

*Multivariate  
Behavioral Research  
Monographs*

PARAMETERS OF PERSONALITY INVENTORY  
CONSTRUCTION AND UTILIZATION: A COMPARISON OF  
PREDICTION STRATEGIES AND TACTICS

BY

LEWIS R. GOLDBERG

MBR Monograph No. 72-2

Published by the Society of Multivariate Experimental Psychology

Copyright 1972 by The Society of Multivariate Experimental Psychology, Inc.

TABLE OF CONTENTS

LIST OF TABLES .....	3
LIST OF FIGURES .....	4
ABSTRACT .....	5
CHAPTER I. INTRODUCTION .....	5
CHAPTER II. A COMPARISON AMONG STRATE- GIES OF INVENTORY CONSTRUCTION .....	10
CHAPTER III. THE EFFECTS OF VARIATIONS IN THE NUMBER OF PREDICTORS .....	25
CHAPTER IV. A COMPARISON AMONG VARI- OUS TYPES OF PREDICTION FUNCTIONS .....	28
CHAPTER V. ADDITIONAL ANALYSES .....	36
CHAPTER VI. DISCUSSION .....	47
APPENDIX A .....	55
REFERENCES .....	56

## LIST OF TABLES

Table 1. Five Scale Construction Strategies and the Resulting Nine Inventories .....	13
Table 2. The Thirteen Criteria .....	17
Table 3. Intercorrelations among the Thirteen Criterion Measures .....	19
Table 4. Average Cross-Validity Coefficients for Each of Nine Inventories: Regression Weights Based upon All 11 Scales in Each Prediction Equation .....	21
Table 5. A Comparison of the Relative Effect Sizes Associated with the Criteria and with the Five Major Strategies.....	23
Table 6. Average Cross-Validity Coefficients for Each of Nine Inventories as a Function of the Number of Predictors In- cluded in the Regression Equation .....	26
Table 7. A Comparison of the Relative Effect Sizes Associated with the Criteria, the Five Major Strategies, and the Num- ber of Predictors in the Regression Equation .....	27
Table 8. The 14 Prediction Functions .....	31
Table 9. A Comparison of the Cross-Validity Coefficients from the 14 Prediction Functions When Five Predictors Were Included in Each Function (Averages across the Five Major Inventories) .....	32
Table 10. A Comparison of the Relative Effect Sizes Associ- ated with the Criteria, Five Major Strategies, Number of Predictors, Weight Type, Weight Precision, Selection Order, and Variable Metric .....	33
Table 11. A Comparison of Original Scores and Orthogonal Components: Average Cross-Validity Coefficients for Each of the Five Major Inventories and Differing Num- bers of Predictors in the Regression Function .....	37
Table 12. The Effect of Employing Nonlinear and Configur- al Terms: The Average Cross-Validity for Each of the Five Major Inventories When Five Predictors Were Included in the Regression Function .....	39
Table 13. Average Cross-Validity Coefficients for 11 Abbrevi- ated MMPI Scales and for the 18 Standard CPI Scales as a Function of the Number of Predictors Included in the Regression Equation .....	40
Table 14. Average Cross-Validity Coefficients for a Set of 61 CPI Scales and for Four Sets of 11 Factor Scores Based on Those Scales .....	41
Table 15. Average Cross-Validity Coefficients for the 240 Odd- Numbered CPI Items Based Upon Six Different Predic- tion Functions .....	43

Table 16. Summary Table: The Highest Average Cross-Valid- ity Coefficients .....	45
Table 17. The Criteria-by-Strategies Interaction Effect: Aver- age Cross-Validity Coefficients When Five Scales Were Included in Each Regression Function .....	52

## LIST OF FIGURES

Figure 1. General Design of the Study .....	11
---	----

# PARAMETERS OF PERSONALITY INVENTORY CONSTRUCTION AND UTILIZATION: A COMPARISON OF PREDICTION STRATEGIES AND TACTICS<sup>1</sup>

LEWIS R. GOLDBERG  
University of Oregon and Oregon Research Institute

## ABSTRACT

This extensive reanalysis of the data originally reported in Hase and Goldberg (1967) was designed to investigate the relative effects of three major sources of variance in multivariate predictions from personality inventories: (a) the strategy of scale construction, (b) the number of predictors included in the prediction functions, and (c) the type of prediction functions utilized. Five strategies of scale construction were used to construct nine different 11-scale inventories from the CPI item pool, five based on three major strategies (External, Internal, and Intuitive) and four based on two control strategies (Stylistic and Random). The average cross-validities of each inventory were compared across 13 criterion indices, and across the seven most predictable criteria, for each of 14 types of prediction functions and for differing numbers of predictors included in each function. In addition, these results were compared with those based upon orthogonal components and factors, nonlinear and configural prediction models, other sets of CPI scales, and empirical item selection tactics. The results of these analyses 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. Moreover, these results were not limited to analyses using any particular number of scales in the prediction functions, nor to analyses based upon the traditional multiple-regression paradigm. 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.

## CHAPTER I

### INTRODUCTION

One of the most salient developments in personality assessment during the past few decades has been the proliferation of multiscale personality inventories (Goldberg, 1971). The basic characteristic of such inventories is that a set of scales, typically more than a dozen, is scored from the same item pool, the items

1. This report was written while the author was a Visiting Professor in the Department of Psychology, and the Institute of Personality Assessment and Research, at the University of California at Berkeley. Data analyses were carried out at the Health Sciences Computing Facility at UCLA, sponsored by NIH Special Research Resources Grant RR-3. The project was supported by Grants MH 12972 and 10822 from the National Institute of Mental Health, U. S. Public Health Service.

Reprints may be obtained from Lewis R. Goldberg at Oregon Research Institute, P. O. Box 3196, Eugene, Oregon, 97403.

often having been published commercially in a printed test booklet. Relatively standard characteristics of such inventories include the use of a single response format for all items, and the test constructor's admonition (through the test manual or other of his publications) that optimal use of the inventory demands the utilization of sets of scales, often by means of multiple-regression procedures.

Ideally, the scale scores from a personality inventory should have implications both for personality theory on the one hand, and for applied prediction problems on the other. In the past it was customary to treat each of these two aspects separately, as though inventories might be constructed solely for one purpose or the other. Today, however, many test constructors attempt to combine both goals in a single instrument; that is, they have made the implicit assumption that theoretically meaningful variables, when reliably measured, can be combined to yield optimal predictions of important social criteria, and conversely that those variables which turn out empirically to be the most useful in the prediction of significant human outcomes are the ones with which personality theory will have to deal. This important assumption permits an evaluation of the adequacy of a personality inventory, and consequently of the comparative usefulness of two or more such inventories.

Specifically, it is the thesis of this monograph that the usefulness of an inventory can be measured by assessing the extent to which it permits valid predictions of important criterion behaviors. Two or more inventories can be compared in terms of the number of such behaviors predictable from the instrument and the validity of these predictions. Viewed in this light, personality inventories can be considered as differing (a) in the *range* of behavioral outcomes to which they apply, and (b) in the *validity* of their predictions within their range. Thus, inventories can be considered as varying in "band-width" and in "fidelity" (Cronbach & Gleser, 1965). Cronbach and Gleser have borrowed these terms from information theory, where in turn they were borrowed from audio-engineering. Band-width has been used to refer to the range of information conveyed to the assessor by an assessment technique, and fidelity to the accuracy with which the information is conveyed. Across varying assessment techniques—as across varying sound systems—these two concepts are negatively correlated: typically, instruments which cover many dimensions (broad-band techniques) tend to be of limited fidelity for

each dimension, while the most accurate measures tend to be focused more sharply (narrow-band techniques). In the present context, an 18-scale inventory in which each scale was constructed to assess some different component of college grade point average would be a more narrow-band inventory than the California Psychological Inventory (CPI), which was constructed to predict college GPA as just one of a large number of criteria within its predictive range.

As long as one is comparing inventories of roughly equivalent range (assuming that they do not vary in other respects, such as testing time and cost, complexity of scoring, etc.), then their value becomes a direct function of their comparative validity in predicting criteria within that range. In the case of general-purpose inventories like the CPI and the Sixteen Personality Factor Questionnaire (16PF), that range is extremely broad; these two inventories, for example, are offered as predictively useful in "schools, colleges, business and industry, clinics and counseling agencies [CPI]" and in "industry, college, and clinics [16PF]"; the traits measured by these inventories are intended to include "concepts which possess broad personal and social relevance . . . (with) wide and pervasive applicability to human behavior . . . personality characteristics important for social living and social interaction [CPI]" and "all the main dimensions along which people can differ . . . the main dimensions that have been found necessary and adequate to cover all the kinds of individual differences of personality found in common speech and psychological literature. (They leave out no important aspect of the total personality.) [16PF]".<sup>2</sup> Consequently, the comparative evaluation of these two inventories, for example, should involve an assessment of their relative validities across a sample of criteria within this very broad range. Although both inventories have been commercially available since 1957 (and the statistical techniques for evaluating them have been available since the turn of the century), few such empirical comparisons have yet been carried out (e.g., Cowden, Schroeder, & Peterson, 1971).

Hopefully, psychometric investigators will soon begin to test the differential validities of such inventories across a broad range of criteria, and test reviews of the sort which appear in the *Mental Measurements Yearbooks* (e.g., Buros, 1970) can then be bolstered by empirical evidence on the comparative validity of the inventory as a whole. However, such evidence will tend to have

2. All quotations are from the inventory manuals (Cattell & Eber, 1957; Gough, 1957).

more practical than methodological value, for inventories such as the CPI and the 16PF differ from each other in at least three critical ways: (a) the nature of their item pools, (b) the number of scales included in each inventory, and (c) the strategy of scale construction utilized by the inventory developer. Consequently, any superiority in the overall validity of one of these inventories could stem from its item pool, from the inclusion of some optimal number of scales, from its scale construction strategy, or from some combination of these three factors. In the comparison of different inventories, no means of unconfounding these variables is available.

However, one method of separating the effects of each of these variables upon inventory effectiveness is the classical univariate experimental design, in which all variables are held constant (controlled) except one, and the effects of this one variable are then observed. In the present case, one could (a) control the number of scales and the scale construction strategy (e.g., use the same strategy and the same number of scales for all inventories) and vary the item pool; (b) control the item characteristics (by using the same item pool) and the number of scales and vary the strategy of scale construction; or, (c) control both the item pool and the scale construction strategy and vary the number of scales to be constructed. The present research project utilized design (b) to discover the comparative effectiveness of each of a number of diverse strategies and tactics for constructing a personality inventory—the value of each experimental inventory being evaluated by its comparative validity across a broad array of criteria.

Some preliminary findings from this project have been presented in four previous reports. The first report (Hase, 1965) was focused upon the construction of new sets of scales from the CPI item pool, each set constructed by a different strategy. The comparative validity of six of these sets, each of 11 scales, was discussed in the second report (Hase & Goldberg, 1967). The third report (Goldberg & Hase, 1967), a preliminary version of the present monograph, included the findings from factor analyses of each of the experimental inventories and from canonical correlations computed among all pairs of inventories and between each inventory and the set of criteria. The usefulness of "stylistic" scales as suppressor and moderator variables was evaluated in the fourth report (Goldberg, Rorer, & Greene, 1970). The present monograph, which should serve to summarize and to extend a number of the analyses presented in the four preliminary reports,

will consider three major sources of variance in predicting non-test criteria from personality inventories: (a) the strategy of scale construction, (b) the number of scales included in the prediction functions, and (c) the type of weights used to derive these functions. Consequently, this monograph will address a number of broad methodological issues in personality assessment. Since the analyses to be reported here are both numerous and complex, and since it may be assumed that most readers will not have read all of the previous reports in this series, let us begin with a summary of the major analyses reported by Goldberg and Hase (1967).

## CHAPTER II

### A COMPARISON AMONG STRATEGIES OF INVENTORY CONSTRUCTION

#### *Experimental Design*

The general design for the first set of analyses reported in this monograph is illustrated in Figure 1. A single item pool of 468 unique items (the CPI item pool) was used to construct 9 different "inventories," each of 11 scales, each inventory being constructed by some variant of five different strategies of inventory construction. Each of these 9 different inventories is represented in the left-hand column of Figure 1. From a population of subjects who had taken the CPI and for whom 13 criterion indices were available, two random samples of 76 subjects each were formed. For each of the 9 inventories, and for each of the 13 criteria, a linear-regression equation using the 11 scales in that inventory was developed in each sample (e.g., the 11 scales in one inventory being combined in an optimal linear manner to predict one criterion for the subjects in Sample A), and the resulting regression function was then cross-validated on the other sample (e.g., Sample B). Conversely, the 9 x 13 regression functions developed on subjects in Sample B were cross-validated on the subjects in Sample A. This process is illustrated in the center section of Figure 1. The *cross-validated* multiple correlations ( $R$ ) were then averaged across both samples to form an average cross-validity for each criterion ( $\bar{R}_1 \dots \bar{R}_{13}$ ), and these, in turn, were averaged across the 13 criteria ( $\bar{R}$ ), a process illustrated on the right-hand side of Figure 1.<sup>3</sup> Consequently, the resultant final  $\bar{R}$  represents the average cross-validity for each inventory across the entire range of the 13 criteria, and the inventories can be compared by the magnitude of these  $\bar{R}$  values. Since the inventories differ only in the strategy of scale construction used to develop them (each inventory was composed of the same number of scales, constructed from the same item pool), any differences in these average cross-validities can be assumed to stem from the strategy utilized.

3. Throughout this monograph, correlations have always been transformed to  $Z$ -values prior to any arithmetic operations. For example, means were computed by averaging the  $Z$ -transformed values and then converting the averages back to correlations.

