DOCUMENT RESUME

ED 061 256 TM 001 158

AUTHOR	Jackson, Douglas N.; And Others
TITLE	An Evaluation of Forced-Choice and True-False Item Formats in Personality Assessment.
INSTITUTION	Educational Testing Service, Princeton, N.J.
REPORT NO	RB-71-67
PUB DATE	Dec 71
NOTE	26p.
EDRS PRICE	MF-$0.65 HC-$3.29
DESCRIPTORS	Behavior Rating Scales; College Housing; *College Students; Comparative Analysis; Correlation; *Forced Choice Technique; Multiple Choice Tests; Peer Relationship; *Personality Assessment; Personality Tests; Response Mode; *Response Style (Tests); Self Evaluation; *Test Bias; Test Reliability; Tests; Test Validity
IDENTIFIERS	*Personality Research Form; PRF

ABSTRACT

In a comparative evaluation of a standard true-false format for personality assessment and a forced-choice format, subjects from college residential units were assigned randomly to respond either to the forced-choice or standard true-false form of the Personality Research Form (PRF). All subjects also rated themselves and the members of their residential units on behavior traits corresponding to the PRF scales. Reliabilities of the scales comprising the true-false form were substantially higher than those in the forced-choice form. Peer rating validities for the true-false and forced-choice forms were in a comparable range, but correlations with self-ratings were higher for the true-false form. Results do not support the contention that for personality scales a forced-choice format is consistently more valid than a standard format. Considering the other advantages of the true-false format, including its freedom from the complicating effects of ipsative scores, the use of this format is recommended for the great majority of applications in personality assessment. (Author)

RB-71-67

RESEARCH BULLETIN

AN EVALUATION OF FORCED-CHOICE AND TRUE-FALSE ITEM FORMATS IN PERSONALITY ASSESSMENT

Douglas N. Jackson
University of Western Ontario

John A. Neill
University of Guelph

and

Ann R. Bevan
Brock University

This Bulletin is a draft for interoffice circulation. Corrections and suggestions for revision are solicited. The Bulletin should not be cited as a reference without the specific permission of the authors. It is automatically superseded upon formal publication of the material.

Educational Testing Service
Princeton, New Jersey
December 1971

An Evaluation of Forced-Choice and True-False Item Formats in Personality Assessment

Douglas N. Jackson
University of Western Ontario

John A. Neill
University of Guelph

and

Ann R. Bevan
Brock University

Abstract

In a comparative evaluation of a standard true-false format for personality assessment and a forced-choice format, subjects from college residential units were assigned randomly to respond either to the forced-choice or standard true-false form of the Personality Research Form (PRF). All subjects also rated themselves and the members of their residential units on behavior traits corresponding to the PRF scales. Reliabilities of the scales comprising the true-false form were substantially higher than those in the forced-choice form. Peer rating validities for the true-false and forced-choice forms were in a comparable range, but correlations with self-ratings were higher for the true-false form. Results do not support the contention that for personality scales a forced-choice format is consistently more valid than a standard format. Considering the other advantages of the true-false format, including its freedom from the complicating effects of ipsative scores, the use of this format is recommended for the great majority of applications in personality assessment.

An Evaluation of Forced-Choice and True-False Item Formats in Personality Assessment[1,2]

Douglas N. Jackson
University of Western Ontario

John A. Neill
University of Guelph

and

Ann R. Bevan
Brock University

The primary purpose of the forced-choice technique in personality assessment, according to its adherents, is to reduce bias in response to items (Edwards, 1953, 1954; Gordon, 1951). In essence, the forced-choice method, as the term is employed in this paper, consists of pairing a self-descriptive statement pertaining to a personality trait with a trait-irrelevant filler statement having a very similar index of favorableness. The subject is asked to choose the statement which is more characteristic of himself.

