VALIDITY IN SELF EVALUATION

SISTER MARY AMATORA, O.S.F.
Saint Francis College

IT has become common practice among users of the various tests and scales in practically all areas, to look for reports of reliability and more specifically of the validity of the instrument contemplated for a specific purpose. Yet, at times users of such tests take at face value a reported figure, or index of such, without considering further possible factors involved nor the conditions and extenuating involvements which may differ in the case of the specific use to which they are to put the scale or test in question.

Validity in self-rating is never easy to establish. Depending upon the trait, characteristic, attitude, or entity evaluated, the self-opinion or self-judgment may defy validation. Only in traits or items in which the self evaluation can be adequately substantiated by bona fide observation and/or recording of overt behavior, can one be assured of an adequate criterion against which he may attempt to validate the self-rating.

Indeed, the literature of the past decade is replete with articles on the computation and the interpretation of indexes of both reliability and validity. Clark and Smith (10), in a validation of the *Bell Adjustment Inventory* and of the *Washburn Social Adjustment Inventory*, computed correlations between faculty evaluation of the adjustment of students and each of these two inventories. They found r's of $-.319$ to .165 for the Bell scale, and r's $-.298$ to .348 for the Washburn instrument. They state that "scores made . . . on paper and pencil personality inventories may be quite fictitious and thus present a false picture of their traits." They conclude that "the *Bell Adjustment Inventory* and the *Washburn Social Adjustment Inventory* are not valid indicators of student adjustment in school situations, such as prevail at the Rochester Athenaeum and Mechanics Institute." On the other hand, Ryans (18) sug-

gests "application of factor analysis techniques . . . as an approach to the validation of behavior measures."

McNemar (14) points out that "individual measurement is subject to two types of error, variable or random and constant or biasing." Again he states that "also in scaling, the use of several questions tends to average out the variable errors." In the same article, McNemar seems to favor the scaling type rather than the inventory type. He states that "the aim of scaling is the development or construction of a 'measuring' device which will distribute individuals along a continuum running from a highly favorable through a neutral to a highly unfavorable attitude." He further stresses "the basic need for reliability, validity and uni-dimensionality for the instruments or devices used to classify or measure individuals with respect to opinions or attitudes."

The effect of position and of context upon personality rating-scale items using a five-point rating scale was studied by Young, Holtzman and Bryant (24). They concluded that "a given rating-scale item may appear early or late in the serial order on the form and not be seriously effected by position." Context effected the item only when it was moved "from a purely positive or negative context to a mixed one." In their study "undesirable personality traits are significantly more susceptible to response shifts with changes in item context than are items referring to desirable traits."

Anastasi (4) asserts that "it is only as a measure of a specifically defined criterion that a test can be objectively validated at all."

"Perhaps one of the most common examples of such an inconsistent treatment of test validity is provided by what we may call the argument of 'extenuating circumstances.'

"It is of fundamental importance, however, to bear in mind that psychological tests do not provide a different *kind* of information from that obtained by any other observation of behavior. The use of such labels as 'intelligence,' 'aptitude,' 'capacity,' and 'potentiality' has probably done much to make test users lose sight of the empirical validation of tests.

"After analyzing various situations, she concludes that "validity should be consistently interpreted with reference to the *specific criteria* against which the given test was validated.

It also follows that validity is not a function of the test but of the use to which the test is put. A test may have high validity for one criterion and low or negligible validity for another.

"The attitude that a good test has 'high validity' and a poor test has 'low validity' is still too prevalent among test users. Tests cannot be validated in the abstract, nor is the usual concept of validity itself universally applicable to psychological testing. . . . It is only when tests are employed for predictive or diagnostic purposes that the correlation with an exterior criterion is relevant at all."

The problem of validity in attitude measurement was attacked by Pace (16). He analyzed the replies of some 2,500 Syracuse University alumni to a 16-page questionnaire and computed Pearsonian *r*'s between scores on activity scales and scores on opinion scales. His correlations ranged from .15 to .20. In another portion of the same study, Pace affirms that "the tetrachoric coefficients tend to be lower than Pearson product-moment coefficients."

He himself states, however, that "the present study is primarily exploratory rather than analytical: it reports relationships in a wide range of fields based on data designed broadly to throw light on the status of the educational product rather than data specific collected to analyze relationships between opinions and actions."

He further states that "advances in the science of measurement will come in proportion to our ability to establish clear relationships between opinion and action. Until we do this, our so called measurements will remain purely descriptive. What we must seek is measurement that is both descriptive and predictive of observable behavior."