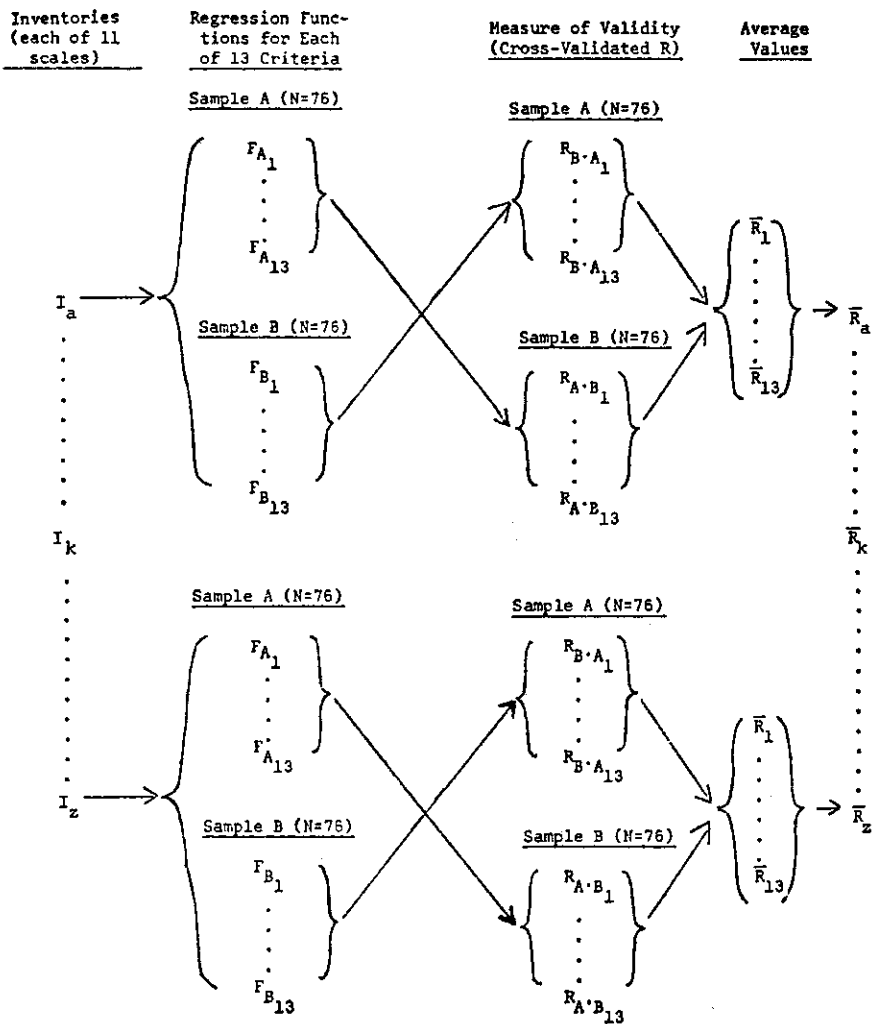


Fig. 1. General design of the study.

### Subjects

The larger sample from which the subjects in this project were recruited consisted of freshman women who were residents of one dormitory at the University of Oregon (Goldberg & Rorer, 1966). Subjects were paid to participate in a six-week program of psychological assessment, and testing sessions of approximately one hour in duration were held on the same evening each week. In addition to the monetary reward for participation in the study, motivation was further enhanced by the promise of extensive feedback on the results of the psychological tests administered. Since well over 90% of the coeds in the dormitory indi-

cated a strong initial interest in participating in the experimental procedures, it can reasonably be assumed that the major source of variance differentiating participants from non-participants stemmed from schedule conflicts (e.g., a class held during the hour when the experimental procedures were being carried out). While larger-sized samples were used for some scale construction purposes and for the analyses reported in Hase and Goldberg (1967), the sample of subjects analyzed in the present monograph included only those 152 subjects for whom all data were available.

### Strategies

A strategy of scale construction may be defined as a systematic procedure for grouping and keying item responses so as to form a composite score. Correspondingly, a strategy of inventory construction may be defined as a systematic procedure for grouping and keying items to form a set of scales from the same item pool. The assumptions underlying the use of any strategy of inventory construction include: (a) the inventory developer has assembled a large set of items which on a priori grounds appear to cover the domain of interest (in the case of a general-purpose personality inventory, a wide array of personality traits); (b) this initial item pool is redundant (e.g., more than one item relates to the same personality trait); and (c) while responses to individual items are quite unreliable, scale scores from sets of items can have considerable reliability. In the case of the CPI, the initial item pool was assembled to cover a broad band of personality characteristics; moreover, the item pool is certainly redundant by almost any criterion; and finally, while responses to single CPI items are indeed unreliable (e.g., Goldberg & Rorer, 1964), scales can be constructed which do possess more satisfactory reliability (Gough, 1957).

While various strategies of scale construction have proliferated over the years under a number of names, the basic strategies can all be divided into three main types, labeled "External," "Internal," and "Intuitive" (see Table 1). The *External* strategy derives its name from the fact that some *non-test* reference groups are used to determine an item's scale membership and direction of keying. This strategy has consequently been referred to as the "criterion-group" or "empirical" strategy (see Meehl, 1945); leading exponents of this strategy of inventory construction have included Starke Hathaway, E. K. Strong, and Harrison Gough. Of the 18 standard CPI scales, 11 were constructed by Gough using

the External strategy; these 11 scales, constructed to assess the traits of Dominance (*Do*), Capacity for Status (*Cs*), Sociability (*Sy*), Responsibility (*Re*), Socialization (*So*), Tolerance (*To*), Achievement via Conformance (*Ac*), Achievement via Independence (*Ai*), Intellectual Efficiency (*Ie*), Psychological-mindedness (*Py*), and Femininity (*Fe*), were included in the present project as an *Empirical* inventory to represent the External strategy. The number of items and both test-retest and KR-20 reliability coefficients for each of these 11 scales can be found in Hase and Goldberg (1967).

Table 1  
Five Scale Construction Strategies and the Resulting Nine Inventories

Strategy	Inventories (each composed of 11 scales)	Scale Developer	Number of Items in the Scales		Reliability ( <i>N</i> = 152)		
			Mean	$\sigma$	Mean Test- Retest <i>r</i> (4 week)	Mean KR20 ( $r_{tt}$ )	Mean $r_{ii}^a$
External	Empirical	Gough	39	9	.82	.65	.06
Internal	Factor	Hase	17	6	.87	.81	.22
	MSA	Lingoes	5	3	.79	.66	.31
Intuitive	Theoretical	Hase	14	3	.81	.62	.12
	Rational	Hase (7) Gough (4)	37	10	.87	.75	.09
Stylistic	Style	Lovell (9) Gough (1) Dicken (1)	32	13	.76	.55	.05
	Random-True	Hase	25	0	.70	.35	.02
Random	Random-I	Hase	25	0	.57	-.10	.00
	Random-II	Hase	25	0	.55	-.08	.00

<sup>a</sup>Internal consistency reliability for the average item:

$$(r_{ii} = \frac{r_{tt}}{n - (n-1)r_{tt}}) \text{—reversal of the Spearman-Brown correction.}$$

The second major strategy of inventory construction has been labeled *Internal*, since the internal structure of the initial item pool is the sole determiner of an item's scale membership and its direction of keying. There have been at least two major variants of the Internal strategy, one aimed at a direct maximization of scale homogeneity or internal consistency (e.g., Loevinger, Gleser, & DuBois, 1953), the other aimed at the construction of scales by means of some factor-analytic procedure (see Cattell, 1957); lead-

ing exponents of the latter position include Raymond Cattell, Hans Eysenck, and J. P. Guilford. In the present project, both variants of the internal strategy were represented. A *Factor* inventory of 11 scales was constructed by a principal components analysis, followed by a Varimax rotation, via the computer program called BIGNV of the BCTRY system for multi-dimensional analysis (Tryon & Bailey, 1970); 179 of the coeds in the sample already described provided the item responses for this analysis, which is discussed in more detail in Hase and Goldberg (1967). In addition, an 11-scale *Multiple Scalogram Analysis (MSA)* inventory was constructed by James Lingoes, using as his basic data the same item responses used to develop the *Factor* inventory. The MSA algorithm tends to produce relatively short, but highly homogeneous, item clusters (see Lingoes, 1963). Lingoes' analysis of the CPI item pool initially produced 28 MSA scales; subjects were scored on each of these scales, the scores were intercorrelated, and the correlations were factor analyzed and rotated orthogonally. Eleven of the most reliable (KR-20) and independent of these scales were selected by Lingoes to form the MSA inventory.

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 personality inventories were constructed from the judgments of a single individual, more recent inventories 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. Proponents of this general strategy of inventory construction, which has often been referred to under such labels as a "rational," "theoretical," or "construct" approach, have included Henry Murray, Gordon Allport, and Allen Edwards. However, in contrast to test developers committed to the External or Internal strategies—and in contrast to the very early users of the Intuitive strategy—present-day scale developers tend to use some mixture of 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).

For the present project, inventories were constructed using each of two major variants of the Intuitive strategy. An 11-scale *Theoretical* inventory was constructed from the consensus of the judgments of three advanced graduate students in clinical psychol-

ogy, each of whom was asked to select those CPI items which would measure 19 "manifest needs," as elaborated by Murray et al. (1938). The resulting 19 scales were reduced to an 11-scale inventory—without recourse to any sort of internal consistency analysis—on the basis of the judgments of the inventory developer (Hase) as to the most significant scales for the prediction of diverse social criteria. The *Theoretical* inventory consisted of scales constructed to measure individual differences in the need for Achievement (*nAc*), Affiliation (*nAf*), Autonomy (*nAu*), Defiance (*nDe*), Dominance (*nDo*), Exhibition (*nEx*), Infravoidance (*nIn*), Nurturance (*nNu*), Order (*nOr*), Play (*nPl*), and Understanding (*nUn*). A more complete description of the *Theoretical* inventory—including the number of items and reliability of each scale—can be found in Hase and Goldberg (1967).

An 11-scale *Rational* inventory was constructed to measure the sort of "folk concepts" Gough has attempted to include in the CPI. Four of these Rational scales—constructed to assess the traits of Social Presence (*Sp*), Self-acceptance (*Sa*), Self-control (*Sc*), and Flexibility (*Fx*)—were developed by Gough and are included in the published version of the CPI. Seven additional Rational scales were constructed from the judgments of one psychologist (Hase), who selected CPI items to measure the traits of Dominance (*Dom*), Sociability (*Soc*), Responsibility (*Res*), Psychological-mindedness (*Psy*), Femininity (*Fem*), Academic achievement (*Ach*), and Conformity (*Con*). After the initial selection of items for these seven scales, each of the scales was "purified" by internal consistency analysis. The preliminary Rational scales were scored on a sample of 108 college women (a different sample from those used for other purposes in this project), and the item responses were correlated with scale scores; items correlating significantly with the total score for each scale were retained to form the final scales for the *Rational* inventory. Additional information on the construction and properties of the scales in this inventory can be found in Hase and Goldberg (1967).

These five inventories (each composed of 11 scales, all constructed from the same item pool) were used to represent the three major approaches to inventory construction, and a comparison of their validities across a broad range of behavioral criteria should serve to provide evidence of the *comparative* value of the strategies they represent. However, the *absolute* value of their validity coefficients is, in part, a function of the characteristics of the initial item pool from which they were constructed. In order to uncon-

found the itemmetric from the strategic sources of variance in the validity coefficients, two "control" strategies—Stylistic and Random—were utilized, each represented by two inventories. Since no one has yet advocated the development of a general-purpose personality inventory by either of these strategies alone, their validities should serve as baseline measures against which to assess the incremental validities of the three major strategies.

In order to construct a Stylistic inventory, Lovell's (1964) nine stylistic CPI scales were supplemented by Gough's (1957) Communitality (*Cm*) scale and Dicken's (1963) Social Desirability (*Dsd*) scale. The resulting *Style* inventory was composed of 11 scales which had been constructed to be relatively pure measures of the three purportedly major dimensions of CPI response variance: "social desirability," "acquiescence," and "response communality." In contrast, another Stylistic inventory was constructed to assess the relative importance of acquiescence response bias alone. Eleven scales were constructed by selecting sets of 25 items from the CPI item pool on a random basis, without replacement of items (i.e., no item overlap); items in these 11 scales were all keyed true—resulting in a set of relatively unsophisticated stylistic scales, labeled the *Random-True* inventory.

To provide an absolute basement for effects due to a particular strategy of scale construction, two inventories were constructed on a completely random basis. A Random-Without-Replacement inventory (*Random-I*) was constructed by randomly keying the items in each of the 11 scales from the *Random-True* inventory. In addition, a parallel Random-With-Replacement inventory (*Random-II*) was constructed by selecting 25 items randomly from the CPI item pool for each of 11 scales, the items being replaced in the pool after each scale was constructed (i.e., item overlap allowed). As with the *Random-I* inventory, all items in the 11 scales in the *Random-II* inventory were keyed randomly.

Table 1 summarizes some psychometric characteristics of the scales in each of the resulting nine inventories. Note that the *Empirical* inventory had the longest scales, while the two inventories constructed to represent the Internal strategy had relatively short scales. Of the two inventories built to represent the Intuitive strategy, one (*Rational*) had relatively long scales while the other (*Theoretical*) had relatively short ones. As expected, scales from the two Internal inventories (*Factor* and *MSA*) had the highest coefficients of internal consistency for the average item ( $r_u$ ), while those from the *Empirical* inventory, as well as from the four

Stylistic and Random inventories, had the lowest. In general, the scales in the *Factor* and *Rational* inventories were slightly more reliable than those in the three other major inventories, and these in turn were superior to those in the four Stylistic and Random inventories.

