabulary definition quantification and a test of Yacorzynski's hypothesis regarding deterioration.

PURPOSE

The purpose of this paper is two fold. First an attempt was made to determine the nature of judgments by a fairly typical group of test administrators of typical vocabulary definitions. More specifically considered are: (1) the degree of agreement expressed by test administrators in rating definitions along a continuum of goodness, and (2) the agreement in rating a given definition as "acceptable" or "not acceptable". Secondly an attempt was made to determine the range and distribution of definitions termed acceptable by the raters.

METHOD

The above purposes necessitated as wide a coverage of possible responses to vocabulary words as practical. To secure this 314

Wechsler-Bellevue records were obtained from the files of the Psychology Clinic at the State University of Iowa. A survey of each record for the definition of each word was made, and all definitions that differed in other than simple wording were listed. The number of definitions for each word varied from three for the word *traduce* to 38 for the word *affliction*. There were certain difficulties in this form of selection which are recognized. The records were from a population showing a marked skewness toward the high IQ levels and a cluster around one age level (18 to 22). The authors feel that extension of this basic survey to include older persons and lower IQ levels would have only increased the number of "not acceptable" definitions without influencing the range or distribution of "acceptable" definitions. A survey of the few records for children in the files of the Psychology Clinic at the State University of Iowa substantiated this assumption by not adding a single definition not included in the basic survey.

A second possible source of bias in the selection method is that the definitions used were those recorded on the Wechsler-Bellevue blanks and hence probably were not verbatim. However they were the definitions on which the test administrator did his scoring so the difficulty would not seem to be serious.

The definitions so selected were rated along a continuum of absolute goodness by a rating population of 40 persons, of whom 36 were graduate students in psychology at the University and 4 were senior psychology majors. The range of experience in administration of the Wechsler varied, but most subjects had given from 10 to 50 Wechslers. The rating scale was a graphic one consisting of 11 points, numbered 0 through 10. The labels "not acceptable" and "best possible" were attached to these respective extremes and the rest of the continuum was numbered 1 through 9 but no written descriptions were attached to them.

Each vocabulary word was defined on the rating sheet before the definitions to be rated were offered. These standard definitions were taken from the Wechsler manual and could be considered as "best possible" definitions.

The Wechsler vocabulary test contains 42 words, hence fatigue and motivation changes would have been important had judges been asked to rate all definitions. Therefore two forms of rating sheets were constructed; however there were definitions of four words over-lapping on both forms to make possible a comparison of the ratings of the two groups. Approximately equal numbers of hard and easy words were included on both forms and about the same number of definitions to be rated. The judges were given several days to complete the ratings in an attempt to reduce fatigue effects further. To avoid a constant motor rating tendency the ends of the graphic scales were randomly alternated. Written and oral instructions were used.

RESULTS

Treatment of Results

Scale Values

Scale values were determined for each definition by computing the median rating given that definition by the judges. It was arbitrarilly decided that any definition receiving 50% or more "not acceptable" ratings would be so considered and given a zero or "not acceptable" value. The median values were weighted according to the number of "not acceptable" ratings given each word; i.e., by (1) computing the median for the judges who rated the word as acceptable, (2) determining the percentage of the total judges who rated it "not acceptable", and (3) multiplying the unweighted median by this percent. Thus if 40% of the judges rated a given definition as "not acceptable" the median would be computed for the remaining judges and that value reduced by 40%.

Quartile deviations were also computed. They serve to express the degree of rater agreement and indicate the confidence one can put on the scale values obtained when using them to study deterioration or for other applications. It is recognized that the magnitude of the Q is influenced by the magnitude of the median, for definitions rated high on the continuum could not have as great a spread as those in the middle.

Two problems of rater reliability were involved. First there was the reliability of the two groups on the definitions of the four overlapping words. Comparison of these shows a tendency for one group to rate slightly higher than the other group. This was a fairly consistent tendency. The group rating the words higher found them about a third of the way through the rating sheet, the group giving lower ratings had the overlap words at the end of the rating sheet, so the differences probably reflect both differences in rating by the groups and fatigue and motivation changes. The small number of subjects probably magnify the differences.

The second problem of reliability was that of how consistently the same judges would rate and rerate a given definition. For some of the words having many definitions to be rated the same definition was included twice on the assumption it would not be noticed. Reports of the judges following completion of their ratings testify to the correctness of this assumption. Accidental destruction of the raw data precluded the possibility of calculating correlation coefficients for all of the six definitions which were repeated. Two of them yielded correlations of .89 and .99 respectively. The medians and Qs of the first and second ratings were compared however and show very slight and entirely inconsistent differences. The definition which had the correlation of .89 showed the greatest rate-rerate difference in medians and so might have been the lowest correlation.

Three judges were found consistently to underrate definitions when compared with other judges. A correction factor was computed for them and all of their ratings adjusted by using it.

DISCUSSION AND RESULTS

Rater agreement

The quartile deviation was the statistical measure of rater agreement used. It was assumed that judgments were rectangularly distributed within an interval and the median and Q computed accordingly. Even though all judges rated a given definition in the same interval it would receive a Q value of .25.

The range of Qs is from .28 indicating almost perfect agreement to 3.28 which represents the greatest amount of disagreement obtained.

