JOURNAL OF RESEARCH IN PERSONALITY 7, 1–20 (1973)

In Response to Jackson's Challenge: The Comparative Validity of Personality Scales Constructed by the External (Empirical) Strategy and Scales Developed Intuitively by Experts, Novices, and Laymen[1]

STEVEN G. ASHTON AND LEWIS R. GOLDBERG[2]
University of Oregon and Oregon Research Institute

Fifteen graduate students in psychology and 15 individuals with no formal psychology training were each paid to construct one 20-item scale to measure Sociability, Achievement, or Dominance. These 30 scales, plus the Personality Research Form and the California Psychological Inventory, were subsequently administered to 168 college females from seven living organizations. Average peer rankings were employed as criteria in order to compare the validity of personality scales constructed by different strategies. While the validity of Intuitive scales constructed by the average nonpsychologist was lower than that of the CPI External scales, the validity of scales constructed by the average psychology student and of the most reliable scales constructed by the nonpsychologists was essentially the same as that of the External scales. Moreover, the most reliable scales constructed by psychology students and the PRF scales were of approximately equal validity, considerably higher than that of any of the CPI scales.

In an important recent paper, Jackson (1971) issued 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. . . . The rules of the game would be that the empirical procedure would employ a heterogeneous item pool like that contained in the CPI or the MMPI, either published or unpublished, whereas the substantive approach would involve

[1] This study was supported by Grant MH 12972 and Grant MH 10822 from the National Institute of Mental Health, U. S. Public Health Service. For their extremely helpful reactions to an earlier draft of this report, we thank Douglas N. Jackson, Warren T. Norman, Jerry S. Wiggins, Paul E. Meehl, Paul Slovic, Jack Block, David P. Campbell, and Donald W. Fiske.

[2] Requests for reprints should be sent to Lewis R. Goldberg, Oregon Research Institute, P. O. Box 3196, Eugene, Oregon 97403.

> two item writers working for two hours each. . . . 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].

Readers unfamiliar with the controversy which has provoked this challenge should see Meehl (1945), Hase and Goldberg (1967), and Goldberg (1972a). The present study provides a first test of the hypothesis that ". . . even an inexperienced item writer would be superior to empirical item selection with a typical heterogeneous item pool."

METHOD

Overview

The concurrent validity of personality scales constructed by the External (or "empirical") strategy was compared with that of scales intuitively developed by nonprofessional item writers, as well as with that of scales carefully developed by an acknowledged test expert (Jackson) using a variety of sophisticated scale-construction tactics. Fifteen graduate students in psychology and 15 individuals with no

TABLE 1
THE EXPERIMENTAL DESIGN, INCLUDING THE STRATEGIES AND SCALES TO BE COMPARED

	Scale-construction strategy	Targeted traits: Sociability (SOC)	Achievement (ACH)	Dominance (DOM)
New scales constructed for this study	Intuitive			
	PPS: Psych.	5	5	5
	PPS: Nonpsych.	5	5	5
Comparison scales, previously constructed	Intuitive			
	CPI Theoretical (H & G)[a]	*nAf*	*nAc*	*nDo*
	Mixed: Intuitive + Internal			
	CPI Rational (H & G)[a]	*Soc*	*Ach*	*Dom*
	PRF (Jackson)	*Af*	*Ac*	*Do*
	External			
	CPI Empirical (Gough)	*Sy*	*Ac*	*Do*

[a] From Hase and Goldberg (1967).

formal training in psychology were each paid to construct one 20-item scale; five persons from each of the two groups constructed scales to measure Sociability, Achievement, and Dominance, respectively. These 30 new Intuitive scales were merged together to form a 600-item inventory, dubbed the Personality Psychological Schedule (PPS). The PPS plus Form AA of Jackson's Personality Research Form (PRF) and Gough's California Psychological Inventory (CPI) were administered to 168 female college students from seven living organizations. Peer rankings and self-rankings were collected for each of the subjects on the three targeted traits, on five broad personality factors (Norman, 1963), and on two control measures (How well known and How well liked). The average peer rankings were employed as criteria, in order to compare the validity of personality scales constructed by the different strategies. This experimental design, including the scales selected as representatives of the various strategies, is presented in Table 1.

The Targeted Constructs

The personality traits of Dominance, Sociability, and Achievement were selected as the constructs to be used for comparing the validity of the strategies. This choice was based upon the following considerations: (a) they have been widely studied in diverse contexts, (b) they are included as structural constructs in a number of personality theories, (c) they can be easily communicated to lay individuals, (d) they have been included as constructs in one or more of the few previous comparative validity studies (Butt & Fiske, 1968; Goldberg, 1972a), and—perhaps of most importance—(e) scales designed to measure these three constructs are included in both the CPI and the PRF, two personality inventories which have been constructed by radically different strategies.

New Scales Constructed for this Study

Thirty intuitively derived scales, including ten each designed to measure Dominance, Sociability, and Achievement, were constructed especially for this study. For each trait, five of the scales were written by graduate students in psychology and five were written by individuals with no background in academic psychology. The nonpsychology item writers ranged from 19–54 yr in age; nine were college students, and the other six were college graduates working in various professional fields, ranging from acting to accounting. There were seven men and eight women. The 15 psychology graduate students ranged from 21–27 yr in age, and their exposure to graduate school ranged from 6 mo to 5 yr. There were ten men and five women. Seven students were in a clinical psychology training program, five were in experimental psychology, and three were in social psychology.