### Criteria

Thirteen criteria were selected to tap a broad range of variables which psychologists have sought to predict from personality inventories. These 13 criterion measures, listed in Table 2, can be grouped into five major areas.

Table 2  
The Thirteen Criteria

Abbreviation	Title	Measure
cSOR	Sorority Joining	Members of a social sorority vs. "independents"
cYLD	Yielding (Conformity)	Degree of yielding to purported group consensus (Goldberg & Rorer, 1966)
cDOM	Dominance	Mean peer ratings by 8-12 close associates
cSOC	Sociability	Mean peer ratings by 8-12 close associates
cRES	Responsibility	Mean peer ratings by 8-12 close associates
cPSY	Psychological-mindedness	Mean peer ratings by 8-12 close associates
cFEM	Femininity	Mean peer ratings by 8-12 close associates
cHWK	How Well Known	Mean peer ratings by 8-12 close associates
cDAT	Heterosexual Popularity (Dating)	Reported number of dates per month
cGPA	College Grade Point Average	First term College GPA
cACH	Over- vs. Under-Achievement	College GPA minus GPA as predicted by SAT + HSGPA
cMAJ	College Major	Liberal Arts majors vs. all other majors
cCDO	College Dropout	Number of years spent at the University

1. *Social conformity.* Two independent indices of social conformity were included in this project: a Sorority Joining index (*cSOR*) and a Yielding index (*cYLD*). The Sorority Joining index was based on the subject's reported interest in sorority affiliation. Coeds indicated whether they: (a) were pledges or members of a sorority, (b) were uncertain as to whether they would join a sorority, or (c) intended *not* to join a sorority. The validity of this index is indicated by the results of a follow-up study on 80 subjects enrolled at the University of Oregon two years later. Eighty-six per cent of the girls in category *a* were members of a sorority two years later, while 96% of the category *c* coeds were not members. Category *b* coeds were approximately evenly split between sorority members and independents. For the *a* and *c* groups alone, the resulting phi coefficient was .83.

The Yielding index, an experimental measure of conformity, is described in detail by Goldberg and Rorer (1966). The measure, obtained from shifts in responses to a double administration of a 45-item Opinion Questionnaire (Hastorf & Piper, 1951) was previously used for the same purpose by Jackson (1964). Subjects were asked to indicate the amount of their agreement or disagreement with each statement on a nine-point scale. On the second administration of the questionnaire, five weeks after the first, each of the 45 questions was followed by a number described as the average response given by the subjects on the first administration of the questionnaire. For the 25 questions with the smallest dispersions of ratings on the first administration, the reported mean value was obtained by shifting the actual mean three points toward whichever end of the scale was more distant. The mean values of the 20 questions with the largest dispersions were reported accurately. The criterion index of yielding indicated the extent to which the subject yielded in relation to her opportunity to yield.

2. *Peer ratings.* All subjects were rated by 8 to 12 peers on five dimensions, using a five-point forced-distribution rating scale. Subjects were rated only by those coeds living on the same floor of their dormitory wing. The traits rated were Dominance (*cDOM*), Sociability (*cSOC*), Responsibility (*cRES*), Psychological-mindedness (*cPSY*), and Femininity (*cFEM*)—all traits for which Gough had developed corresponding Empirical scales for the CPI. The rating instructions were adopted from Gough's CPI scale descriptions. Each subject's mean rating on each of the five traits was utilized as a criterion index.

3. *Personal popularity.* Two criteria of personal popularity,

How Well Known, and Heterosexual Popularity, were utilized. The criterion of How Well Known (*cHWK*) was the mean rating made by the same peers who provided the other peer ratings, using the same type of rating scale. The criterion of Heterosexual Popularity was based on the subject's reported average number of dates per month (*cDAT*).

4. *Academic achievement.* Two measures of academic achievement were utilized: first-term Grade Point Average (*cGPA*), and Over- vs. Under-Achievement (*cACH*)—actual GPA minus GPA predicted by the University admissions office from a college aptitude test plus high school grades.

5. *Academic interest and perseverance.* Choice of Major (*cMAJ*), a dichotomy between Liberal Arts vs. non-Liberal Arts majors, was included in the present study. The final criterion, College Dropout (*cCDO*), was a follow-up measure providing an index of the subject's diligence in pursuing her academic career. Thirty-six per cent of the original subjects left the University after their freshman year; an additional 21% left the University after their sophomore year, while 43% were still enrolled in their junior year. The College Dropout criterion indicated whether the coed had enrolled in the University for one, two, or three academic years.

The intercorrelations among these 13 criterion measures are presented in Table 3, along with the means and standard deviations of each of the two samples used in the double cross-validation design. Note that the mean scores of the two random samples were very similar on each of the 13 criteria, and that the criterion measures were, in general, relatively uncorrelated. Only five of the 78 pairs of criteria had as much as 25% of their variance in common: (a) GPA and Achievement, (b) Dominance and Sociability, (c) Sociability and How Well Known, (d) Responsibility and Psychological-mindedness, and (e) Psychological-mindedness and How Well Known. For these five pairs of criteria, the correlations were all in the direction expected by logical analysis of these criterion measures. Further information, including estimates of reliability, for each of the 13 criterion measures can be found in Hase and Goldberg (1967). While it would be difficult to argue that these 13 criteria represent a comprehensive set of all that might be of significance, they are meant to be a representative sample of important and diverse criteria which have been of interest to many investigators.

### Results of the Regression Analyses

The major findings from this phase of the project are summarized in Table 4. Listed in the columns of Table 4 are the average *cross-validated* multiple correlations for each inventory for each of the 13 criteria, as well as the mean values for the five inventories constructed by the major strategies, and for the four inventories constructed by the control strategies. The 13 criteria are ordered in Table 4 by their average predictability across the five major inventories. The most predictable criterion was peer-rated Sociability (*cSOC*), with an average cross-validity of nearly .50. Six other criteria—peer ratings of How Well Known (*cHWK*), Femininity (*cFEM*), Dominance (*cDOM*), and Responsibility (*cRES*), plus Dating Frequency (*cDAT*) and Sorority Joining (*cSOR*)—each produced average cross-validities around .35. On the other hand, College Grade Point Average (*cGPA*), with an average cross-validity around .20, was only a bit more predictable than peer-rated Psychological-mindedness (*cPSY*), Yielding (*cYLD*), and College Major (*cMAJ*), which produced average values around .15. Virtually unpredictable by any of the inventories in this study were two criteria, Over- vs. Under-Achievement (*cACH*) and College Dropout (*cCDO*). For a more complete discussion of the comparative predictability of these 13 criteria, in-

Table 3  
Intercorrelations among the Thirteen Criterion Measures ( $N = 152$ )

	cSOR	cYLD	cDOM	cSOC	cRES	cPSY	cFEM	cHWK	cDAT	cGPA	cACH	cMAJ	cCDO
Sorority Joining (cSOR)													
Yielding (cYLD)	-.01												
Dominance (cDOM)	.17	-.20											
Sociability (cSOC)	.37	-.03	.60										
Responsibility (cRES)	-.15	.04	.24	.19									
Psychological-mindedness (cPSY)	-.14	.03	.24	.34	.63								
Femininity (cFEM)	.28	.15	.38	.39	.32	.29							
How Well Known (cHWK)	.08	.10	.20	.62	.32	.50	.27						
No. Dates per Month (cDAT)	.23	.05	.16	.30	.18	-.10	.16	.24					
Grade Point Average (cGPA)	-.14	-.20	.17	-.15	-.24	.15	-.02	-.09	-.16				
Achievement (cACH)	-.06	-.14	.17	-.08	.09	.02	-.01	-.03	.06	.03			
Choice of Major (cMAJ)	-.14	-.25	.17	.01	-.14	-.07	-.20	.00	.03	.14	.13		
College Dropout (cCDO)	-.21	.10	-.10	-.08	-.05	.13	-.05	.12	-.02	-.21	-.08	.06	
Mean:													
Sample A	2.1	.3	3.1	3.3	3.1	3.1	3.2	3.2	6.5	2.4	1.0	.4	2.1
Sample B	2.1	.3	3.1	3.2	3.1	3.0	3.1	3.1	7.8	2.4	1.0	.5	2.1
$\sigma$ : Sample A	.8	.2	.6	.7	.6	.5	.7	.4	5.1	.6	.4	.5	.9
$\sigma$ : Sample B	.8	.2	.6	.7	.6	.5	.7	.5	6.0	.6	.4	.5	.9

Table 4  
Average Cross-Validity Coefficients for Each of Nine Inventories:  
Regression Weights Based upon All 11 Scales in Each Prediction Equation

Criterion	Major Strategies					Control Strategies					Mean
	External Empirical	Internal Factor	MSA	Intuitive		Style	Stylistic		Random		
				Theoretical	Rational		Style	Rand. T	R-I	R-II	
cSOC*	.37	.61	.43	.55	.50	.25	.21	.15	.12	.18	
cHWK*	.26	.45	.28	.43	.41	.18	.15	.23	-.05	.13	
cFEM*	.34	.40	.31	.38	.36	.25	.03	.22	-.04	.11	
cDOM*	.31	.29	.24	.44	.41	.20	.20	.21	.12	.18	
cDAT*	.21	.25	.49	.38	.39	.08	.22	.32	-.13	.12	
cRES*	.32	.33	.34	.36	.36	.21	.22	-.05	-.04	.08	
cSOR*	.25	.46	.36	.24	.32	.28	.23	.09	.01	.15	
cGPA	.22	.23	.26	.07	.28	.05	.34	-.04	-.04	.08	
cPSY	.21	.11	.04	.24	.20	.06	.05	.00	.01	.03	
cYLD	.28	.03	.10	.13	.20	.04	-.13	-.13	.15	-.02	
cMAJ	.27	.16	.06	.02	.12	.01	.01	.21	.09	.08	
cACH	.10	-.03	.08	-.03	.05	-.11	.22	-.12	.19	.04	
cCDO	.06	-.01	.02	.04	-.01	-.01	-.04	.11	-.13	.05	
Mean (13) <sup>a</sup>	.25	.26	.24	.26	.28	.12	.14	.09	.04	.10	
Mean (7) <sup>b</sup>	.29	.40	.35	.40	.39	.21	.18	.17	.00	.14	

<sup>a</sup>Averaged over all 13 criteria.

<sup>b</sup>Averaged over the 7 most predictable criteria (\*).

cluding an analysis of the CPI scales most highly correlated with each of them, see Hase and Goldberg (1967).

The bottom two rows of Table 4 present the mean cross-validity coefficients for each of the nine inventories, averaged across (a) all 13 criterion measures, and (b) the seven most predictable criteria. Note that the mean cross-validities for the five major strategies were appreciably higher than those for the four control strategies, while differences among the four control strategies were rather small. Moreover, differences among the five major strategies—especially for the coefficients based upon all 13 criteria—were also quite small. While it was this latter finding which was emphasized in the preliminary publication of this project (Hase & Goldberg, 1967), it is now appropriate to estimate more completely the relative sizes of the effects on cross-validity associated with the criteria, the strategies, and their interaction. To do this, the fixed-model analysis of variance (ANOVA) has been employed in its descriptive, rather than its inferential, role. Each of the two cross-validity coefficients (one from each of the two samples) was first normalized via the Z-transformation, and these pairs of Z-transformed values were then used as the dependent variable in an ANOVA designed to compare three orthogonal effects, due in turn to (a) the 13 criteria (or, alternatively, the seven most predictable criteria), (b) the five major strategies (or, alternatively, all nine strategies), and (c) the criteria-by-strategies interaction. While the intent of these analyses—and those to follow—is to provide some basis for comparing the relative strength of the effects of the various strategies and tactics investigated in this project, it is important that the reader bear in mind that these comparisons are limited to this particular sample of subjects, strategies, and criteria, and that the generality of any conclusions from these analyses must be established empirically.

Table 5 summarizes the findings from an analysis of the five major strategies; the results of analyses using the four control strategies and those using all nine strategies are available from the author. The values presented in the left half of Table 5 are based upon all 13 criteria, while those in the right half are based upon the seven most predictable criteria. Included in the table are the sums of squares (*SS*), degrees of freedom (*df*), and mean squares (*MS*) for each effect, along with a gross index of effect size, namely the proportion of the total sum of squares associated with that effect. Note that for the analyses based upon all 13 criteria, virtually all of the variance in the cross-validity coeffi-

ents appears to have stemmed from only two sources, namely the main effect due to the criteria (i.e., the *general* predictability of each of the 13 criteria) and the criteria-by-strategies interaction effect. As the reader may note from inspection of Table 4, this latter effect is due largely to differences between the *Empirical* inventory and the other four major inventories under comparison. Specifically, of the five major inventories, the *Empirical* inventory provided the least valid predictions for the most predictable criteria (e.g., Sociability [*cSOC*] and How Well Known [*cHWK*]) while simultaneously providing the most valid predictions for the least predictable criteria (e.g., College Dropout [*cCDO*], Under- vs. Over-Achievement [*cACH*], College Major [*cMAJ*], and Yielding [*cYLD*]). For the six most predictable criteria, the *Empirical* inventory produced cross-validities around .30, as compared to values around .40 for the four other major inventories; for the six least predictable criteria, on the other hand, the *Empirical* inventory produced cross-validities around .20, as compared to values around .10 for the four other inventories. Consequently, the *Empirical* inventory appears to be a broader band-width and lower fidelity instrument than the other major inventories.

