1991

1. Computer-Assisted Personality Test Interpretation: The Dawn of Discovery

Douglas N. Jackson
University of Western Ontario

Follow this and additional works at: http://digitalcommons.unl.edu/buroscomputerdecision
Part of the Educational Assessment, Evaluation, and Research Commons, and the Educational Methods Commons

http://digitalcommons.unl.edu/buroscomputerdecision/3

This Article is brought to you for free and open access by the Buros-Nebraska Series on Measurement and Testing at DigitalCommons@University of Nebraska - Lincoln. It has been accepted for inclusion in The Computer and the Decision-Making Process by an authorized administrator of DigitalCommons@University of Nebraska - Lincoln.
My aim in this chapter is to outline some of the substantive and psychometric bases on which we can build a science of assessment that takes advantage of the enormous potential inherent in the digital computer and in artificial intelligence. Some of these foundations are within the traditions of classical assessment. But others represent urgently needed areas of explicition and research.

It is my view, in the tradition of Cronbach (1954), that developers of computer software for testing should listen to what psychometricians say, and, as well, psychometricians should be sensitive to new research ideas waiting to be solved that arise out of the experience of preparing software for test interpretation. This is particularly true because some of classical test theory based on fixed sets of items is rendered obsolete by the prospect of adaptive testing. The fact that psychometricians and authors of interpretive software are rarely prone to listen to one another brings to mind a quotation from the world-weary French novelist and philosopher, André Gide, cited by Block (1978): “It has all been said before, but you must say it again, since nobody listens.”

**SOME PRECONDITIONS FOR VALID COMPUTER-ASSISTED TEST INTERPRETATION**

Accurate test interpretations depend on valid data. Stated another way, the validity of the score data set an upper bound for the accuracy of test interpretations. This sounds like such a truism as to appear almost trivial. But surprisingly little attention has been directed at this issue by those who write and write about computer software for test interpretation. For example, in a recent book devoted
to computer-based test interpretation (Butcher, 1987) there is scant attention directed at fundamental questions about the reliability of scores or indexes forming the bases for interpretations.

I would like to outline five preconditions for valid computer-assisted test interpretations and to discuss each in turn. These preconditions point both to the traditional wisdom of testing that can be incorporated appropriately into thinking about test interpretations, and, as well, to areas of needed research. Let me list the five: (1) Interpretations should, in general, be built around constructs of broad import; (2) Interpretations should bear an explicit substantive relationship to the constructs underlying the measures employed; (3) Where predictions are made about specific behaviors, both the reliability of the assessment data and the reliability of the criterion to be predicted should be taken into account; (4) The implications of evaluative biases both in the assessment situation and in outcomes need to be given explicit attention; and (5) Attention needs to be directed to base rates, both in the assessment situation and in outcome situations. I would like to discuss each of these points in turn.

The Usefulness of Personality Constructs

With regard to the importance of theory-based constructs, I do not know whether I should say a great deal or very little. There is a substantial literature in personality and social judgment bearing on this topic. But there is an unfortunate tendency for psychologists to consider new areas such as computerized test interpretation in isolation as if little were to be gained from treating it as part of a larger assessment endeavor. But there is something to be learned from the knowledge and controversies of personality and assessment. One of the most controversial issues in the personality literature over the past two decades is the question of whether or not there are broad personality traits or dispositions. One of the strongest advocates of the position that there are not is Walter Mischel, who has argued forcefully that what appear to be broad behavioral consistencies are in fact illusory. The evidence proffered in support of this position and its implications for computerized assessment warrant careful examination.