Item parameters based on both desirability scale values and item popularities have been used as the favorableness index for matching purposes. Heineman (1953) and Edwards (1954, 1957), for example, have preferred matching items on the basis of desirability scale values, while Jackson and Payne (1963) preferred matching items on item popularities. The rationale for the former is that if a subject is forced to choose between items matched on desirability scale value, he cannot respond in terms of the desirability of items, and therefore is more likely to respond to the content of items. The result should be a reduction in the influence of desirability response style, increased resistance to faking, and consequent higher scale validity. Matching on item popularity, in addition to reducing the influence of response styles, has the added advantage that the expected popularity of each forced-choice item should be close to .50, no matter how extreme the popularities of the original statements were. Consequently, the matching procedures produce an increase in item and scale variance with a subsequent increase in scale reliability (see Magnusson, 1967, pp. 53-77). Because of the relatively high correlation between the item popularity and desirability scale value (Edwards, 1953), however, the two methods of pairing probably yield scales with similar properties, although the research has indicated that forced-choice scales often do have higher reliabilities than their nonforced counterparts; for example, Jackson and Payne (1963) reported that reliability increased from .81 to .96 when a forced-choice format was used instead of a standard single stimulus format.

On the subject of validity, the research indicates that neither nonforced nor forced-choice items have a clear advantage (cf. Borislow, 1958; Izard & Rosenberg, 1958; Krug, 1958; Longstaff & Jurgensen, 1953; Maher, 1959; Mais, 1951; Norman, 1963; Rusmore, 1956; Scott, 1968; Waters & Wherry, 1962; Winters, Bartlett, & Leve, 1965). Furthermore, it has become very clear that matching statements on desirability scale value does not prevent people from reliably judging one member of the forced-choice pair to be more desirable than the other (Corah, Feldman, Cohen, Gruen, Meadow, & Ringwall, 1958; Edwards, Wright, & Lunneborg, 1959; Feldman & Corah, 1960; Saltz, Reece, & Ager, 1962). Apparently placing statements in the forced-choice context accentuates subtle differences in the desirability of items (see Corah et al., 1958; Feldman & Corah, 1960; La Pointe & Auclair, 1961).

Although many studies have been undertaken to compare true-false and forced-choice item formats, they have been fraught with difficulties (Scott, 1968). One problem in assessing the existing research comparing true-false and forced-choice formats is that many of the instruments have been composed of unselected samples of items, a procedure which is hardly justified in view of modern developments in test theory and computer analysis (Neill & Jackson, 1970). Rather, recent recommendations would emphasize the usefulness of a variety of strategies for selecting items to maximize content saturation and minimize sources of bias. Thus, an appropriate investigation within this perspective would be to evaluate the advantages of a forced-choice format after self-descriptive statements had been carefully selected for content saturation and freedom from bias. If the role of desirability bias has already been greatly reduced in item selection, the question remains as to the extent to which the forced-choice format might enhance validity. Another problem that makes existing research difficult to assess is that comparisons have often been made between forced-choice and true-false scales that did not contain the same items. In such cases, differences between scales could at least partially be attributed to differences in samples of items. A further problem in assessing existing research is that the studies have often been conducted on a very narrow range of content, often on only one or two dimensions of personality.

The present study seeks to remedy these problems by comparing a set of forced-choice scales with a parallel set of true-false scales covering a large range of personality dimensions. For each dimension the statements are identical in the true-false version and in the forced-choice version. Furthermore, unlike many previous studies, which have limited their comparisons to reliability, the present study extends the comparisons between scales into the area of their validity with respect to behavior rating criteria. In the course of this investigation we will have occasion to examine properties of the behavior rating criteria, particularly the effects of degree of acquaintance of another person upon the validity and differentiation of ratings of that person.

Method

Subjects

Subjects, a total of 216 female university student volunteers, were drawn from 13 residential groups, each consisting of one wing of a large women's residence. Twenty-six women lived in each wing; the number of volunteer participants from each wing ranged from 13 to 23. Subjects in 12 of the 13 units had been living together for at least seven months, and in the remaining unit for three and one-half months.

Experimental Measures

Personality Research Form. Form AA (Jackson, 1967) of this personality questionnaire consists of 440 self-report statements yielding scores for 20 personality traits in the tradition of Murray (1938), as well as for two validity scales, infrequency and desirability. The standard instructions used in the present study indicated that the subject was to decide whether or not each item was characteristic of her, and then to answer true or false on a separate answer sheet.