Each scale developer was given a description of the trait to be measured, along with examples of items from a scale constructed to measure a different trait and some suggestions as to what helps make a "good" item (e.g., short, concise statements) and a "good" scale (e.g., an equal number of True and False items). The three trait descriptions, which were derived by combining the pairs of corresponding scale descriptions from the manuals of the CPI and PRF, are presented below:

Sociability

Definition: outgoing, sociable, with a participative temperament.

Behavioral aspects: enjoys being with friends and people in general; accepts people readily; makes efforts to win friendships and maintain associations with people.

Terms used to describe individuals high in sociability: outgoing, enterprising, ingenious, neighborly, loyal, warm, amicable, good-natured, competitive, forward, friendly, companionable, genial, affable, original and fluent in thought, cooperative, gregarious, hospitable, good-willed.

Achievement

Definition: setting high goals and working hard to accomplish them.

Behavioral aspects: aspires to accomplish difficult tasks, maintaining high standards and a willingness to work toward distant goals; responds to competition; willing to put forth effort to attain excellence.

Terms used to describe individuals high in achievement: mature, forceful, strong, foresighted, striving, capable, purposeful, of superior intellectual ability and judgment, efficient, attaining, industrious, aspiring, enterprising, organized, self-improving, productive, driving, ambitious, stable, responsible, persistent, resourceful, competitive.

Dominance

Definition: a tendency to lead others, be persistent, and have social initiative.

Behavioral aspects: attempting to control one's environment and to influence or direct other people; expressing opinions forcefully; enjoying the role of being a leader; assuming leadership spontaneously.

Terms used to describe individuals high in dominance: aggressive, confident, persistent, planful, governing, controlling, commanding, domineering, persuasive, verbally fluent, self-reliant, influential, forceful, independent, having leadership potential, ascendant, leading, directing, assertive, authoritative, powerful, supervising.

Each scale developer was instructed to write at least 25 items within a time limit of 2 hr. Each was paid $5.00, and promised an additional $25.00 if he produced the best scale. A secretary removed any sexually or racially offensive items, plus any items calling for criminal admission, from each item list; if there were more than 20 items remaining, the first 20 were selected for each scale. The resulting 30 scales, all of which are included in Ashton (1973), were interspersed in the PPS booklets so that every third item was targeted for the same trait.

Comparison Scales, Previously Constructed

The new PPS scales are clear illustrations of the straightforward use of the Intuitive strategy of scale construction, since no empirical data were used to select the items and/or determine their direction of keying. However, while each of the 30 PPS scales was devised by a single individual, the pooled judgments of a number of persons could have been substituted in order to attenuate any idiosyncratic viewpoints. For example, using the CPI item pool, Hase and Goldberg (1967) developed an Intuitive inventory based upon the consensus of the judgments of three advanced graduate students in clinical psychology as to which CPI items appeared to measure each of the "manifest needs" described by Murray *et al.* (1938). The resulting 11-scale Theoretical inventory, which is described in detail in Hase and Goldberg (1967), includes three scales—need Affiliation (*nAf*), need Achievement (*nAc*), and need Dominance (*nDo*)—that were constructed as measures of the targeted traits. These three scales are included in the present study as additional representatives of the Intuitive strategy.

Both the new PPS scales and the CPI Theoretical scales were developed by purely Intuitive procedures, without recourse to any data on their internal consistency.

However, the developers of many recent personality inventories have used a mixture of the Intuitive and the Internal strategies, beginning scale construction by the intuitive assembly and keying of items, and then refining the resulting preliminary scales through one or more types of internal consistency (or homogeneity) analysis. For example, again using the CPI item pool, Hase and Goldberg (1967) developed their 11-scale Rational inventory by such a mixed strategy. First, CPI items were grouped and keyed purely on judgmental grounds, and later empirical homogeneity data were used to refine the scales. Specifically, only those items that correlated significantly with the total preliminary scale scores were retained in the final scales. Of these 11 Rational scales, three—Sociability (*Soc*), Achievement (*Ach*), and Dominance (*Dom*)—were constructed as measures of the targeted traits, and they are therefore included here as representatives of a mixed (Intuitive + Internal) strategy of scale construction.

Such a mixture of the Intuitive and the Internal strategies lies at the heart of the far-more-elaborate sequential procedures used by Jackson (1970) in developing the PRF. First, the literature relating to each of the "manifest needs" originally posited by Murray *et al.* (1938) was reviewed, and a large set of items intuitively relating to each such construct was developed. Subsequently, this preliminary item pool was administered to samples of college students. Biserial correlations were computed between each item and the total provisional scale of which it was a member, for related scales, and for a set of items scaled for social desirability. Items were retained for further analysis only if they showed higher correlations with the scale for which they were written than with any other scale. Items with extreme endorsement proportions ($\leq .05$ or $\geq .95$) were eliminated. The items were ranked by a "differential reliability index" (roughly, the difference between an item's correlation with the targeted scale and its correlation with the social desirability scale). The 20 top-ranking items were then selected for inclusion in the final scales, subject to the provision that half be keyed True and half False. Three of the PRF scales—Affiliation (*Af*), Achievement (*Ac*), and Dominance (*Do*)—were constructed as measures of the targeted traits and are therefore included in the present study. A comparison between the validity of these PRF scales and that of the new PPS intuitive scales should be most illuminating, since the developers of these two inventories vary in expertise from complete laymen (PPS: nonpsychologists) to novices (PPS: psychology graduate students) to experts (PRF: Jackson).

