*Reprinted from*
JOURNAL OF PERSONALITY ASSESSMENT
Vol. 36 1972 No. 6

# Some Recent Trends in Personality Assessment[1]

LEWIS R. GOLDBERG
University of Oregon and Oregon Research Institute

*Summary:* Some past and current research in personality assessment is discussed and evaluated, including such topics as (*a*) the exploitation of the natural language to construct descriptive personality taxonomies, (*b*) the search for trait-by-treatment interaction effects, (*c*) the relative validity of different strategies of personality inventory construction, (*d*) the comparative utility of nonlinear prediction schemes, (*e*) the quest for models of the judgments of expert decision makers, and (*f*) the utility of substituting judgmental models for human judges in applied contexts.

Individuals obviously differ in a host of ways, not all of which are of equal importance to themselves or to others. Tom can wiggle his ears like hummingbird wings, while elaborate electronic sensors detect no movement in Mary's. While Tom and Mary may fall at the top and bottom percentile of the world's population on ear-wiggling skill, this particular difference between them is not nearly as important as the fact that Mary is female and Tom is not. The first basic goal of psychological assessment is to identify the most important individual differences from the enormous variety which are manifested.

While most of us might agree that sex is a more important human difference than skill at ear-wiggling, it is far less likely that we will agree on the relative importance of many other sorts of individual differences. The concept of importance implies a value hierarchy, and scientists – no less than nonscientists – have widely differing ones. However, any evaluation of the relative importance of various individual differences ultimately depends on one's intentions. What do we want to *do* with our measures of these differences?

It is traditional in psychological discourse to distinguish between two apparently disparate uses of assessment tools. Assessment constructs (and the resulting instrumentation) are typically viewed as having rather unique significance either for the development of psychological theories ("basic" science) or for the forecasting of critical outcomes demanded by societal decision-makers ("applied" science). Fortunately, these two uses of assessment instruments are not as disparate as was once believed; today many psychologists make at least an implicit assumption that theoretically meaningful variables, when reliably measured, can be used to predict important societal criteria, and conversely that those individual differences which turn out empirically to be the most useful in the prediction of significant human outcomes are the variables which an eventual theory of individual differences will have to include.

Viewed in this light, then, it is apparent that trait discovery is intimately connected with test utilization – via the process of measurement. Once one has decided on the particular individual differences one wants to tap, one is faced with the more technical (psychometric) problem of measuring these differences in as precise and reliable a fashion as possible. The resulting measures must then be combined in some manner so as to generate nontest predictions, which in turn must be compared with expectations from some theory or validated against some set of external criteria. The results of these operations bear on the construct validity of the entire system: the choice of particular individual differences, the subsequent measurement of these differences, and the combination function which links the measures in some overall predictive expression.

[1] Invited lecture presented at the meetings of the American Psychological Association, Washington, D. C., September 1971, sponsored jointly by the Division of Personality and Social Psychology and by the Society for Personality Assessment. The preparation of this paper was supported by Grant MH 12972 from the National Institute of Mental Health, U.S. Public Health Service.

The field of psychological assessment, then, has three major goals: (*a*) the discovery of the most important individual differences, (*b*) the optimal measurement of these critical dimensions, and (*c*) the most effective utilization of the resulting measures for both theoretical and applied purposes. I will discuss some recent trends in personality assessment by focusing on each of these three goals in turn.

## Why Measure *That* Trait?

While the most critical question we can ask of any test constructor is why he chose to measure *those particular attributes,* it is a rare test manual that includes any rationale for this choice. Two psychologists, Raymond Cattell and Harrison Gough, stand out as the major exceptions to this generalization. Since each has propounded a general rationale for selecting the most important individual differences from the much larger total set, let us briefly consider their positions.

Both rationales share a core assumption, namely that *those individual differences which are of the most significance in the daily transactions of humans with each other will become encoded into the natural language as single-word trait-descriptors.* This powerful theoretical position, which was once propounded by Gordon Allport, has been stated more recently by Warren Norman:

> Attempts to construct taxonomies of personality characteristics have ordinarily taken as an initial data base some set of perceptible variations in performance and appearance between persons .... By far the most general efforts to specify the domain of phenomena on which to base such a system have proceeded from an examination of the natural language. The argument in its essential form has been that perceptible differences between persons in their characteristic appearance or manner of behaving ... become codified as a subset of the descriptive predicates of the natural language in the course of its development [Norman, 1963; p. 574].

That is, the "importance" of an individual difference is given operational definition as its probability of occurrence in the natural language. The more important the difference, the more will people notice it, wish to talk of it, and eventually invent a word for it. Since one can see this process exemplified in the evolution of nouns in the natural language (e.g., snow, of more importance to Eskimos than to Englishmen, has led to more words in Eskimo dialects than in English), one can assume that a similar process operates for adjectives, including those describing differences among individuals. Moreover, one might also assume that the more important individual differences within any linguistic culture would have more synonyms associated with them (thereby providing more nuances of meaning), and that these synonyms in turn would be shorter and more phonetic than those associated with the less important individual differences. To return to an earlier example, there is no single word in English which refers to "ear-wiggling skill," while there are a number of words which refer to gender.

While both Cattell and Gough have started with the natural language, their positions diverge from this point on. To Gough, the most important individual differences are those he calls "folk concepts," which he has defined as: "variables used for the description and analysis of personality in everyday life and in social interaction. It is theorized that such folk concepts, viewed as emergents from interpersonal behavior, have a kind of immediate meaningfulness and *universal relevance* ... [Gough, 1965; p.295]."