In addition, a special experimental forced-choice form of the PRF (Form C) was constructed from the statements comprising Form AA to measure the same 20 traits. A statement from each of the 20 scales was paired with a second statement from one of the other scales, with the restriction that no more than two pairs of statements were comprised of statements from the same two scales. For almost every trait 19 of its statements were paired with statements representing the other 19 traits, one for each scale. Statements were paired on the basis of similar endorsement proportions, the difference between proportions of paired statements being in almost every case no greater than .02. Statements not paired with other keyed statements were paired with one of 50 irrelevant filler statements. Because of the special nature of the characteristics measured by the Infrequency and Desirability scales, positively keyed items for each of these scales were paired with negatively keyed items from the same scale. This procedure yielded a total of 247 item pairs. The subject was instructed to choose which statement of each pair was more characteristic of her, and to indicate her choice (A or B) on a separate answer sheet.

Although most statements representing a given personality dimension were paired with statements keyed on other scales, the item keying was not strictly ipsative. The analytical problems uncovered for ipsative measures (Radcliffe, 1963, 1965; Stricker, 1965) will thus not hold for these items. One of the purposes of the present investigation is to evaluate the extent to which partial ipsatization will allow analytical treatment of forced-choice results.

Behavior rating questionnaire. Subjects were requested to complete a behavior rating schedule with respect to 20 behavior traits on which they rated themselves, as well as every member of their residential group. Each trait was designated by an adjective and an accompanying definition selected carefully to represent each of the 20 scales on the PRF. The technique used was a refinement of one adapted by Jackson (1967), Jackson and Guthrie (1968), and Kusyszyn and Jackson (1968) from the work of Campbell, Miller, Lubetsky, and O'Connell (1964). A nine-point scale was used for all ratings, ranging from "nine" (extremely characteristic), through "5" (neutral), to "1" (extremely uncharacteristic). In order to appraise the effects of degree of acquaintance, a rating on a nine-point scale of how well a subject knew each member of her group was obtained, with a rating of "9" defined as knowing the resident "extremely well" and a "1" as "don't know her at all."

Procedure

Subjects within each residential group were randomly divided into two sets, the first to be administered PRF Form AA, and the second PRF Form C. Form AA was completed by 98 subjects and the forced-choice form by 118. Upon completion of the PRF, the behavior rating questionnaire was distributed and completed. The full session lasted about two hours.

Data Reduction and Analysis

The PRF data were scored in the usual fashion, by counting for each subject the number of responses in the keyed direction for a scale (Jackson, 1967). This yielded 20 content scores per subject. The 20 self-ratings per subject were in their final form, requiring no further reduction.

Reduction of the peer rating data was more complex, since each subject had rated 13 to 22 of her peers. Working with the data of one intact residential group at a time, a set of 20 mean peer ratings was computed for each subject.

The foregoing procedure produced 62 scores per subject, 22 PRF scores, 20 self-ratings, and 20 mean peer ratings. The 62 scores from the sample of subjects who took the PRF Form AA and from the sample who took Form C were intercorrelated separately to produce two multitrait-multimethod matrices (Campbell & Fiske, 1959).

For purposes of computing the reliability of the peer ratings, ratings pertaining to a given subject were alternately used in computing two additional mean ratings per trait. Essentially the new mean ratings so formed were random split-half mean ratings, based on two separate subsets of judges. Therefore, the correlation between them was corrected for double the number of raters by the Spearman-Brown formula, giving the reliability coefficient of the peer ratings based on its generalizability to a population of judges using a fixed rating scale.

Results

In the following paragraphs, the properties of the true-false and forced-choice scales are examined in the context of analysis of multitrait-multimethod matrices. Each matrix involves the measurement of 20 traits by each of three methods, the PRF, self-ratings, and peer ratings. A separate matrix was computed for the true-false sample and for the forced-choice sample. Comparisons are made between forced-choice and true-false scales in terms of the usual scale properties of internal consistency and convergent validity, but emphasis is placed on examination of discriminant reliability and validity. In addition, an examination is made of differences between forced-choice and true-false scale means. Finally, the effects on scale properties of degree of rater acquaintance with the ratee were examined.

Peer Ratings

In the present study peer ratings and self-ratings on the 20 traits corresponding to the 20 PRF content scales were used as criteria for assessing the relative validity of true-false and forced-choice scales. Therefore, it is appropriate to present the reliabilities of the peer ratings. The reliabilities, based on means for each subject derived from split halves of judges, were within an acceptable range, ranging from .58 to .92, with a median of .85 in one sample and .86 in the other sample.

