Construction Strategies for Multiscale Personality Inventories

Matthias Burisch
University of Hamburg

In a replication of the well-known Goldberg (1972) study, sets of inventory scales were constructed from a common item pool, using variants of what are here called the Inductive, Deductive, and External strategies, respectively. Two additional scales were also written. Peer ratings for 21 traits, employing two different scale formats, served as criteria. Subjects were 138 students of both sexes. In spite of a number of procedural differences, most notably a concentration on "trait-relevant" criteria and the use of zero-order correlations as indices of validity, Goldberg's main finding stood unchallenged: Very little variation in validity was attributable to construction strategies. It is pointed out, however, that the Deductive strategy accomplishes its feats with much less effort and considerably shorter scales. Further work with this methodology is urged, as is increased attention to the measurement of criterion variables.

That validity coefficients of approximately $r = .30$ are not altogether atypical for personality scales is not a new finding. Judging from the public embarrassment, however, that followed Mischel's (1968) provocative book, one might infer collective repression of this information on the part of many. While the ensuing discussion concentrated on the issues of "consistency of personality" and problems of validation research (e.g., Alker, 1972; Bem, 1972), there was hardly any mention of another suspicion—namely, that something might be wrong with the way personality inventories are usually constructed.

But there were some (e.g., Fiske, 1971) who questioned the prevalent beliefs that a good inventory had to be:

1. global at the conceptual level (Peterson, 1965);
2. heterogeneous and "disguised" in item content (e.g., Cattell & Tsujioka, 1964); and
3. developed either by means of some elaborate multivariate technique, preferably factor analysis, or by the "empirical" criterion-group approach (e.g., Meehl, 1945).

And the evidence was not all that negative. By methods that laid maximum stress on clearly defined constructs and content saturation of items, Jackson (1967, 1970) had achieved median validity coefficients of approximately $r = .50$ for his 22-scale Personality Research Form, using peer ratings as criteria. Though still not satisfactory compared to absolute standards, these figures were encouraging, indeed.

What could be expected from modifying construction strategies? Clearly, this is an empirical question; but systematic studies were painfully wanting.

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 2, No. 1 Winter 1978 pp. 97-111
© Copyright 1978 West Publishing Co.

The Goldberg Study

The important studies of Goldberg and his associates (Hase & Goldberg, 1967; Goldberg, 1972) were undertaken to fill this gap. Only the roughest outline can be presented here; and the comparison of multivariate prediction techniques, contained in the monograph by Goldberg (1972), will be omitted completely.

Using various strategies or sub-strategies, 5 sets of 11 scales each were (or had been) constructed from the common 468-item pool of the California Psychological Inventory (CPI). Four additional "control" sets (e.g., random scales) will be neglected here.

The analyses in Goldberg (1972) used the data from 152 female students living together in one dormitory. In addition to the CPI protocols from which all scales were scored, 13 criterion variables were measured for each subject. These ranged from mean peer ratings for five traits (such as "dominance") to a measure of "academic survival" (one, two, or three years spent in college).

For each of the five 11-scale inventories and each of the 13 criteria, a multiple regression equation was developed, using only one-half of the sample. The resulting equation was then cross-validated on the other half. Repeating the process in the reverse direction yielded two cross-validity coefficients. These were averaged to furnish a validity index for each particular inventory-criterion combination.

The most outstanding result was obtained when the validity coefficients were again averaged across the 13 criteria: not only was the general level very low (grand mean of .26), but the range was only from .24 to .28. This finding—that all of the strategies produced inventories with almost identical validities—should have surprised many, whose expectations, it must be admitted, may have varied. Those favoring the "content" approach might find some comfort in Goldberg's result that a 5-scale subset of his "Rational" inventory outdid all other subsets in the study, with a mean cross-validity (averaged over all 13 criteria) of $r = .39$.

The Goldberg study must be regarded as a milestone in the history of personality assessment. Nonetheless, it is open to criticisms which limit generalizations from its results, as the author himself concedes in welcome frankness. Since these objections provided part of the motivation behind the present replication, some of them are discussed below.

Criticisms of the Goldberg Study

Criterion variables. Under what circumstances can a variable Y be called a criterion for a test X? One informal rule is that Y and X should be measures of the same target construct and that, given conflicting information, any practical decision should be based on Y rather than on X. This definition obviously pertains to theory-guided research.

Admittedly, practical necessities often require some complex behavior to be predicted by whatever information happens to be at hand. The differences between these two research contexts should be reflected in terminology: "criteria" can be distinguished from "target variables." The (cross-validated) correlation of a broad-bandwidth battery with a target of the latter sort might then be termed its "effectiveness" rather than its "validity."

From this point of view almost all but the six peer ratings (five traits plus "how well known") seem to fall into the "target" category. To expect that the scales constructed to measure 11 of Murray's (1938) "needs" should predict the time a student will drop out of college may be asking a little too much. If one inventory excels another with respect to "effectiveness," that may be because it contains more pertinent scales—not necessarily better ones. And, if several inventories yield equally unsatisfactory predictions, one might surmise that most of their target variables were equally out of reach.

Validity coefficients. A related objection can be raised against the use of multiple correlations as validity coefficients. If the sociability scale in an inventory does not correlate with a peer rating of sociability, while scales for dominance and "intellectual efficiency" do, the multiple correlation between the whole set and the rating can be quite high. Should it then be claimed that "validity" has been demonstrated? In fact, as Tables 6 and 7 in Hase and Goldberg (1967) show, some criteria were most highly correlated with the "wrong" predictors.