Finally, in direct response to Jackson's challenge, all of these scales must be compared to a set constructed by the External strategy. In constructing External scales, the responses of some nontest reference groups are used to determine each item's scale membership and keying direction; consequently, this strategy has often been called the "empirical" or "criterion-group" strategy. Of the 18 standard CPI scales, 11 were constructed by Gough using the External strategy; three of these—Sociability (*Sy*), Achievement via Conformance (*Ac*),[3] and Dominance (*Do*)—can be viewed as measures of the targeted traits, and are thus included here as representatives of the External strategy. The *Do* scale includes items that correlated significantly with peer ratings of Dominance; the criterion for the *Sy* scale was degree of participation in

[3] Two other CPI scales, *Ai* and *Ie*, are logically related to the Achievement criterion, and both were initially included in these comparisons. However, since the correlation with the mean peer ranking of Achievement was considerably higher for *Ac* ($r = .30$) than for either *Ai* ($r = .16$) or *Ie* ($r = .12$), only the results based on *Ac* are presented here.

high school extracurricular activities;[4] and the *Ac* scale is based on items which correlated significantly with high school grade point average. Over the years, these three CPI scales have generated a rich variety of research, and they are often cited as outstanding examples of present-day External scales (e.g., Megargee, 1972).

Nontargeted Criteria

While it is clearly important to gather data on the concurrent validity of personality scales against those specific criteria for which they are targeted, most personality inventories are used to predict a host of criterial behaviors, the vast majority of which were never considered in the development of the inventory scales. Typically, scores from multiscale inventories are used via multiple-regression procedures to make these predictions, and their comparative validity across a wide range of such criteria provides the data base for an assessment of the differential utility of the strategies used in their construction (Goldberg, 1972a). The inclusion of five nontargeted criteria in the present study permits a partial replication of the comparative validity findings provided by Hase and Goldberg (1967).

Norman's (1963) five peer-rating factors of Surgency (SUR), Agreeableness (AGR), Conscientiousness (CON), Emotional Stability (STA), and Culture (CUL) were selected as a set of additional criteria. Based on earlier work by Allport and Odbert (1936), Cattell (1947, 1957), and Tupes and Christal (1961), these replicated peer-rating factors resulted from the preliminary systematic distillation of all trait names in standard English. Scores on each of the five factors were derived by summing the peer rankings on the four scales used to measure each factor.

Subjects

Initially, 182 single women (mean age 20.1 yr), from seven different campus living organizations, were paid to participate in this 5 hr experiment. Data from 14 subjects were discarded, including seven who did not complete all of the inventories, and seven more who scored above 2 on the Infrequency scale of the PRF. All analyses are based upon the remaining 168 subjects.

Procedures

The subjects were first asked to complete (a) the five unipolar peer rankings on How well known, Sociability, Achievement, Dominance, and How well liked; and then (b) the 20 bipolar peer-ranking scales for the five Norman (1963) factors. The descriptions of Dominance, Sociability, and Achievement which the subjects used in evaluating each other were identical to those used by the item writers. All of the peer rankings were filled out in the presence of one experimenter, simultaneously by all of the participating residents of a single living organization. Each subject ranked herself together with 8 to 15 fellow residents whose rooms were located nearest to her own.

[4] "Participation in extracurricular activities served as the original criterion for one scale, so Gough named it *Social Participation*. Later research showed that high scorers were seen as more sociable by their friends and acquaintances and that this sociability was a more salient characteristic than their willingness to participate in other activities. Consequently it was renamed *Sociability (Sy)*" [Megargee, 1972; pp. 26–27].

The peer-ranking forms were based upon a procedure developed in an industrial context (Esso Standard, 1962), which has been proposed as a means of increasing the reliability and validity of job-performance evaluations. Each subject was instructed to decide which member of her peer group was *best* described by the personality trait; then, out of the remaining peers on the list, which one was *least* well described. Next, out of the remaining names, she was instructed to repeat the alternation ranking process until the list of names was exhausted. Each ranker was also instructed to include her own name in the rankings, yielding rank-order self-rankings. After averaging across raters and adjusting the values for group size, the mean peer rankings served as the criteria in this study.

After completing the peer rankings, each subject was given the CPI and the PRF, plus the specially constructed Personality Psychological Schedule. These three inventories were administered in a counterbalanced order, so as to control for any possible order-of-presentation effects. The subjects were allowed to complete these three inventories on their own, but were required to complete them all within a week. To help ensure candid answers, the subjects were instructed that all responses would be held in complete confidentiality, that payment was contingent upon prompt and honest performance by all participants, and that the experiment was of great scientific importance.

Analyses

Three types of analyses were conducted. First, means, standard deviations, and KR-20 reliability coefficients were computed for each of the scales. Second, all of the scales and self-rankings were correlated with one another and with the ten average peer rankings. Third, each of the four 11-scale inventories from the CPI (Hase & Goldberg, 1967) was regressed, in a stepwise fashion, upon each of five nontargeted peer-rating factors (Norman, 1963), and the average cross-validity across the five factors of each of these inventories was compared with that of the self-rankings and that of two sets of PRF scales.

RESULTS

Analyses of the Criteria Themselves

Table 2 presents the intercorrelations among the ten peer rankings and the eight meaningful self-rankings (self-rankings on How well known and How well liked are not included in this and subsequent tables). As expected (Goldberg, 1972b), peer rankings on Dominance and Sociability were substantially related ($r = .59$), Dominance and Achievement somewhat less highly related ($r = .48$), and Sociability and Achievement virtually unrelated ($r = .12$). Similarly, the intercorrelations among the five nontargeted peer rankings were almost identical to those reported in Norman's (1963) original study. Specifically, two pairs of factors were moderately related, namely (a) Agreeableness and Emotional Stability, and (b) Conscientiousness and Culture.