Although reliabilities were substantial, the judges illustrated poor discrimination among the various traits. The fact that many of the mean peer ratings were highly intercorrelated indicates that the judges were basing their ratings on fewer than the 20 dimensions involved. The extent of the problem is illustrated by the fact that 10 per cent of the correlations among peer ratings were equal to or greater than .60. The discriminant properties of peer ratings were considerably improved when calculations were based only on the ratings of peers who indicated a higher than average degree of acquaintance with the ratee, as is indicated below. Nevertheless, the overall weak evidence for the discriminant reliability of the peer ratings should be borne in mind in considering the validity of the questionnaire data.

Comparison of True-False and Forced-Choice Scales

Summary statistics and reliability. When statements are paired on item popularity, the resulting forced-choice item has an expected popularity of .50. Therefore, the expected mean for a 20-item scale is 10. Furthermore, in a fully ipsatized set of scales, a given subject must have a mean score of 10 across scales. This combination of conditions, therefore, was expected to restrict the range of means for the scales in the partially ipsatized forced-choice version. In fact, Form C had a mean of 9.6 across scale means, close to the expected mean of 10.0, and a range of mean scores from 7.6 for Dominance to 12.8 for Nurturance. The true-false form, Form AA, had a mean across scales of 10.6, but had a larger range of mean scale scores, ranging from 5.2 for Aggression to 16.6 for Affiliation. The scales with the highest and lowest means in Form C were different from the scales with the highest and lowest means respectively in Form AA, although there was a substantial correlation between the two sets of means.

The KR-20 reliabilities for the 20 scales in Form AA and the 20 scales in Form C are presented in Table 1. For 19 of the 20 scales the reliability was higher for Form AA than for Form C. In Form AA the mean reliability was .75 with a range of .44 to .86 (somewhat lower than those reported for Form AA in the PRF Manual), while in Form C the mean reliability was .53 with a range from .39 to .71. The range of reliabilities is smaller for the forced-choice form, a fact probably attributable to the partial ipsatization procedure. The marked differences in reliability between the forced-choice and true-false forms bears very directly on the interpretation of differences between the validities of the respective scales.

Insert Table 1 about here

In order to interpret the reliabilities, one must look not only at the absolute size of the reliabilities, but at the size of the reliabilities relative to the correlations among the scales in the respective forms (Campbell & Fiske, 1959). For both forms there was a good degree of discriminant reliability. In Form C only two scales had reliabilities which were reached or exceeded by correlations with other scales. Form AA had none.

Relative validity of true-false and forced-choice scales. Table 1 lists the correlations between all scales and the corresponding peer ratings. Examination of Table 1 reveals that there were no clearcut differences in validity between the Form AA scales and Form C scales. For Form AA, 12 of the 20 scales were significantly correlated ($p < .05$) with peer ratings, while 15 from Form C were significantly correlated with peer ratings. In both Form AA and Form C the range of peer rating validities was from 0 to .54, with means of approximately .30.

The situation with self-ratings was slightly different. On Form AA all 20 scales were significantly correlated with self-ratings; on Form C, 19 scales were significantly correlated with self-ratings. The mean self-rating validity coefficients for Form AA and Form C were .47 and .35, respectively. Eighteen out of the 20 scales had higher self-rating validity coefficients on Form AA than on Form C.

It would appear that the two formats of the PRF are essentially similar in terms of the uncorrected validity coefficients found in this study, and that there is little basis for choosing one or another based on their ability to predict the behavior ratings. It should be remembered, however, that the reliabilities were lower for the forced-choice scales. It follows, therefore, that if the reliabilities of the forced-choice scales could be experimentally raised to equal those of the true-false scales, the forced-choice scales might be expected to be more valid than the true-false scales.