Construction of inventory scales. It would be unfair to judge the scales in a study of this type by the most taxing of standards. On the other hand, a comparison of strategies is meaningful only insofar as each of them is given a chance to "do its best." The Theoretical and the Rational inventories, variants of what Goldberg calls the "Intuitive" strategy, would not be expected to operate at their upper bound of quality. Of the 11 Rational scales, 7 were compiled by just one author; the rest were standard CPI scales. For both sets the implicit definitions of the concepts to be measured were apparently very broad. The construct of "Social Presence" was represented by no less than 56 items, almost one-eighth of the total pool. The item "I like to listen to symphony orchestra concerts on the radio" was scored for "Academic Achievement." The item analysis eliminated only items correlating below .19 with their own scale; correlations with other scales were not inspected. In fact, even item overlap among the scales was tolerated.

The Theoretical scales included items which at least two of three graduate students agreed were relevant for one of 11 Murray needs. No empirical data were used in the process. Both facts may account for the item consistency (r_{it}) coefficients (Hase & Goldberg, 1967, Table 1), which could have been higher for such relatively short scales.

Sample effects. Goldberg admits that the degree of optimization differed among the various inventories. The Factor and Multiple Scalogram Analysis inventories were constructed from practically the same set of data as that used for validation purposes. Thus, there was hardly a chance for these scales to come "unglued" internally. In contrast, the Empirical and Rational inventories were developed from different samples; and the Theoretical inventory was constructed data-free.

If one accepts the above criticisms of criterion variables and construction methodology and neglects the problem of validity coefficients for the moment—since no analyses of discriminant validity were published—then the Theoretical inventory seems to deserve the prize. Regarding only the six peer ratings as criteria, mean cross-validated correlations were Theoretical .40, Rational .38, Factor .38, Empirical .30, and Multiple Scalogram Analysis .28.

While it is true that the Rational and Factor sets are close runners-up, a global comparison should take into consideration the fact that the Theoretical inventory not only had the second shortest scales, but also was constructed with a minimum of effort. On the basis of just this one study, test constructors would probably opt for one of the "Intuitive" strategies.

It might be speculated that an even clearer contrast would have emerged had the latter strategies not been handicapped in what is potentially their strongest point—unfortunately by a necessary feature of the design. With the content approach, the first step is a precise definition of the concept to be measured (see Jackson, 1971). Subsequently, there is an attempt to formulate items that "hit the core" of the construct; thus, the term "deductive" is preferred for the strategy. In the reverse process, that of selecting items which seem to "fit" a particular concept (or even to cluster items first and label them later), there is always the temptation to accept too many items, thereby making the scale "fuzzy." However, granted that the item pool had to be the same for all strategies, this problem could not be circumvented.

In view of the above considerations, a replication seemed worthwhile—if only to investigate the matter in a different cultural setting.

Method

Inventory Scales

The Freiburger Persönlichkeitsinventar (FPI, Fahrenberg & Selg, 1970; Fahrenberg, Selg, &

Hampel, 1973) provided both the common item pool for this study and a set of scales representing the "Internal" strategy of scale construction (which is here called "Inductive"). Although content for its 212 items was borrowed from such sources as the MMPI, MPI, EPI, and 16 PF questionnaires, the FPI is one of the few inventories that was not simply translated into German. Of its 12 standard scales, 9 were formed (with some exceptions) according to the highest absolute item loading in a principal components analysis, using the test protocols of 630 subjects. Only the first 8 of these scales (hereinafter referred to as STD scales) were included in this project. (STD 9 has the function of a lie scale, while STD 10 Extraversion, STD 11 Neuroticism, and STD 12 Masculinity are superordinate scales constructed by quasi-rational or external methods, respectively.)

Quite a few items apparently have little to do with the names of their scales. For example, the items "I feel almost constantly hungry" and "I used to dream rather often" appear in STD 2 Aggressiveness. As a matter of fact, in this study roughly one quarter of all items correlated highest with some STD scale other than its own. Although derived from an orthogonal factor rotation, several scales intercorrelate substantially (maximum $r = .61$).

In an attempt to approximate the Deductive strategy, the author and three students of psychology independently searched the FPI for trait constructs that seemed to be represented in its items. In one case an STD concept (Excitability) could be carried over, while others were more specific subconstructs (e.g., Offensiveness as the spontaneous variety of Aggressiveness). Still others were unique (e.g., Daydreaming). Working definitions were set up for 10 variables that appeared promising. Next, each team member independently collected items that (for him/her) conformed to the trait definition. Items were either included in a provisional scale or rejected only after intensive group discussions.

FPI protocols of 186 first-term psychology students provided the data base for an item analysis. Items correlating higher with an alien scale than with their own (so-called "anomalies") were discarded and not reused for other scales. (Four anomalous items were retained because of their content; of these only two remained anomalous in the main sample.)

Those eight Deductive scales that showed the best internal properties—high KR20 (r_{tt}), Loevinger's homogeneity coefficient (H_t), mean item-scale correlation, and low correlations of items with alien scales—were kept for the final inventory, abbreviated DED. As with the STD scales, there was no item overlap. After exclusion of 35 items during item analysis, 74 remained.

In terms of the above internal properties and with respect to scale correlations, even the ad hoc provisional scales equalled or surpassed the much more elaborately constructed STD scales. (The one exception was the KR20 coefficient, which heavily depends on test length. This coefficient is reported in spite of its well-known disadvantages, e.g., Loevinger (1947) and Lumsden (1976), chiefly because the custom is so deeply rooted.)

The last-minute decision to write two additional Deductive scales was prompted by the discovery that a second questionnaire, to be given for other purposes, did not quite fill its two pages in print. An ostensibly easy-to-measure construct (Spontaneous Aggression) and an ostensibly hard one (Depressive Mood) were chosen, and in about two hours seven items were written for the first (scale A′) and five items for the second (scale D′). No further editing was possible, nor could any quantitative analyses be undertaken at the time.