Indeed, the ten peer rankings can be grouped into three major clusters: (a) Sociability, Dominance, Surgency, and How well known; (b) Achieve-

TABLE 2

INTERCORRELATIONS AMONG THE PEER RANKINGS AND SELF-RANKINGS (N = 168)[a]

		Peer rankings										Self-rankings						
		Targeted			Nontargeted					Control		Targeted			Nontargeted			
		SOC	ACH	DOM	SUR	AGR	CON	STA	CUL	HWK	HWL	SOC	ACH	DOM	SUR	AGR	CON	STA
Peer rankings	ACH	.12																
	DOM	.59	.48															
	SUR	.83	−.06	.61														
	AGR	.51	.20	.00	.21													
	CON	.04	.78	.29	−.24	.31												
	STA	.25	.31	.14	.01	.51	.28											
	CUL	.27	.63	.38	.06	.31	.57	.25										
	HWK	.59	.26	.54	.59	.18	.10	−.06	.35									
	HWL	.59	.32	.27	.34	.58	.28	.32	.42	.61								
Self-rankings	SOC	.46	−.09	.38	.51	.06	−.12	−.01	−.05	.21	.03							
	ACH	−.13	.48	.17	−.23	−.07	.44	.17	.28	−.03	−.09	.04						
	DOM	.40	.32	.66	.38	−.06	.23	.07	.28	.38	.08	.42	.30					
	SUR	.47	−.24	.39	.66	−.10	−.31	−.12	−.18	.25	.01	.62	−.20	.39				
	AGR	.26	.07	.13	.17	.37	.14	.12	.08	.03	.14	.43	.06	.07	.25			
	CON	−.15	.41	.12	−.25	.00	.58	.03	.22	.01	−:04	.12	.60	.26	−.11	.24		
	STA	.08	.03	.14	.03	.15	.06	.43	.02	−.14	−.06	.16	.06	.13	.13	.34	.05	
	CUL	−.09	.11	.04	−.08	−.18	.07	−.02	.33	.04	−.08	.04	.33	.23	.05	.07	.29	.08

[a] Note: Correlations ≥ .15 and ≥ .20 are significantly greater than zero at $p \leq .05$ and $p \leq .01$, respectively.

ment, Conscientiousness and Culture; and (c) Agreeableness and Emotional Stability. Although there was much the same pattern of interrelationships among the self-rankings, the magnitude of the correlations was somewhat lower, presumably due to the fact that average peer rankings are more reliable than individual self-rankings. As might be expected, peer rankings on How well known and How well liked were substantially related to one another ($r = .61$), although the two patterns of correlation with other peer rankings were not identical. Specifically, Dominance and Surgency related more highly to being well known than to being well liked, while Agreeableness and Emotional Stability related more highly to being well liked than to being well known. It is not surprising that individuals are more noticed if they are seen as dominant and/or exhibitionistic, but better liked if they are seen as stable and/or agreeable.

Analyses against the Targeted Criteria

Table 3 presents the intercorrelations within each of the three monotrait sets of ten PPS scales, as well as the reliability (KR-20) and validity coefficients for each of these 30 new Intuitive scales. Of the 135 convergence correlations, all but one were positive. Among the scales developed by the psychology students, all of the monotrait correlations were significantly larger than zero beyond the .05 level, all but one beyond the .01 level. Moreover, for two of the three traits (Sociability and Dominance), all such correlations among the scales developed by nonpsychologists were also significant beyond the .01 level. The set of Sociability scales developed by psychology students was extraordinarily convergent, all of these intercorrelations being greater than .60.

For 20 item scales, each developed in less than 2 hr by such a wide variety of novices and laymen, these findings are quite remarkable. And, for such short scales, the reliability coefficients are also quite respectable: 7 of the 30 KR-20 coefficients were greater than .70, half of them were greater than .60, and 23 of the 30 were greater than .50. Moreover, all of the 30 validity coefficients were positive, most of them significantly so. For the scales developed by psychology students, 12 of the 15 validity coefficients were greater than .20, and 9 of the 15 were greater than .30. In contrast, only 6 of the 15 scales developed by nonpsychologists produced validities greater than .20, and only 3 were greater than .30.

The results of two analyses of variance, based on the 30 Z-converted reliability and validity coefficients, are presented in Table 4. Note that there were no significant differences in either the reliability or the validity of the PPS scales across the three targeted traits. On the other hand, there was a highly significant ($p < .01$) difference in the validity of scales con-

TABLE 3
INTERCORRELATIONS AMONG THE NEW PPS SCALES SPECIALLY CONSTRUCTED FOR THIS STUDY (N = 168)[a]

		Psychology students					Nonpsychologists				
		1	2	3	4	5	6	7	8	9	10
Sociability	1	—	.64	.64	.61	.64	−.08	.06	.55	.62	.09
	2		—	.61	.63	.72	.12	.25	.64	.58	.36
	3			—	.77	.63	.08	.14	.73	.64	.18
	4				—	.63	.02	.18	.74	.66	.30
	5					—	.07	.25	.65	.59	.28
	6						—	.12	.06	.14	.24
	7							—	.13	.17	.20
	8								—	.64	.30
	9									—	.17
	KR-20	.70	.53	.81	.78	.71	.42	.15	.66	.68	.62
	Validity	.34	.32	.36	.30	.28	.06	.06	.26	.30	.11
Achievement	1	—	.38	.18	.27	.41	.36	.35	.31	.44	.26
	2		—	.32	.42	.58	.42	.42	.40	.50	.32
	3			—	.48	.25	.37	.39	.30	.41	.23
	4				—	.40	.49	.55	.34	.44	.35
	5					—	.38	.43	.34	.42	.34
	6						—	.57	.38	.55	.26
	7							—	.33	.64	.34
	8								—	.55	.27
	9									—	.38
	KR-20	.40	.49	.52	.56	.57	.60	.68	.38	.70	.58
	Validity	.15	.36	.12	.31	.28	.21	.26	.12	.18	.16
Dominance	1	—	.56	.68	.51	.70	.74	.78	.19	.57	.41
	2		—	.56	.31	.49	.48	.50	.24	.34	.39
	3			—	.37	.60	.68	.68	.13	.53	.37
	4				—	.46	.42	.45	.11	.32	.34
	5					—	.65	.58	.21	.51	.45
	6						—	.73	.22	.59	.41
	7							—	.20	.57	.36
	8								—	.22	.28
	9									—	.35
	KR-20	.72	.31	.69	.54	.64	.72	.65	.18	.50	.51
	Validity	.39	.23	.31	.17	.35	.31	.37	.00	.15	.14