Table 5  
A Comparison of the Relative Effect Sizes Associated with  
the Criteria and with the Five Major Strategies

Effect	Analyses Based Upon All 13 Criteria				Analyses Based Upon the 7 Most Predictable Criteria			
	SS	df	MS	% SS	SS	df	MS	% SS
Criteria ( <i>C</i> )	2.92	12	.24	.74	.28	6	.05	.30
Strategies ( <i>S</i> )	.03	4	.01	.01	.16	4	.04	.17
<i>C</i> × <i>S</i>	.79	48	.02	.20	.36	24	.02	.39
Residual	.22	65	.00	.06	.14	35	.00	.14
Total	3.97	129		1.00	.94	69		1.00

However, these initial analyses oversimplify the total experimental paradigm under scrutiny in this project. For, the values presented in Tables 4 and 5 stem from (a) multiple-regression equations, (b) using all 11 scales in each function. However, previous analyses of these and other data (e.g., Burket, 1964; Goldberg & Hase, 1967) have shown that 11 may be too large a number of variables to include in prediction functions for samples of this size. Moreover, there is even some doubt whether regression weights are as powerful as other types of weights for small sample cross-validation designs (Marks, 1966). Consequently, the con-

clusions drawn from these initial ANOVA analyses might well be limited not only to the particular set of items, strategies, and samples utilized in the project, but also to the use of regression functions based upon all 11 scales. To discover whether these findings have any wider generality, it is necessary to unconfound the variance in cross-validity due to strategies from that due to the use of a particular number of predictors and the use of a particular type of prediction function.

CHAPTER III

THE EFFECTS OF VARIATIONS IN THE NUMBER OF PREDICTORS

Within the sample used to derive the regression weights, any increase in the number of predictors in the regression equation will never decrease its validity. On the other hand, the inclusion of additional predictors may serve to fit too closely the chance peculiarities of a particular derivation sample and thus may lead to decreased *cross-validity*. All other factors being equal, the optimal number of predictors in a regression function is directly related to *N*, the size of the sample used to derive the regression weights; as *N* increases in size, the optimal number of predictors also increases. Consequently, were psychological studies always based upon thousands of subjects, this problem might be moot. Unfortunately, however, most investigations—including the present one—employ relatively small samples, where the problem may be of some importance. Since it is not uncommon for investigators initially to sample around 200 subjects and to end up with complete data on about 150 individuals, the findings from the present project should be generalizable to a host of studies employing samples of this approximate size.

Let us now turn to some analyses of the effects of varied values of *n*, the number of predictors included in each regression function. While it is certainly possible that the optimal number of predictors might differ somewhat for different criteria, the following analyses all utilize the same specified values of *n* for each of the 13 criteria. Table 6 presents the cross-validity coefficients (averaged across all 13 criteria and across the seven most predictable criteria) for the same nine inventories compared in Table 4 and for each of six values of *n*. Note that while the cross-validities for *n* = 11 were not radically different from those based on smaller values, generally the largest cross-validities occurred for values of *n* between three and seven.

To compare the relative effect sizes of these three major parameters—criteria, strategies, and *n* (and the various interactions among them)—the pairs of Z-transformed cross-validity coefficients were used to analyze seven orthogonal effects, due in turn to (a) the 13 criteria (or, alternatively, the seven most predictable criteria), (b) the five major strategies, (c) the six values of *n*, (d) the criteria-by-strategies interaction, (e) the criteria-by-*n* interaction, (f) the strategies-by-*n* interaction, and finally (g) the criteria-by-strategies-by-*n* interaction.

Table 6  
Average Cross-Validity Coefficients for Each of Nine Inventories as a Function of the Number of Predictors Included in the Regression Equation

Number of Predictors Included in the Regression Equation	Major Strategies					Control Strategies				
	External Empirical	Internal Factor		Intuitive Theoretical		Stylistic Style	Random		Mean	
		MSA	R-I	R-II	Rand. T		R-I	R-II		
1	.21	.23	.21	.23	.29	.08	.07	.06	.02	.06
3	.27*	.28	.22	.25	.32*	.12	.12	.09	.04	.09
5	.27	.28*	.23	.26	.31	.11	.13	.10*	.03	.09
7	.25	.27	.23	.26*	.29	.12*	.14*	.10	.04	.10*
9	.25	.26	.24	.26	.28	.12	.13	.10	.04	.10
11	.25	.26	.24*	.26	.28	.12	.14	.09	.04*	.10
1	.30	.35	.30	.36	.39	.16	.11	.10	-.02	.09
3	.34*	.42*	.34	.38	.44*	.21	.17	.16	-.01	.13
5	.32	.42	.34	.40	.42	.21	.18	.18*	-.02	.14
7	.30	.40	.34	.40*	.40	.23*	.19*	.18	-.01	.15*
9	.30	.40	.35*	.40	.39	.21	.18	.17	-.01	.14
11	.29	.40	.35	.40	.39	.21	.18	.17	.00*	.14

Note.—Values in the upper half of the table are averaged over all 13 criteria, while those in the lower half are averaged over the 7 most predictable criteria (see Table 5).  
\*Highest value for each inventory.

Table 7 summarizes the findings from these analyses. The values presented in the left half of the table are based upon all 13 criteria, while those in the right half are based upon the seven most predictable criteria. As Table 7 indicates, neither the main effect due to  $n$ , nor either of the two-way interactions involving this variable, was of any appreciable size. Virtually all of the variance in the cross-validity coefficients appears to have stemmed from the same two sources noted in Table 5, namely the main effect due to the criteria (i.e., the general predictability of each of the 13 criteria) and the criteria-by-strategies interaction effect. Consequently, it now seems safe to assert that the conclusions drawn from the initial analyses are generalizable across this entire set of  $n$  values, at least for analyses based upon multiple-regression functions.

Table 7  
A Comparison of the Relative Effect Sizes Associated with the Criteria, the Five Major Strategies, and the Number of Predictors in the Regression Equation

Effect	Analyses Based Upon All 13 Criteria				Analyses Based Upon the 7 Most Predictable Criteria			
	SS	df	MS	% SS	SS	df	MS	% SS
Criteria ( $C$ )	17.77	12	1.48	.72	1.68	6	.28	.28
Strategies ( $S$ )	.44	4	.11	.02	.82	4	.21	.14
No. of Predictors ( $n$ )	.10	5	.02	.00	.10	5	.02	.02
$C \times S$	3.53	48	.07	.14	1.85	24	.08	.31
$C \times n$	.18	60	.00	.01	.08	30	.00	.01
$S \times n$	.07	20	.00	.00	.09	20	.00	.01
$C \times S \times n$	.86	240	.00	.03	.40	120	.00	.07
Residual	1.85	390	.00	.07	1.01	210	.00	.17
Total	24.82	779		1.00	6.02	419		1.00

## CHAPTER IV

### A COMPARISON AMONG VARIOUS TYPES OF PREDICTION FUNCTIONS

Of the many potential sources of variance in the validity of multivariate predictions from personality inventories, the two already discussed in this monograph have been alluded to in the psychometric literature for at least a decade. In contrast, discussion of a third issue, the use of differing prediction functions for combining scale scores, has largely been confined to unpublished reports (e.g., Marks, 1966) and to scientific scuttlebutt. In fact, the use of the multiple-regression model is so ubiquitous in applied psychology that many investigators may never consider the use of any other method of weighting scale scores than by their regression coefficients. Yet many of these same investigators would never consider using regression coefficients for weighting the *items* used to construct an empirical scale; unit weights based upon the item validity coefficients are typically elected for this latter purpose. But regression weights could be applied to items, and unit weights to scales; in fact, a variety of weighting schemes could be used for both purposes. The issue is: given the parameters of the prediction problem as previously defined in this monograph, what types of weights are most useful? Or, alternatively, can the previous conclusions regarding scale construction strategies be generalized beyond the multiple-regression paradigm?

In order to answer these questions, it will first be necessary to describe the major types of linear prediction functions. All such functions can be viewed as variants of the following general linear equation:

$$[1] \quad \hat{Y}_j = \sum_{i=1}^k a_i X_{ij} + c$$

where:

- $\hat{Y}_j$  = the predicted score on the criterion (dependent) variable for individual  $j$ ;
- $X_{ij}$  = the score obtained by individual  $j$  on the  $i^{\text{th}}$  predictor (independent) variable;
- $k$  = the number of predictor (independent) variables in the prediction function;
- $c$  = a constant, used to equate the mean values on both sides of the equation; and,
- $a_i$  = the "weight" of the  $i^{\text{th}}$  predictor (independent) variable in the prediction function.

Given the traditional statistical criterion ("least squared errors"), the multiple-regression coefficient will always provide the most accurate predictions *in the sample used to derive these weights*. On the other hand, it has been argued (e.g., Marks, 1966) that the use of regression weights may lead to lower predictive accuracy upon *cross-validation* than would the use of some alternative types of weights. For each regression coefficient represents an optimal amalgamation of four types of information: (a) the validity coefficient of the predictor with which it is associated, (b) the intercorrelations between that predictor and all other predictors included in the regression function, (c) the validities of each of these other predictors, and (d) the variance of the scores on the predictor relative to that of the criterion. Any change in one or more of these parameters from the derivation to the cross-validation sample can serve to attenuate the cross-validity of the prediction function. On the other hand, should these parameters remain unchanged in the two samples, multiple-regression weights will be more valid than any other type of weighting scheme, and the size of the validity coefficient will not shrink in the new sample. In fact, of course, empirical samples vary on these parameters (thus providing the impetus for cross-validation), and therefore weights which do not focus so finely on the vagaries of a particular sample may sometimes provide higher cross-validity coefficients than regression weights. The argument here is directly analogous to one raised earlier, namely that concerning the optimal number of variables to include in a prediction function; while any increase in the number of predictors can never serve to decrease validity in the derivation sample, increases in the number of predictors beyond some optimal size may well serve to fit too finely the idiosyncrasies of a derivation sample and thus lead to decreased cross-validity.

The weights which one could use in prediction functions (the  $a_i$  values in Equation 1) may utilize all of the information incorporated in the regression weight, or only some of it. Moreover, in either case the resulting weight may be specified quite precisely or only approximately. In addition, within any stepwise prediction scheme, potential functions differ in their order of prediction selection. And finally, the independent variables (the  $X_{ij}$  scores in Equation 1) may be used in their unaltered (raw score) form or they may be first transformed in some fashion. Let us briefly consider each of these possibilities.

(a) *The type of weights.* The two most common forms of

predictor weights are regression (or  $b$ ) coefficients, on the one hand, and validity (or  $r$ ) coefficients, on the other hand. The former ( $b$  coefficients) incorporate information about predictors other than that with which the weight is associated, while the latter ( $r$  coefficients) contain no such information. Thus, regression coefficients epitomize an *interdependent* approach to prediction, since information from many predictors is used simultaneously; in contrast, validity coefficients epitomize an *independent* focus, since the validity coefficient is in no way affected by the inclusion (or exclusion) of other predictors.

(b) *The precision of the weights.* Regardless of whether  $b$  or  $r$  coefficients are used, the actual values entered into the prediction function may vary from the "double-precision" figures afforded by some computers (e.g., 16 or more decimal places) at one extreme, through *integer* weights (0, 1, 2, etc.), to *unit* weights (0 vs. 1). For convenience, it is useful to contrast *precise* weights (e.g., four decimal places) with the most extreme (and convenient) alternative, namely *unit* weights.

(c) *The order of predictor selection.* While there are a number of predictor selection schemes available (e.g., Anderson & Fruchter, 1960; Garside, 1965; Horst & MacEwan, 1960; Linhart, 1960; Lubin & Summerfield, 1951; Lunneborg, 1967; Madden & Bottenberg, 1963; McDonald, 1968; Sclove, 1968; Summerfield & Lubin, 1951; Toops, 1941; Wherry & Gaylord, 1946), the most common methods involve selecting either by the size of the original validity coefficients or by the size of the residual validity coefficients after the variance associated with all predictors previously selected has been partialled out of the criterion scores. The former method (here termed *validity* ordering) has been traditionally used only with items, while the latter method (here termed *residual* ordering) is common to most of the computer algorithms for stepwise multiple-regression analysis. For any fixed number of predictors, the first and last step from both methods will produce the same results. Each of the intermediate steps, however, may produce functions with quite different cross-validities.

(d) *The use of original vs. transformed predictor variables.* Since each of the predictors is typically scaled in some different metric, these scores will differ in their means and variances. While the traditional regression weights are compensated for these scaling differences, validity weights are not. Consequently, one might elect to transform the predictors to a common metric, thereby equating the predictor score distributions. Since the most com-

mon method of equating such distributions is by transforming them to standard scores, predictor functions based upon *original* scores can be contrasted with analogous functions based upon *standardized* scores.

As should be apparent from this discussion, there are hosts of different prediction functions available for applied use; in addition to the 2<sup>4</sup> or 16 possibilities based upon the above four dichotomies, one could include a number of other stepwise ordering algorithms (including some which permit predictor deletion). On the other hand, of the 16 major prediction functions both *original* and *standardized* scores will provide the same results when precise *b* weights are used. Consequently, one can distinguish at least 14 unique prediction functions, and these are listed in Table 8. Of the 14 major functions, only Function 1 (precise *b* weights with residual ordering) and Function 13 (unit weights based upon, and ordered by, validity coefficients) have enjoyed any substantial popularity; the former functions are ubiquitously used with scale scores, while the latter functions are traditionally used with items.