To see how an inductive grouping method would fare with the 74 items selected for inventory DED, these were subjected to a principal components analysis, using the same data as for the item analyses. An effort was made to imitate the steps taken for the construction of the standard FPI: phi coefficients were used and five factors, accounting for 32% of the variance, were extracted and rotated by varimax. Items were assigned to factors according to highest absolute loading, again with a few exceptions. Factor scales were labeled in the customary impressionistic fashion. The resulting inventory was abbreviated FAC.

A set of External scales (abbreviated EXT) was constructed after collection of the validation data. For this purpose, items were selected which correlated maximally with the mean peer ratings that served as the primary criteria. It was decided to match the length of parallel External and STD or DED scales sharing the same criterion rating. Thus, since scale STD 1 Somatic Lability had 34 items, those 34 items out of the total pool most valid for the Somatic Lability rating were picked to form scale EXT-S1. The same procedure was followed for the rest of the STD scales and for DED 1 through DED 8. Two corresponding versions were developed for each scale, one from one-half of the sample which was validated in the other, and vice-versa. This made up a total of 32 External scales, many of them overlapping, of course. The 2 × 8 scales aiming at STD ratings were designated EXT-S, while the 2 × 8 scales paralleling DED comprised subinventory EXT-D.

Criterion Variables

As the best compromise solution to the criterion problem, trait ratings by close acquaintances were chosen. Self-ratings were also obtained.

Two rating scale formats were employed. One was the familiar horizontal 9-point ("unanchored") scale, as shown in Figure 1.

In an attempt to reduce the much-criticized ambiguity of such ratings, example-anchored scales were painstakingly constructed after the methods of Taylor et al. (1970, 1972). These were vertical 100-point scales anchored by 8 to 12 statements illustrating various degrees of the trait in question. For instance, a rating of 91 on an aggressiveness scale was illustrated by the phrase "Sometimes enjoys tormenting others," while "Gets into arguments very rarely" was printed next to point 31. The statements paraphrased content appearing in the corresponding inventory scales. Their positions on the rating scales were obtained from a rank-order scaling (cf. Taylor, 1968).

Subjects

Subjects were 78 male and 60 female students of the University of Hamburg, excluding psychology students. They were contacted in their dormitories and offered a small amount of money in addition to feedback of their test results if they participated. Each subject was required to find two friends who knew him or her well and were willing to act as raters. Raters were paid the same sum of money.

Group testing sessions of about 60 to 90 minutes in duration were held in the dormitories. Subjects anonymously filled in the FPI and a second questionnaire containing scales A′ and D′, interspersed with other items. They then rated themselves on the same scales that were used for the peer ratings.

Concurrently, but in separate rooms, the criterion raters received a booklet containing 21 unanchored and 21 example-anchored scales, together with written instructions.

The entire procedure was repeated after two weeks. Raters of the same subject worked inde-

Figure 1

Example of Unanchored Rating Scale

Aggressiveness

aggressive									unaggressive
inconsiderate	9	8	7	6	5	4	3	2	1	considerate
offensive									peacable

pendently during each session. Coordination of the various data sets was accomplished by means of code numbers.

Analyses

The principal method of analysis consisted of intercorrelating all variables in the study. A few special procedures are described below.

Criterion variables. Before computing retest stabilities and correlations for the peer ratings, these were averaged over each ratee's pair of raters. No further transformations of raw data were used.

Inter-rater agreement was appraised by the coefficient r_k^*. This coefficient is a slightly modified variant of the standard coefficient r_k (see Winer, 1962, sect. 4.5), yielding an unbiased estimate for the reliability of k raters' average rating, as defined by Equation 1.

$$r_k^* = 1 - \left(\frac{n-3}{n-1}\right)\left(\frac{MS_e}{MS_b}\right), \quad (n>3) \quad [1]$$

with

$$MS_e = \frac{SS_b + SS_r}{n\ (k-1)} \quad [2]$$

and

n = number of ratees
k = number of raters
MS_e = mean square error
MS_b = mean square between raters
SS_b = sum of squares between raters
SS_r = sum of squares for residual

Inventory scales. Internal properties were analyzed in a variety of ways. In addition to familiar coefficients (such as retest stability, internal consistency, and Loevinger's homogeneity coefficient), item consistency and mean item-scale correlation (part-whole corrected biserials) were also computed. The coefficient of item consistency, r_{ii}, is simply a scale's internal consistency estimated for the fictitious scale length "1 item" by the Spearman-Brown formula. The number of "anomalies" per scale was also counted, i.e., instances where an item correlated more highly with one or more other scales than with its own.

All figures given for external scales represent averages of coefficients computed separately for the two versions of each scale, using the respective cross-validation half of the data only. Therefore the N in these cases if 69. (By mistake three subjects did not receive scales A′ and D′, reducing that N to 135.)

Since all the validities were based on a mixed-sex sample, one might wonder to what extent they are comparable to results from studies using subjects of either sex only, e.g., Goldberg's. To control for possible inflation or deflation of coefficients due to sample heterogeneity, sex was partialled out of all validities, using point-biserials for correlations with that variable. The strongest single effect produced in this way was the decrease of a coefficient from .40 to .35. The mean validity of an inventory was reduced by .02 at most. The same was true for validities of self for peer ratings: mean shrinkage of .01 with unanchored scales and of .01 (first session) or .02 (second session) with example-anchored scales. It seemed justified, therefore, to report only uncorrected correlations computed from the total sample.

Results

Criterion Variables

The time-consuming construction of example-anchored (EA) rating scales (from which considerable improvement had been expected) must be regarded as a failure. In no respect were they appreciably superior to the simple unanchored (UA) scales.