"The goal ... is to measure those traits of character which arise directly and necessarily from interpersonal life, and which should therefore be relevant to the understanding and prediction of social behavior in any and all situations and *in any culture* ... 'folk concepts' are *culturally universal* [Gough & Sandhu, 1964; p.544]." (Italics added) Thus, to Gough, "importance" gains operational currency by reference to the set of all natural languages: the more languages which have one or more trait-descriptive terms for a particular kind of individual difference, the more universal, and hence more important, is that difference. While

Gough did not utilize any cross-cultural linguistic studies in order to initially select the traits he included in his *California Psychological Inventory,* it is clear that this might have been a logical starting place for the development of such an inventory. On the other hand, of all inventory constructors, Gough has been among the most active in carrying out cross-cultural studies of his inventory scales, after their construction.

To Cattell, on the other hand, the natural language merely provides a starting point – rather than an ultimate destination. Allport and Odbert (1936) originally culled 17,954 trait names from *Webster's Second Unabridged Dictionary,* some 4,504 of which they deemed descriptive of "real traits." Cattell (1950, 1957) further reduced the set to 171 terms, by eliminating words which he considered synonomous with others in the set, and he collected peer ratings based upon these 171 terms. A cluster analysis of these ratings yielded 36 clusters or "surface traits," ostensibly the most important phenotypic individual differences in mankind. Factor analyses from a series of peer-rating studies led Cattell to the conclusion that between 15 and 20 distinct factors were necessary to account for the covariance among the surface traits; these latter became the "primary personality factors" in Cattell's taxonomic schema. Since these factors were themselves oblique, Cattell was able to continue factoring and thus arrive at four broad second-order factors. Consequently, one can view the distillation process as progressing from roughly 18,000 concepts to 4,500, and then from 170, to 35, to 15, and finally to 4.

Recently, the original Allport-Odbert list of trait descriptors has been expanded by Warren Norman, on the basis of a comprehensive survey of *Webster's Third Unabridged Dictionary.* A new pool of approximately 40,000 trait-descriptive terms has been identified and classified by Norman, and a subset of approximately 2,800 terms, hopefully providing comprehensive coverage of the stable and specific "biophysical" traits encoded in the English language, is presently under investigation (Norman, 1967).

In contrast to Gough and Cattell, most other test developers have never discussed their rationales for trait selection. Consequently, to uncover some of the rationales implicit in the work of other psychologists, I have recently reviewed the history of personality scales and inventories (Goldberg, 1971). Most of the published personality measures have been developed as a response to applied societal pressures, namely to forecast (*a*) personal or social adjustment, (*b*) satisfaction and success in vocational choice, or (*c*) academic achievement. Many of the remaining measures, which were less directly stimulated by applied demands, include those scales and inventories directed at two extraordinarily popular targets for structured measurement, namely (*d*) introversion-extroversion and (*e*) masculinity-femininity, and at two influential "theories" of individual differences, namely (*f*) Spranger's (1928) schema for classifying "personal values" and (*g*) Murray et al.'s (1938) classification of "manifest needs."

During the past few decades, the intuitive taxonomies devised by Freud, Jung, Rosanoff, Spranger, Murray, and Allport have been supplemented by the empirical schemas developed by Thurstone, Cattell, Guilford, Eysenck, Comrey, and Edwards. Moreover, future research based upon the Norman (1967) list of 2,800 trait descriptors should encourage another round of taxonomic ventures. However, if the future is anything like the past, new personality measures are at least as likely to be targeted upon constructs arising out of applied societal pressures as upon any new theoretical schemes.

In fact, the most potent factor in the determination of the targets for past personality measures has been sheer historical precedence. For better or for worse, psychologists have tended to utilize those constructs already identified by their predecessors. Moreover, this general "law of least effort" has also led inventory developers to borrow heavily from past item pools. Items devised around the turn of the century may have worked their way

via Woodworth's *Personal Data Sheet,* to Thurstone and Thurstone's *Personality Schedule,* hence to Bernreuter's *Personality Inventory,* and later to the *Minnesota Multiphasic Personality Inventory,* where they were borrowed for the *California Psychological Inventory,* and then injected into the *Omnibus Personality Inventory* – only to serve as a source of items for the new *Academic Behavior Inventory.* As a result of the widespread practice of item borrowing, there is substantial item overlap between a number of present inventories (one result of which is that convergent validity coefficients computed between scales from two inventories, generally lamented as being too low, may in fact be spuriously high).

Moreover, among those inventory developers who have eschewed past constructs or past item pools, another trend is equally clear. Each original individual difference has been gradually split into smaller and smaller constructs. As an example, Introversion-Extroversion was later divided into three components, one of which was social extroversion; the latter, in turn, has been fractionated into at least five components, one of which was dominance; and recently, dominance has been shattered into 30 to 40 "facets" by Butt and Fiske (1968, 1969). Adjustment was once just that – a single global construct; over the years, the construct has been shredded so finely that Jackson and Messick's new *Differential Personality Inventory* purports to measure some 28 varieties of maladjustment. Analogously, anxiety has been dichotomized into general anxiety and test anxiety, and the former, which was construed as five independent factors in the *16 Personality Factor Questionnaire* (16 PF), has more recently detonated into myriads of "person-by–situation interactions" in the hands of Endler, Hunt, and Rosenstein (1962).

This last effort reflects an evergrowing tendency to minimize and/or belittle the transsituational generality of individual differences (e.g., Mischel, 1968). The recent explosion of interest in the newest paradigm in the personality arena, namely social learning theory, has led some psychologists to switch "disciplines" – in the language of Cronbach's (1957) classic APA presidential address, from correlational psychology to experimental psychology – from a focus upon individual differences to a focus on situational influences on behavior. In the name of science, an enormous amount of poppycock has recently been expressed to the effect that (*a*) all behavior is "situational" in character, and/or (*b*) that psychometricians and/or trait theorists have never considered situational influences on human behavior. In fact, the classic psychometric position has been that situations "constrain" individual differences – that they profoundly affect both the mean and the variance of these differences, though the rank order of individuals on the "trait" should remain relatively invariant across those situations which permit sizeable trait variation to occur. The most extreme form of the social learning viewpoint not only posits that situations are moderator variables (affecting the rank order of individuals on the trait across situations) but that the correlations across individuals in all pairs of situations are near zero for all classes of behaviors. Such an extreme S-R viewpoint, which was decried by Cronbach (1957) and before him Dashiell (1939), seems patently absurd in 1971.

