Bulletin of the Psychonomic Society
1987, 25 (4), 277-279

Student-generated personality scales

JOHN B. CAMPBELL
Franklin and Marshall College, Lancaster, Pennsylvania

A laboratory exercise, based on Jackson's (1975) procedure, was developed in which four groups of undergraduates each generated a rationally derived, 10-item personality scale to measure one of four personality traits. Student responses on these and corresponding scales from Jackson's (1974) Personality Research Form (PRF) were correlated with self-ratings and peer ratings of the students. The median validity (i.e., scale-rating correlation) for the student-generated scales was .42, compared with .55 for the PRF scales; however, the PRF scale accounted for an appreciably larger proportion of the criterion rating variance than did the corresponding student scale in only 6 of the 16 comparisons. The results generally were consistent with past work demonstrating the validity of student-generated, rationally derived scales of personality. Furthermore, the exercise proved useful in illustrating principles of item analysis and scale validation for the participating students.

Hase and Goldberg (1967) reported that strategy of scale construction had little effect on the validity of personality scales. This conclusion prompted a number of investigations of the relative merits of alternative strategies for the construction of personality tests. Thus, Jackson (1971) advocated the rational or "construct" approach to item selection, as opposed to purely empirical item selection. That is, he argued that items should be selected for inclusion in a scale based on the test constructor's judgment of the items' relevance to the definition of the construct being measured, rather than on the correlation of items with some criterion measure of the characteristic. Jackson went so far as to "fully expect under cross-validation that even an inexperienced item writer would be superior to empirical item selection with a typical heterogeneous item pool" (1971, p. 238). Ashton and Goldberg (1973) provided results largely consistent with Jackson's claim. Rationally derived scales produced by nonpsychologists yielded validities (i.e., correlations of targets' scores with peer ratings of the targets) lower than those for the corresponding empirically derived scales from the California Psychological Inventory (CPI; Gough, 1975). The validities for the rationally derived scales constructed by graduate students in psychology, however, were equivalent to those for the CPI scales. Furthermore, validities for the most reliable scales constructed by psychology students were "considerably higher than that of any of the CPI scales" (Ashton & Goldberg, 1973, p. 1). Jackson (1975), using a different set of personality constructs and undergraduate psychology students as item writers, found that validities (i.e., correlations of scale scores with self-ratings and roommate ratings) for 16-item, rationally constructed student scales were much higher than those for the corresponding CPI scales and almost as high as those for scales from the Jackson Personality Inventory. Thus, there is evidence that rationally derived scales from naive item writers can be relatively effective predictors of self- and peer ratings (see also Burisch, 1984, and Paunonen, 1984).

The laboratory component of our undergraduate introductory personality course focuses on the principles underlying construction and validation of personality tests. As such, it provides a natural forum for the teaching and testing of the work described above, and Jackson's (1975) procedure prompted development of an exercise for that course. In this exercise, groups of students constructed and validated their own tests. The exercise was designed both to replicate past work on the validity of student-generated scales and to introduce students to basic principles of test construction, item analysis, and correlational techniques. The substantive prediction was that the results of the exercise would be consistent with Jackson's (1975) finding of substantial validity for rationally constructed student scales.

Note that the present integration of research and instruction provides an example of the "'incidental' data collection" strategy advocated by Carlson (1971, pp. 214-215). Consistent with that orientation, the exercise was presented to the students with the strong caveat that test construction is a complicated procedure that should not be undertaken casually. Test proliferation has been associated with problems in the field (e.g., ad hoc procedures and conclusions, tests that overlap to an undetermined extent, and absence of theoretical context), and the students were urged to apply these same principles to the evalution and interpretation of existing scales.

Method

Subjects

The subjects were 21 undergraduates enrolled in the Personality and Individual Differences course during the spring 1985 term and 18 undergraduates enrolled in the same course during the spring 1986 term.

Requests for reprints should be sent to John B. Campbell, Department of Psychology, Franklin and Marshall College, Lancaster, PA 17604-3003.

Procedure

The exercise was conducted in two 2-h sessions, 1 week apart, although the final discussion continued into a third session. The first session began with a brief lecture on two topics. First, the students were given a description of some dos and don'ts in writing items (e.g., use simple language, avoid double negatives, avoid double-barreled questions, avoid items with heavy social desirability loadings or probable extreme response proportions, use equal numbers of positively and negatively keyed items). Second, the students were given a description of the rational, internal consistency, and empirical methods of scale construction, including examples of each and evaluating each in terms of the ease of demonstrating reliability and validity (see, e.g., chapter 17 in Anastasi, 1976; chapters 2 and 3 in Walsh & Betz, 1985; or chapter 9 in Wiggins, 1973).