Table 1 contains information on retest stability over the two-week interval (r_{tt}), inter-rater agreement (r_k^*), and intercorrelations of ratings within the same scale format. Figures below the main diagonal of the correlation matrix repre-

Table 1

Retest Stability (r_{tt}), Inter-Rater-Agreement (r^*_k), and Intercorrelations of the Peer Ratings, Separately for Both Scale Formats

		r_{tt}		r^*_k		Intercorrelations[a]																				
	Scale	UA	EA	UA	EA	S1	S2	S3	S4	S5	S6	S7	S8	D1	D2	D3	D4	D5	D6	D7	D8	F1	F2	F3	F4	F5
STD 1	Somatic Lability	77	80	43	63	73	22	55	33	02	-44	08	16	36	28	52	30	-01	55	03	-14	33	00	-21	46	15
STD 2	Aggressiveness	82	85	60	62	21	77	40	73	-11	-36	63	11	11	72	37	74	-02	18	45	-07	77	01	10	25	74
STD 3	Depressiveness	77	82	47	59	58	26	82	58	-29	-64	17	49	68	24	83	52	-31	54	-08	-36	56	-27	-43	79	23
STD 4	Excitability	80	87	66	63	38	66	42	75	-12	-62	47	20	34	59	51	79	-08	29	25	-09	82	-04	-17	41	55
STD 5	Gregariousness	81	79	61	60	-06	-12	-35	05	78	25	07	-57	-37	02	-36	-07	80	-25	23	68	-16	84	40	-44	-08
STD 6	Calmness	69	75	28	41	-47	-25	-65	-31	28	60	-23	-36	-58	-32	-55	-58	28	-45	01	29	-62	25	50	-46	-22
STD 7	Striving for Dominance	75	81	46	55	11	57	-06	46	23	19	73	-18	-09	52	18	56	07	00	63	11	57	10	28	04	55
STD 8	Inhibition	66	73	48	39	19	05	44	02	-61	-46	-30	68	64	-01	53	17	-53	32	-38	-43	22	-52	-54	55	04
DED 1	Anxiousness	74	76	50	39	42	12	64	14	-45	-62	-23	61	74	02	65	26	-36	48	-30	-41	39	-37	-69	68	-02
DED 2	Offensiveness	72	79	39	57	14	54	02	48	16	09	57	-10	-17	72	21	56	05	14	44	06	62	11	13	05	72
DED 3	Depression Proneness	84	84	62	58	59	30	80	33	-38	-55	-01	41	57	06	83	50	-39	47	-07	-46	52	-34	-41	75	27
DED 4	Excitability	82	82	55	65	38	70	41	84	06	-34	48	05	16	51	33	75	00	19	38	-06	82	04	-01	33	52
DED 5	Sociability	69	81	50	52	-08	-10	-35	06	77	28	17	-61	-44	18	-34	-04	74	-26	25	63	-12	84	46	-46	-11
DED 6	Fatigue	67	81	44	43	43	16	45	16	-28	-39	-13	33	40	-14	37	13	-32	70	-19	-33	25	-27	-41	60	14
DED 7	Unyieldingness	69	71	49	40	-05	45	-18	32	20	27	65	-30	-21	52	-07	35	18	-23	68	27	27	29	45	-23	42
DED 8	Enterprise	74	75	48	27	02	06	-21	22	66	17	31	-46	-34	26	-20	15	60	-11	22	65	-14	72	42	-47	03
FAC 1	Lack of Control	85	86	56	62	42	69	48	81	-01	-49	40	10	30	41	40	83	00	18	29	14	83	-11	-17	38	58
FAC 2	Gregariousness	80	83	52	52	-08	-07	-35	08	75	27	23	-53	-40	23	-30	-01	76	-20	21	76	-02	79	46	-43	-02
FAC 3	Boldness	58	74	39	25	-23	09	-51	07	48	57	44	-64	-63	31	-45	06	45	-39	43	45	-07	45	71	-53	17
FAC 4	Unstable Personality	76	81	68	57	51	33	78	39	-41	-58	-07	49	62	01	81	35	-39	46	-14	-23	42	-32	-44	79	13
FAC 5	Aggressiveness	82	80	63	43	15	81	19	62	-10	-16	55	11	06	62	25	68	-14	09	43	02	60	-07	11	27	66
	Mean	76	80	51	51																					

[a] To save space in column headings, the abbreviations STD, DED, and FAC are further abbreviated to S, D, and F, respectively. Values in the lower triangular matrix represent correlations of UA scales, values in the upper triangular are correlations of EA scales. The underlined figures in the main diagonal are the correlations of corresponding UA and EA scales. Decimal points are omitted throughout.

sent UA correlations, while EA intercorrelations are above the main diagonal. Only the results of Session 1 are given, since no marked changes occurred in Session 2.

Although the average stability coefficients were quite adequate, inter-rater agreement was disappointing. This failure to obtain well-defined criterion information may be the chief weakness of the present study. Increasing the number of raters per subject would have raised the size of these coefficients, but it might also have served to reduce the validity of the ratings as more remote acquaintances would have had to be accepted.

The pattern of correlations within the UA and EA scales, respectively, is quite similar and in general makes good sense semantically. Correlations of parallel UA and EA scales (the main diagonal elements) ranged from .60 to .83, with a mean of .74. Apparently the anchor statements, added on EA scales to clarify trait definitions, did not substantially change raters' interpretations of what should be rated.

There are some hunches as to why the example-anchored scales did not live up to expectations. Single statements were employed as anchors in this study, whereas Taylor et al. (1972) used groups of statements clustering at certain scale positions. The latter approach may introduce confusion when some statements in a cluster apply to a ratee, while others do not. But it may also help to convey images of personality "types" which can serve as reference points for the raters. With the technique used in the present paper, different anchor statements, sometimes far apart on a scale, often seemed equally applicable. In addition, item content was reflected in the statements, in order to guarantee the same degree of "fairness" to all inventories and strategies. This resulted in sets of statements covering a wide spectrum of behavior at very different levels of abstraction.