Cronbach (1957) has probably expressed the issue most articulately:

> A true federation of the disciplines is required. Kept independent, they can give only wrong answers or no answers at all regarding certain important problems. It is shortsighted to argue for one science to discover the general laws of mind or behavior and for a separate enterprise concerned with individual minds, or for a one-way dependence of personality theory upon learning theory [p.673].
>
> In both applied work and general scientific work, psychology requires combined, not parallel, labors from our two historic disciplines. In this common labor, they will almost certainly become one, with a common theory, a common method, and common recommendations for social betterment [p.683].

While the arguments advanced by Cronbach have inspired others to begin the search for trait-by-treatment interac-

tion effects, the journey has barely begun. For example, in my own review of literally hundreds of adjustment scales and inventories (Goldberg, 1971), I could uncover very few instances in which adjustment measures were explicitly constructed to predict differential responses to different treatments – ostensibly the *raison d'etre* for clinical diagnosis. Moreover, of the dozens of vocational inventories developed over the years, few – if any – have been focused on the prediction of differential responses to different types of work settings. It is only in the area of scholastic prediction – and then only recently – that psychologists have begun to develop measures to produce trait-by-treatment interaction effects (Siegel & Siegel, 1964, 1965, 1967). And in this type of setting, the road has proved incredibly rocky. Recent reviews of trait-by-treatment interaction research by Bracht (1969), Cronbach and Snow (1969), and Goldberg (1972b) have revealed very few replicated effects.

In one large-scale investigation of trait-by-treatment interactions in a college setting (Goldberg, 1972b), we found that for four different instructional treatments and three different classes of criteria, when the best of the general predictors were compared with the best of the differential predictors, the general predictors produced by far the largest effects. That is, if one were to make differential predictions for these three criteria using the most significant interaction effects, in no case would one's resulting predictions be as valid as simply using a single general predictor and thus ignoring all experimental variations in teaching methods. About twice as much criterion variance was predictable by the best of the general predictors as by the best of the differential predictors, in spite of the fact that the most powerful general predictors of one of the criteria were not included in these analyses. These poignant findings suggest that the significant interactions we discovered in this project – even if replicated at the very same strength in future studies – are unlikely to lead to differential predictions which are more valid than those achievable by general predictors alone. Clearly, one of the most tantalizing problems in the field of psychological assessment is the discovery of those traits and those situations which do produce reliable interaction effects. In the interim, Cronbach's clear vision of a "common theory" still appears hyperopic.

### *How* Measure That Trait?

Let us turn now to the second major goal of personality assessment, the optimal measurement of those individual differences previously discovered or conceptualized. The assessment enterprise differs from such other endeavors as astrology, palm reading, and tea-leaf gazing – all of which also try to predict important human outcomes – by its reliance on scientific methods of verification, and by its use of samples of human behavior as raw data. The constraints placed on such behavior by the psychologist can obviously vary from none to high. At one extreme of this continuum are all relatively "unobtrusive" measures (Webb, Campbell, Schwartz, & Sechrest, 1966), including behavior observations and sociometric techniques, while at the other extreme are highly structured tests, scales and inventories. In the middle of the distribution are the procedures most favored by many clinicians, namely the interview and the projective techniques.

One important recent trend in the personality assessment literature has been an enormous shift in emphasis away from the middle and towards the two extremes of this distribution. Specifically, a host of studies stemming from the "social learning" paradigm, as well as those from the "ecological" movement stimulated by Roger Barker and his students, have redirected research attention to *L*-data (Cattell, 1957), to behavior which is relatively unconstrained by any interventions of the psychologist. And, at the other extreme, there has been a rapid proliferation of research on structured scales and inventories (Goldberg, 1971). While interviews and projective techniques have been gradually losing favor among psychometric researchers (Buros, 1970; Damarin, 1971), if not among clinical practitioners,

this trend probably represents a reaction to the transducer involved. For raw behavior, of whatever type and under whatever degree of constraint, must be transduced – either automatically or through some human cognitive processing – before measurement may be said to have occurred. In the coding or scoring of behavior observations, interviews, and projective techniques, another man operates as the transducer; in personality scales and inventories, the transduction is automated, and man is not needed in this role. Some problems endemic to the use of man as transducer have been widely discussed of late, and I will return to this issue in a later section of this paper.

In a sense, all psychometric problems can be divided into two types, namely those concerned with the specification of optimal test *stimuli* or the most appropriate classes of *responses* to be recorded or coded, and those concerned with the selection of an optimal strategy for grouping these responses to produce an aggregate test or scale *score*. At this point, I will focus solely on the second class of psychometric problems, namely those concerned with strategies and tactics of personality scale construction.

While various strategies of scale construction have proliferated over the years under a number of names, they can all be divided into three main types, which can be labeled *External, Internal,* and *Intuitive.* The *External* strategy derives its name from the fact that some nontest reference groups are used to determine an item's scale membership and direction of keying, and this strategy has consequently been referred to as the "criterion-group" or "empirical" strategy. The second major strategy of inventory construction has been labeled *Internal,* since the internal structure of the item pool is the sole determiner of an item's scale membership and its direction of keying. The third major strategy of inventory construction, labeled *Intuitive,* relies on the cognition of the test developer for judgments regarding the suitability of an item for inclusion (and direction of keying) in a scale. While the earliest Intuitive scales were constructed from the judgments of a single individual, more recent ones have sometimes been based on the pooled judgments of a number of individuals, in an attempt to attenuate the idiosyncratic features of any single judge. Moreover, recent proponents of this general strategy of inventory construction – which has often been referred to under such labels as the "rational" or "theoretical" approach – have tended to use a mixture of two strategies, typically beginning scale construction by the intuitive assembly and keying of items, then later "purifying" the resulting scales through internal consistency analysis (e.g., discarding items with low, or negative, correlations with a priori scale scores). Let us look at the origins of these and other strategies.