[a] Note: Correlations $\geq .15$ and $\geq .20$ are significantly greater than zero at $p \leq .05$ and $p \leq .01$, respectively.

TABLE 4
RESULTS OF THE ANALYSES OF VARIANCE: RELIABILITY AND VALIDITY OF SCALES DEVELOPED BY GRADUATE STUDENTS IN PSYCHOLOGY VERSUS INDIVIDUALS NOT TRAINED IN PSYCHOLOGY[a]

Source	SS	*df*	MS	*F*
Reliability (KR-20)				
Three criteria (SOC-ACH-DOM)	.0846	2	.0423	.851
Two types of item writers (Psych. vs nonpsych.)	.0652	1	.0652	1.313
Interaction	.2429	2	.1214	2.444
Error	1.1927	24	.0497	
Total	1.5854	29		
Validity				
Three criteria (SOC-ACH-DOM)	.0055	2	.0028	.258
Two types of item writers (Psych. vs nonpsych.)	.0933	1	.0933	8.737[b]
Interaction	.0150	2	.0075	.704
Error	.2563	24	.0107	
Total	.3701	29		

[a] Note: The dependent variable is the *Z*-converted validity (or reliability [KR-20]) coefficient for each of the 30 new PPS scales.
[b] $p < .01$.

structed by the psychology students ($\bar{r} = .29$) as compared to those constructed by nonpsychologists ($\bar{r} = .18$). While none of the psychology students had ever received formal training in scale construction, and the vast majority had never even taken a graduate course in psychological assessment, as a group they demonstrated considerable superiority over the nonpsychologists in constructing personality scales.

Nonetheless, as Table 3 indicates, there was considerable variation in the validity of Intuitive scales developed by different individuals, even among those scale developers with graduate training in psychology. From any sizable set of such Intuitive scales, it would normally be important to cull out the least promising scales, in the absence of empirical evidence on their comparative validity. Classical test theory provides one rationale for such a selection process, namely to select those scales with the highest reliabilities. A scatterplot of the bivariate distribution of reliability (KR-20) and validity coefficients for the 30 PPS scales indicated that their relationship was linear, and the correlation between these two in-

dices was .71. This is a most heartening finding, since it suggests that if one selected among a set of Intuitive scales solely on the basis of their reliability coefficients, one would tend to increase the validity of the resulting subset.

However, there is no guarantee that the scale with the highest reliability will always possess the highest validity. While the relationship between reliability and validity was quite substantial among those PPS scales constructed to measure Sociability and Dominance, this relationship was around zero among scales targeted for Achievement. Specifically, within the ten Sociability and the ten Dominance scales, the scale with the highest reliability did turn out to be the most valid; on the other hand, the Achievement scale with the highest reliability was considerably less valid than a number of other similarly targeted scales. For purposes of comparing these new PPS scales with those constructed by other strategies, both the average validity of all scales in a set, and the validity of the most reliable single scale, will always be presented. Moreover, since the preceding analyses have shown a decided superiority in the validity of scales constructed by psychology students over those constructed by laymen, the scales developed by these two groups will subsequently be analyzed separately.

TABLE 5

AVERAGE ITEM HOMOGENEITY COEFFICIENTS (r_{ii}) AS A FUNCTION OF SCALE-CONSTRUCTION STRATEGY (N = 168)[a]

	Scale-construction strategy	Targeted traits			
		SOC	ACH	DOM	Average
New PPS scales	Intuitive				
	Average psych.	.12	.05	.08	.08
	Average nonpsych.	.06	.07	.06	.06
	Most reliable psych.	.17	.06	.12	.12
	Most reliable nonpsych.	.10	.11	.11	.11
Comparison scales	Intuitive				
	CPI Theoretical (H & G)[b]	.11	.09	.27	.16
	Mixed: Intuitive + Internal				
	CPI Rational (H & G)[b]	.13	.06	.12	.10
	PRF (Jackson)	.12	.11	.21	.15
	External				
	CPI Empirical (Gough)	.06	.05	.08	.06

[a] Note: The original KR-20 reliability coefficients have been converted via the Spearman–Brown formula to provide reliability estimates for a single item.

[b] From Hase and Goldberg (1967).

Table 5 presents some summary data on the reliability of scales constructed by diverse strategies. The scales used in the analyses presented in this and the following table are the targeted scales listed in Table 1. Since the various CPI and PRF scale sets differ in the number of items included in each scale, all KR-20 reliability values have been converted via the Spearman–Brown formula to reliability estimates for a single item (r_{ii}). These results indicate that the targeted scales from the PRF $\bar{r}_{ii} = .16$) and the Theoretical (Hase & Goldberg, 1967) CPI scales ($\bar{r}_{ii} = .15$) possess quite substantial internal consistency, while the scales developed by the External strategy (Gough's CPI Empirical scales) and the PPS Intuitive scales developed by nonpsychologists have considerably lower homogeneity values ($\bar{r}_{ii} = .06$).