Table 8  
The 14 Prediction Functions

Function	Type	Precision	Order	Metric
1	<i>b</i>	Precise	Residual	(either)
2	<i>b</i>	Precise	Validity	(either)
3	<i>b</i>	Unit	Residual	Original
4	<i>b</i>	Unit	Residual	Standardized
5	<i>b</i>	Unit	Validity	Original
6	<i>b</i>	Unit	Validity	Standardized
7	<i>r</i>	Precise	Residual	Original
8	<i>r</i>	Precise	Residual	Standardized
9	<i>r</i>	Precise	Validity	Original
10	<i>r</i>	Precise	Validity	Standardized
11	<i>r</i>	Unit	Residual	Original
12	<i>r</i>	Unit	Residual	Standardized
13	<i>r</i>	Unit	Validity	Original
14	<i>r</i>	Unit	Validity	Standardized

In the present project, the cross-validities of each of these 14 types of prediction functions were compared, using each of the five major inventories and all values of *n* from 1 to 11.<sup>4</sup> Table 9 presents these average cross-validities for each of the 14 functions, which are rank ordered in the table from most to least valid.

4. To compute the standardized scores, the means and standard deviations in each of the two derivation samples were used in the calculation of standard scores for all subjects in the cross-validation samples.

The values listed in Table 9 are averages across (a) both cross-validation samples, (b) the five major inventories, and (c) all 13 criteria (or, alternatively, the seven most predictable criteria). Perhaps the most striking feature of Table 9 is the remarkably small difference in average cross-validity coefficients associated with these 14 different prediction functions. These values ranged from .22 to .28 when based upon all 13 criteria, and from .32 to .39 when based upon the seven most predictable criteria.

Table 9  
A Comparison of the Cross-Validity Coefficients from the 14 Prediction Functions When Five Predictors Were Included in Each Function (Averages across the Five Major Inventories)

Function	Precision	Type	Order	Metric	Average Cross-Validities	
					All 13 Criteria	Most Predictable 7 Criteria
7	P	<i>r</i>	R	O	.28	.39
8	P	<i>r</i>	R	S	.28	.39
9	P	<i>r</i>	V	O	.27	.38
10	P	<i>r</i>	V	S	.27	.38
1	P	<i>b</i>	R	-	.27	.38
2	P	<i>b</i>	V	-	.27	.38
13	U	<i>r</i>	V	O	.26	.36
14	U	<i>r</i>	V	S	.25	.36
11	U	<i>r</i>	R	O	.24	.34
3	U	<i>b</i>	R	O	.24	.34
12	U	<i>r</i>	R	S	.24	.34
4	U	<i>b</i>	R	S	.23	.33
5	U	<i>b</i>	V	O	.23	.32
6	U	<i>b</i>	V	S	.22	.32
All 13 Criteria	P = .27 U = .24	<i>r</i> = .26 <i>b</i> = .24	R = .25 V = .25	O = .25 S = .25		
Most Predictable 7 Criteria	P = .38 U = .34	<i>r</i> = .37 <i>b</i> = .35	R = .36 V = .36	O = .36 S = .36		

Moreover, the average values presented at the bottom of Table 9 demonstrate that virtually all of this small variation in cross-validity can be accounted for by only two of the four facets by which these 14 functions were generated. The best functions were those employing precise correlational (*r*) weights, regardless of selection order and variable metric. While these results are strikingly confirmatory of earlier Monte Carlo findings presented by Marks (1966), it is important to note that the typical multiple-regression weights (Function 1) produced cross-validities which were virtually identical to those produced by the best functions

based upon correlation weights (Functions 7 and 8). In general, then, the results presented in Table 9 suggest that the very best functions for small sample analyses may be either *r* or *b* weights, computed precisely. At the other extreme, though not remarkably less valid, are the unit weighted functions, the worst being unit weights based upon regression (*b*) coefficients with validity ordering (Functions 5 and 6).

Let us now return to the original focus of this monograph and compare the relative effect sizes associated with the criteria, the strategies, and the number of predictors (*n*), as well as those associated with the four facets which generated the 14 different prediction functions. Such analyses, which should establish whether the results presented in previous sections of this monograph are limited to multiple-regression functions, should also highlight any important interactions among the parameters under study. The results from two such analyses, the first based upon all 13 criteria and the second based upon the seven most predictable cri-

teria, are summarized in Table 10. Included in this table are the sums of squares (*SS*), degrees of freedom (*df*), mean squares (*MS*), and proportion of *SS* associated with the following seven parameters: (*a*) the 13 (or the seven most predictable) criteria, (*b*) the five major strategies, (*c*) *n* values of 3, 5, 7, and 9, (*d*) the weight type (*b* vs. *r*), (*e*) the precision of the weights (precise vs. unit), (*f*) the selection order (residual vs. validity), and (*g*) the predictor metric (standardized vs. original). In addition, the largest of the 21 two-way and 35 three-way interactions are also tabled.

The results of these analyses suggest that the most potent sources of variation in cross-validity coefficients once again stem from the main effect due to criteria (i.e., the general predictability of the criteria) and to the criteria-by-strategies interaction effect. At the other extreme, effects due to the number of predictors included in the function, the weight type, the selection order, and the predictor metric—as well as most of the interaction effects—appear to be of minimal importance in affecting these average cross-validities. Moreover, since none of the two-way interactions involving scale construction strategies—other than the strategies-by-criterion interaction—showed any sizeable effects, it is now reasonable to assume that the earlier findings are not limited to the traditional multiple-regression paradigm.

Let us pause at this point to summarize some tentative conclusions. The results of the analyses presented thus far suggest that:

(*a*) The findings reported in Hase and Goldberg (1967)—that inventories constructed by the three major strategies of scale construction produced quite similar *average* cross-validities—were supported by more extensive analyses of these data. However, these later analyses indicated the existence of 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.

(*b*) Moreover, these results are not limited to analyses using all eleven scales in the prediction functions. While the optimal number of predictors for samples of this size appears to lie between three and seven, the number of predictors in each function (*n*) did not have a sizeable effect on the average cross-validities of the various inventories. In fact, neither the main effect due to *n*, nor any of the interaction effects involving this variable, were of any appreciable size.

(*c*) In addition, these results do not appear to be limited to

Table 10  
A Comparison of the Relative Effect Sizes Associated with the Criteria,  
Five Major Strategies, Number of Predictors, Weight Type,  
Weight Precision, Selection Order, and Variable Metric

Effect	Analyses Based Upon All 13 Criteria				Analyses Based Upon the 7 Most Predictable Criteria			
	<i>SS</i>	<i>df</i>	<i>MS</i>	% <i>SS</i>	<i>SS</i>	<i>df</i>	<i>MS</i>	% <i>SS</i>
Main Effects (7)	169.51	23	7.37	.70	23.37	17	1.37	.37
Criteria ( <i>C</i> )	163.92	12	13.66	.68	13.64	6	2.27	.21
Strategies ( <i>S</i> )	2.34	4	.59	.01	6.29	4	1.57	.10
No. Predictors ( <i>n</i> )	.29	3	.10	.00	.34	3	.11	.01
Weight Type ( <i>T</i> )	.32	1	.32	.00	.24	1	.24	.00
Precision ( <i>P</i> )	2.62	1	2.62	.01	2.84	1	2.84	.04
Selection Order ( <i>O</i> )	.01	1	.01	.00	.00	1	.00	.00
Metric ( <i>M</i> )	.01	1	.01	.00	.01	1	.01	.00
2-way								
Interactions (21)	28.62	178	.16	.12	14.52	112	.13	.23
<i>C</i> × <i>S</i>	23.70	48	.49	.10	11.53	24	.48	.18
<i>C</i> × <i>P</i>	1.25	12	.10	.01	.40	6	.07	.01
3-way								
Interactions (35)	9.23	646	.01	.04	5.25	370	.01	.08
<i>C</i> × <i>S</i> × <i>n</i>	2.26	144	.02	.01	.85	72	.01	.01
<i>C</i> × <i>S</i> × <i>T</i>	1.93	48	.04	.01	.94	24	.04	.01
4-way								
Interactions (35)	5.38	1229	.00	.02	3.05	665	.00	.05
5-way								
Interactions (21)	1.60	1267	.00	.01	.96	661	.00	.01
6- and 7-way								
Interactions (8)	.23	816	.00	.00	.16	414	.00	.00
Residual	26.46	4160	.01	.11	16.66	2240	.01	.26
Total	241.02	8319		1.00	63.96	4479		1.00

analyses based upon the traditional multiple-regression paradigm. Specifically, similar results were found for the 14 different prediction functions generated by variations in weight type ( $b$  vs.  $r$  coefficients), weight precision (precise vs. unit weights), predictor selection order (validity vs. residual orders), and predictor metric (standardized vs. original scores). Of these 14 prediction functions, the best were precise correlation weights with residual selection order (regardless of metric); however, the traditional precise regression weights were of approximately equal validity. In general, weight precision and weight type showed the largest effects—and selection order and predictor metric the least—though differences in cross-validity between the best and the worst of these 14 functions were not great.

## CHAPTER V ADDITIONAL ANALYSES

### *Orthogonal Components vs. Original Scales*

A number of psychometricians (e.g., Burket, 1964; Cureton, 1951; Herzberg, 1969; Horst, 1941; Lunneborg, 1967; McCornack, 1970) have advocated a two-stage approach to multiple prediction. First, the original set of predictors is factor analyzed to form a smaller set of component scores (each new component orthogonal to all of the others); these orthogonal components are then treated as the independent variables in the prediction function. Since there are fewer components than original predictors, since the components are uncorrelated with each other, and since the components should be more reliable than the original predictors, many of the traditional interpretive and empirical problems associated with multiple-regression analysis might be attenuated by their use.

To test this hypothesis, the 11 scales in each of the five major inventories were factor analyzed in turn, using the entire (combined) sample of subjects and a standard principal components solution; factor scores for each subject, based on each of the first five unrotated principal components, were then included in stepwise multiple-regression analyses, using the same design employed in previous analyses. The cross-validities associated with each of these five sets of components, one set from each inventory, were compared with the analogous values based upon the original scale scores from the same inventory. These comparisons are summarized in Table 11. The results are presented separately for each of the five major inventories and for each of five values of  $n$  (the number of predictors in the regression function). Consequently, each cell of the table provides a comparison between components and scores from the same inventory and for the same value of  $n$ ; the highest coefficient within each comparison pair has been indicated by an asterisk. Analyses based on all 13 criteria are presented in the upper half of the table, while those based on the 7 most predictable criteria are presented in the lower half.

In general, the results displayed in Table 11 indicate no great difference in the *average* cross-validity between scores and components (roughly half of the comparisons favoring each kind of procedure). However, two sorts of interaction effects stand out in the table: (a) For low values of  $n$ , scales fared better than

Table 11

A Comparison of Original Scores and Orthogonal Components:  
Average Cross-Validity Coefficients for Each of the Five Major Inventories  
and Differing Numbers of Predictors in the Regression Function

No. of Predictors ( <i>n</i> )	Kind	Inventory					Average
		EMP	FAC	MSA	THE	RAT	
1	Scales	.21*	.23*	.21*	.23	.29*	.23*
	Components	.19	.18	.19	.24*	.22	.20
2	Scales	.26*	.27*	.22	.24	.31*	.26*
	Components	.19	.24	.23*	.27*	.25	.24
3	Scales	.27*	.28*	.22	.25	.32*	.27*
	Components	.24	.26	.25*	.29*	.28	.26
4	Scales	.27*	.28*	.23	.25	.32*	.27
	Components	.25	.27	.25*	.30*	.30	.27*
5	Scales	.27*	.28*	.23	.26	.31	.27
	Components	.25	.28	.25*	.30*	.31*	.28*
1	Scales	.30*	.35*	.30*	.36*	.39*	.34*
	Components	.25	.23	.28	.36	.32	.29
2	Scales	.34*	.39*	.33	.38	.42*	.37*
	Components	.25	.33	.34*	.40*	.37	.34
3	Scales	.34*	.42*	.34	.38	.44*	.38*
	Components	.33	.36	.37*	.42*	.40	.38
4	Scales	.32	.42*	.35	.39	.42*	.38
	Components	.34*	.38	.36*	.43*	.42	.39*
5	Scales	.32	.42*	.34	.40	.42	.38
	Components	.34*	.39	.36*	.43*	.43*	.39*

Note.—Values based on all 13 criteria are presented in the upper half of the table, while those based on the 7 most predictable criteria are presented in the lower half. The highest value within each comparison pair is indicated by an asterisk (\*).

components, while at higher values of *n*, the reverse was true. For example, at *n* = 1, 90% of the comparisons favored the scales, while at *n* = 5, 80% of the comparisons favored the components. (b) An even stronger interaction occurred for strategies. For the *Factor* inventory, 100% of the comparisons—and for the *Empirical* and *Rational* inventories, 80% of the comparisons—favored the scales. On the other hand, for the *Theoretical* and *MSA* inventories, 90% and 80% (respectively) of the comparisons favored the components. Moreover, a series of ANOVA analyses of these data revealed the presence of some additional strong interaction effects. For example, the components were relatively superior in predicting peer ratings of Responsibility (*cRES*), Psychological-mindedness (*cPSY*), and Femininity (*cFEM*), and the original scores in predicting Dating Frequency (*cDAT*), Sorority Affiliation (*cSOR*), and peer-rated Sociability (*cSOC*). Moreover, the former effect was particularly striking for the *Theoretical* inventory, while the latter effect was even more dramatic for the

*Factor* inventory. These complex interactions among criteria, strategies, and kinds of predictors—coupled with the general equality of the original scales and the orthogonal predictors—demonstrate the difficulties which may be involved in trying to decide a priori on the relative efficacy of using original scales vs. orthogonal components in future studies.

