Journal of Counseling Psychology
1967, Vol. 14, No. 5, 462–472

Importance of Test Item Content:
An Analysis of a Corollary of the Deviation Hypothesis[1]

LEWIS R. GOLDBERG[2] AND PAUL SLOVIC
University of Oregon *Oregon Research Institute*

As a corollary to his deviation hypothesis, Berg has proposed that content is an unimportant source of variance in personality scales and that "contentless" item pools like the Perceptual Reaction Test are of equal quality to pools containing face valid verbal items. The present study provided a direct test of this assumption by investigating the relationship between the face validity and the empirical validity of diverse kinds of personality inventory items, both verbal and nonverbal. The results indicated that against 5 of 6 criteria, only scales constructed from items of high face validity had significant cross-validity, although all scales had almost equally high validities in the derivation samples. These findings provide the grounds for a strong argument against Berg's corollary.

In a series of articles dating back to 1955, Irwin Berg has repeatedly stated what he calls the deviation hypothesis (e.g., Berg, 1955, 1957, 1959, 1961). A "corollary" of this hypothesis has been formulated by Berg (1957) as follows:

> Stimulus patterns of any type and of any sense modality may be used to elicit response patterns; thus particular stimulus content is unimportant for measuring behaviors in terms of the Deviation Hypothesis. This means that ... we should be able to produce a Bernreuter Personality Inventory, an MMPI, a Strong VIB, etc., by using sights, sounds, tastes, smells, etc., in any combination for item content [p. 160].

"Indeed, any content which produces deviant response patterns will serve, judging from the available evidence.... Accordingly for personality and similar tests a particular item content is unimportant [Berg, 1959, p. 95]."

To test this proposition, Berg and his co-workers developed the Perceptual Reaction Test (PRT), which consists of 60 abstract designs of the sort that could be drawn with ruler and compass; subjects are required to choose one of four response options for each design: "like much," "like slightly," "dislike slightly," or "dislike much." The PRT has been presented as a "contentless" item pool, from which personality scales can be empirically constructed.

A number of studies using the PRT have been reviewed by Berg (1961). They have shown that some PRT items discriminate between such grossly dissimilar criterion groups as (*a*) psychiatric patients and normals (Adams, 1960; Barnes, 1955; Hesterly & Berg, 1958; House, 1960); (*b*) mental retardates and normals (Cieutat, 1960); and (*c*) children, young adults, and the elderly (Boozer, 1961; Hawkins, 1960; Hesterly, 1960; Hesterly & Berg, 1958; Roitzsch & Berg, 1959). Of questionable success were attempts to discriminate tubercular and cardiac patients from nonpatients (Berg, 1961; Engen, 1959). Admittedly unsuccessful were attempts to discriminate delinquents from nondelinquents (Berg, 1961), children from neurotic adults (Roitzsch & Berg, 1959), as well as attempts to discriminate the degree of

[1] This study was supported by Grants G-25123 and GS-429 from the National Science Foundation to Lewis R. Goldberg at Oregon Research Institute and by Grant MH 12972 from the United States Public Health Service. The authors wish to thank Leonard G. Rorer for his invaluable help and Jacob Kind, Patricia Taylor, Richard Hammersley, William Johnson, and George McCarger, who served as research assistants on this project. The authors are also indebted to Warren T. Norman for his thoughtful comments on an earlier draft of this article.

[2] Also at Oregon Research Institute.

schizophrenia (Harris, 1958) and the degree of emotional disturbance (House, 1960). None of these reports present validity coefficients between PRT scales and a criterion index. Instead, they present only the number of PRT responses that differentiated between the criterion groups at some level of statistical significance. In addition, the fact that almost half of the studies are in need of cross-validation (and few, if any, have been replicated in other settings) adds to the difficulty of comparing the PRT with other inventories.