As is to be expected the Q value decreases as the Md increases By the weighting system used the converse of this might not be expected to follow. Illustrative of this tendency is the following table.

TABLE 1

MEAN QUARTILE DEVIATIONS OF DEFINITIONS WITH VARYING MEDIANS

Median Range	Mean Q
3.00-3.99	1.69
4.00-4.99	1.70
5.00-5.99	1.50
6.00-6.99	1.39
7.00-7.99	1.32
8.00-8.99	.98

The mean Q values of the definitions of the more difficult words are smaller than those of the definitions of the easy words, as is to be expected from the fact that the median values of the definitions of the more difficult words are higher than of the easy words.

It is hoped that examination of the definitions and degree of rater agreement will yield classes of definitions which show little rater agreement. One illustration of this would seem to be definitions which involve technical words or classificatory names. The following illustrate this well, as all have uniformly high Q values.

	Q Value
Apple—Pome fruit of the tree genus malus	2.84
Donkey—A member of the family Eohippus with with long ears	2.05
Fur—Outgrowth of cornified ectoderm	2.97

A second aspect of the problem of rater agreement is the consistency of the acceptance or rejection of a definition as a credit-value definition. To indicate this the mean number of "not acceptable" ratings for definitions of different median interval ranges is presented in the following table.

TABLE 2

MEAN NUMBERS "NOT ACCEPTABLE" RATINGS GIVEN DEFINITIONS WITH VARYING MEDIANS

Unweighted Median Range	Mean Number of "Not Acceptable" Ratings given each definition
.00- .99	7.00
1.00-1.99	5.88
2.00-2.99	3.56
3.00-3.99	2.39
4.00-4.99	1.63
5.00-5.99	.81
6.00-6.99	.41
8.00-8.99	.08

It is apparent that as the unweighted median value increases, the number of "not acceptable" ratings falls off sharply. Those definitions which received many "not acceptable" ratings are rated low by other judges who, however, find them acceptable definitions.

It is felt that rater agreement can be considered as fairly high as expressed by the magnitude of the Q values. However the need for some form of objective scoring would seem to be suggested when one rater will classify a definition as "best possible" and another will rate it as "not acceptable." Such cases are few, but it is seen from the above table that for judges who found a definition acceptable and rated it about half way along the continuum, one judge would rate it as not acceptable. With 20 judges that means that about 5% rated as "not acceptable" the same definition which the other 95% rated half way along the continuum of absolute goodness.

It is important to know if the same degree of rater agreement would be found by judges who had received their training under diverse conditions and instructors. The problem of reliability of the use of a quantitative scale once constructed remains. In general however it is felt that the degree of rater agreement obtained in this initial study warrants further investigation of a quantitative scale of qualitative differences.

Scale Values

The second purpose of the paper was to determine the range and distribution of definitions along a qualitative continuum. The median values obtained represent the place along the continuum the definition falls. The highest possible median would have been 9.50, the highest obtained was 9.42 and the lowest acceptable definition rating was .45 although there were a number rated not acceptable.

The medians of the definitions of the difficult words was higher than that of the easy words. The means of the medians of definitions of the first, middle and last six words in the scale are presented below:

Place	Mean Median Value
First six words	5.20
Middle six words	5.44
Last six words	6.63

It would seem that if the subject knew the difficult word at all he could define it adequately, but subjects might "know at" the definitions of easy or middle range words without being able to give a good definition of them. If this is so we might expect a greater qualitative range of easy words than of difficult ones. Standard deviations of the medians of the definitions of the first few words and the last few words show this to be the case. The definitions of difficult words cluster closely around a mean high up on the scale. Definitions of easy words spread out more and the middle of the distribution is about in the middle of the continuum. The distribution of the definitions of the first six words was found to be symmetrical with a mean about 5. For the last six words it was markedly negatively skewed with a mean above 6.5, and for the total vocabulary was almost symmetrical, with a slight negative skew and a mean about 5.5. The range of definitions of the difficult words was smaller than that of the easy words, the latter being as large as the range for the total vocabulary definition distribution.

These ratings strongly suggest that there exists a true continuum of goodness of definition and definitions are distributed along it in a symmetrical bell-shaped form. The range is from near "best possible" to "not acceptable". To give only full or no credit or in some instances half credit to definitions would seem to be obscuring valuable data to the practicing and research-minded clinician and to be distorting the measures obtained. The high degree of rater reliability found is encouraging for it would seem to indicate the practicability of a quantified vocabulary scale.

DEPARTMENT OF PSYCHOLOGY,
GRINNELL COLLEGE, GRINNELL, IOWA.
AND
DEPARTMENT OF PSYCHOLOGY,
WASHINGTON UNIVERSITY.

BIBLIOGRAPHY

Green, H. J. 1931. A qualitative method for scoring the vocabulary test of the new revision of the Stanford-Binet, Unpublished masters thesis. Stanford University Library.

Guilford, J. P. 1936. Psychometric methods. New York, McGraw-Hill.

Terman, L. M. and Merrill, M. A. 1937. Measuring Intelligence. New York, Houghton Mifflin.

Wechsler, D. 1941. The Measurement of Adult Intelligence. Baltimore, The Williams and Wilkens Co.

Yacorzynski, G. K. 1941. An evaluation of the postulates underlying the Babcock deterioration test. Psych. Rev. 48:261-67.