The earliest means for gauging the extent to which an individual manifested some phenotypic trait was to ask him for a self-estimate. Since the form of the question and the conditions under which the question was asked might affect the reply, psychologists began to develop rating scales in order to standardize the process of self-estimation. However, early investigators quickly noted some characteristics of self-ratings that appeared to limit their usefulness. In the first place, such ratings turned out to be only moderately reliable when the same individuals were assessed on two or more occasions. In addition, it seemed likely that subjects might have difficulty estimating their status on any rather complex or global trait, since they would not know how much to weight each of the trait elements in order to arrive at a composite rating; moreover, it is probable that individuals would differ in the weights they assigned.

To solve these problems, early investigators attempted to break up the self-rating task into more molecular units, and the Intuitive scale construction strategy was born. The burden of proof that the scale measured the trait fell squarely on the shoulders of the test constructor, who ideally would have to demonstrate that *(a)* all of the items in the scale were related to the trait, *(b)* no set of items tapping important elements of the trait were not included in the scale, and *(c)* the method of combining or weighting items to ob-

tain a scale score was appropriate for the trait (Loevinger, 1957).

By the late 1930s, a number of psychologists began to argue that psychology had not yet reached the stage where trait relevance could be reliably and validly intuited. Therefore, they argued, only the empirically-determined effectiveness of each item should legitimately influence the decision as to whether it belonged in a scale. Moreover, if one could locate two groups of subjects, each of whom logically could be seen as falling at one of the two poles of a trait, then the differential item response frequency of the two criterion groups could provide a nonsubjective index of item validity (Meehl, 1945). So was born the External strategy, and with it two of today's most popular personality inventories, the MMPI and the CPI.

Over the years, personality scales began to proliferate so hardily that they threatened to outnumber the available supply of people. Clearly, some method of birth control seemed called for, and factor analysis appeared to some psychologists as the final solution to this problem. For example, Cattell (1950, 1957) warned that there was virtually no limit either to the number of traits that personality theorists could invent or to the number of criteria psychologists might be asked to predict, and, therefore, that if a separate scale had to be devised for every trait and every criterion, the public would be swamped in a sea of test booklets. To solve this dilemma, Cattell proposed a systematic search for the most salient and important individual differences in mankind. Cattell's goal has been to provide a comprehensive battery of factor scales which could be used empirically via multiple-regression techniques to predict any trait or criterion of interest.

In assessment controversies, as in wars, there are always hostile armies waiting to ravage both sides, and in the 1950s a new force entered the fray. Many psychologists have long had a dim view of personality scales constructed by any strategy, since all self-report measures seemed too easily amenable to various forms of impression management. While test constructors have long sought methods of controlling dissimulation and image enhancement (e.g., Meehl & Hathaway, 1946), only recently have such tendencies – now reconceptualized as "social desirability response set" – been considered to account for the major portion of the variance in Intuitive, Internal, and External scales (e.g., Edwards, 1957). What scale variance remains has been viewed as being largely determined by another such bias, namely "acquiescence response style" (e.g., Jackson & Messick, 1958). As one might guess, it was not long before some investigators sought to measure these putative biases directly (e.g., Jackson & Messick, 1961, 1962) – and so was born the Stylistic strategy of scale construction.

While no one has yet argued for the use of Stylistic scales as direct predictors of important societal criteria, their proponents have advocated the elimination of various types of response bias, either during the original scale construction process (e.g., Jackson, 1971) or through the addition of Stylistic scales as potential suppressor or moderator variables in prediction functions which include scales constructed by other strategies. Interestingly, the seemingly plausible hypothesis that the use of Stylistic scales might improve the validity of other measures, either in a suppressor or a moderator role, has never been confirmed. In fact, none of the investigators of this issue has as yet discovered any Stylistic scale which generally served either as a suppressor variable in multivariate prediction functions (e.g., Dicken, 1963; Goldberg, Rorer, & Greene, 1970), or as a moderator of the validity of other sorts of personality scales (e.g., Goldberg, Rorer, & Greene, 1970).

While the relative merits of the various strategies have been heatedly argued over the years, only recently have any empirical comparisons among strategies been reported (Butt & Fiske, 1968; Hase & Goldberg, 1967). In my own comparative validity project (Goldberg, 1972a), five strategies of scale construction were used to construct nine different 11-scale "inven-

tories" from the CPI item pool, five inventories based on the three major strategies (*External, Internal,* and *Intuitive*) and four inventories based on two control strategies (*Stylistic* and *Random*). The average cross-validities of each inventory were compared across 13 criterion indices. My findings suggest that while the inventories constructed by the three major strategies produced quite similar average cross-validities, there was a sizeable criteria-by-strategies interaction effect. Specifically, the External strategy appeared to produce a broader band-width but lower fidelity inventory than did either the Internal or the Intuitive strategies. However, a subset of five Rational scales provided average cross-validities at least as high as those produced by any of the other strategies and tactics under study.

This later finding suggests that some of the strongest criticisms of the Intuitive strategy may have been unfounded. For, of the three major strategies, it is only the Intuitive which does not capitalize on sample-specific characteristics, and it may be for this reason that the Intuitive inventories performed as validly as they did in our project. That is, the very characteristic of both the External and Internal strategies which gives them their power also provides their Achilles heel: namely, their dependence upon – and vulnerability to – characteristics of the particular samples used in their construction. On the other hand, the validity of inventories constructed by intuitive procedures is dependent upon the wisdom of the particular judge, or the sample of judges, used to construct the scales. In the past, the sampling of judges has generally been considered to be more crucial than the sampling of subjects, and thus the Intuitive strategy has lost some favor in the psychometric community. One of the main lessons from recent assessment research may be that such judgmental biases are not as critical as has previously been believed.