Jackson's (1974) Personality Research Form (PRF) served as a useful vehicle for this latter discussion in that items were written or selected for the original PRF item pools based on their "conceptual link" with definitions of Murray's (1938) 20 needs. That is, the initial item selection for the 20 scales was based on rational or substantive considerations. Items then were retained for a scale if they demonstrated higher correlations with their own scale than with irrelevant scales (including social desirability) and if they were intermediate in endorsement proportion. That is, item retention was based on internal consistency or structural considerations. Finally, all scales were validated against trait and behavior ratings in a variety of samples. Thus, scale validation was based on external criteria.

After the lecture, the class was divided into four groups, and each was assigned one of the four traits to be measured: affiliation, dominance, impulsivity, and order. Each group was given the definition of the appropriate trait from the PRF manual and was allowed approximately 30 min to construct a 10-item, yes-no measure of that trait containing five positively (i.e., "yes" indexes presence of the trait) keyed items and five negatively (i.e., "no" indexes presence of the trait) keyed items. Each group handed in one legible copy of its scale and retained one copy with the scoring indicated on it. The students then received a form on which they were asked to provide a single self-rating for each of the four traits. The response format was a 1-7 scale where "1 indicates a *very low* general tendency to exhibit that sort of behavior and 7 signifies a *very high* tendency to act that way." Finally, all students were given two copies of the same four-question rating form, told to write their own name in as the target person, and asked to have 2 friends rate them anonymously on all four traits. The friends were to mail the completed form to the instructor through campus mail or to return it in a sealed envelope to the student target, who would return it to the instructor by the next week's lab period.

The second session was held during lab period the following week. Each student completed each of the four student-generated scales and each of the four corresponding 16-item PRF Form E scales. The students identified themselves by code numbers to maintain anonymity. Each of the four groups then used its own answer key and a PRF answer key to score all the completed PRF and student scales for its trait (e.g., the order group scored its own scale and the PRF order scale for all students in the class). During this procedure, each group was given a table containing the self-ratings and the peer ratings for each student, identified by student code number. Each group then constructed a table containing the scores of 14 variables relevant to their trait for each student in the class: the score on each of the 10 items in the student scale (where 1 indicated *selecting the keyed response* and 0 indicated *answering with the nonkeyed response*), the total score on the student scale, the score on the PRF scale, the self-rating, and the mean of the two peer ratings. Each group then used a general-purpose statistics package on a Macintosh microcomputer to calculate and print descriptive statistics and the intercorrelation matrix for the 14 variables. An overhead transparency of these results for each of the four traits provided the basis for the ensuing class discussion.

RESULTS

The intercorrelations of the final four variables for each trait provided the basis for a comparison of the student and the PRF scales. These correlations are provided in Table 1 for both samples. As a prelude to this comparison, the students used the item and scale variances to calculate coefficient alpha reliability for each of the student scales. These ranged from .42 (impulsivity) to .78 (order) for the 1985 class, with a median of .58, and from .00 (impulsivity) to .72 (order) for the 1986 class, with a median of .39. These reliabilities are lower than those reported for the 16-item PRF scales, which range from .67 to .89, with a median of .76 (see Table 21 in Jackson, 1974). This disparity results in part from the differences in scale length, and the lower reliabilities for the student scales clearly also attenuate the resulting validities.

The median correlation between the corresponding student and PRF scales across both samples was .63, with only the correlations for dominance and impulsivity in the 1986 sample failing to reach significance. As a further indication of the overlap of the student and PRF scales, each correlation between a corresponding student and PRF scale was compared both with the correlations of the student scale with the peer and self-ratings and with the correlation of that PRF scale with the peer and self-ratings, separately for each trait within each sample. This produced a total of 16 comparisons within each sample (e.g., the correlation between the student and PRF affiliation scales compared with the correlation between the student affiliation scale and the self-rating of affiliation). In 24 of the 32 comparisons, the student scale–PRF scale correlation exceeded the scale–rating correlation.

Comparison of the student and PRF scale correlations with the "criteria" of self-rating and mean peer rating were of the greatest interest (see Table 1; see Table 9-6 in Jackson, 1974, for corresponding correlations for the 20-item PRF Form AA scales). The study included four traits and two ratings on each trait in two separate samples; thus, there were 16 validity correlations for the stu-

Table 1
Intercorrelations of Test Scores and Ratings

Trait		SS	PRF	SR	PR
Affiliation	SS	—	.69*	.59*	.46*
	PRF	.57*	—	.52*	.20
	SR	.37	.64*	—	.66*
	PR	.34	.58*	.75*	—
Dominance	SS	—	.68*	.48*	.34
	PRF	.25	—	.51*	.33
	SR	−.11	.68*	—	.86*
	PR	.25	.25	−.04	—
Impulsivity	SS	—	.46*	.07	−.01
	PRF	.41	—	.15	.13
	SR	.60*	.66*	—	.88*
	PR	.18	.26	−.18	—
Order	SS	—	.82*	.73*	.72*
	PRF	.73*	—	.85*	.73*
	SR	.71*	.85*	—	.77*
	PR	.51*	.75*	.58*	—