Mischel and Peake (1982) presented evidence that they believed failed to support the existence of broad traits of conscientiousness and friendliness. They intercorrelated behaviors purportedly representing each of these traits and interpreted mean intercorrelations of the order of .13 as evidence indicative of doubt about the existence of broad traits. But their analyses and interpretations are illustrative of the sort of ad hoc theorizing that is tempting when constructing computerized-based test interpretation systems. Mischel and Peake merely assumed that certain behaviors were linked to the traits of conscientiousness and friendliness without providing any explicit bases in the form of definitions or classification rules for their categorization. Nor did they fully consider the importance of aggregating data prior to inferring broadly based personality dispositions. Jackson and Paunonen (1985) undertook a reconceptualization and re-
analysis of the Mischel and Peake data on conscientiousness, distinguishing separate dimensions of studiousness, punctuality, and academic diligence by conceptual and empirical means. We estimated reliabilities for 20 behaviors relevant to our reinterpreted dimensions of .93, .95, and .86, respectively. A major import of these findings is that in drawing inferences about behavior from sample observations, the steps in construct validation (Jackson, 1971; Loevinger, 1957; Wiggins, 1973) do not only apply to tests, but apply equally to other formal and informal assessment situations, such as might be involved in combining behavioral “signs” in a computerized interpretation. The whole assessment procedure should be evaluated. A number of our conclusions (Jackson & Paunonen, 1985) have special relevance to automated test interpretations. First, in drawing an inference about a respondent based on the magnitude of a score representing a trait or disposition, a crucial aspect of construct validation is the explicit definition of traits and of situations, including their theoretical and empirical implications, and their differentiation from other related traits. Second, the structure of behavioral representations of traits and of different situations should be evaluated in a multidimensional framework. For example, if the bases for linking predicted behaviors to scores on a test is expert clinical judgment, it would be fitting to provide expert judges with a set of construct-based trait definitions and to instruct them to perform a multidimensional scaling of these traits and a larger set of predicted behavioral exemplars. Third, a crucial step in the appraisal of the predictability of behavior is its evaluation in a multitrait-multimethod context in which situations are also carefully defined and empirically studied. As an initial step in such an undertaking it is appropriate to employ scales or scores that possess appropriate levels of convergent and discriminant validity. If differential predictions are to be made on the basis of scale scores, or if profile shape is the basis for classification, it can be demonstrated that predictions or classifications will be more accurate if the constituent scales are minimally intercorrelated and discriminantly valid. This is often difficult to achieve because many measures of personality, particularly those of psychopathology, share a large common component reflecting general psychopathology or self-evaluation. The presence of such a large elevation component, while perhaps facilitating the classification of the person’s results into a global category of psychopathology, militates against accuracy in differential prediction, for example, of specific manifestations of psychopathology. The simple implication of the foregoing is that good automated test interpretation systems depend on good tests, a point to which I shall return.

Linking Interpretations to Constructs

The point that interpretations should bear a substantive link to the constructs underlying the measures employed, like the remaining points, can be considered as special cases of the first point on the importance of broadly based constructs. In the construction of personality tests, at least for those whose scales are de-
signed to represent constructs, items are selected so that they show higher degrees of association with the factor underlying their own scale than with those underlying irrelevant scales or response biases. It is reasonable to require that behavioral exemplars external to the testing situation show a similar pattern of association. I have already suggested that expert judgments might be used to evaluate the substantive links between constructs and external behaviors. Here I am suggesting that interpretations be validated empirically. It might be objected that some types of behavioral predictions, for example, “likely to engage in assaultive behavior when ridiculed,” are not easily evaluated empirically. This is true, but if one employs a conceptualization of constructs as encompassing domains containing related behaviors, then it is possible to sample relevant behaviors that are easier to elicit and manage under controlled conditions. Tendencies to engage in monetary risk taking, for example, might be assessed by observing the person gamble large sums of money in the real world, but might equally be represented by a person’s indicating a preference for the job of commodity trader, by volunteering for an experiment involving monetary risk, or by evaluating a person’s behavior when in the role of an economic decision-maker in an Internation Simulation to make or not to make long-term investments in research and development (Jackson, Hourany, & Vidmar, 1972). Thus, by a process of exemplar sampling, the underlying construct may be validated, and the validation may be generalized to other exemplars not actually observed.