Analysis of Degree of Acquaintance

Respondents were assigned two sets of behavior rating scores for each of the 20 traits, one based on the average ratings received by that respondent from all judges whose rating of degree of acquaintance for this subject was above the mean rating of degree of acquaintance; and the second score was the mean rating of the judges rating this subject as below the mean in degree of acquaintance. These scores were then correlated with the corresponding 20 scores for the PRF. The resulting sets of correlations represented the convergent validities of the PRF scores, permitting a comparison of their relative validity for the two levels of degree of acquaintance. For Group I (Form AA), 18 of the 20 scales showed a higher PRF validity for judges high in degree of acquaintance ($p < .01$ by sign test), while for Group II (Form C), 17 of the 20 scales showed higher validities for the high degree of acquaintance judges ($p < .01$). PRF scales were divided into two groups of 10 on the basis of their mean validities in the present study, and the average validity coefficient was plotted as a function of degree of acquaintance. From Figure 1 it can be seen that the role of degree of acquaintance operates not only for scales showing substantial validity, but for scales showing lesser validity, this trend being equally apparent in the two distinct groups which were administered different forms of a personality questionnaire. These results suggest strongly that judgmental accuracy varies as a function of degree of acquaintance, and they lend credence to the hypothesis that behavior ratings are based on discriminant information about ratees.

Insert Figure 1 about here

There is another way in which degree of acquaintance might operate to affect peer judgments; by serving to simplify the factor structure of the monomethod correlations. The traits defining the PRF have generally not been found to intercorrelate excessively when measured by personality items. Nevertheless, the trait ratings showed many high intercorrelations. When the entire set of ratings was intercorrelated, a total of 40 exceeded the rather high value of .60. However, when just those raters indicating above average degree of acquaintance were separated for each subject, only 24 correlations in the matrix exceeded .60, suggesting that the simplification that usually takes place in judgments about personality seems to decrease when acquaintance is higher. This would seem to be at variance with the findings of Passini and Norman (1966), who found no greater differentiation among well-acquainted subjects.

One further analysis was undertaken, namely, an investigation of the extent to which ratings of degree of acquaintance by individual judges correlated with their ratings of substantive traits. Some rather dramatic correlations were uncovered, as, for example, a correlation of .73 between a rating of high degree of acquaintance and a rating of "sociable." The pattern of these correlations was such as to suggest that there was some systematic distortion in the ratings for substantive traits, depending upon the degree of acquaintance between the judge and the ratee. The pattern of relationships seemed to suggest further that the distortion was marked for some types of traits, but not for others. Acquaintance ratings were associated with trait ratings representing Affiliation and Exhibition, for example, but not Achievement. In order to test the hypothesis that such findings might be linked to the degree to which judges may tend to overestimate the presence of salient traits in individuals they know well, we separated the personality dimensions into two groups: those within the PRF correlating highly with the dominance scale, and those correlating less highly or negatively with dominance. These were designated <u>salient</u> and <u>nonsalient</u> traits, respectively. It should be noted that this separation, being based upon PRF intercorrelations, was entirely independent of the results obtained with the trait ratings.

Figure 2 presents the rather dramatic relationship between the salience of traits and their correlation with degree of acquaintance. It appears that

Insert Figure 2 about here

judges are very prone to attribute characteristics linked to sociability, play, dominance, impulsivity, and even thrill-seeking to individuals whom they know well, and to attribute the lack of these, or their opposites, to individuals whom they know less well. It is tempting to speculate that the causation might go the other way; that assertive individuals might be more likely to be well known. However, it should be remembered that these results pertain to every subject, and that the judges, not the subjects, have been distinguished in terms of degree of acquaintance, with scores for every subject based on the two sets of judges differentiated in terms of their acquaintance with him. Indeed, the mean degree of acquaintance for a particular subject was found to possess generally low correlations with heteromethod information about personality traits.

Discussion

Some important results emerged from the present study. Validities for Form AA and Form C were very similar, while scales on Form AA had higher reliabilities than did the corresponding scales on Form C. Form AA was found to be superior to Form C in predicting self-ratings.

This study was different from other studies in that the true-false and forced-choice scales being compared were composed of identical self-descriptive statements, while most previous studies have compared scales composed of different statements (Scott, 1968). Furthermore, unlike most previous studies, the present study examined scales covering a broad range of personality dimensions.