In an important theoretical article, Jackson (1971) has forcefully made this point by issuing the following provocative challenge:

> For any trait for which substantive definition is possible, let the most elaborate empirical item-selection procedures using criterion groups be pitted against two hours of work by a couple of good item writers . . . . One might extend this challenge even further. It might even be possible to use unselected item writers. It might be interesting, for example, to have an introductory class of psychology students write one item each with regard to a defined dimension, with perhaps just a bit of screening for substantive cogency and clarity of style, and conduct the comparison on that basis. The comparison proposed would be, of course, that of the empirical validity against a criterion relevant to the construct in question. The author would fully expect under cross-validation that even an inexperienced item writer would be superior to empirical item selection with a typical heterogeneous item pool [pp. 237-238].

While the findings from my own comparative validity project should not encourage those empiricists who would leap to take up Jackson's gauntlet, it is important to realize that Jackson's challenge was specifically directed at scale *fidelity,* and that he made no specific claims for *band-width.* Yet, the most surprising aspect of my own work is the finding that the External strategy produced an inventory of slightly broader band-width than those produced by either the Internal or Intuitive strategies. This tantalizing finding is, at first blush, majestically counter-intuitive. One might well predict that inventories produced by the External strategy should be relatively valid solely for those target criteria used to develop the scales (and thus be relatively narrow in band-width across a range of nontarget criteria), while the more homogeneous and independent sets of scales produced by the Internal and Intuitive strategies should possess relatively wider band-width when these scales are combined via multiple-regression procedures. Yet, our findings suggest that the less homogeneous scales produced by the External strategy may include some personologically relevant variance which is not included in those constructed by the Internal and Intuitive strategies, and that this type of variance may permit slightly higher cross-validities against precisely those criteria

which are generally the least predictable. Clearly this finding now demands replication in other settings.

## How *Use* That Measure?

Measures of individual differences are developed in order that they be used. Specifically, just as item responses must be amalgamated to produce scale scores, so scale and test scores must be combined in some manner to generate optimal predictions for the multidimensional types of criteria psychologists are typically called upon to forecast. And, in general, this score combination process can be carried out in one of two ways, either actuarially (i.e., mechanically, statistically) or via the use of a human as an information processer. As virtually all psychologists are now aware, the relative merits of these two modes of data processing have been hotly debated of late, and the resulting "clinical vs. statistical prediction controversy" (e.g., Gough, 1962; Meehl, 1954; Sawyer, 1966) has produced a flurry of recent experimental studies.

I can summarize this ever-growing body of literature by pointing out that over a rather large array of judgment tasks, rather simple actuarial formulae have typically performed at a level of validity no lower than that of the human expert. Consequently, it now seems safe to assert rather dogmatically that when acceptable criterion information is available, the proper role of the human in the decision-making process is that of a scientist: discovering or identifying new cues which will improve predictive accuracy, and constructing new sorts of systematic procedures for combining predictors in increasingly more optimal ways. Let us now examine some recent research bearing on, in turn, (*a*) actuarial models, (*b*) human judgments, and (*c*) the amalgamation of these two data processing modes.

### *Actuarial Models*

One of the most central questions to be addressed by investigators in applied settings concerns the nature of the mathematical prediction function they will utilize. Specifically, they must ascertain whether some nonlinear or configural function will provide more valid predictions than the classical linear regression equation, and if so, which of the many varieties of nonlinear functions should be so employed. Over the years, a number of nonlinear and configural techniques have been proposed for psychometric use, but only rarely has anyone assessed their incremental utility over the linear model. I conducted one such comparison a few years ago (Goldberg, 1969). It was inspired by a prediction by Paul Meehl (1956) that the relationships between MMPI scores and the diagnostic classification of psychosis vs. neurosis should be highly configural in character, and, therefore, that no linear combination of MMPI scores should be able to differentiate neurotic from psychotic patients as accurately as configural actuarial techniques. After ten years of research on this question, I can now assert that neither moderated regression analyses, profile typologies, the Perceptron algorithm, density estimation procedures, Bayesian techniques, nor sequential analyses – when cross-validated – have been able to improve on a simple linear function.

However, one might justifiably remain skeptical of any single set of empirical findings, if they stood alone. Yet, over the past few years a number of extensive, systematic, and methodologically sophisticated attempts to uncover configural relationships between predictors and *other* criteria have been reported, and virtually all of them have presented a similar tale. For example, Stilson and Astrup (1966) reported a comparison between linear and nonlinear methods for long-term prognosis among psychotic patients. They used large samples, good clinical data, and reasonably clean criteria. They concluded,

> The gains resulting . . . from the use of a nonlinear procedure are almost entirely lost in cross-validation. This indicates that the simple additive formula based on the number of symptoms will prove to be about as good as the nonlinear procedures . . . if prognoses are to be made for new patients [p.472].

This study should be examined carefully, since pattern analysis, profile cod-

ing, and other nonlinear classification techniques are presently so fashionable in clinical circles.

As one further example, Lunneborg and Lunneborg (1967b) reported a study of a randomly-selected group of 3,000 high school seniors. A series of aptitude, interest, and achievement tests was used to predict academic success in four college areas – English, mathematics, foreign language, and the physical sciences – as well as a dichotomous criterion, satisfactory vs. unsatisfactory progress toward a degree. A number of sophisticated pattern-analytic systems were investigated. The Lunneborgs concluded their report with the following disturbing passage:

> The failure of pattern information to aid prediction confirms the negative results of earlier attempts to use patterns and is all the more poignant a failure because of the attention paid to the selection of differentiated criteria and relevant predictors. There are even well-developed ideas throughout the educational literature as to the different configurations of abilities intuited behind different achievement criteria. In response to similar speculations in the counseling literature regarding patterns of personality needs associated with academic achievement, a study of reliable, frequent EPPS need patterns demonstrated the same lack of predictive stability (Lunneborg & Lunneborg, 1967a). Given the content similarity in the present study between predictors and criteria, the use of only reliable patterns, and the refinement of criteria, there would seem to be small room for continuing the conjecture that patterns can go above and beyond prediction from simple linear functions of original variables (Lunneborg & Lunneborg, 1967b, [p.951]).