More research is needed to identify the crucial factors. Training raters prior to obtaining their judgments would also appear promising where it is technically possible. One lesson from these results is that methods of criterion measurement deserve much more attention than they normally receive in validity studies.

Inventory Scales

Internal properties. Table 2 lists some basic information for the 21 scales of the Standard (STD), Deductive (DED), and Factor (FAC) inventories, as well as for scales A′ and D′ and the 16 scales of the External (EXT) inventory.

Column NI gives the number of items per scale. STD (and EXT-S) had the longest scales and DED (and EXT-D) had the shortest. Note that the scale lengths of inventories STD and EXT-S and of inventories DED and EXT-D were matched. In spite of wide variations in scale length, mean retest stabilities, listed in column r_{tt}, ranged only from a low of .82 (EXT-D) to a high of .88 (EXT-S). Mean internal consistency (KR20) was much more affected, varying from .62 and .64 (EXT-D) to .80 and .83 (STD). Controlling for the effects of scale length (coefficient r_{ii}) brought DED to the top and sent EXT-S to the bottom of the rank order. The r_{it} values reflect the somewhat higher specificity of the DED scales; so does Loevinger's homogeneity coefficient H_t. A similar pattern was found for mean item-scale correlations, presented in the next two columns under the heading r_{it}. Finally, in the "Anom." columns are the numbers of items anomalously correlating more highly with at least one other scale than with their own. More than one-quarter of the STD items fell into this category, as compared to less than 15% of the DED and FAC items. (With 8 DED and 5 FAC scales, chance probabilities are unequal.) Because of large item overlap, anomalies were not counted for EXT scales.

In summary, wherever the "sameness" of the items in a scale was the quality to be assessed, the DED scales were generally best, followed closely by the FAC inventory. In contrast, the EXT and STD scales, and particularly the subset EXT-D, proved to be more heterogeneous statistically. As with the validity coefficients,

Table 2

Internal Properties of All Inventory Scales, Separately for Sessions 1 and 2[a]

				KR20		r_{ii}		H_t		$\bar{r}_{it}$		Anom.	
	Scale	NI	r_{tt}	1	2	1	2	1	2	1	2	1	2
STD 1	Somatic Lability	34	87	84	90	13	21	19	30	47	62	8	3
STD 2	Aggressiveness	26	82	67	70	07	08	12	13	31	36	14	15
STD 3	Depressiveness	28	91	87	92	20	28	27	37	55	67	4	2
STD 4	Excitability	20	87	86	87	23	25	33	35	62	64	4	3
STD 5	Gregariousness	28	92	88	89	20	23	30	33	57	62	5	2
STD 6	Calmness	20	89	82	84	19	21	26	29	53	56	7	9
STD 7	Striving for Dominance	20	76	64	66	08	09	12	13	34	37	11	11
STD 8	Inhibition	20	87	83	86	20	24	27	33	52	59	6	7
STD	Mean	24.50	87	80	83	16	20	23	28	49	57	30%	27%
DED 1	Anxiousness	13	87	80	84	23	29	30	37	56	64	2	3
DED 2	Offensiveness	11	75	56	67	11	16	17	25	32	46	3	2
DED 3	Depression Proneness	10	85	81	83	30	33	45	53	65	69	1	1
DED 4	Excitability	11	87	85	85	33	34	49	50	72	75	0	0
DED 5	Sociability	12	90	82	86	28	34	38	45	63	71	0	0
DED 6	Fatigue	5	80	60	68	23	29	38	45	50	58	1	2
DED 7	Unyieldingness	6	72	55	59	17	20	32	33	42	47	0	1
DED 8	Enterprise	6	80	60	59	20	19	28	27	43	42	2	2
DED	Mean	9.25	83	70	74	23	27	35	39	56	63	12%	15%
FAC 1	Lack of Control	15	88	82	83	23	25	37	39	62	65	1	1
FAC 2	Gregariousness	16	91	84	87	25	29	35	40	60	67	1	1
FAC 3	Boldness	12	81	75	79	20	23	29	34	51	58	3	3
FAC 4	Unstable Personality	18	90	84	89	22	30	32	44	57	69	3	1
FAC 5	Aggressiveness	13	76	59	68	10	14	16	22	37	43	3	2
FAC	Mean	14.80	86	77	81	20	24	30	36	54	62	15%	11%
EXT-S1	Somatic Lability	34	92	88	91	18	23	27	34	54	63		
EXT-S2	Aggressiveness	26	82	73	72	10	09	14	14	40	39		
EXT-S3	Depressiveness	28	91	88	90	21	23	30	34	59	62		
EXT-S4	Excitability	20	84	83	83	21	22	30	31	58	59		
EXT-S5	Gregariousness	28	91	82	85	14	18	20	25	50	56		
EXT-S6	Calmness	20	89	87	87	26	25	35	35	63	64		
EXT-S7	Striving for Dominance	20	77	49	52	06	05	08	08	23	23		
EXT-S8	Inhibition	20	92	81	82	17	19	24	27	51	56		
EXT-S	Mean	24.50	88	79	80	16	18	23	26	50	53		
EXT-D1	Anxiousness	13	89	78	80	21	23	29	31	55	58		
EXT-D2	Offensiveness	11	74	35	32	06	05	09	08	20	17		
EXT-D3	Depression Proneness	10	84	81	81	29	29	41	42	66	67		
EXT-D4	Excitability	11	82	77	79	26	28	35	36	61	63		
EXT-D5	Sociability	12	89	72	77	19	22	24	31	51	57		
EXT-D6	Fatigue	5	81	52	61	19	24	23	32	40	48		
EXT-D7	Unyieldingness	6	76	58	55	19	18	24	24	41	40		
EXT-D8	Enterprise	6	77	47	50	13	16	19	23	33	40		
EXT-D	Mean	9.25	82	62	64	19	21	25	28	47	50		
A'	Spontaneous Aggression	7	75	61	64	19	20	37	40	38	44		
D'	Depressive Mood	5	73	60	55	23	20	67	52	39	43		

[a] For STD, DED, and FAC N = 138; for EXT N = 138; for A' and D' N = 135. See text for explanations. Scales EXT-S1 through EXT-S8 share the same criterion rating (and length) with scales STD 1 through STD 8; so do scales EXT-D1 through EXT-D8 with scales DED 1 through DED 8. Anomalies were not counted for the EXT scales which overlap mutually.

there are no appropriate tests of significance, since just one sample was used. Moreover, all inventories overlap mutually. From a practical standpoint, however, the similarities are far more impressive than the differences.