Table 6, which presents the validity coefficients for each of the targeted scales compared in this study, summarizes the major findings on the comparative validity of scales constructed by the different strategies. The results based upon the various CPI scale sets are remarkably similar to those presented in Hase and Goldberg (1967) and Goldberg (1972a), showing that the average validity of scales constructed by the Intuitive

TABLE 6
VALIDITY COEFFICIENTS AS A FUNCTION OF SCALE-CONSTRUCTION STRATEGY ($N = 168$)[a]

	Scale-construction strategy	Targeted traits			
		SOC	ACH	DOM	Average
New PPS scales	Intuitive				
	Average psych.	.32	.24	.29	.29
	Average nonpsych.	.16	.19	.19	.18
	Most reliable psych.	.36[c]	.28	.39[c]	.34
	Most reliable nonpsych.	.30	.18	.31	.26
Comparison scales	Intuitive				
	CPI Theoretical (H & G)[b]	.21	.27	.38	.29
	Mixed: Intuitive + Internal				
	CPI Rational (H & G)[b]	.19	.23	.34	.26
	PRF (Jackson)	.29	.35[c]	.39	.35[c]
	External				
	CPI Empirical (Gough)	.15	.30	.38	.27
	Self-rankings	.46	.48	.66	.53

[a] Note: Correlations $\geq .15$ and $\geq .20$ are significantly greater than zero at $p < .05$ and $p < .01$, respectively.

[b] From Hase and Goldberg (1967).

[c] Most valid scale for this criterion (not including self-rankings).

(Theoretical) strategy ($\bar{r}$ = .29), the External (Empirical) strategy ($\bar{r}$ = .27), and a mixed Intuitive + Internal (Rational) strategy ($\bar{r}$ = .26) are essentially identical, all being dramatically inferior to self-rankings ($\bar{r}$ = .53). The major innovation in the present study over the earlier ones, however, lies in the inclusion of the new PPS Intuitive scales, plus the carefully developed PRF scales. And, the results based upon these scales paradoxically serve both to confirm, and simultaneously to refute, the claims made by Jackson (1971) in the challenge that inspired this study. For clearly the average validity of "unselected item writers"—in this case, the average nonpsychology scale developers ($\bar{r}$ = .18)—was considerably lower than that of the External (Empirical) scales ($\bar{r}$ = .27). On the other hand, the average validity of all scales constructed by psychology students ($\bar{r}$ = .29) and of the most reliable scales constructed by nonpsychologists ($\bar{r}$ = .26) was essentially the same as that of the External scales. Moreover, and of far greater potential significance, the average validity of the most reliable scales constructed by psychology students ($\bar{r}$ = .34) was virtually identical to that of the targeted PRF scales ($\bar{r}$ = .35), both sets being substantially more valid than *any* of the sets of CPI scales. For two of the three targeted traits (Sociability and Dominance), the highest validity coefficient (not including the self-rankings) was achieved by the most reliable scale constructed by psychology students. For the other targeted trait (Achievement), the most valid scale came from the PRF.

Analyses against the Nontargeted Criteria

The cross-validity coefficients of four major 11-scale inventories developed from the CPI item pool, each constructed on the basis of a different strategy, were compared when these various scale sets were used in multiple-regression analyses to predict each of the five nontargeted peer rankings. In addition, each of these CPI scale sets was compared with two sets of scales from the PRF. Specifically, the average cross-validities of the following six inventories were compared: (a) Empirical (11 CPI scales, all constructed by Gough using the External strategy); (b) Factor (11 CPI scales, constructed by Hase and Goldberg [1967] using the Internal strategy); (c) Theoretical (11 CPI scales, constructed by Hase and Goldberg [1967] using the Intuitive strategy); (d) Rational (11 CPI scales, four constructed by Gough and seven by Hase, all developed by a mixed strategy, starting with the Intuitive assembly and keying of items, followed by subsequent scale refinement through Internal [homogeneity] analyses); (e) PRF-A (the subset of 15 PRF scales, from the set of 22 included in Form AA, that make up the PRF short form, labeled Form A); and (f) PRF-AA (all 22 scales included in Form AA of the PRF).

Since self-rankings were also collected on each of the five nontargeted peer-ranking indices, the average cross-validities of these five self-rankings were also compared with those of the six personality inventories.

For this purpose, the sample of 168 subjects was randomly divided into

TABLE 7

STEPWISE MULTIPLE-REGRESSION ANALYSES WITH THE NONTARGETED CRITERIA: CROSS-VALIDITY COEFFICIENTS—AVERAGED ACROSS ALL 5 PEER-RATING FACTORS (NORMAN, 1963)—AS A FUNCTION OF THE NUMBER OF PREDICTORS INCLUDED IN THE REGRESSION EQUATIONS[a]

No. of predictors included in the regression equations	Scale-construction strategy						
				Mixed			
	Ext. 11 Emp.[b]	Intern. 11 Fac.[c]	Intuit. 11 The.[c]	11 Rat.[c]	15 PRF[d]	22 PRF[e]	5 Self-rankings
1	.09	.13	.11	.15	.23	.27	.48
2	.12	.15	.21	.22	.24	.30	.51[f]
3	.11	.14	.23	.22	.24	.30	.50
4	.12	.17[f]	.23	.23	.26	.29	.48
5	.14[f]	.14	.25[f]	.22	.28	.31	.48
6	.14	.15	.24	.22	.28	.30	
7	.14	.16	.24	.23	.28	.31[f]	
8	.14	.16	.23	.23	.28	.31	
9	.13	.17	.23	.24	.28	.27	
10	.13	.16	.22	.24[f]	.29[f]	.30	
11	.13	.17	.23	.24	.28	.29	
12					.29	.29	
13					.28	.28	
14					.28	.27	
15					.28	.28	
16						.28	
17						.27	
18						.27	
19						.27	
20						.27	
21						.27	
22						.27	

[a] Note: The tabled values are average cross-validity coefficients, based upon a double cross-validation design, with 84 subjects in each of the two derivation and each of the two cross-validation samples.