While the deviation hypothesis, itself, has been shown to have very limited utility on strictly logical grounds (Norman, 1963a; Sechrest & Jackson, 1963), Berg's corollary asserting the unimportance of item content has such great implications for personality assessment that it certainly demands a careful empirical examination. If the MMPI could have been constructed using 550 abstract designs as items, then two decades of experimental work aimed at the detection and prevention of faking on personality inventories would have been unnecessary and advice to cheat on personality inventories (e.g., Whyte, 1957) would be muted. While the MMPI item pool was explicity assembled from statements reflecting psychiatric symptoms, Berg's corollary seems to assert that a miscellaneous collection of statements (or nonverbal stimuli) would have produced an equally effective inventory.

Berg's statements of his deviation hypothesis are ambiguous enough that a number of interpretations are logically possible (Norman, 1963a; Sechrest & Jackson, 1963). One interpretation, for instance, is that the *type* of stimulus is "unimportant," that scales could be constructed equally well using nonverbal or verbal items, though within any type of stimuli only a subset of "content specific" items are valid. Recently, however, data illustrating the relative invalidity of nonverbal items have been presented by Norman (1963a), who found that scales empirically developed from the Welsh Figure Preference Test had virtually no cross-validity for predicting peer ratings of personality characteristics, while scales constructed from adjectives that had at least moderate face validity had statistically significant cross-validities for the same criteria.

Another interpretation of Berg's corollary would posit that our present knowledge of the underlying relationships between item content and item validity is so sketchy that one type of content can substitute equally well for another type as an *initial* item pool from which scales may be developed. That is, Berg may simply be questioning the relationship between the "face validity" of an item and its empirical validity. While no psychologist would argue that this relationship is perfect, Berg seems to have implied that there is no relationship at all!

In a recent study, Duff (1965) investigated the relationship between the face validity and the empirical validity of the items in the *Hy*, *Pd*, and *Sc* scales of the MMPI. As an index of empirical validity, Duff used the item's discriminating efficiency in separating psychiatric patients from normal controls; the pooled judgments of 58 advanced graduate students in psychology provided an index of face validity. Duff found that the correlations between face validity and empirical validity were positive and statistically significant in all three item pools ($Hy = .48$; $Pd = .38$; $Sc = .22$). McCall (1958) carried out a similar study of the items in the *D* scale of the MMPI. Using a group of 41 depressive patients and a matched group of nondepressive psychotic patients, McCall found that 26 *D* scale items previously classified as "face valid" were considerably more discriminating than 22 items classified as "congruent" and that items from both of these sets were significantly more valid than 12 items previously classified as "irrelevant." Both Duff's (1965) and McCall's (1958) studies complement an earlier study by Brozek and Erickson (1948), who investigated the effects of experimental semistarvation upon responses to items in the *Hs*, *D*, and *Hy* scales of the MMPI. Brozek and Erickson found that items classified as "subtle" tended to show less response change under the ex-

perimental conditions than did all other items.[3]

The results of these three studies, when considered in connection with the findings of Norman (1963a) and the relative success of rational scales (and self-ratings) in the Hase and Goldberg (1967) study, suggest the following "revision" of Berg's corollary: The greater the face validity of the items included in an initial item pool, the smaller will that pool have to be in order to provide the stimuli for the development of scales with some fixed level of external validity. Conversely, the more "subtle" the items included in an initial item pool, the larger must that pool be. If the "efficiency" of an initial item pool is defined as the proportion of its items which will add to the validity of a scale constructed from it, then face valid item pools should be more efficient than all pools of similar size made up of items lacking face validity.

The present study was designed as a direct test of this proposition. The experimental design permits an evaluation of the relative importance of various kinds of item content, verbal and nonverbal, for the prediction of nonpathological criteria. Thus, this study provides an empirical test of two competing viewpoints: Berg's corollary asserting the unimportance of item content versus the present suggestion relating face validity to predictive efficiency.

The need to understand more clearly the nature of the relationship between face validity and actual validity is a vital one for at least two reasons: (*a*) Theoretically, such knowledge is critical for the establishment of a "rational psychometrics" (e.g., Loevinger, 1957); (*b*) practically, such knowledge is of great importance for test constructors in the choice of initial item pools. For if item content is unimportant (in either of the senses of Berg's corollary), then subtle items (such as those in the PRT) would typically be preferable, since such items would obviate the need for concern about response dissimilation and/or image enhancement (e.g., Edwards, 1957).