Scales on Form AA consistently showed higher reliabilities than the corresponding scales on Form C. The relatively lower reliabilities on the forced-choice form might be due to the fact that when two highly reliable statements with almost identical popularities are paired, a subject may choose one of these because of the salience of one statement or because of rejection of the other. For example, if an affiliation and an achievement statement were paired, a subject may have chosen the affiliation statement either because she considered the affiliation statement to be particularly self-descriptive or because she wished to avoid endorsing the achievement item. Therefore, the decision to choose or not choose the one alternative may be based at least partially upon irrelevant considerations, namely, the presence or absence of a second trait. Thus, if a subject in the example endorsed an affiliation item, not because of her level of affiliation but primarily because she wished to avoid endorsing an achievement item, this would add to the unreliability of the affiliation scale. The systematic pairing of reliable items from diverse scales would thus tend to attenuate the reliability of each scale. Of course, the procedure of requiring only one response to yield information about two items also reduces the reliability by essentially halving the number of item responses.

An alternative strategy for constructing forced-choice items would be to pair each statement with an irrelevant filler statement, but this would require twice as many statements as are contained in the true-false version, an extremely inefficient procedure. Yet another strategy would involve pairing two oppositely-keyed items from the same scale (Jackson & Minton, 1963). But this strategy, while avoiding acquiescence bias, would not ordinarily permit the incorporation of the major presumed advantage of the forced-choice procedure, namely, its suppression of favorability or communality bias. This is the case because it is not possible for most personality traits to develop item pools with symmetric distributions of desirability or popularity values around a neutral point for positively and negatively keyed items.

It was mentioned that the validity of the forced-choice scales might be improved if their reliabilities could be improved. It should be clear, however, that experimentally increasing the reliability of the forced-choice scales would be fraught with practical difficulties. This is not to say that the finding is not important. The forced-choice scales may indeed be more valid than the true-false scales in conditions where other factors might lower the reliabilities of the true-false scales; for example, in situations where subjects are prone to acquiesce. Another situation where true-false scales could be expected to be less reliable than forced-choice scales is in the measurement of psychopathological traits where the expected endorsement proportions of true-false items are very small or very large. In such cases where the popularities are extreme, the restricted item variance attenuates reliability. However, when such items are paired on popularity, the expected popularity of the resulting forced-choice item is .50. But it may not make good psychological sense to force highly skewed distributions into a normal distribution. At the item level pairing items from a hallucination scale and from a delusion scale would force a respondent to endorse one of these even if these dispositions were absent in his behavior.

Previous studies have found the true-false format to be more susceptible to desirability bias than the forced-choice format, a fact which may account for the higher correlations between Form AA scales and their respective self-ratings than the correlations between Form C scales and their respective self-ratings. Desirability bias was probably operating in the self-ratings of both samples. It is possible that desirability bias was operating in a similar manner in Form AA, while not operating, or operating to a lesser degree, in Form C. Thus desirability bias could account for the higher self-rating validities of the true-false scales.

The implications of the results bearing on degree of acquaintance are important. Degree of acquaintance in studies utilizing behavior ratings or peer judgments is a variable of critical importance, both for understanding the accuracy of these judgments, and for identifying a form of systematic bias in these judgments. This bias creates a distortion causing judges to ascribe certain kinds of traits to ratees with whom they are well acquainted. This form of method variance tends to be specific to judgments, and might therefore ultimately be useful as a suppressor variable, should validities be high enough to warrant the use of suppressors.

In conclusion, in the absence of clearcut evidence for superior properties for scales using one or the other item format, decisions must be based on other considerations such as the simplicity and the nonipsative nature of the true-false form. Thus, the true-false form will likely be the method of choice for some time to come.

References

Borislow, B. The Edwards Personal Preference Schedule (EPPS) and fakability. Journal of Applied Psychology, 1958, 42, 22-27.

Campbell, D. T., & Fiske, D. W. Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin, 1959, 56, 81-105.

Campbell, D. T., Miller, N., Lubetsky, J., & O'Connell, E. J. Varieties of projection in trait attribution. Psychological Monographs, 1964, 78, No. 15(Whole No. 592), 1-33.

Corah, N. L., Feldman, M. J., Cohen, I. S., Gruen, W., Meadow, A., & Ringwall, E. A. Social desirability as a variable in the Edwards Personal Preference Schedule. Journal of Consulting Psychology, 1958, 22, 70-72.

Edwards, A. L. The relationship between the judged desirability of a trait and the probability that the trait will be endorsed. Journal of Applied Psychology, 1953, 37, 90-93.

Edwards, A. L. Manual for the Edwards Personal Preference Schedule. New York: Psychological Corporation, 1954.