[b] From Gough (1957).

[c] From Hase and Goldberg (1967).

[d] Form A.

[e] Form AA.

[f] Highest value in each column.

two subsamples, each of 84 subjects, and stepwise multiple-regression analyses were carried out using the same double cross-validation design previously employed in the analyses reported by Hase and Goldberg (1967) and Goldberg (1972a). The average cross-validity coefficients (averaged across the two cross-validation samples and across the five peer-rating factor scores [Norman, 1963]) are presented—for each step in the regression analyses—in Table 7. The findings presented in this table are primarily important as an indication of the optimal number of predictors to include in such regression functions, for samples of this size. The results displayed in Table 7 are highly congruent with those addressed to this same issue in Goldberg (1972a), based on slightly smaller subsamples. In the present study, the optimal number of predictors for various scale sets varied from two to ten. In both studies, the optimal number of predictors, over all scale sets, averaged about five.

The average cross-validity coefficients for each of the scale sets, when five scales were included in each regression function, are presented in Table 8. The average values near the bottom of the table (the column means) indicate the general predictability of each of the five criteria

TABLE 8

STEPWISE MULTIPLE-REGRESSION ANALYSES WITH THE NONTARGETED CRITERIA: AVERAGE CROSS-VALIDITY COEFFICIENTS FOR EACH OF THE 5 PEER-RATING FACTORS WHEN FIVE PREDICTORS WERE INCLUDED IN THE REGRESSION EQUATIONS[a]

Scale-construction strategy	Factor					Average
	SUR	AGR	CON	STA	CUL	
External						
11 Empirical (Gough)	.10	.07	.38	.02	.11	.14
Internal						
11 Factor (H & G)	.16	.15	.29	.06	.07	.14
Intuitive						
11 Theoretical (H & G)	.27	.28	.38	.16	.15	.25
Mixed						
11 Rational (H & G)	.15	.32[b]	.39	.22[b]	.01	.22
15 PRF (Form A)	.42	.28	.49[b]	.14	.05	.28
22 PRF (Form AA)	.43[b]	.24	.41	.16	.29[b]	.31[b]
Average	.25	.22	.39	.13	.11	.22
5 Self-rankings	.66	.46	.57	.44	.27	.48

[a] Note: The tabled values are average cross-validity coefficients, based upon a double cross-validation design, with 84 subjects in each of the two derivation and each of the two cross-validation samples.

[b] Highest value in each column (self-rankings excluded).

across these different sets of scales. Peer rankings of Conscientiousness were consistently quite predictable by all of the scale sets, the average cross-validities in the present study being virtually identical to those based upon peer ratings of Responsibility in Goldberg (1972a). Peer rankings of Surgency were quite predictable from both the self-rankings and the PRF scales, though all of the CPI scale sets provided considerably lower cross-validities, values which were surprisingly lower than those based on peer ratings of Sociability in Goldberg (1972a). While the predictability of the present peer rankings of Agreeableness were quite similar to the earlier analyses of Psychological-mindedness, the peer rankings of Emotional Stability and Culture were virtually unpredictable from any of the scale sets under study here. However, Emotional Stability, unlike Culture, did correlate substantially with the self-rankings.

The average values presented in the last column of Table 8 (the row means) indicate the overall cross-validity across all five criteria of each of the scale sets. Since these criteria were generally not as predictable as the peer ratings employed in Goldberg (1972a), the average values for each of the CPI scale sets are not directly comparable between the two studies. Moreover, since the present study did not include as wide a variety of diverse criterion indices as those included in Hase and Goldberg (1967), these results should not be construed as bearing on the differential bandwidth of scales constructed by the various strategies (Goldberg, 1972a). Nonetheless, the results presented in Table 8 suggest that, against these peer-ranking criteria, the CPI-based inventories constructed by either the External (Empirical) or Internal (Factor) strategies were not as valid as those constructed by the Intuitive (Theoretical) and the mixed (Rational) strategies. Moreover, the mixed-strategy PRF scale sets generally outperformed all of the CPI inventories. For two of the five criteria (Agreeableness and Emotional Stability), the highest cross-validities were obtained from the CPI Rational inventory; for the other three criteria, the highest cross-validities were obtained from the PRF. Finally, the self-rankings generally produced substantially higher cross-validities than did the inventories constructed by any of the strategies under study.

DISCUSSION

Certainly the most striking finding of the present study was that the average graduate student in psychology—in 2 hr or less—was capable of producing personality scales of equal reliability and validity to those developed by the far more expensive and time-consuming External strategy. Obviously, additional studies are now required to ascertain the generality of these findings to other samples of item writers, targeted traits, and subject populations. However, if these findings are replicated and

accepted, then an era opened by Meehl's (1945) stirring empirical manifesto may have been closed. For, while the results of the present investigation did not support the more expansive aspects of Jackson's (1971) challenge—that even an "unselected" item writer can produce scales of superior validity to ones constructed by the External strategy—clearly the major thrust of his argument was confirmed.