[3] For additional studies of the concept of face validity, see Fricke, 1957; Gough, 1954; Kimber, 1947; Mehlman and Rand, 1960; Seeman, 1952, 1953; Stone, 1964, 1965; and Wiener, 1948.

Method

Subjects

From a freshman dormitory at the University of Oregon, 173 coeds volunteered as paid participants for this study. The mean age of the *S*s was 18.0 years, with a standard deviation of .4 years.

Criteria

Two broad classes of criteria were used in this study, academic performance in college and social affiliation, each measured by a number of criterion indexes. For a more detailed description of the criteria, see Hase and Goldberg (1967).

Grade-point average (GPA). The *S*s were divided on the basis of first semester GPA into high-GPA ($n = 88$) and low-GPA ($n = 79$) criterion groups. GPAs were not available for six *S*s.

Achievement (ACH). A prediction equation using high school GPA and aptitude test scores is used to predict academic achievement at the University of Oregon. The *S*s in the present study were divided into criterion groups of high achievers ($n = 83$), defined as those with high GPA relative to predicted GPA, and low achievers ($n = 70$), defined as those with low GPA relative to predicted GPA. GPAs and/or predicted GPAs were not available for 20 *S*s.

Sorority joining (SOR). The first criterion of social affiliation contrasted sorority girls with "independents." The criterion groups consisted of 68 coeds who belonged to or were pledging a social sorority and 72 coeds who had not joined a sorority and had indicated that they did not intend to join one. Uncertain as to whether they would join a sorority, 33 *S*s were not classified on this variable.

Yielding (YLD). A criterion of social conformity was obtained from shifts in responses to a double administration of a 45-item Opinion Questionnaire (Hastorf & Piper, 1951) previously used for the same purpose by Jackson (1964). The *S*s were asked to indicate the amount of their agreement or disagreement with each statement on a nine-point scale. For the second administration of the Opinion Questionnaire, each of the 45 questions was followed by a number described as the average response given by the *S*s on the first administration. For the 25 questions with the smallest dispersions of group ratings on the first administration, the reported mean value was obtained by shifting the actual mean three points toward whichever end of the scale was most distant. The mean values of the 20 questions with the largest dispersions were reported accurately. During the second administration, 5 weeks after the first, the *S*s filled out the questionnaire using the same rating scale used previously. The criterion index of yielding was computed by averaging (across the 25 items for which spurious mean values were reported) the discrepancy between the *S*'s average

distance from the reported mean at Administration 1 and her average distance at Administration 2. The reliability of this index was .79 (a split-half reliability estimate corrected by the Spearman-Brown formula). A more complete description of this yielding index is given in Goldberg and Rorer (1966). By dichotomizing *Ss* on the yielding index, criterion groups of 78 "yielding" and 82 "nonyielding" coeds were obtained. Yielding scores were not available for 13 *Ss*.

Sociometric status. Since all *Ss* in this experiment were coeds living in the same dormitory, it was possible to get sociometric ratings of each girl from her peers. Each *S* rated all of the coeds on her floor (approximately 15 girls), whether they were in the study or not; consequently, for each *S* the mean rating of 8–12 close associates could be obtained. Each girl rated her peers and herself on six personality traits, two of which (sociability and dominance) were relevant to the criterion of social affiliation. Ratings were made on a five-point scale, with instructions to assign one-third of the targets to Categories 1 and 2, one-third to the middle category, and one-third to Categories 4 and 5. The raters were given detailed descriptions of each personality trait and were instructed to rate all the girls on one trait before proceeding to the next. The mean ratings of each of the 173 *Ss* on sociability (SOC) and dominance (DOM) were dichotomized at the median for each trait in order to produce approximately equal-sized criterion groups of (*a*) sociable versus retiring and (*b*) dominant versus submissive coeds.