One of the great challenges of future assessment work must be to show how – and under what conditions – these conclusions are wrong.

*Human Judgment*

Interestingly, the picture remains much the same when we turn to recent research on judgmental processes. Just as the linear model has proved remarkably robust as an actuarial tool, so the same model has proved equally powerful in predicting, or representing, human judgments themselves. Over the years, the research focus among judgmental investigators has changed dramatically, from the early studies of judgmental accuracy (e.g., Holtzman & Sells, 1954) to more recent attempts to simulate (or "model" or "capture the policies of") professional decision-makers (Goldberg, 1968).

An investigator of the clinical judgment process might express his aims through the following question: By what mathematical model can one use the data available to a judge so as to simulate most accurately the judgments he actually makes? To answer this question, one must (*a*) discover some formal (i.e., specifiable) model, which (*b*) uses as its "input" the information (data, cues, symptoms, etc.) initially presented to the judge, and (*c*) combines the data in some optimal manner, so as to (*d*) produce as accurately as possible a copy of the responses of the judge – (*e*) regardless of the actual validity of those judgments themselves. Note that such a model is always an intraindividual one; that is, it is intended as a representation of the responses of a single judge, and the test of the model is how well it predicts these judgments.

What sort of judgmental model should one try? Since introspective accounts describe the judgment process as curvilinear, configural, and sequential (e.g., McArthur, 1954; Meehl, 1954, 1960; Parker, 1958), one possible strategy is to begin with fairly complex models, perhaps with an eye to seeing how they may eventually be simplified. The research of investigators at two major centers for research on human inference – Oregon Research Institute and the Behavior Research Laboratory of the University of Colorado – has proceeded from a diametrically opposite strategy (Hammond, Hursch, & Todd, 1964; Hoffman, 1960), namely to start with a linear regression model and then to proceed to introduce complications only so far as is necessary to reproduce the responses of a particular judge.

Since experts generally describe their cognitive processes as complex ones involving the curvilinear, configural, and sequential utilization of cues, one might ex-

pect that the linear model would provide a rather poor representation of their actual judgments. Consequently, we might anticipate the need to introduce new terms into the model to represent these more complex processes. While the introduction of such terms can never serve to decrease accuracy in the sample of judgments used to derive the regression weights, these extra terms may simply serve to explain the vagaries of the particular judgments from the derivation sample and thus can severely attenuate the accuracy of the resulting model upon its cross-validation in another sample of judgments. However, when the judge is actually using the cues in a curvilinear or in a configural manner, then the introduction of the mathematical approximations of these processes should serve to improve the model.

In study after study, however, the accuracy of the linear model has been at approximately the same level as the reliability of the judgments themselves, and – no doubt because of this – the introduction of more complex terms into the basic equation has rarely served to increase the cross-validity of the more complex model. Hammond and Summers (1965) and Slovic and Lichtenstein (1971) have reviewed a series of studies in which the same general finding has emerged: for a number of different judgmental tasks and across a considerable range of judges, the simple linear model appears to predict the judgmental responses quite adequately, in spite of the reports of the judges that they are using cues in a highly configural manner. This is not to say that human judges behave like linear data processors, but only that the power of the linear regression model is so great that it serves to obscure most of the configural processes in judgment.

*Amalgamating Man and Machine*

There is an old psychometric axiom that validity is constrained by reliability, specifically that the correlation between a measure and any criterion cannot be higher than the square root of the reliability of that measure. While we routinely use that axiom to guide our thinking about test construction procedures, the same axiom can also be applied to the predictions from human judgments. For we know that the expert is not a machine. While he possesses his full share of human learning and hypothesis-generating skills, he lacks the machine's reliability. He "has his days." Boredom, fatigue, illness, situational and interpersonal distractions all plague him, with the result that his repeated judgments of the exact same stimulus configuration are not identical. He is subject to all those human frailties which keep the reliability of his judgments below unity. And, if the judge's reliability is less than perfect, there must be error in his judgments – error which can serve no other purpose than to attenuate his accuracy. If we could remove some of this human unreliability by eliminating the random error in his judgments, we might thereby increase the validity of his predictions. The problem, then, is to separate the expert's judgmental unreliability from his – hopefully, somewhat valid – judgmental strategy.

As I have already noted, ten years of research on the judgment process have demonstrated that for many types of common clinical decisions and for many sorts of expert judges, a simple linear regression equation can be constructed which will predict the responses of a judge at approximately the level of his own reliability (Hoffman, 1960; Hammond, Hursch, & Todd, 1964; Naylor & Wherry, 1965; Goldberg, 1968). While the regression model has been utilized (probably inappropriately) to explain the manner in which experts combine cues in making their diagnostic and prognostic decisions (Green, 1968; Hoffman, 1968), there is little controversy about its power as a predictor of these judgments. In addition, of course, such a model possesses at least one asset which humans typically lack: perfect reliability.

Now, how would such models fare as predictors themselves? That is, if the set of regression weights generated from an analysis of the judgments of an expert were used to make predictions for each target individual, would these predictions be more valid, or less valid, than the origi-

nal judgments from which the weights were derived? To the extent that the model fails to capture valid nonlinear variance in the judge's decision processes, it should perform less creditably than the judge. To the extent that it eliminates the random error component in human judgments, it should perform more validly than the judge. Which of these counteracting factors is more important in typical clinical decision-making? Can we construct a mathematical representation of a judge – without any recourse to criterion information – which is more valid as a decision maker than the human we have used as a model?