Aggregation, Reliability, and Validity

The effects of aggregation on reliability have been recognized by psychometricians at least since the time of Spearman just after the turn of the century. Much recent literature has reminded us of this important requirement for assessment (Epstein, 1983; Rushton, Jackson, & Paunonen, 1981). But many psychologists—even those who write interpretive software systems—act as if this matter is only the concern of psychometricians. (An exception is Roy Schafer [1954] who cautioned that for Rorschach interpretation an important principle is that there should be “sufficient evidence” for the interpreted tendency, since Rorschach responses, like other responses, are multiply determined.) But aggregation and reliability also have implications for preparing automated test interpretations. For example, basing interpretive statements on responses to single critical items is fraught with error. If a 90-year span of experience with psychological testing has taught us anything, it is that individual episodic events are inherently difficult to predict. As exemplars of an item universe, they suffer from the possibility of being unrepresentative, unstable over time, and subject to error variance from a number of sources. Given the well-known relationship between predictor and criterion reliability and validity, validity inevitably will suffer if measures are not dependable. However, in many areas of psychological prediction we can produce very creditable results if the criterion that is being
predicted is aggregated. It follows then that interpretive statements are more likely to be accurate if reference is made to probabilities within a specified domain rather than if predictions of specific events are attempted. An aggression scale will predict aggression as a probabilistic series of events, but will not do well at allowing one to state with accuracy that person X will kick his or her dog on a certain day.

Evaluative Biases and Base Rates
In regard to taking into account base rates and desirability in preparing interpretative reports, I believe the situation is rather poorly understood in spite of the very extensive literature on the subject. But even though the situation is more complex than the first papers in this area in the 1950s and 1960s would have us believe, I do not think that it should be ignored. Psychometrically, there is a very serious problem if all or most scales in the psychopathy area correlate very substantially with a marker scale for undesirable responding. Ideally, personality scales should be developed in such a way as to avoid undue saturation with a general desirability factor. However, some item pools are so saturated with evaluative bias that it is very difficult to construct homogeneous scales that are free from desirability responding. For example, Reddon, Marceau, and Jackson (1982) found that five of six factors identified in an item factor analysis of the MMPI had items showing higher correlations with desirability scales than with their own factors, even on the derivation sample. Many people argue that psychopathology is inherently undesirable and the best way to deal with this problem is to ignore it. But since we now have capabilities for recognizing the multidetermined nature of psychological responses, it is possible to partition variance on scales into variance associated with content unique to the scale and variance associated with general factors such as those attributable to response bias. For example, multivariate regression procedures can be used to identify component scores with sources of response bias statistically removed and treated as a separate component score. Jackson and Reddon (1987) have recently shown that by transforming MMPI scale scores so that they are mutually uncorrelated, a new set of scores can be produced that are relatively free from desirability variance but nevertheless correlate substantially with the original scores. Even though raw scores have confounded content and stylistic variance, computer programs for interpreting scores can first unconfound these distinct sources of variance. Where desirability variance is elevated, for example, under conditions of impression management, appropriate statistical means are available to weigh this elevation in generating interpretations.

But desirability variance and variance associated with what Wiggins terms hypercommunality do not only represent invalid variance. Under certain circumstances knowledge of this from a respondent may increase one’s ability to predict accurately the respondent’s behavior. Indeed, although the “Barnum effect” of
simply making high base rates statements in an interpretative report is to be avoided, knowledge of how a particular respondent conforms with societal norms is useful in enhancing the accuracy of statements made about that person.

THE DAWN OF DISCOVERY

After paying homage to some traditional concerns in assessment as they apply to test interpretation, it is appropriate now to suggest some ways in which we can do better with computer-assisted test administration and interpretation. Again, I will focus my remarks on the personality assessment area, although many of these apply as well to other kinds of assessment.