Weinland (22) is of a like opinion. He argues for "a rating scale for the use of others, and an inventory on the same attributes, to be filled out by the subject himself." He thus proposes to check the validity of the instruments "by comparing the measurements obtained." Although he does not state the method by which his *r*'s were computed, nor whether he used a formula for correction for attenuation, he reports correlations ranging from .15 to .70 for the eight traits measured on 57 subjects.

In the present study, the author has borne this point par-

ticularly in mind. Not only were the self-ratings of the individuals secured, but also ratings of their classmates were obtained. Whether children's ratings of peers are accurate is a moot question. One thing, however, is certain. The children rate their companions as the latter are observable to them. In other words, the behavior of the child-as-seen-by-his-companion is the behavior that is being rated. Whether this behavior is evaluated by peers in the same light as it would be evaluated by teachers or other adults is not the question in point.

Ordinarily, adults judge others by adult standards or by an adult's point of view as to what child-behavior should be. This could differ widely from a child's point of view. It seemed to the writer that a child's view of a child might be a more adequate criterion of a child's self-evaluation than an adult's opinion. It is for this reason that teacher evaluation for the same pupils is not used for the present study as a criterion for the validity of the self-evaluation of children.

The effect of test-retest on personality questionnaires was studied by Windle (23). He examined 41 studies of test-retest with various personality ratings and inventories. His cumulative table shows the research results of retests with various personality inventories. He states that "41 studies were found in which the authors did not postulate, either in prospect or in retrospect, that an external variable was operating to effect a test-retest difference." Windle further states that "differences between test and retest will probably tend to be reported more often than will lack of differences, but will also tend to be attributed more often to the influence of possible variables. In the absence of any basis for evaluating the effects of such counterbalancing reporting-biases, it will be assumed in this review that this sample of the literature is representative of research results of retested personalities in general."

He concludes that "although the evidence in the literature indicates that the usual retest change is slightly increased adjustment in test scores, this finding is by no means universal. One important factor accounting for the rather large variance in retest change appears to be the interval between tests. However, until more of these factors have been identified no general retest control can be safely assumed."

On the basis of this and other findings, the possibility of

using retests for self-evaluation in the present study was ruled out. It was deemed that a retest study for purposes of validity would be rather futile.

In certain areas of opinion measurement internal consistency or reliability is often used in lieu of the possibility of securing other measures of validity. In his study, Dudek (12) also states that "the reliability coefficient has meaning only in relation to the way in which it is computed and the manner in which it is interpreted." He further states that "we must recognize that some traits may not be capable of accurate prediction because of instability of the trait itself. In such instances it seems relatively futile to increase testing time with the idea of obtaining a precise measure of an individual's performance for purposes of predicting some future score. Also, 'correcting' a coefficient, merely to raise it, and then implying that predictions based on the scores will be as accurate as indicated by the corrected coefficient may lead to a false sense of security."

"Whether the computed coefficient should be 'corrected' will depend upon the particular question being put to the data and on being able to satisfy certain assumptions."

He concludes that "coefficients should be interpreted in a relatively specific manner."

One might say that if such a response is reliable, it is valid, from the principle of internal consistency. Yet Dudek carefully distinguishes between "test reliability" and "individual consistency" in performance, which "is a function of the instrument used, the variability of the subjects, and the stability of the trait being measured." It is true, for example, that Johnny's peers may not view his behavior with the same point of reference as he views it, yet it does lend some degree of objectivity over and above a self-opinion when a particular situation would imply the need for such.

With all the above in mind the writer set up the present investigation for comparing the self-evaluation of 200 boys and 200 girls of grades four through eight with the judgments of their peers. In the latter case, the opinion of eight classmates, both boys and girls, was averaged. *The Child Personality Scale* (2) which provides separate scales, each on a ten-point continuum, for twenty-two personality traits was used.

Pearsonian *r*'s were computed for each of the twenty-two

TABLE 1

Self Evaluation versus Peer Evaluation

Boys		Girls	
Courtesy	.67**	Interests	.62**
Popularity	.65**	Dependability	.58**
Neatness	.60**	Intelligence	.58**
Intelligence	.59**	Disposition	.49**
Religiousness	.55**	Sense of Humor	.48**
Generosity	.44**	Religiousness	.47**
Entertaining-Boring	.40**	Boisterous-Quietness	.45**
Persistence	.25**	Punctuality	.45**
Patience	.22**	Neatness	.41**
Dependability	.22**	Entertaining-Boring	.41**
Boisterous-Quietness	.20**	Popularity	.38**
Honesty	.19**	Persistence	.38**
Punctuality	.19**	Honesty	.35**
Sociability	.19**	Pep	.34**
Nervous-Calmness	.17*	Thoughtfulness of Others	.28**
Thoughtfulness of Others	.16*	Patience	.20**
Interests	.16*	Cooperation	.20**
Pep	.15*	Sociability	.17*
Goodsportsmanship	.15*	Generosity	.16*
Disposition	.13	Courtesy	.15*
Sense of Humor	.12	Nervous-Calmness	.12
Cooperation	.10	Goodsportsmanship	.11

* Significant at the five per cent level.
** Significant at the one per cent level.

scales separately for boys and for girls. Table I gives the resultant *r*'s.