Somewhat clearer contrasts emerge when within-inventory correlations are compared. Mean absolute intercorrelations of scales were: STD .36; DED .24; FAC .26; EST-S .67; EXT-D .43. The higher level of independence among the DED and the FAC scales cannot be attributed to their lower reliability. Double correction for attenuation, using retest stabilities, gave these average figures: STD .42; DED .30; FAC .31. Disattenuated correlations could not be averaged for the heavily overlapping EXT scales since quite a few of them exceeded 1.0 in absolute value.

Validity. Correlations of all inventory scales with corresponding peer ratings are presented in Table 3, separately for both rating scale formats and both testing sessions. Mean coefficients, averaged over both formats and sessions, are given in the column "PR $\bar{r}$." Number of items is repeated in the "NI" column.

There were no special rating scales for A′ and D′. It seemed appropriate, however, to use the ratings of Aggressiveness (STD 2), Offensiveness (DED 2), and Aggressiveness (FAC 5) as targets for A′ and those of Depressiveness (STD 3), Depression Proneness (DED 3), and Unstable Personality (FAC 4) for D′.

The righthand column of Table 3 ("OL %") gives the percentage of overlapping items included in both versions of an EXT scale. The two EXT-S1 scales, for instance, had 12 items (35.3%) in common. Because there was no within-inventory overlap among the scales of STD, DED, and FAC, the figures there are zero.

Looking at these results in summary fashion, there is mostly support for the null hypothesis. For inventories STD, DED, and FAC, mean validity coefficients of .39, .39, and .41 render tests of significance unnecessary. The scales of EXT-S were slightly better than their STD counterparts; EXT-D, with scales equally as short as DED, was less valid. The most plausible interpretation of this divergency is in terms of item overlap. In choosing items for EXT-D, a higher degree of selectivity was possible because fewer items were needed. But in a new sample these items regressed more markedly towards the mean. Upon cross-validation, they were typically *not* among the most valid ones. In fact, with the 6-item scales of EXT-D7 and EXT-D8, there was no overlap at all.

Particularly gratifying, of course, was the performance of scales A′ and D′. Although by far the shortest in their respective content areas, A′ clearly performed best of all "aggression" scales, while D′ reached the same level as the other "depression" scales.

Some additional findings cast further doubt on the traditional wisdom concerning test length and validity. In a few cases, DED or FAC scales were almost or fully contained within longer STD scales. For example, DED 5 and DED 6 are subsets of STD 5 and STD 1, respectively. The validities of the longer and the shorter scales proved to be practically undistinguishable. The mean correlation of STD 1 (34 items) with STD 1 peer ratings was .40, but DED 6 (5 items) achieved almost the same value (.38). DED 6 peer ratings correlated .34 with both DED 6 and STD 1. This was typical for all such comparisons.

While one important aspect of test quality is discriminant validity, there is no standard procedure to appraise it. To obtain a rough index, incidences were counted where an inventory scale possessed an equal or higher correlation with some "alien" peer rating scale than with its target. The proportion of such "stray" cases to the maximum possible number was 18% for STD, 11% for DED, 8% for FAC, 24% for EXT-S, and 18% for EXT-D. Although none of these figures is entirely satisfactory—taking into account the often small distance between "right" and second best "wrong" validities—inventory EXT-S appears least acceptable on these grounds. This is explained mainly by the heavy overlap among different scales within the same

Table 3

Validities of Inventory Scales for Peer Ratings (PR) on Unanchored (UA) and Example-Anchored (EA) Rating Scales, Separately for Sessions 1 and 2[a]