Note—For each trait, values above diagonal are from the 1985 class, values below diagonal are from the 1986 class. SS = student scale score, PRF = Jackson's (1974) Personality Research Form scale score, SR = self-report, PR = peer report. *$p < .05$, two-tailed.

dent scales and 16 corresponding validity correlations for the PRF scales. The median validities were .42 for the student scales and .55 for the PRF scales. In 7 of the 16 possible comparisons between corresponding student and PRF validities, both the student and PRF validities were statistically significant (e.g., the correlations of the student and PRF affiliation scales with self-ratings in the 1985 sample). In 1 comparison only the student scale validity was significant (i.e., correlations of the affiliation scales with peer ratings in the 1985 sample), and in 3 comparisons only the PRF scale validity reached significance (e.g., affiliation with self-ratings in the 1986 sample). In 5 of the comparisons, neither of the validities reached significance (e.g, dominance with peer ratings in the 1985 sample). In the 11 instances in which at least one of the validities reached significance, proportion of variance accounted for (r^2) was approximately the same (arbitrarily defined as within .08) 4 times, was appreciably larger for the student scale 1 time, and was appreciably larger for the PRF scale 6 times. Finally, the lack of relationship between the self-ratings and peer ratings of both dominance and impulsivity within the 1986 sample, coupled with the consistently low correlations with peer ratings in these instances, suggests that these peer ratings were somehow inadequate.

During class, the results quickly led to discussions of previously abstract concepts, such as the relationship between endorsement rate and item-scale correlation and the general effects of variability and restricted range on correlation coefficients. The validities of the student scales, both in absolute terms and relative to the PRF, led to discussions of the effects of scale length on reliability and of reliability on validity, the effects of aggregation, reasons for the varying difficulty of measuring different traits, the limits of rating data, and the criterion problem in general. Student discussion of the data also focused on an item analysis of the student scales. The item analysis involved examining the mean response (i.e., endorsement rate) and the item-scale correlation for all 10 items on each of the student scales. Effectiveness of the items varied widely in their performance, and it was obvious to the students which items had performed very well (and therefore would be retained if subsequent revisions were made) and which had performed very poorly (and therefore should be replaced). For example, item-scale correlations for affiliation items in the 1986 sample ranged from −.22 to .77, whereas endorsement rates ranged from .11 to .99. It was a revelation for the students to see how low the interitem correlations were, even for "good" items.

DISCUSSION

The procedure in the present study departed from the original Jackson (1975) procedure in that the student scale writers also were the targets, rather than an independent sample. Despite this modification, the results were consistent with those reported by Jackson (1975) in demonstrating the validity of student-generated, rationally constructed personality scales. The present results do not suggest that such student scales are equivalent to those in a rigorously developed inventory such as the PRF. For a more adequate comparison, however, the student scales should be of the same length as those in the comparison scales.

Finally, the pedagogic value of the present incidental data collection should not be ignored. The research design afforded students direct experience in generating, interpreting, and evaluating personality scales. The students' involvement in, enthusiastic response to, and apparent benefit from the exercise argue strongly that it was successful in achieving the second goal of teaching principles of test construction and evaluation to students with no previous experience in this domain.

REFERENCES

Anastasi, A. (1976). *Psychological testing* (4th ed.). New York: Macmillan.

Ashton, S. G., & Goldberg, L. R. (1973). In response to Jackson's challenge: The comparative validity of personality scales constructed by the external (empirical) strategy and scales developed intuitively by experts, novices, and laymen. *Journal of Research in Personality*, **7**, 1-20.

Burisch, M. (1984). Approaches to inventory scale construction: A comparison of merits. *American Psychologist*, **39**, 214-227.

Carlson, R. (1971). Where is the person in personality research? *Psychological Bulletin*, **75**, 203-219.

Gough, H. G. (1975). *Manual for the California Psychological Inventory*. Palo Alto, CA: Consulting Psychologists Press.

Hase, H. D., & Goldberg, L. R. (1967). The comparative validity of different strategies of deriving personality inventory scales. *Psychological Bulletin*, **67**, 231-248.

Jackson, D. N (1971). The dynamics of structured personality tests: 1971. *Psychological Review*, **78**, 229-248.

Jackson, D. N. (1974). *Personality Research Form manual*. Port Huron, MI: Research Psychologists Press.

Jackson, D. N. (1975). The relative validity of scales prepared by naive item writers and those based on empirical methods of personality scale construction. *Educational & Psychological Measurement*, **35**, 361-370.

Murray, H. A. (1938). *Explorations in personality*. Cambridge: Harvard University Press.

Paunonen, S. V. (1984). Optimizing the validity of personality assessments: The importance of aggregation and item content. *Journal of Research in Personality*, **18**, 411-431.

Walsh, W. B., & Betz, N. E. (1985). *Tests and assessment*. Englewood Cliffs, NJ: Prentice-Hall.

Wiggins, J. S. (1973). *Personality and prediction*. Reading, MA: Addison-Wesley.

(Manuscript received for publication February 17, 1987.)