Fortunately, the answer to this last question appears to be "Yes." I described the mathematics of the situation in a recent paper (Goldberg, 1970), and presented the results of an illustrative study comparing man and his model. Specifically, I compared the validity of 29 clinical psychologists with their linear models as predictors of the diagnostic classification, psychosis vs. neurosis, from the MMPI profiles of 861 psychiatric patients. I found that, in general, models of the judges were slightly more valid than the judges themselves. Moreover, this slight incremental validity of model over man persisted even when the models were constructed on a small set of cases, and then man and model competed on a completely new and much larger set. Dawes (1971) has extended the generality of these conclusions by demonstrating that a linear model of the decisions made by the graduate admissions committee of a university psychology department is a more valid predictor of graduate achievement than the committee's decisions themselves. An even more impressive demonstration of the same phenomenon has been reported by Wiggins and Kohen (1971), who asked 98 graduate students in psychology to predict the grade point average of 100 other psychology graduate students. The average judge obtained a validity coefficient of .33 on this task, as compared to a value of .50 for the average model. Moreover, for every one of the 98 judges, the model of the judge was more valid than were the judgments themselves!

Proponents of actuarial prediction have repeatedly emphasized the lower cost of such procedures when compared to the cost of human experts. Analogously, one should realize that the use of judgmental models is inherently less costly than the use of human experts, for after a judge has been used to derive his model he is free to perform other activities. Therefore, if cost is a factor in deciding between the use of men or their models, then in many situations judges would have to be substantially more valid than their models before the overall utilities would favor the continued use of expensive professional time. So far, however, there have been no reports of any substantial incremental validity of man over his model. Consequently, if these findings can be generalized to other sorts of judgmental problems, it would appear that only rarely – if at all – will the utilities favor the continued employment of a man over an actuarial – or a judgmental – model.

## A Concluding Note

While the three central questions in psychological assessment (Why measure *that* trait?, *How* measure that trait?, How *use* that measure?) are still far from answered, a common thread in all future answers might well involve a reconception of the optimal roles for man's judgment and for his empirical techniques. In the past, it has been customary to pit "reason" against "facts," to debate the relative virtues of (*a*) the intuitive taxonomies (e.g., Freud) vs. the empirical ones (e.g., Cattell), (*b*) the Intuitive strategy of scale construction vs. the empirical strategies (External and Internal), and (*c*) the use of clinical intuition vs. actuarial modes of information processing. However, "intuition" and "facts" must both be deployed, although at different stages of the scientific process. Intuition is an absolute necessity at the earliest stages (e.g., perception and concept formation), while empirical analyses are equally necessary in later stages. To deny the roles of either intuition or data is to court disaster. To establish their proper roles in trait discovery, psychometrics, and test usage

is perhaps the greatest unfinished business of the next decade.

## REFERENCES


Allport, G. W., & Odbert, H. S. Trait-names: A psycho-lexical study. *Psychological Monographs,* 1936, *47,* (1, Whole No. 211).

Bracht, G. H. The relationship of treatment tasks, personological variables, and dependent variables to aptitude-treatment interactions. Unpublished doctoral dissertation, University of Colorado, 1969.

Buros, O. K. (Ed.) *Personality tests and reviews.* Highland Park, N. J.: Gryphon Press, 1970.

Butt, D. S., & Fiske, D. W. Comparison of strategies in developing scales for dominance. *Psychological Bulletin,* 1968, *70,* 505-519.

Butt, D. S., & Fiske, D. W. Differential correlates of dominance scales. *Journal of Personality,* 1969, *37,* 415-428.

Cattell, R. B. *Personality: A systematic theoretical and factual study.* New York: McGraw-Hill, 1950.

Cattell, R. B. *Personality and motivation: Structure and measurement.* New York: World Book, 1957.

Cronbach, L. J. The two disciplines of scientific psychology. *American Psychologist,* 1957, *12,* 671-684.

Cronbach, L. J., & Snow, R. E. Individual differences in learning ability as a function of instructional variables. Office of Education Final Report No. OEC 4-6-061269-1217, March 1969.

Damarin, F. A special review of Buros' *Personality tests and reviews. Educational and Psychological Measurement,* 1971, *31,* 215-241.

Dashiell, J. F. Some rapprochements in contemporary psychology. *Psychological Bulletin,* 1939, *36,* 1-24.

Dawes, R. M. A case study of graduate admissions: Application of three principles of human decision making. *American Psychologist,* 1971, *26,* 180-188.

Dicken, C. Good impression, social desirability and acquiescence as suppressor variables. *Educational and Psychological Measurement,* 1963, *23,* 699-720.

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

Endler, N. S., Hunt, J. McV., & Rosenstein, A. J. An S-R inventory of anxiousness. *Psychological Monographs,* 1962, *76* (17, Whole No. 536).

Goldberg, L. R. Simple models or simple processes? Some research on clinical judgments. *American Psychologist,* 1968, *23,* 483-496.

Goldberg, L. R. The search for configural relationships in personality assessment: The diagnosis of psychosis vs. neurosis from the MMPI. *Multivariate Behavioral Research,* 1969, *4,* 523-536.

Goldberg, L. R. Man vs. model of man: A rationale, plus some evidence, for a method of improving on clinical inferences. *Psychological Bulletin,* 1970, *73,* 422-432.

Goldberg, L. R. A historical survey of personality scales and inventories. In P. McReynolds (Ed.), *Advances in psychological assessment: Volume Two.* Palo Alto, Calif.: Science and Behavior Books, 1971.

Goldberg, L. R. Parameters of personality inventory construction and utilization: A comparison of prediction strategies and tactics. *Multivariate Behavioral Research Monographs,* 1972, *7,* No. 72-2.(a)

Goldberg, L. R. Student personality characteristics and optimal college learning conditions: An extensive search for trait-by-treatment interaction effects. *Instructional Science,* 1972, *1,* 153-210.(b)

Goldberg, L. R., Rorer, L. G., & Greene, M. M. The usefulness of "stylistic" scales as potential suppressor or moderator variables in predictions from the CPI. *Oregon Research Institute Research Bulletin,* 1970, *10* (No. 3).