#### *Linear vs. Nonlinear and Configural Terms*

All of the preceding analyses have been based on the general linear model, the basic assumptions of which are that (a) the bivariate relationships between the scales and the criteria can best be approximated by a straight line, and (b) the multivariate relationships between sets of scales and each criterion can best be approximated by a flat or planar surface. On the other hand, it seems intuitively obvious that at least *some* scale vs. criterion relationships should be curvilinear in character, and, moreover, there ought to be *some* interactive or configural relationships involving two or more scales. To test whether the addition of nonlinear and/or configural terms would significantly increase predictive accuracy, the 11 scales from each major inventory were augmented by a set of *nonlinear* terms, namely their squares and their square roots, plus a set of *configural* terms, namely the 55 cross-products from all possible pairs of scales. For each inventory in turn, a series of stepwise multiple-regression analyses was carried out, each employing one of the following four sets of terms: (a) the 11 original scale scores, (b) a set of 33 linear and nonlinear terms (the 11 original scales, the 11 squares, and the 11 square roots), (c) a set of 55 configural terms (the 55 cross-products), and (d) all 88 terms. A comparison of the cross-validities from these four sets of terms, when five predictors were included in each regression equation, is presented in Table 12. As in previous tables, values averaged across all 13 criteria are listed in the upper half of the table, while those averaged across the seven most predictable criteria are listed in the lower half. The highest coefficient within each comparison set is indicated by an asterisk.

The results displayed in Table 12 are illustrative of other findings using different numbers of predictors (i.e., values of *n* from one to eleven) and different types of prediction functions. In general, there were no substantial gains in cross-validity for any inventory achieved by the inclusion of either nonlinear or configural terms (or both), and the average cross-validities across all

Table 12  
The Effect of Employing Nonlinear and Configural Terms:  
The Average Cross-Validity for Each of the Five Major Inventories  
When Five Predictors Were Included in the Regression Function

	<i>k</i>	Inventory					Average
		EMP	FAC	MSA	THE	RAT	
Original Scales	11	.27*	.28*	.23*	.26	.31*	.27*
Scales + Nonlinear	33	.26	.25	.19	.23	.28	.24
Configural Terms	55	.25	.26	.16	.27*	.29	.25
All Terms	88	.24	.25	.19	.25	.28	.24
Original Scales	11	.32	.42*	.34*	.40	.42*	.38*
Scales + Nonlinear	33	.33*	.39	.30	.38	.38	.36
Configural Terms	55	.31	.37	.28	.41*	.39	.35
All Terms	88	.30	.38	.32	.40	.39	.36

Note.—Values based on all 13 criteria are presented in the upper half of the table, while those based on the 7 most predictable criteria are presented in the lower half. The highest coefficient within each comparison set is indicated by an asterisk (\*).

five major inventories were typically slightly higher for the 11 original scales than for any of the three comparison sets of terms. Thus, it is apparent that any gains in initial validity stemming from nonlinear and/or configural relationships in these data are overshadowed by the increased shrinkage on cross-validation produced by the addition of new parameters.

However, it might be argued that these general analyses might have obscured some significant increase in cross-validity for one or two criteria. To test this hypothesis, the average cross-validities associated with each criterion were examined carefully, for each of the major inventories in turn, and at every value of *n*. Again, the results were negative: no criterion was consistently better predicted by the functions which included nonlinear or configural terms than by the analogous functions based upon the original scale scores. Moreover, a series of ANOVA analyses of these data uncovered no criteria-by-models interaction effects of any appreciable size. In summary, then, these findings corroborate the results from previous attempts to discover nonlinear or configural relationships for other problems; the more complex models, while extracting more variance than the linear model from derivation samples, result in so much greater shrinkage upon cross-validation that they end up providing no incremental validity over the simpler linear model (Goldberg, 1965, 1968, 1969; Lunneborg & Lunneborg, 1967a, 1967b; Stilson & Astrup, 1966; Ward, 1954).

#### Other Sets of Scales

All of the preceding analyses have employed the five major

11-scale inventories developed for this project. At this point, it should be instructive to compare the average cross-validities produced by these experimental inventories with those from some other sets of CPI scales. The results from stepwise multiple-regression analyses for two such comparison sets are presented in Table 13. Included in the table are the cross-validated correlations (averaged across all 13 criteria and across the seven most predictable criteria) for (a) a set of 11 abbreviated MMPI scales, and (b) the 18 scales included in the published version of the CPI (the 11-scale *Empirical* inventory, four scales from the *Rational* inventory, and three stylistic scales). Since the CPI includes 178 items taken directly from the MMPI, plus 35 more MMPI items which were revised slightly, it is possible to score abbreviated versions of a number of MMPI scales from the CPI item pool (see Rodgers, 1966). Specifically, the following 11 MMPI scales were scored (the figures in parentheses indicate the proportion of items in each of the original MMPI scales which were included in the corresponding CPI scoring keys): *K* (43%), *Hs* (30%), *D* (43%), *Hy* (43%), *Pd* (60%), *Mf* (47%), *Pa* (42%), *Pt* (44%), *Sc* (36%), *Ma* (48%), and Barron's (1953) Ego-strength scale (34%).

Table 13  
Average Cross-Validity Coefficients for 11 Abbreviated MMPI Scales  
and for the 18 Standard CPI Scales as a Function of the Number of  
Predictors Included in the Regression Equation

Number of Predictors Included in the Regression Equation	Based Upon All 13 Criteria		Based Upon the 7 Most Predictable Criteria	
	11 MMPI	18 CPI	11 MMPI	18 CPI
1	.11	.20	.18	.30
3	.16	.24*	.25	.33*
5	.17	.23	.28	.33
7	.18	.23	.29*	.31
9	.17	.22	.28	.30
11	.18*	.22	.28	.31

\*Highest value in the column.

The results displayed in Table 13 indicate that this set of abbreviated MMPI scales produced considerably lower cross-validities than did any of the five major inventories used in the present project (see Table 6). Moreover, the employment of all 18 standard CPI scales served to decrease cross-validity over that achievable by the set of 11 Empirical scales alone, as can be seen by comparing the values in Table 13 with those in the first column of

Table 6. While this result may initially appear paradoxical, the finding is easily explainable using the same rationale discussed earlier in this monograph. Any procedure which serves to fit too finely the vagaries of a particular derivation sample may result in decreased cross-validity; in the present case, any increase in validity stemming from the addition of seven scales to the 11-scale *Empirical* inventory was more than compensated by the increased shrinkage of these terms upon cross-validation.

This same effect can be seen in the last two columns of Table 14, which summarizes some stepwise regression analyses based upon an initial set of 61 CPI scales. This set included all 55 scales from the five major inventories, four Stylistic scales (the three included in the published version of the CPI, plus Dicken's [1963] Social Desirability scale), and two scales constructed by Nichols and Schnell (1963) to mark the first two CPI factors. For comparison, the results for the *Rational* inventory (from Table 6) are also presented. As in previous tables, the cross-validities averaged across all 13 criteria are presented in the upper half of the table, while those averaged across the seven most predictable criteria are presented in the lower half. Note that, for all values of  $n$ , the 11-scale *Rational* inventory was slightly more valid than the entire 61-scale set, a set in which the 11 *Rational* scales were embedded. In general, the inclusion of 50 additional scales to the

Table 14  
Average Cross-Validity Coefficients for a Set of 61 CPI Scales  
and for Four Sets of 11 Factor Scores Based on Those Scales

Number of Predictors Included in the Regression Equation	Factor Scores Based Upon 61 Scales				Original 61 Scales	11 <i>Rational</i> Scales
	Unrotated		Rotated			
	Comp.	Factors	Comp.	Factors		
1	.22	.20	.21	.22	.26	.29*
3	.29	.30	.30	.30	.28	.32*
5	.31	.33*	.30	.30	.27	.31
7	.33	.33*	.32	.32	.28	.29
9	.33*	.33	.32	.33	.27	.28
11	.33	.33*	.33	.33	.26	.28
1	.29	.27	.28	.30	.38	.39*
3	.40	.43	.41	.40	.38	.44*
5	.43	.45*	.41	.42	.35	.42
7	.44	.45*	.43	.44	.36	.40
9	.44	.45*	.43	.44	.36	.39
11	.44	.44*	.44	.44	.35	.39

Note.—Values based upon all 13 criteria are presented in the upper half of the table, while those based upon the 7 most predictable criteria are presented in the lower half.

\*Highest value in each row.

initial predictor pool served to decrease the average cross-validity about .03 across all 13 criteria and about .04 across the seven most predictable criteria.

However, as noted earlier in this monograph, one procedure for counteracting the increased shrinkage associated with large numbers of parameters is to transform the original predictors into a smaller set of orthogonal components. While previous analyses showed that the use of such components was not unambiguously successful when applied to each of the 11-scale inventories, such a procedure should be far more effective when applied to this larger set of predictors. To test this hypothesis, the correlations among these 61 scales were factor analyzed, using the entire sample of 152 subjects and using each of four different factor-analytic procedures: (a) an unrotated principal components solution (i.e., unities in the diagonals of the correlation matrix), extracting 11 orthogonal components, (b) a Varimax rotation of these 11 components, (c) an unrotated principal factors solution (i.e., communalities in the diagonal), again extracting 11 factors, and (d) a Varimax rotation of these 11 factors. Four sets of factor scores were computed for each subject, one set based upon each of these four procedures. The results of using each of the four sets of 11 orthogonalized scores in stepwise multiple-regression analyses are presented in the first four columns of Table 14.

As expected, differences in average cross-validity among the four sets of scores produced by the different factor-analytic tactics were negligible, since each procedure merely involves a redistribution of the same scale covariance. While the Unrotated Factors generally produced the highest values, their incremental validity over the other three procedures was minimal. Though all four of the orthogonalized scores produced higher cross-validities than any of the original scale sets for all regression analyses based upon five or more predictors, the original scales provided higher cross-validities for  $n$  values between one and three. Moreover, the *Rational* inventory produced cross-validities at  $n = 3$  which were approximately as high as those produced by the best of the orthogonalized scores for higher values of  $n$ .

#### Combining Items vs. Combining Scales

Perhaps the most fundamental issue of inventory usage concerns the comparative validity of *any* inventory with some other method of prediction. Specifically, some exponents of the External strategy of scale construction have suggested that one might

best develop a new scale for each criterion of interest, rather than constructing a new prediction function from a standard set of scales. Such an approach is based upon the implicit assumption that linear combinations of items will generally be more valid than will linear combinations of scales. To test this hypothesis, the 480 CPI items were divided into two equal sets (the 240 odd-numbered vs. the 240 even-numbered items), and a series of stepwise analyses were carried out on each set, using the same design previously described (see Fig. 1) and using six different prediction functions (Functions 1, 3, 7, 9, 11, and 13 in Table 8). The average cross-validities, based upon the 240 odd-numbered CPI items, for scales of various lengths (i.e.,  $n$  values of 1, 3, 5, 7, 9, 11, 18, 30, 45, 61, 90, and 150) are presented in Table 15. Corresponding values for the 240 even-numbered items, which are available from the author, surprisingly produced average cross-validities which

Table 15  
Average Cross-Validity Coefficients for the 240 Odd-Numbered  
CPI Items Based Upon Six Different Prediction Functions

No. of Items (Scale Length)	Correlation ( $r$ ) Weights				Regression ( $b$ ) Weights	
	Validity Order Precise	Unit	Residual Order Precise	Unit	Residual Order Precise	Unit
1	.16	.16	.16	.16	.16	.16
3	.20	.20	.20*	.19	.20	.19
5	.23*	.23	.20	.17	.18	.17
7	.25*	.25	.21	.19	.19	.18
9	.27*	.27	.20	.17	.18	.16
11	.28*	.27	.21	.18	.19	.18
18	.30*	.30	.22	.17	.19	.17
30	.30*	.29	.24	.19	.17	.14
45	.31*	.30	.25	.19	.16	.12
61	.31*	.30	.27	.20	.16	.12
90	.32*	.30	.29	.21	.18	.12
150	.32*	.30	.30	.22	.18	.09
1	.24	.24	.24	.24	.24	.24
3	.33*	.33	.29	.27	.29	.27
5	.35*	.34	.30	.27	.27	.27
7	.36*	.35	.32	.27	.27	.27
9	.38*	.38	.30	.25	.26	.25
11	.39*	.38	.31	.26	.26	.25
18	.42*	.41	.32	.23	.26	.24
30	.43*	.41	.33	.25	.26	.22
45	.44*	.43	.36	.27	.26	.21
61	.43*	.41	.37	.29	.25	.21
90	.44*	.41	.41	.32	.27	.19
150	.44*	.41	.42	.34	.27	.15

Note.—Values based on all 13 criteria are presented in the upper half of the table, while those based on the 7 most predictable criteria are presented in the lower half. The highest value in each row is indicated by an asterisk (\*).

were generally about .06 lower than those based upon the odd-numbered items.

The results displayed in Table 15, as well as those based upon the 240 even-numbered items, indicate that correlation ( $r$ ) weights were dramatically better than regression ( $b$ ) weights for combining items. Among the four varieties of correlation ( $r$ ) weights, the highest cross-validities were produced by precise weights ordered by validity coefficients (Function 9). The traditional type of weights employed with items, unit weights based on validity ordering (Function 13), produced cross-validities only slightly higher than those produced by precise weights based upon residual ordering (Function 7). The worst of the four correlation weights, though still superior to either of the regression weights, were unit weights based upon residual ordering (Function 11). Perhaps of greatest importance, scale length was generally monotonically related to cross-validity for all four of the correlation weights (the longer the scales, the higher the validities). On the other hand, even the 150-item scales produced average cross-validities no higher than those produced by the *Rational* inventory (see Table 6).