			UA		EA		PR	
	Scale	NI	1	2	1	2	$\bar{r}$	OL %
STD 1	Somatic Lability	34	43	40	40	37	40	0.0
STD 2	Aggressiveness	26	17	13	23	17	18	0.0
STD 3	Depressiveness	28	46	50	42	48	47	0.0
STD 4	Excitability	20	51	44	55	52	51	0.0
STD 5	Gregariousness	28	51	54	45	57	52	0.0
STD 6	Calmness	20	29	49	42	45	42	0.0
STD 7	Striving for Dominance	20	26	20	27	24	24	0.0
STD 8	Inhibition	20	38	38	31	39	37	0.0
STD	Mean	24.50	38	39	39	41	39	0.0
DED 1	Anxiousness	13	30	49	37	45	41	0.0
DED 2	Offensiveness	11	14	18	29	16	19	0.0
DED 3	Depression Proneness	10	54	44	48	45	48	0.0
DED 4	Excitability	11	53	48	49	50	50	0.0
DED 5	Sociability	12	52	46	44	52	49	0.0
DED 6	Fatigue	5	34	34	35	32	34	0.0
DED 7	Unyieldingness	6	23	31	25	31	28	0.0
DED 8	Enterprise	6	38	37	38	39	38	0.0
DED	Mean	9.25	38	39	38	39	39	0.0
FAC 1	Lack of Control	15	55	45	57	43	50	0.0
FAC 2	Gregariousness	16	44	39	52	52	47	0.0
FAC 3	Boldness	12	38	36	33	48	39	0.0
FAC 4	Unstable Personality	18	53	51	47	42	48	0.0
FAC 5	Aggressiveness	13	12	22	23	25	21	0.0
FAC	Mean	14.80	41	39	43	42	41	0.0
EXT-S1	Somatic Lability	34	44	48	42	42	44	35.3
EXT-S2	Aggressiveness	26	38	30	42	31	35	26.9
EXT-S3	Depressiveness	28	51	49	47	48	49	25.0
EXT-S4	Excitability	20	48	43	50	50	48	45.0
EXT-S5	Gregariousness	28	52	49	40	53	49	42.9
EXT-S6	Calmness	20	32	51	47	41	43	30.0
EXT-S7	Striving for Dominance	20	36	34	24	34	32	20.0
EXT-S8	Inhibition	20	46	46	43	47	46	25.0
EXT-S	Mean	24.50	43	44	42	43	43	31.3
EXT-D1	Anxiousness	13	28	37	32	39	34	7.7
EXT-D2	Offensiveness	11	23	37	34	34	32	18.2
EXT-D3	Depression Proneness	10	50	40	45	42	44	30.0
EXT-D4	Excitability	11	53	49	48	53	51	45.5
EXT-D5	Sociability	12	43	39	39	46	42	33.3
EXT-D6	Fatigue	5	33	37	31	32	33	20.0
EXT-D7	Unyieldingness	6	18	19	08	20	16	0.0
EXT-D8	Enterprise	6	36	36	25	34	33	0.0
EXT-D	Mean	9.25	36	37	33	38	36	19.3
A'	Spontaneous Aggression	7	45 34 42	52 49 53	44 43 39	52 49 45	45	0.0
D'	Depressive Mood	5	51 55 45	47 49 51	47 52 46	44 47 46	48	0.0

[a] See footnote Table 2 and text for explanations.

sub-inventory version; there were some intercorrelations beyond .90. This, in turn, may partly be due to the comparatively small item pool of the FPI; a few items were among the most valid for many criteria.

A final remark concerns the correlations of inventory scales with corresponding self-ratings. Reflecting largely the similarity between the test constructor's explicit and the subjects' implicit conceptualizations of a trait, these coefficients may be viewed as measures of "face validity." Mean correlations averaged across scales, scale formats, and sessions were: STD .58, DED .60, FAC .62, EXT-S .59, EXT-D .55. These figures corroborate the findings as to validity. The good showing of the EXT scales is even more remarkable here because they contained many remote items with little or no manifest relationship to scale labels. No explanation can be offered, but the phenomenon will be studied further.

Self-Ratings

To compare the usefulness of direct self-ratings against the more elaborate inventory scales, subjects had also been asked to judge themselves, using the same scales as had been used for the peer ratings. Table 4 presents their retest stabilities (r_{tt}) and validities for peer ratings (r_{sp}) separately for both scale formats and both sessions.

The average stability was comparatively low here—one coefficient was even below .50—but these self-ratings took only a fraction of the time required by most of the inventory scales. Again, there was no advantage whatsoever for either scale format.

Peer ratings were predicted slightly better by direct self-ratings than by inventory scales, as was also found by several previous investigators, e.g., Carroll (1952); Wetzel (1963; cf. Peterson, 1965); Hase & Goldberg (1967); Norman (1969). This effect is not altogether surprising, since there must be interdependencies between the image held by a person's peers and his or her self-image; if both are measured by identical instruments, "nuisance effects" due to different operationalizations of a trait are minimized. In view of the simplicity and economy of rating scales, however, the question can be raised, "Why continue to construct inventories if self-ratings do a better job more inexpensively?" (cf. also Taylor et al., 1972).

There is no hard and fast answer to this question. To the author's knowledge, the superiority of self-ratings has only been demonstrated in research with anonymous subjects. In real-life settings, such as a psychiatric clinic, the outcome might be different. Admittedly, any verbal assessment technique is susceptible to impression management. Laymen's ratings must be more or less restricted to "folk concepts," which may be the most meaningful ones, anyway. Trait labels may convey even more disparate denotations in representative samples of subjects than items do; and there may be difficulties with subjects low on verbal comprehension, but both these hunches may turn out wrong. Thus, in the absence of reliable evidence, it would seem wisest to push both lines of research.

Discussion

In spite of the best of intentions, there are a number of weak points that limit interpretations in much the same way as was pointed out for the Goldberg study. Most notable is the use of insufficiently focused criterion ratings. Two factors that could not be adequately controlled are similarity between the derivation and the validation samples and the size of the former. The greatest amount of data went into the construction of STD, but these subjects were probably not fully comparable with the student groups. The opposite is true for DED and FAC, and even more so for EXT. The two random halves of the main sample can be expected to be the most similar pair of subsamples; nevertheless, an N of only 69 may not permit stable correlation estimates.

The external strategy was also handicapped by a limited item pool; however, so was the de-