Predictor Items

All *Ss* were administered the PRT and a 180-item inventory, the Statement Reaction Test (SRT). To insure comparability of the formats of the PRT and SRT, *Ss* were instructed to respond to each item in the SRT using one of the following options: "agree much," "agree slightly," "disagree slightly," "disagree much." The SRT included three sets of 60 items, each set presumably tapping a different area of verbal content. Items from each of the following three content areas were randomly arranged in the SRT.

Achievement. Sixty items, which on rational grounds reflected content relating to college achievement, were included. Some achievement items were selected from the alternatives scored on the n Achievement scale of the Edwards Personal Preference Schedule (EPPS), and others were rewritten from the Oregon Instructional Preference Inventory (Goldberg, 1963; Shiman, 1966). Additional items were written especially for this study. Some examples of achievement items are "I expect to get very good grades in college," "I work much harder in courses that I like," "I do most of my schoolwork just before it is due."

Affiliation. A second group of 60 items was selected to tap the broad dimension of social affiliation. This item pool included the alternatives scored in the n Affiliation, n Abasement, n Autonomy, and n Dominance scales from the EPPS. Some examples of affiliation items are "I like to be loyal to my friends," "I like to be self-sufficient," "In matters of conduct, I conform to custom."

Content-irrelevant verbal items. The remaining 60 SRT items were chosen at random from the Kuder Preference Record. Kuder items were chosen as examples of content-irrelevant verbal items, to be contrasted with the 60 PRT items characterized as content-irrelevant nonverbal items. Consequently, this experimental design makes possible a comparison between content-relevant and content-irrelevant items, and among the latter between verbal and nonverbal items.

Examination of the items from the SRT indicated that some of the items in the achievement pool could also have relevance for the affiliation criteria (e.g., "I would rather be a good student than have an active social life") and vice versa. Similarly, some Kuder items could be construed as having subtle affiliation or achievement themes. Therefore, an evaluation of each of the SRT items was carried out in order to allow for the possibility that an item might be relevant to more than one criterion or that a Kuder item might have high face validity. From an undergraduate course in general psychology, 29 male and 27 female judges were each given a list of the 180 SRT items and the six criterion variables, with detailed descriptions of the latter. Each judge was then asked to indicate for each item and each criterion variable whether he thought the two criterion subgroups would have responded differently to that item; in this way he categorized each item as either a valid or a nonvalid predictor for each of the six criteria. An index of face validity was obtained for each item and each criterion by calculating the proportion of judges who thought the item would be a valid predictor for that criterion. Since the correlations between the face validity scores calculated from the males and the scores calculated from the females averaged .92 across the six criteria, the male and female judgments were combined into a single index. The 180 SRT items were divided, for each criterion, into three pools containing, with respect to face validity, the 60 highest, 60 middle, and 60 lowest items.

Analyses

The two contrasting groups for each criterion were split into random halves for purposes of a double cross-validation study. Empirical scoring keys were developed separately from each of the four 60-item subsets (high, middle, and low face validity and PRT) to predict each of the six criteria. Keys derived from Sample A were cross-validated on Sample B, and keys derived from Sample B were cross-validated on Sample A.

The method used to key items was identical to the one recommended by Berg (1961, pp. 342–349) for evaluating the validity of the PRT and the deviation hypothesis. Each of the 60 items in a subset had four response options, and the responses were tabulated separately for each item and each option. Item options that discriminated between a

TABLE 1
Percentage of SRT Items from Each Content Pool in Each of the Three Face Validity Categories