Edwards, A. L. The social desirability variable in personality assessment and research. New York: Dryden, 1957.

Edwards, A. L., Wright, C. E., & Lunneborg, C. E. A note on "Social desirability as a variable in the Edwards Personal Preference Schedule." Journal of Consulting Psychology, 1959, 23, 558.

Feldman, M. J., & Corah, N. L. Social desirability and the forced-choice method. Journal of Consulting Psychology, 1960, 24, 480-482.

Gordon, L. V. Validation of the forced-choice and the questionnaire methods of personality measurement. Journal of Applied Psychology, 1951, 35, 407-412.

Heineman, C. E. A forced-choice form of the Taylor Anxiety Scale. *Journal of Consulting Psychology*, 1953, *17*, 447-454.

Izard, C. E., & Rosenberg, N. Effectiveness of a forced-choice leadership test under varied experimental conditions. *Educational and Psychological Measurement*, 1958, *18*, 57-62.

Jackson, D. N. *Personality Research Form*. Goshen, New York: Research Psychologists Press, 1967.

Jackson, D. N., & Guthrie, G. M. Multitrait-multimethod evaluation of the Personality Research Form. In *Proceedings of the 76th Annual Convention of the American Psychological Association*, 1968, *3*, 177-178.

Jackson, D. N., & Minton, H. A forced-choice adjective preference scale for personality assessment. *Psychological Reports*, 1963, *12*, 515-520.

Jackson, D. N., & Payne, I. R. Personality scale for shallow affect. *Psychological Reports*, 1963, *13*, 687-698.

Krug, R. E. The effect of specific selection sets on a forced-choice self-description inventory. *Journal of Applied Psychology*, 1958, *42*, 89-92.

Kusyszyn, I., & Jackson, D. N. A multimethod factor analytic appraisal of endorsement and judgment methods in personality assessment. *Educational and Psychological Measurement*, 1968, *28*, 1047-1060.

La Pointe, R. E., & Auclair, G. A. The use of social desirability in forced-choice methodology. *American Psychologist*, 1961, *16*, 446. (Abstract)

Longstaff, H. P., & Jurgensen, C. E. Fakability of the Jurgensen Classification Inventory. *Journal of Applied Psychology*, 1953, *37*, 86-89.

Magnusson, D. *Test theory*. Reading, Mass.: Addison-Wesley, 1967.

Maher, H. Studies of transparency in forced-choice scales: I. Evidence of transparency. Journal of Applied Psychology, 1959, 43, 275-278.

Mais, R. D. Fakability of the Classification Inventory scored for self-confidence. Journal of Applied Psychology, 1951, 35, 172-174.

Murray, H. A. Explorations in personality. Cambridge, Mass.: Oxford University Press, 1938.

Neill, J. A., & Jackson, D. N. An empirical evaluation of item selection strategies in personality scale development. Educational and Psychological Measurement, 1970, 30, 647-661.

Norman, W. T. Personality measurement, faking, and detection: an assessment method for use in personnel selection. Journal of Applied Psychology, 1963, 47, 225-241.

Passini, F. T., & Norman, W. T. A universal conception of personality structure? Journal of Personality and Social Psychology, 1966, 4, 44-49.

Radcliffe, J. A. Some properties of ipsative score matrices and their relevance for some current interest tests. Australian Journal of Psychology, 1963, 15, 1-11.

Radcliffe, J. A. Review of Edwards Personal Preference Schedule. In O. K. Buros (Ed.), The Sixth Mental Measurements Yearbook. Highland Park, New Jersey: Gryphon Press, 1965. Pp. 195-200.

Rusmore, J. T. Fakability of the Gordon Personal Profile. Journal of Applied Psychology, 1956, 40, 175-177.

Saltz, E., Reece, M., & Ager, J. Studies of forced-choice methodology: individual differences in social desirability. Educational and Psychological Measurement, 1962, 22, 365-370.

Scott, W. A. Comparative validities of forced-choice and single-stimulus tests. _Psychological Bulletin_, 1968, _70_, 231-244.

Stricker, L. J. Review of Edwards Personal Preference Schedule. In O. K. Buros (Ed.), _The Sixth Mental Measurements Yearbook_. Highland Park, New Jersey: Gryphon Press, 1965. Pp. 200-207.