I would like to review with you a few of the possibilities that are beginning to be realized in computer-aided test administration and interpretation. It is fortunate, I believe, that we are now in a position to go beyond the old traditions of testing. We can now avoid the mold of being constrained to a particular response format and a fixed set of items. I also see much hope in our potential for developing systems that transcend the human frailties of memory in, for example, only being able to distinguish a small number of types of personality or of ability constellations. I see at least five areas that show considerable potential: (1) branching; (2) the evaluation and use of explicit models for the processes underlying responding; (3) the development of more sophisticated methods for detecting invalid or nonpurposeful responding; (4) expansion in the use of different stimulus materials and response formats; and (5) the development and refinement of prototypes to aid in interpretation.

Adaptive Testing by Computer

Much has been written about adapting the difficulty level of items to the respondent's ability level as estimated from previous responses. It has been shown in the ability area that only approximately half the number of items is required to arrive at a level of reliability comparable with that of the longer scale. I am now happy to report that this finding also appears to hold even more strongly for personality scales in the area of psychopathology. Richard MacLennan, working in my laboratory, has been able to demonstrate that he can get 4 items to do the work of 20 if they are appropriately chosen to be consistent with the individual's level of psychopathology as measured by a particular scale. Of course, the method for branching depends on the question that one wishes to address. As long ago as 1969, if you can believe it, I undertook a study to see how few items were required to rule out the possibility that a given scale for psychopathology was elevated beyond two standard deviations. Our conclusion, at that time unpublished (I believed then that no one was interested in the result), was that four items were all that were required. Wayne Velicer (personal communication) has
informed me that he came to the same conclusion on mathematical grounds, although I have not seen his reasoning in this regard. This sort of finding raises interesting questions about the nature of the items and in what order they should be presented. Ideally we would like items that are highly differentiating, but, as well, items that have a sufficiently high level of variance that they provide useful information. In 1969 I developed an index to permit an optimal item ordering based on information derived from endorsement proportions and content saturation, but further empirical work is needed to show that this index indeed is optimal.

Whereas in the ability area branching has traditionally served to identify more accurately and more efficiently an individual's location on a single underlying dimension, the problem in the domains of psychopathology and of vocational interests is the question of which dimensions are descriptive of the person. Even for psychiatrically hospitalized individuals, most scales of psychopathology will reveal scores for most patients in the normal ranges. Of course it is inefficient to focus on areas that have little probability of yielding evidence of elevated scores for that person. Thus, branching can also operate hierarchically. I am now in the process of undertaking a large scale study of psychiatric patients, using an item pool of approximately 5,000 items and developing an algorithm to identify the best 300 to 400 items for the purpose of identifying critical dimensions for a particular individual. If the person, for example, responds to a general scale of somatic complaints, then it is appropriate to probe more deeply into areas such as hypochondriasis and imaginary symptoms and to seek to identify the focus of the somatic complaints, as well as to investigate related disorders, such as headaches proneness, dietary habits, health concern, loss of energy, and similar dimensions. For other people for whom there is little evidence of somatic concern, this area will be touched over lightly and the time can be used to probe more extensively in areas that are relevant to the person. This provides a basis for computer interpretative reports that are more relevant to the individual patient or respondent and more reliable. This is possible because items can more optimally be assigned to areas of greater concern.