Since these results were all based upon only one-half the total CPI item pool, some potentially powerful items may have been excluded, thereby biasing the comparisons between items and scales. To check on this possibility, all 480 CPI items were included in a series of stepwise analyses, using the two prediction functions which had previously turned out to be most useful at the item level (Functions 9 and 13: precise and unit correlational weights, based upon validity ordering). The results were virtually identical to those found earlier, though the 480-item set produced average cross-validities which were about .01 lower than those based upon the 240 odd-numbered items.

#### A Final Comparison

Previous analyses in this monograph have indicated that few of the myriad strategies and tactics under study have provided average cross-validities any higher than those produced by the 11-scale *Rational* inventory. Moreover, Goldberg and Hase (1967) have shown that if the least reliable scale (*Fem*) is eliminated from the seven *Rational* scales constructed by Hase, the resulting six-scale subset (Sociability [*Soc*], Responsibility [*Res*], Achievement [*Ach*], Conformity [*Con*], Dominance [*Dom*], and Psychological-mindedness [*Psy*]) produced average cross-validities which were slightly higher than those produced by the entire *Rational*

inventory. Moreover, a five-scale Rational subset (*Soc, Res, Ach, Con, and Dom*) produced average cross-validities of .36 across all 13 criteria, the highest value from any of the prediction sets employed in this project; this value was virtually identical to those produced by the six-scale and one four-scale (*Soc, Res, Ach, and Con*) subsets, both of which produced average cross-validities of .35.

The results based on the five-scale Rational subset are presented in the top row of Table 16, which serves to summarize some of the major analyses from this monograph. Sixteen different predictor sets are ordered in Table 16 by their average cross-validity across all 13 criteria. The tabled values represent the *highest* average cross-validities produced by that set, and included in the table are both the particular type of prediction function and the particular value of *n* (the number of predictors included in the function) which produced that value. Consequently, Table 16 allows a summary appraisal of the optimal kind and number of CPI predictors, as well as the optimal type of prediction function, for criteria of this sort and for samples of this size.

Table 16  
Summary Table: The Highest Average Cross-Validity Coefficients

	Prediction Function				Average Cross-Validity Coefficients	
	<i>n</i>	Number <sup>a</sup>	Type	Order	All 13	7 Most
					Criteria	Predictable
5 Rational Scales	5	1	<i>b</i>	R	.36	.44
11 Unrotated Factors	7	9	<i>r</i>	V	.34	.46
11 Rational Scales	3	7	<i>r</i>	R	.32	.44
240 Odd-numbered Items	150	9	<i>r</i>	V	.32	.44
61 Scales	15	7	<i>r</i>	R	.32	.42
11 Theoretical Scales	9	9	<i>r</i>	V	.29	.40
11 Factor Scales	5	7	<i>r</i>	R	.28	.42
11 Empirical Scales	4	7	<i>r</i>	R	.28	.34
11 MSA Scales	10	7	<i>r</i>	R	.28	.39
240 Even-numbered Items	150	9	<i>r</i>	V	.27	.38
18 Standard CPI Scales	5	7	<i>r</i>	R	.27	.36
11 Abbreviated MMPI Scales	7	1	<i>b</i>	R	.18	.29
11 Random-True Scales	6	1	<i>b</i>	R	.15	.19
11 Stylistic Scales	4	9	<i>r</i>	V	.13	.19
11 Random-I Scales	4	7	<i>r</i>	R	.10	.18
11 Random-II Scales	3	7	<i>r</i>	R	.04	.00

<sup>a</sup>From Table 8.

The results displayed in Table 16 suggest that the five-scale subset from the *Rational* inventory, the 11 Unrotated Factors (based upon factor analysis of all 61 CPI scales), the 11-scale *Rational* inventory, the 240 odd-numbered CPI items, and all 61 CPI scales provided the highest average cross-validity coefficients. Only slightly less valid were the other four major inventories, the 240 even-numbered items, and the standard 18 CPI scales. Considerably less valid, however, were the 11 abbreviated MMPI scales and the four inventories constructed by the two "control" strategies, Stylistic and Random. For the 16 sets of predictors, 13 of the highest values were produced by correlation (*r*) weights (8 ordered by residual coefficients and 5 ordered by validity coefficients). In all three cases where regression (*b*) weights produced the highest cross-validities, residual ordering was employed. The optimal values of *n* ranged roughly from three to seven for most of the shorter scale sets, though the highest cross-validity for the 61-scale set was achieved when *n* was 15. For items, on the other hand, the optimal values of *n* were dramatically higher.

*Strategies of Inventory Construction*

A brief examination of the origins of the strategies may help to put the findings from this project into some perspective; for a more complete historical survey, see Goldberg (1971). 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 was excluded from the scale, and (c) the method of combining or weighting items to obtain a scale score was appropriate for the trait (see Loevinger, 1957).

By the late 1930's, a number of psychologists began to argue that psychology had not yet reached the stage where trait relevance could be reliably and validly intuited (e.g., Meehl, 1945). 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 non-

subjective index of item validity. 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 difference factors 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 1950's a new force entered the fray. Many psychologists have long had a dim view of personality inventories constructed by any strategy, since all scales seemed to be far 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” (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). Moreover, in the present project, the Stylistic inventories performed at the same level as the two Random inventories, all of them being considerably less valid than the five inventories constructed from the three major strategies. However, it is the relative performance among the five major inventories that most merits discussion, since some of these results run counter to prevailing psychometric lore.

For example, the *Empirical* inventory might have been expected to perform more validly than the others, since criteria providing direct cross-validation targets for six Empirical scales were included in this study. On the other hand, the scales developed for the *Empirical* inventory were constructed from different samples than those used for purposes of cross-validation, while the scales developed for the *Factor* and *MSA* inventories were constructed from a larger sample of the same subjects used in this project. Consequently, any idiosyncratic characteristics of Gough's original samples could have served to reduce the validity of the *Empirical* inventory under cross-validation conditions. The validity of the two Internal inventories, on the other hand, may have been increased slightly, due to the fact that they were constructed from essentially the same sample used for cross-validation purposes.

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 two Intuitive inventories performed as validly as they did in the present study. 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. The Intuitive strategy, in contrast, is minimally dependent on sample-specific characteristics; only at the stage of scale "purification" (e.g., discarding items with low correlations with scale scores) do sample characteristics have any chance to enter the scale con-

struction process. 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 critical than the sampling of subjects, and thus the Intuitive strategy has lost some favor in the psychometric community. One of the main lessons from the present project may be that such judgmental biases are not as crucial as has previously been believed.

In an important theoretical article, Jackson (1971) forcefully made this very 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 results reported in the present monograph, as well as those described in Hase and Goldberg (1967), 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*. Of the relatively few comparative studies of scales constructed by different strategies, most have either focused exclusively on internal psychometric properties of the scales (e.g., Neill & Jackson, 1970; Pearson, 1970; Poppleton & Pilkington, 1963) or have employed but one or two external criteria (e.g., Alumbaugh, Davis & Sweney, 1969; Butt & Fiske, 1968, 1969; DuBois, Loevinger, & Smith, 1956; Heilbrun, 1962; Hermans, 1969; Yee & Kriewall, 1969). Only the studies by Crewe (1967) and Schaie (1963) employed a multidimensional set of criteria, and in neither of these two studies were the External and Intuitive strategies compared. Yet studies which include a diverse array of criteria may well be necessary if the most surprising aspect of the present project is to be replicated, namely the finding that the External strategy produced an inventory of slightly broader band-

width than those produced by either the Internal or Intuitive strategies.

And this tantalizing finding certainly *must* be replicated, since it 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 possess narrow band-width across a range of non-target criteria), while the more homogeneous and independent sets of scales produced by the Internal and Intuitive strategies should be of relatively wider band-width when these scales are combined via multiple-regression procedures. Yet the findings from the present project suggest that the less homogeneous scales from the *Empirical* inventory may include some personally relevant variance which is not included in the inventories developed 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.

The five middle columns of Table 17 display this criteria-by-strategies interaction effect; presented in the table are the average cross-validities for each criterion and for each of the five major inventories, when five scales were included in each regression function. The five inventories are ordered in the table by their cross-validity averaged across all 13 criteria, and the criteria are ordered by their general predictability averaged across all of the inventories. The means and standard deviations in Table 17 were originally computed from the Z-transformed multiple correlations and later transformed back to the correlation metric. Note that, of the five major inventories, the *Empirical* inventory produced the smallest amount of variation in its cross-validities; the standard deviation of these values across the 13 criteria was about .10 for the *Empirical* inventory, as compared to values around .20 for the *Factor* and *Theoretical* inventories. As an alternative method of viewing this interaction effect, for each inventory in turn, the average cross-validity based upon the six least predictable criteria was subtracted from the analogous average based upon the seven most predictable criteria; on this index, the *Empirical* inventory produced a correlational difference of .13, as compared to analogous values of .34 for the *Factor* and *Theoretical* inventories and .26 for the *Rational* and *MSA* inventories. Moreover, these results are consistent with other (untabulated) analyses based upon differing values of *n* (the number of predictors in the prediction function) and different types of functions.

Table 17  
The Criteria-by-Strategies Interaction Effect:  
Average Cross-Validity Coefficients when Five Scales  
Were Included in Each Regression Function

Criteria	Five Rational Scales	Major 11-Scale Inventories					Mean
		RAT	FAC	EMP	THE	MSA	
Sociability ( <i>cSOC</i> )	.52	.52	.60*	.38	.55	.41	.50
How Well Known ( <i>cHWK</i> )	.49*	.41	.44	.30	.41	.30	.40
Femininity ( <i>cFEM</i> )	.40	.42*	.41	.39	.37	.29	.38
Dominance ( <i>cDOM</i> )	.46*	.42	.33	.38	.41	.24	.38
Dating Freq. ( <i>cDAT</i> )	.35	.45	.26	.24	.38	.51*	.37
Responsibility ( <i>cRES</i> )	.45*	.35	.37	.32	.39	.29	.36
Sorority Affil. ( <i>cSOR</i> )	.39	.34	.49*	.24	.28	.35	.35
G.P.A. ( <i>cGPA</i> )	.44*	.32	.21	.18	.08	.23	.24
Psychol.-mindedness ( <i>cPSY</i> )	.29*	.21	.15	.25	.25	.07	.21
Yielding ( <i>cYLD</i> )	.28	.19	.08	.30*	.15	.10	.19
College Major ( <i>cMAJ</i> )	.25	.15	.20	.27*	.02	.11	.17
Achievement ( <i>cACH</i> )	.26*	.15	-.05	.10	-.06	.01	.07
College Dropout ( <i>cCDO</i> )	.04	-.01	.01	.07*	.01	-.02	.02
Mean	.36*	.31	.28	.27	.26	.23	.28
$\sigma$	.14	.16	.21	.11	.21	.17	.17

\*Highest value in each row.

Since the major methodological finding from the present project is this criteria-by-strategies interaction effect, future research should certainly seek to discover its generality. On the other hand, as the first column of Table 17 indicates, the subset of five Rational scales may serve quite adequately for a host of practical assessment problems. This five-scale mini-inventory achieved higher cross-validities than any of the scale sets under study in this project, including all nine 11-scale inventories, the 18 scales included in the commercially-published version of the CPI, and the total set of 61 CPI scales (see Table 16). Moreover, even when based upon the six least predictable criteria, the five Rational scales produced relatively high average cross-validities. Thus, the five-scale set emerged as both a relatively broad band-width *and* a relatively high fidelity instrument, with this sample of subjects and criteria. Again, as with the criteria-by-strategies interaction effect, this finding *must* be replicated in diverse settings. With this goal in mind, the CPI items included in each of the five Rational scales are presented in Appendix A.

#### *Tactics of Multivariate Prediction*

Investigators A and B are independently seeking to predict the same criterion, and unknown to each other they have both

used the same sample of 100 subjects. Investigator A administers a good, short personality inventory (say, six scales) to the sample, while Investigator B administers a much longer and more comprehensive instrument (say, all 53 scales from the new Edwards Personality Inventory). Both investigators derive a five-scale weighted composite via stepwise multiple-regression procedures. Which of these two composites is most likely to produce the higher cross-validity when Investigator C compares them in a new sample? If you place your bets on Investigator B's composite (the best five-scale subset of 53 scales), your intuition is similar to that of most psychologists—and virtually all graduate students. Nonetheless, the results from this project suggest that Investigator A's composite would likely show the highest cross-validity, even though he neglected to measure at least 47 potentially important scales which were included in Investigator B's battery.

Moreover, experimental and Monte Carlo studies of different prediction models (e.g., Burket, 1964; Forgy, 1962; Halinski & Feldt, 1970; Lawshe & Patinka, 1958; Lawshe & Schucker, 1959; Marks, 1966; McCornack, 1970; Rock, Linn, Evans, & Patrick, 1970; Schmidt, 1971; Shine, 1966; Wesman & Bennett, 1959) have generally produced results similar to those found in the present project. While psychologists are generally aware of the effect upon cross-validity of differing numbers of predictors in the prediction function ( $n$ ), it is easy to forget more subtle sources of capitalization on chance. Whether one is adding new scales to an initial battery, using large numbers of items instead of a much smaller number of scales, including nonlinear and/or configural terms in one's prediction function, or even using such dynamic prediction models as multiple regression, one is always sacrificing precious degrees of freedom (see Cureton, 1951).

Unfortunately, such findings present the psychometrician with a most difficult dilemma. Since present-day psychological tests and scales are not all that potent as predictors of important criteria, the search for new measures must continue. On the other hand, the proliferation of predictors leaves the applied investigator with an *embarras de richesses*, since the more predictors he includes in his initial battery, the more likely he is to decrease the cross-validity of his predictions in new samples. The lesson here may be a tough one: investigators are going to need as much as an order of magnitude increase in typical sample size if they are to utilize optimally the large predictor sets, the huge item pools, or the multi-parameter prediction models now available.