Gough, H. G. Clinical versus statistical prediction in psychology. In L. Postman (Ed.), *Psychology in the making.* New York: Knopf, 1962.

Gough, H. G. Conceptual analysis of psychological test scores and other diagnostic variables. *Journal of Abnormal Psychology,* 1965, *70,* 294-302.

Gough, H. G., & Sandhu, H. Validation of the CPI socialization scale in India. *Journal of Abnormal and Social Psychology,* 1964, *68,* 544-547.

Green, B. F., Jr. Descriptions and explanations: A comment on papers by Hoffman and Edwards. In B. Kleinmuntz (Ed.), *Formal representation of human judgment.* New York: Wiley, 1968.

Hammond, K. R., Hursch, C. J., & Todd, F. J. Analyzing the components of clinical inference. *Psychological Review,* 1964, *71,* 438-456.

Hammond, K. R., & Summers, D. A. Cognitive dependence on linear and nonlinear cues. *Psychological Review,* 1965, *72,* 215-224.

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

Hoffman, P. J. The paramorphic representation of clinical judgment. *Psychological Bulletin,* 1960, *57,* 116-131.

Hoffman, P. J. Cue-consistency and configurality in human judgment. In B. Kleinmuntz (Ed.), *Formal representation of human judgment.* New York: Wiley, 1968.

Holtzman, W. H., & Sells, S. B. Prediction of flying success by clinical analysis of test protocols. *Journal of Abnormal and Social Psychology*, 1954, *49*, 485-490.

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

Jackson, D. N., & Messick, S. Content and style in personality assessment. *Psychological Bulletin*, 1958, *55*, 243-252.

Jackson, D. N., & Messick, S. Acquiescence and desirability as response determinants in the MMPI. *Educational and Psychological Measurement*, 1961, *21*, 771-790.

Jackson, D. N., & Messick, S. Response styles on the MMPI: Comparison of clinical and normal samples. *Journal of Abnormal and Social Psychology*, 1962, *65*, 285-299.

Loevinger, J. Objective tests as instruments of psychological theory. *Psychological Reports*, 1957, *3*, 635-694.

Lunneborg, C. E., & Lunneborg, P. W. EPPS patterns in the prediction of academic achievement. *Journal of Counseling Psychology*, 1967, *14*, 389-390. (a)

Lunneborg, C. E., & Lunneborg, P. W. Pattern prediction of academic success. *Educational and Psychological Measurement*, 1967, *27*, 945-952. (b)

McArthur, C. Analyzing the clinical process. *Journal of Counseling Psychology*, 1954, *1*, 203-208.

Meehl, P. E. The dynamics of "structured" personality tests. *Journal of Clinical Psychology*, 1945, *1*, 296-303.

Meehl, P. E. *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence.* Minneapolis: University of Minnesota Press, 1954.

Meehl, P. E. Clinical versus actuarial prediction. *In Proceedings of the 1955 Invitational Conference on Testing Problems.* Princeton: Educational Testing Service, 1956.

Meehl, P. E. The cognitive activity of the clinician. *American Psychologist*, 1960, *15*, 19-27.

Meehl, P. E., & Hathaway, S. R. The K factor as a suppressor variable in the MMPI. *Journal of Applied Psychology*, 1946, *30*, 525-564.

Mischel, W. *Personality and assessment.* New York: Wiley, 1968.

Murray, H. A., et al. *Explorations in personality.* New York: Oxford, 1938.

Naylor, J. C., & Wherry, R. J., Sr. The use of simulated stimuli and the "JAN" technique to capture and cluster the policies of raters. *Educational and Psychological Measurement*, 1965, *25*, 969-986.

Norman, W. T. Toward an adequate taxonomy of personality attributes: Replicated factor structure in peer nomination personality ratings. *Journal of Abnormal and Social Psychology*, 1963, *66*, 574-583.

Norman, W. T. 2800 personality trait descriptors: Normative operating characteristics for a university population. Office of Research Administration No. 08310-1-T, 1967, University of Michigan, NIMH Grant No. MH-07195.

Parker, C. A. As a clinician thinks . . . *Journal of Counseling Psychology*, 1958, *5*, 253-262.

Sawyer, J. Measurement *and* prediction, clinical *and* statistical. *Psychological Bulletin*, 1966, *66*, 178-200.

Siegel, L., & Siegel, L. C. The instructional Gestalt: A conceptual framework and design for educational research. *AV Communication Review*, 1964, *12*, 16-45.

Siegel, L., & Siegel, L. C. Educational set: A determinant of acquisition. *Journal of Educational Psychology*, 1965, *56*, 1-12.

Siegel, L., & Siegel, L. C. A multivariate paradigm for educational research. *Psychological Bulletin*, 1967, *68*, 306-326.

Slovic, P., & Lichtenstein, S. Comparison of Bayesian and regression approaches to the study of information processing in judgment. *Organizational Behavior and Human Performance*, 1971, *6*, 649-744.

Spranger, E. *Types of men (Lebensformen).* Halle: Niemeyer, 1928.

Stilson, D. W., & Astrup, C. Nonlinear and additive methods for long-term prognosis in the functional psychoses. *Journal of Nervous and Mental Disease*, 1966, *141*, 468-473.

Webb, E. J., Campbell, D. T., Schwartz, R. D., & Sechrest, L. B. *Unobtrusive measures: A survey of nonreactive research in social science.* Chicago: Rand McNally, 1966.

Wiggins, N., & Kohen, E. S. Man versus model of man revisited: The forecasting of graduate school success. *Journal of Personality and Social Psychology*, 1971, *19*, 100-106.

Lewis R. Goldberg
Box 3196
Eugene, Oregon 97403

Received: November 24, 1971
Revised: February 16, 1972