Waters, L. K., & Wherry, R. J., Jr. The effect of intent to bias on forced-choice indices. _Personnel Psychology_, 1962, _15_, 207-214.

Winters, S., Bartlett, C. J., & Leve, R. Instructional and response style factors with forced-choice response. Paper presented at the meetings of the American Psychological Association, Chicago, Illinois, 1965.

Footnotes

[1] Reprints are obtainable from Douglas N. Jackson, Department of Psychology, University of Western Ontario, London 72, Ontario, Canada.

[2] Supported in part from Research Grant No. 397 from the Ontario Mental Health Foundation. This paper was completed when Douglas N. Jackson was Visiting Scholar, Personality and Social Behavior Research Group, Division of Psychological Studies, Educational Testing Service, Princeton, New Jersey.

Table 1

Reliability and Validity of True-Fals.

and Forced-Choice Forms

Scale	True-False (N=98)			Forced-Choice (N=118)		
		Validity			Validity	
	KR-20	Self-Ratings	Peer Ratings	KR-20	Self-Ratings	Peer Ratings
Abasement	58	32	-05	47	19	10
Achievement	77	61	45	44	39	46
Affiliation	75	63	37	54	45	23
Aggression	71	43	21	42	37	21
Autonomy	67	48	41	53	36	32
Change	69	41	09	39	22	20
Cognitive Structure	76	24	10	40	36	06
Defendence	54	33	13	42	27	29
Dominance	79	48	41	68	42	54
Endurance	73	46	24	46	28	19
Exhibition	75	50	42	61	38	45
Harm Avoidance	84	55	31	71	52	46
Impulsivity	71	57	34	47	31	30
Nurturance	72	50	17	56	35	11
Order	86	78	54	66	63	50
Play	66	56	29	49	34	21
Sentience	44	28	17	54	32	14
Social Recognition	81	53	18	67	49	20
Succorance	72	45	43	61	43	18
Understanding	60	31	17	51	39	09

Note:--Decimals have been omitted. For Form AA, the .05 and .01 significance levels are .20 and .26, respectively; for Form C they are .18 and .23, respectively.

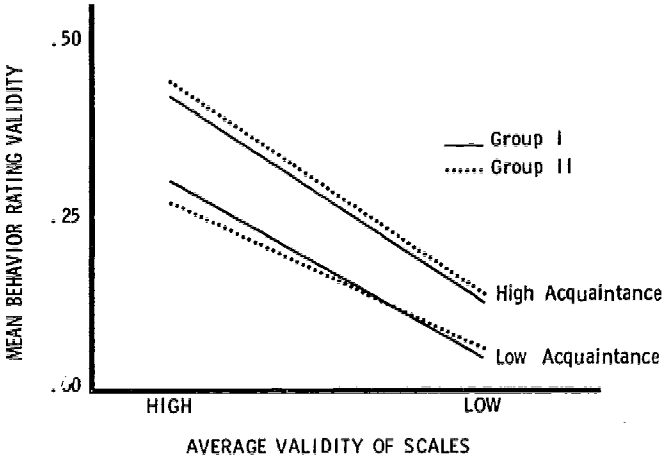

Fig. 1. Relation of degree of acquaintance of ratee to behavior rating validity of personality scales.

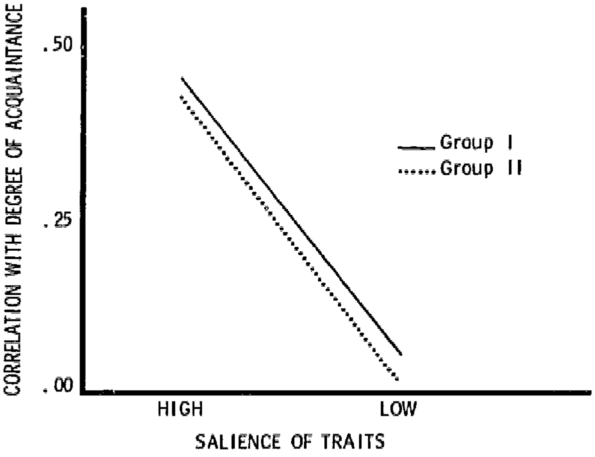

Fig. 2. Relation of rated degree of acquaintance with ratings for salient and nonsalient personality traits.