Moreover, investigators of the comparative validity of personality inventories are plagued with an even more difficult problem: they must simultaneously try to maximize the size of their sample of subjects and the size of their sample of criteria, if their findings are to be generalizable across criterion situations and across subject samples. Given a finite amount of money and testing time, one would normally opt either to increase his sample size at the cost of decreasing the number of criteria available for each subject, or conversely to increase the range of criteria by the intensive study of a relatively small sample of subjects. In the first case, one may be expected to lose generality across criteria, and in the second, generality across subject samples. In either case, the basic endeavor—that of comparing the validity of personality inventories—must suffer.

#### *A Paradigm for the Experimental Investigation of Personality Inventories*

Clearly, the findings from the present project can be generalized only with caution. The present investigation, limited to one item pool and one method of item presentation, casts no light on the validity of inventories where other response formats and other item content are employed. Moreover, the 13 criteria utilized in this study can hardly be seen as comprehensive. Psychologists are often called upon to forecast job success, personal and marital happiness, differential diagnosis, and a host of other criteria not included in the present investigation. While no single study will include them all, some replications of the present work must be carried out in other kinds of settings, and with different sorts of subjects.

The function, then, of the present monograph is not to provide answers to all of the critical questions on inventory construction, but rather to point the way whereby these questions can begin to be answered. As Cattell (1964) has so aptly noted: "the *genuine* advances in a science can often be evaluated, despite the smoke screens of pretentious theorizing, by the amount of actual technical control which ensues. . . . For the advances in *technical* precision and predictive control . . . will quickly offer evidence on *theoretical* soundness, convincing even for those relatively unversed in the theory [p. 70]." The present monograph, meant to provide a step along this path, hopefully offers a paradigm for gauging "the amount of technical control" offered by present strategies and tactics of personality inventory construction.

## APPENDIX A

### The Items in Five Rational CPI Scales

#### Sociability (Soc)

True: 1, 21, 52, 77, 102, 108, 119, 143, 163, 167, 208, 218, 242, 251, 280, 287, 346, 395  
 False: 38, 40, 57, 74, 83, 109, 111, 124, 134, 156, 159, 182, 227, 236, 252, 284, 285, 286, 334, 416, 418, 461

#### Responsibility (Res)

True: 14, 51, 112, 149, 162, 181, 195, 221, 234, 235, 260, 278, 312, 323, 380, 389, 442, 473  
 False: 43, 49, 73, 101, 117, 120, 139, 145, 185, 203, 253, 262, 275, 297, 307, 331, 374, 388, 420

#### Achievement (Ach)

True: 6, 50, 61, 84, 95, 103, 122, 140, 166, 181, 204, 222, 224, 228, 246, 256, 260, 269, 283, 292, 391, 408  
 False: 54, 94, 99, 116, 121, 145, 169, 185, 230, 326, 331, 352, 422, 436, 450, 456

#### Conformity (Con)

True: 7, 58, 88, 127, 165, 198, 212, 223, 229, 255, 260, 263, 276, 290, 304, 305, 314, 348, 385, 387, 462, 478  
 False: 29, 170, 250, 268, 275, 302, 339

#### Dominance (Dom)

True: 6, 37, 50, 53, 81, 102, 179, 180, 200, 202, 224, 239, 256, 267, 319, 320, 346, 355, 359, 376, 403, 412, 448, 476  
 False: 7, 11, 13, 25, 31, 111, 134, 177, 227, 258, 272, 335, 369, 379, 383, 385, 418, 429, 443, 452, 462

## REFERENCES

- Alumbaugh, R. V., Davis, H. G., & Sweney, A. B. A comparison of methods for constructing predictive instruments. *Educational and Psychological Measurement*, 1969, 29, 639-651.
- Anderson, H. E., Jr., & Fruchter, B. Some multiple correlation and predictor selection methods. *Psychometrika*, 1960, 25, 59-76.
- Barron, F. An ego-strength scale which predicts response to psychotherapy. *Journal of Consulting Psychology*, 1953, 17, 327-333.
- Burket, G. R. A study of reduced rank models for multiple prediction. *Psychometric Monographs*, 1964, No. 12.
- Buros, O. K. (Ed.) *Personality Tests and Reviews*. Highland Park, N. J.: Gryphon, 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 Books, 1957.
- Cattell, R. B. Objective personality tests: A reply to Dr. Eysenck. *Occupational Psychology*, 1964, 38, 69-86.
- Cattell, R. B., & Eber, H. W. *Handbook for the Sixteen Personality Factor Questionnaire*. 1957 Ed. with 1964 Supplement. Champaign, Ill.: Institute for Personality and Ability Testing.
- Cowden, J. E., Schroeder, C. R., & Peterson, W. M. The CPI vs. the 16 PF at a reception center for delinquent boys. *Journal of Clinical Psychology*, 1971, 27, 109-111.
- Crewe, N. M. Comparison of factor analytic and empirical scales. *Proceedings of the 75th Annual Convention of the American Psychological Association*, 1967, 367-368.
- Cronbach, L. J., & Gleser, G. C. *Psychological tests and personnel decisions*. 2nd Ed. Urbana, Ill.: University of Illinois Press, 1965.
- Cureton, E. E. Approximate linear restraints and best predictor weights. *Educational and Psychological Measurement*, 1951, 11, 12-15.
- Dicken, C. Good impression, social desirability and acquiescence as suppressor variables. *Educational and Psychological Measurement*, 1963, 23, 699-720.
- DuBois, P. H., Loevinger, J., & Smith, T. L., Jr. Evaluation of methods of keying psychological tests for prediction of external criteria. Research Report AFPTRC-TN-56-65. Personnel Research Laboratory, Lackland Air Force Base, Texas, 1956.
- Edwards, A. L. *The social desirability variable in personality assessment and research*. New York: Dryden, 1957.
- Forgy, E. W. A consistent bias in predictions made by multiple regression weights. Paper presented at the meeting of the Western Psychological Association, April 1962.
- Garside, M. J. The best sub-set in multiple regression analysis. *Applied Statistics*, 1965, 14, 196-200.
- Goldberg, L. R. Diagnosticians vs. diagnostic signs: The diagnosis of psychosis vs. neurosis from the MMPI. *Psychological Monographs*, 1965, 79, (9, Whole No. 602).
- 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. 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., & Hase, H. D. Strategies and tactics of personality inventory construction: An empirical investigation. *Oregon Research Institute Research Monograph*, 1967, Vol. 7, No. 1.
- Goldberg, L. R., & Rorer, L. G. Test-retest item statistics for the California Psychological Inventory. *Oregon Research Institute Research Monograph*, 1964, Vol. 4, No. 1.
- Goldberg, L. R., & Rorer, L. G. The use of two different response modes and repeated testings to predict social conformity. *Journal of Personality and Social Psychology*, 1966, 3, 28-37.
- 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, Vol. 10, No. 3.
- Gough, H. G. *Manual for the California Psychological Inventory*. Palo Alto, Calif.: Consulting Psychologists Press, 1957.
- Halinski, R. S., & Feldt, L. S. The selection of variables in multiple regression analysis. *Journal of Educational Measurement*, 1970, 7, 151-158.
- Hase, H. D. The predictive validity of different methods of deriving personality inventory scales. Unpublished doctoral dissertation, University of Oregon, 1965.
- Hase, H. D., & Goldberg, L. R. The comparative validity of different strategies of deriving personality inventory scales. *Psychological Bulletin*, 1967, 67, 231-248.
- Hastorf, A. H., & Piper, G. W. A note on the effect of explicit instructions on prestige suggestions. *Journal of Social Psychology*, 1951, 33, 289-293.
- Heilbrun, A. B., Jr. A comparison of empirical derivation and rational derivation of an affiliation scale. *Journal of Clinical Psychology*, 1962, 18, 101-102.
- Hermans, H. J. M. The validity of different strategies of scale construction in predicting academic achievement. *Educational and Psychological Measurement*, 1969, 29, 877-883.
- Herzberg, P. The parameters of cross-validation. *Psychometrika Monograph Supplement*, 1969, 34, No. 16.
- Horst, P. *The prediction of personal adjustment*. *Social Science Research Council Bulletin* 48, 1941.
- Horst, P., & MacEwan, C. Predictor elimination techniques for determining multiple prediction batteries. *Psychological Reports*, 1960, 7, 19-50.
- Jackson, D. N. Stylistic response determinants in the California Psychological Inventory. *Educational and Psychological Measurement*, 1960, 20, 339-346.
- Jackson, D. N. Desirability judgments as a method of personality assessment. *Educational and Psychological Measurement*, 1964, 24, 223-238.
- 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.
- Lawshe, C. H., & Patinka, P. J. An empirical comparison of two methods of test selection and weighting. *Journal of Applied Psychology*, 1958, 42, 210-212.
- Lawshe, C. H., & Schucker, R. E. The relative efficiency of four test weighting methods in multiple prediction. *Educational and Psychological Measurement*, 1959, 19, 103-114.
- Lingoes, J. C. Multiple scalogram analysis: A set-theoretic model for analyzing dichotomous items. *Educational and Psychological Measurement*, 1963, 23, 501-524.
- Linhart, H. A criterion for selecting variables in a regression analysis. *Psychometrika*, 1960, 25, 45-58.
- Loevinger, J. Objective tests as instruments of psychological theory. *Psychological Reports*, 1957, 3, 635-694.
- Loevinger, J., Gleser, G. C., & DuBois, P. H. Maximizing the discriminating power of a multiple-score set. *Psychometrika*, 1953, 18, 309-317.
- Lovell, V. R. Components of variance in two personality inventories. Unpublished doctoral dissertation, Stanford University, 1964.
- Lubin, A., & Summerfield, A. A square root method of selecting a minimum set of variables in multiple regression: II. A worked example. *Psychometrika*, 1951, 16, 425-437.
- Lunneborg, C. E. Developing prediction weights by matching battery factorings. *Psychometrika*, 1967, 32, 133-141.
- 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)
- Madden, J. M., & Bottenberg, R. A. Use of an all possible combination solution of certain multiple regression problems. *Journal of Applied Psychology*, 1963, 47, 365-366.
- Marks, M. R. Two kinds of regression weights which are better than betas in crossed samples. Paper presented at the meeting of the American Psychological Association, New York, September 1966.
- McCornack, R. L. A comparison of three predictor selection techniques in multiple regression. *Psychometrika*, 1970, 35, 257-271.
- McDonald, R. P. A unified treatment of the weighting problem. *Psychometrika*, 1968, 33, 351-381.
- Meehl, P. E. The dynamics of "structured" personality tests. *Journal of Clinical Psychology*, 1945, 1, 296-303.
- Meehl, P. E., & Hathaway, S. R. The K factor as a suppressor variable in the MMPI. *Journal of Applied Psychology*, 1946, 30, 525-564.
- Murray, H. A., et al. *Explorations in personality*. New York: Oxford, 1938.
- Neill, J. A., & Jackson, D. N. An evaluation of item selection strategies in personality scale construction. *Educational and Psychological Measurement*, 1970, 30, 647-661.
- Nichols, R. C., & Schnell, R. R. Factor scales for the California Psychological Inventory. *Journal of Consulting Psychology*, 1963, 27, 228-235.
- Pearson, P. H. Relationships between global and specified measures of novelty seeking. *Journal of Consulting and Clinical Psychology*, 1970, 34, 199-204.
- Poppleton, P. K., & Pilkington, G. W. A comparison of four methods of scoring an attitude scale in relation to its reliability and validity. *British Journal of Social and Clinical Psychology*, 1963, 2, 36-39.
- Rock, D. A., Linn, R. L., Evans, F. R., & Patrick, C. A comparison of predictor selection techniques using Monte Carlo methods. *Educational and Psychological Measurement*, 1970, 30, 873-884.
- Rodgers, D. A. Estimation of MMPI profiles from CPI data. *Journal of Consulting Psychology*, 1966, 30, 89.
- Schaie, K. W. Scaling the scales: Use of expert judgment in improving the validity of questionnaire scales. *Journal of Consulting Psychology*, 1963, 27, 350-357.
- Schmidt, F. L. The relative efficiency of regression and simple unit predictor weights in applied differential psychology. *Educational and Psychological Measurement*, 1971, 31, 699-714.
- Selove, S. L. Improved estimators for coefficients in linear regression. *Journal of the American Statistical Association*, 1968, 63, 596-606.

- Shine, L. C. The relative efficiency of test selection methods in cross-validation of generated data. *Educational and Psychological Measurement*, 1966, 26, 833-846.
- 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.
- Summerfield, A., & Lubin, A. A square root method of selecting a minimum set of variables in multiple regression: I. The method. *Psychometrika*, 1951, 16, 271-284.
- Toops, H. A. The *L*-method. *Psychometrika*, 1941, 6, 249-266.
- Tryon, R. C., & Bailey, D. E. *Cluster analysis*. New York: McGraw-Hill, 1970.
- Ward, J. H., Jr. An application of linear and curvilinear joint functional regression in psychological prediction. Research Bulletin AFPTRC TR-54-86. Personnel Research Laboratory, Lackland Air Force Base, Texas, 1954.
- Wesman, A. G., & Bennett, G. K. Multiple regression vs. simple addition of scores in prediction of college grades. *Educational and Psychological Measurement*, 1959, 19, 243-246.
- Wherry, R. J., & Gaylord, R. H. Test selection with integral gross score weights. *Psychometrika*, 1946, 11, 173-183.
- Yee, A. H., & Kriewall, T. A new logical scoring key for the Minnesota Teacher Attitude Inventory. *Journal of Educational Measurement*, 1969, 6, 11-14.