Process Models and Response Latencies

Psychometricians have been accused, perhaps fairly, of studying response outcomes, namely black marks on answer sheets, to the exclusion of the processes entering into respondents' decisions. Latency data and explicit formulations of the response process provide a framework for investigating other facets of responding than the outcome. For example, Fekken and Holden (1988; Holden & Fekken, 1987) following up earlier work begun at the University of Western Ontario, have reported a series of studies investigating latencies for items with different characteristics. Long response times were associated with items in which responses prove to be unstable. One of the models investigated was the
threshold model for responding. This model involves an individual operating characteristic in which items are scaled for a particular characteristic and individuals show different levels of sensitivity to and threshold for responding in the keyed direction. As expected, latencies are greater for items near the individual’s threshold. Of special interest are the data related to the validity of latencies. For scales on which respondents receive high scores, they are quicker to endorse relevant items and slower to reject them. This finding holds also when an external criterion instead of the scale score is used. There is even evidence that latencies contribute incrementally to validities based on scale scores. Fekken and Holden are now investigating the use of latencies to items on particular scales to predict psychiatric classification with some very promising results. Another investigator working at the University of Western Ontario, Edward Helmes (1978), pursued this line of work with a multidimensional model employing content scale values and permitting the separation of response determinants due to general desirability and to content. The implications for computer-aided administration and interpretation are that these kinds of data may serve to enhance and corroborate data from traditional sources.

Identifying Nonpurposeful Responding

A number of approaches are possible for identifying records that contain nonpurposeful responses. One approach is to compute a kind of person reliability by summing an individual’s responses to odd-numbered items in a set of personality subscales and even-numbered items in the same set. This yields pairs of values consisting of odd and even responses to each of a number of scales. These may be correlated, using as \( N \) the number of scales. The resulting correlation coefficient may be interpreted as indicating the consistency to which an individual has responded over several scales. The individual reliabilities so obtained have a central tendency of about .85 for a well-constructed test and show excellent separation from responses that are generated randomly. A number of other techniques are possible for unobtrusive assessment of the consistency of responding, for example, in the correlation of an individual’s pattern of responses with frequency of endorsement values for each of a large number of items. Atypical response latencies might also be diagnostic of motivated distortion or random responding.

A Game-like Approach to Assessment

One nice feature of computer presentation is that one is not limited to stationary figures and the true–false response. At the moment we are doing two or three things in this area but perhaps the most interesting is the development of game-like stimuli which capture both the accuracy of judgment, speed of response, and
some psychomotor and perceptual skills (Jackson, Vernon, & Jackson, 1988). Our findings indicate that performance levels on such a task correlate as highly with general intelligence as do standard intelligence subtests while capturing new factors not measured by traditional IQ tests, one in which cognitive styles may become apparent.

Prototypes

Finally, there is the possibility of employing prototypes. We have conducted a series of studies using a technique called modal profile analysis in which similar profile types have been grouped analytically. Using such a procedure, we discovered that occupational group vocational interest profiles could be classified cogently—all physician groups formed one cluster, as did various types of salespeople, merchandisers, and educators. We extended this approach to alcoholics, psychiatric patients, university students, and military personnel, and found that whereas there was not one, but 16, alcohol personality profiles, many of these same types were also identified among the psychiatric patients and university students (Jackson, 1983). To investigate the degree to which these types were cogent exemplars of a class of people, we conducted a series of studies (e.g., Reed & Jackson, 1975) in which judges were asked to predict a pattern of responses to a particular type, described in a few sentences. Judges showed very high reliability. Then we identified a number of patients who had the characteristics described and asked our judges again to predict their pattern of responses. When components of the judgments were separated, and we took account of desirability and base rate, as well as content, judges proved to be highly accurate in their estimates. The implication is that knowledge of salient characteristics implies membership in a type, which, in turn, permits accurate identification of response probabilities. But not any old type will do. The evidence is that arbitrary types do not yield meaningful results.

Overview

With accelerating advances in computer technology, including the advent of touch screens, voice recognition, rapid access to massive stored data, and the like, we have the capability at hand to do justice to the complexity of personality in computerized interpretation. But to achieve this promise, our conceptualizations of personality, understanding of the process of responding, and implementation of this knowledge in computer software must keep pace. This is a large, labor-intensive undertaking, but if the dawn of discovery is to be realized, such implementation is essential.
REFERENCES

Loevinger, J. (1957). Objective tests as instruments of psychological theory. Psychological Reports, 3(Monograph No.9), 635–694.