When self-ratings are highly associated with the criterion indices, as they were in this study, then appropriately targeted scales constructed by the Intuitive strategy should likewise obtain at least moderate convergent validities. This situation will occur when subjects have nothing to gain by deception and/or when they generally accept the role of surrogate guinea pig in a scientific investigation. Under such conditions of experimenter–subject mutual trust, self-ratings provide personal assessments of the subject's mean behavior across time and settings, and consequently there should be a significant relation between an item's face validity and its empirical (concurrent) validity (Goldberg & Slovic, 1967).

Yet, even under such idyllic conditions, we should not expect an automatic correspondence between *all* self-reports and the appropriate peer reports or other nontest behaviors (Buchwald, 1961). Clearly, Meehl (1945) was correct when he argued that the nontest correlates of self-reports must be discovered by empirical means. However, the response to a single item now appears to be too frail a limb on which to hang that empirical enterprise. For the crux of the problem is not whether empirical analyses are necessary, but rather whether they should be carried out on (a) the responses to single items or (b) the average values across many responses to content-coherent sets of items. The former option leads to the construction of External scales. The latter demands the prior construction of reasonably homogeneous Intuitive scales.

The relative advantages of the latter approach have become clearer during the past decade. First of all, scales constructed by the External strategy have been shown to be alarmingly vulnerable to contamination from the idiosyncratic characteristics of the samples on which they are constructed, potentially introducing a host of sources of nuisance variance (Jackson, 1971; Meehl, 1972). Moreover, the content homogeneity of good Intuitive scales provides a less ambiguous sample of self-report than is found in most External scales (Norman, 1972), and consequently the empirical linkages between self-reports and other important behavioral patterns can be described more clearly and conceptualized more simply with sets of Intuitive scales than with sets of External ones.

Finally, and most significantly, it has now been shown that linear combinations of Intuitive scale scores provide as valid, or more valid, predictions of nontest criteria than do linear combinations of single item re-

sponses (Goldberg, 1972a). And, Hase and Goldberg (1967) have shown that predictions based on linear combinations of External scale scores are no more valid than those based on linear combinations of Intuitive scale scores. In the present study, inventories constructed by the Intuitive strategy from the CPI item pool produced average cross-validities against nontargeted criteria at least equal to those from the corresponding inventory constructed by the External strategy, a direct replication of the earlier results. Borgen (1972) has provided another replication of these findings in an entirely different context, based on a comparison between the occupational (External) scales and the Basic Interest (Intuitive) scales from the Strong Vocational Interest Blank in differentiating among individuals in diverse careers.

This is not to say that there is no place in assessment for empirical item analyses of the sort that Meehl (1945) once advocated so persuasively. Clearly, such analyses are often necessary in the context of discovery, in identifying particular types of item content that relate to the construct under investigation. Nonetheless, as Meehl (1972) himself has now agreed, scale development should rarely stop at this stage.

REFERENCES

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

ASHTON, S. G. In response to Jackson's challenge: Item writers vs. the correlation coefficient. Unpublished doctoral dissertation, University of Oregon, 1973.

BORGEN, F. H. Predicting career choices of able college men from occupational and basic interest scales of the Strong Vocational Interest Blank. *Journal of Counseling Psychology,* 1972, **19,** 202–211.

BUCHWALD, A. M. Verbal utterances as data. *In* H. Feigl & G. Maxwell (Eds.), *Current issues in the philosophy of science.* Pp. 461–468. New York: Holt, 1961.

BUTT, D. S., & FISKE, D. W. Comparison of strategies in developing scales for dominance. *Psychological Bulletin,* 1968, **70,** 505–519.

CATTELL, R. B. Confirmation and clarification of primary personality factors. *Psychometrika,* 1947, **12,** 197–200.

CATTELL, R. B. *Personality and motivation structure and measurement.* Yonkers-on-Hudson: World Book, 1957.

Esso Standard. How to use alternation ranking. Report no. III-7.11 in *Social Science Research Reports, Vol. III: Performance Review and Evaluation.* Standard Oil Company (New Jersey), 1962.

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

GOLDBERG, L. R. Review of the California Psychological Inventory. *In* O. K. Buros (Ed.), *The seventh mental measurements yearbook.* Pp. 94–96. Highland Park, NJ: Gryphon, 1972. (b)

GOLDBERG, L. R., & SLOVIC, P. Importance of test item content: An analysis of a

corollary of the Deviation Hypothesis. *Journal of Counseling Psychology,* 1967, **14,** 462–472.

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

JACKSON, D. N. A sequential system for personality scale development. *In* C. D. Spielberger (Ed.), *Current topics in clinical and community psychology.* Vol. 2, Pp. 61–96. New York: Academic Press, 1970.

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

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

MEEHL, P. E. Reactions, reflections, projections. *In* J. N. Butcher (Ed.), *Objective personality assessment: Changing perspectives.* Pp. 131–189. New York: Academic Press, 1972.

MEGARGEE, E. I. *The California Psychological Inventory handbook.* San Francisco: Jossey-Bass, 1972.

MURRAY, H. A., *et al. Explorations in personality.* New York: Oxford University Press, 1938.

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

NORMAN, W. T. Psychometric considerations for a revision of the MMPI. *In* J. N. Butcher (Ed.), *Objective personality assessment: Changing perspectives.* Pp. 59–83. New York: Academic Press, 1972.

TUPES, E. C., & CHRISTAL, R. E. Recurrent personality factors based on trait ratings. *USAF ASD Technical Report,* 1961, No. 61–97.