It will be noted that the *r*'s between the boys' self-evaluation and that of their peers are statistically significant at the one per cent level of confidence for fourteen of the 22 scales used, and at the five per cent level of confidence for five more of the scales.

Girls' self-evaluation compared with that of their peers achieved statistical significance at the one per cent level of confidence on seventeen of the scales, and at the five per cent level on three more scales.

On only three scales for boys and two scales for girls are the correlations too low for the five per cent level. When girls are rating themselves on nervousness and on good sportsmanship they differ from the judgment of their classmates. Boys' self-evaluation does not agree with that of their classmates on disposition, sense of humor, and cooperation.

It would seem that on the personality traits measured by the other scales the self-evaluations are valid, when opinion of classmates is the criterion. At least the child's own view of himself is in sufficient agreement with his overt behavior as judged by his peers.

REFERENCES

1. Amatora, S. M. "A Twenty-Two Trait Personality Scale." *Journal of Psychology*, XVIII (1944), 3–8.
2. Amatora, S. M. *Child Personality Scale*. Cincinnati: C. A. Gregory Co., 1951.
3. Amatora, S. M. "Factor Analysis of Children's Personality Scale." *Journal of Psychology*, XVIII (1944), 197–201.
4. Anastasi, Anne. "The Concept of Validity in the Interpretation of Test Scores." EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, X (1950), 67–78.
5. Argelander, Annelies. "The Personal Factor in Judging Human Character." *Character and Personality*, V (1937), 285–295.
6. Bonney, M. E. "The Constancy of Sociometric Scores and Their Relationship to Teacher Judgments of Social Success and to Personality Self-Ratings." *Sociometry*, VI (1943), 409–424.
7. Booth, Gotthard C. "Objective Techniques in Personality Testing." *Archives of Neurology and Psychiatry*, XLII (1939), 514–530.
8. Brogden, Hubert E. "A Factor Analysis of Forty Character Tests." *Psychological Monographs*, LII (1940), 35–39.
9. Champney, Horace and Marshall, Helen. "Optimal Refinement of the Rating Scale." *Journal of Applied Psychology*, XXIII (1939), 323–331.
10. Clark, William A. and Smith, Leo F. "Further Evidence on the Validity of Personality Inventories." *Journal of Educational Psychology*, XXX (1942), 81–91.
11. Cronbach, L. J. "Test 'Reliability': Its Meaning and Determination." *Psychometrika*, XII (1947), 1–16.
12. Dudek, Frank J. "Concerning "Reliability" of Tests." EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, XII (1952), 293–299.
13. Guilford, J. P. "Trends in Personality Research." *Education*, LXI (1941), 636–642.
14. McNemar, Quinn. "Opinion-Attitude Methodology." *Psychological Bulletin*, XLIII (1946), 289–374.
15. McQuitty, L. L. "Conditions Affecting the Validity of Personality Inventories." *Journal of Social Psychology*, XV (1942), 33–52.

16. Pace, Robert C. "Opinion and Action: A Study in Validity of Attitude Measurement." EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, X (1950), 411–419.
17. Preston, Malcolm G. "Note on the Reliability and the Validity of the Group Judgment." *Journal of Experimental Psychology*, XXII (1938), 462–471.
18. Ryans, David G. "A Note on Methods of Test Validation." *Journal of Educational Psychology*, XXX (1939), 315–319.
19. Spencer, Douglas. "Frankness of Subjects on Personality Measures." *Journal of Educational Psychology*, XXIX (1938), 26–35.
20. Thouless, R. H. "Test Unreliability and Functional Fluctuation." *British Journal of Psychology*, XXVI (1935–1936), 325–343.
21. Weiss, L. A. "Rating Scales, with Special Reference to the Field of Child Development," *Psychological Bulletin*, XXX (1933), 185–208.
22. Weinland, J. D. "Use of Rating Scales and Personal Inventories to Check Each Other." *Journal of Applied Psychology*, XXXII (1948), 631–635.
23. Windle, Charles. "Test-Retest Effect on Personality Questionnaires." EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, XIV (1954), 617–633.
24. Young, H. H., Holtzman, W. H. and Bryant, N. D. "Effects of Item Context and Order on Personality Ratings." EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, XIV (1954), 499–517.