Table 4

Analyses of the Self Ratings: Retest Stability (r_{tt}) and Correlations with Corresponding Peer Ratings (r_{sp}), Separately for Both Scale Formats and Sessions. N = 138

	r_{tt}		r_{sp}			
			UA		EA	
	UA	EA	1	2	1	2
STD 1 Somatic Lability	76	84	50	54	47	44
STD 2 Aggressiveness	65	69	44	32	43	36
STD 3 Depressiveness	76	76	49	54	37	47
STD 4 Excitability	64	76	47	40	53	52
STD 5 Gregariousness	83	76	47	52	47	53
STD 6 Calmness	65	68	31	43	30	33
STD 7 Striving for Dominance	66	54	42	31	34	34
STD 8 Inhibition	64	67	35	36	43	50
DED 1 Anxiousness	77	66	39	46	43	43
DED 2 Offensiveness	68	58	40	39	25	19
DED 3 Depression Proneness	77	75	52	50	53	53
DED 4 Excitability	69	58	56	57	52	47
DED 5 Sociability	78	79	56	45	41	51
DED 6 Fatigue	74	77	41	42	31	34
DED 7 Unyieldingness	56	60	19	31	33	26
DED 8 Enterprise	77	75	41	35	39	31
FAC 1 Lack of Control	69	74	56	50	52	49
FAC 2 Gregariousness	68	77	44	38	42	50
FAC 3 Boldness	47	64	43	41	32	41
FAC 4 Unstable Personality	81	76	48	55	46	47
FAC 5 Aggressiveness	68	60	45	41	36	29
Mean	71	71	44	44	41	42

ductive construction. The good showing of inventories DED and FAC, which used precisely the same items in somewhat different arrangements, lends some support to the notion that it is the intrinsic qualities of an item (e.g., comprehensibility, subjective relevance to subjects) that count, rather than its assignment to the "correct" construct. However, there will be limits to this rule.

Considering the points above and the many procedural differences, the replication of Goldberg's main results is remarkably clear: no dramatic strategy effects when a common item pool is used. Equally well substantiated and representing even more of the facts is a somewhat different formulation: With "economy-class" criteria, typical validity coefficients of inductively- or externally-constructed inventories can be achieved by much simpler deductive methods and considerably shorter scales. This is in line with the recent findings of Ashton and Goldberg (1973) and Jackson (1975), employing relatively inexperienced students as item writers.

The next task, then, should be to aim at some real advancements, using more serious criteria and optimization of the deductive methodology.

References

Alker, H. A. Is personality situationally specific or intrapsychically consistent? *Journal of Personality*, 1972, *40*, 1–16.

Ashton, S. G., & Goldberg, L. R. 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. *Journal of Research in Personality*, 1973, *7*, 1–20.

Bem, D. J. Constructing cross-situational consistencies in behavior: Some thoughts on Alker's critique of Mischel. *Journal of Personality*, 1972, *40*, 17–26.

Carroll, J. B. Ratings on traits measured by a factored personality inventory. *Journal of Abnormal and Social Psychology*, 1952, *47*, 626–632.

Cattell, R. B., & Tsujioka, B. H. The importance of factor-trueness and validity, versus homogeneity and orthogonality, in test scales. *Educational and Psychological Measurement*, 1964, *24*,3–30.

Fahrenberg, J., & Selg, H. *Das Freiburger Persönlichkeitsinventar*. Göttingen: Hogrefe, 1970.

Fahrenberg, J., Selg, H., & Hampel, R. *Das Freiburger Persönlichkeitsinventar* (2nd ed.). Göttingen: Hogrefe, 1973.

Fiske, D. W. *Measuring the Concepts of Personality*. Chicago: Aldine, 1971.

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

Hase, H. D., & Goldberg, L. R. Comparative validity of different strategies of constructing personality inventory scales. *Psychological Bulletin*, 1967, *67*, 231–248.

Jackson, D. N. *Personality Research Form Manual*. Goshen, NY: Research Psychologists Press, 1967.

Jackson, D. N. A sequential system for personality scale development. In C. D. Spielberger (Ed.), *Current Topics in Clinical and Community Psychology* (Vol. 2). New York: Academic Press, 1970.

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

Jackson, D. N. The relative validity of scales prepared by naive item writers and those based on empirical methods of personality scale construction. *Educational and Psychological Measurement*, 1975, *35*, 361–370.

Loevinger, J. A systematic approach to the construction and evaluation of tests of ability. *Psychological Monographs*, 1947, *61* (Whole no. 285).

Lumsden, J. Test theory. *Annual Review of Psychology*, 1976, *27*, 251–280.

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

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

Murray, H. A. *Explorations in Personality*. Cambridge: Harvard University Press, 1938.

Norman, W. T. "To see oursels as ithers see us!": Relations among self-perceptions, peer-perceptions and expected peer-perceptions of personality attributes. *Multivariate Behavioral Research*, 1969, *4*, 417–443.

Peterson, D. R. Scope and generality of verbally defined personality factors. *Psychological Review*, 1965, *72*, 48–59.

Taylor, J. B. A brief ranking method as an alternative to Thurstone scaling procedures. *Perceptual and Motor Skills*, 1968, *26*, 533–534.

Taylor, J. B., Haefele, E., Thompson, P., & O'Donoghue, C. The reliability of example-anchored scales under conditions of rater heterogeneity and divergent behavior sampling. *Educational and Psychological Measurement*, 1970, *30*, 301–310.

Taylor, J. B., Ptacek, M., Carithers, M., Griffin, C., & Coyne, L. Rating scales as measures of clinical judgment. III: Judgments of the self on personality inventory scales and direct ratings. *Educational and Psychological Measurement*, 1972, *32*, 543–557.

Wetzel, L. C. *Personality traits and cognitive meanings*. Master's thesis, University of Illinois, 1963.

Winer, B. J. *Statistical Principles in Experimental Design*. New York: McGraw-Hill, 1962.

Acknowledgements

Most data analyses for this project were carried out at the University of Hamburg Computer Center. The author is indebted to Ansa Burisch, Ingrid Harbeck, Barbara Hochdörfer, Barbara Knauf, Ute Oster-

walder, Barbara Rindermann, and Renate Rudolph, who cooperated at various stages of the investigation. Special thanks are due Lewis R. Goldberg, upon whose groundwork this study builds. His stimulation, encouragement, and good advice is gratefully acknowledged—which certainly does not imply his responsibility for the final product.

Author's Address

Matthias Burisch, Universität Hamburg, Psychologisches Institut II, von-Melle-Park 5, Hamburg 13, Germany.