Criterion	Category	Initial content pool		
		Achievement[a]	Affiliation[a]	Kuder[a]
GPA	Highest 60	85	0	15
	Middle 60	15	38	47
	Lowest 60	0	62	38
ACH	Highest 60	80	10	10
	Middle 60	18	32	50
	Lowest 60	2	58	40
SOR	Highest 60	22	68	10
	Middle 60	45	25	30
	Lowest 60	33	7	60
YLD	Highest 60	22	75	3
	Middle 60	46	20	34
	Lowest 60	32	5	63
SOC	Highest 60	10	62	28
	Middle 60	18	38	44
	Lowest 60	72	0	28
DOM	Highest 60	28	60	12
	Middle 60	30	28	42
	Lowest 60	42	12	46
College performance[b]	Highest 60	83	5	12
	Middle 60	16	35	49
	Lowest 60	1	60	39
Social affiliation[b]	Highest 60	21	66	13
	Middle 60	34	28	38
	Lowest 60	45	6	49

[a] $n = 60$.
[b] Average.

pair of criterion groups at a prescribed level of statistical significance were selected for the key for that particular criterion. Empirical keys were first developed by selecting options that discriminated at the .10 level of significance or better; a second set of keys employed options discriminating at the .20 level or better. Since the stringency of the significance level had no important effect upon the size of the validity coefficients, the results to be reported below are based upon the average cross-validities for both significance levels.

Results

Table 1 presents the percentage of items from each of the three 60-item SRT content subpools falling into each of the three face validity categories for each of the six criteria. Note that, in general, most of the achievement items had relatively high face validity for the GPA and achievement criteria, while few of the affiliation items fell in the high face validity categories for these two criteria. Conversely, most of the affiliation items had relatively high face validity for the four affiliation criteria, while the achievement items tended to fall in the low face validity categories for these criteria. As expected, most Kuder items fell in the middle or lowest categories of face validity for all six criteria. Interestingly, Kuder items were perceived as having their highest face validity for the sociability (SOC) criterion and their lowest face validity for the yielding (YLD) criterion.

The average cross-validated point-biserial correlations for the three face validity pools of the SRT and the average cross-validities for the PRT are shown in Table 2. For five of the six criterion variables, only scales built from items of the highest

TABLE 2
Average Validity Coefficients for Scales Constructed from Three SRT Face Validity Categories and the PRT

Item pool	Criteria						*M*
	College performance		Social affiliation				
	GPA	ACH	SOR	YLD	SOC	DOM	
High FV SRT[a]	.29 (.58)	.14 (.66)	.35 (.68)	.11 (.62)	.35 (.61)	.18 (.56)	.24 (.62)
Med. FV SRT[a]	−.12 (.57)	−.13 (.61)	.06 (.65)	.14 (.70)	.08 (.58)	.03 (.54)	.01 (.61)
Low FV SRT[a]	.12 (.63)	−.05 (.62)	.12 (.60)	.15 (.57)	.05 (.65)	.02 (.45)	.07 (.59)
PRT[a]	.04 (.62)	.04 (.62)	.08 (.62)	−.04 (.58)	.07 (.59)	−.04 (.47)	.02 (.56)

Note.—The average cross-validated coefficients are based on the mean of the two cross-validation samples; average coefficients from the derivation samples are in parentheses. Abbreviated: FV = face validity, SRT = Statement Reaction Test, PRT = Perceptual Reaction Test.
[a] $n = 60$.

face validity had significant cross-validity, although all scales had almost equally high correlations with the criteria in the derivation samples. The only criterion not best predicted by the most face valid items was the YLD index, which was predicted about equally poorly by all three categories of verbal items and not predicted at all by the PRT items. It is important to note that the average derivation correlations shown in parentheses in Table 2 were quite high and did not differ across item pools. Had this study used Berg's method of keying responses, *without* benefit of cross-validation, the results would have mistakenly implied that the PRT was a good predictor of the six criterion indexes and that item content was unimportant.

The results presented above have focused on the relationship between face validity and predictive validity for scales composed of sets of items. In order to understand more fully the nature of this relationship, additional analyses were carried out at the level of the individual item. The validity of each of the 180 SRT items against each of the six criteria was assessed. All validity coefficients were based on the 2 × 4 table (two criterion groups by four response options) for all subjects. A high validity coefficient indicated that the distribution of responses to that item differed between the two criterion groups.

Six indexes of item validity were computed for each of the 180 SRT items: χ^2, $\sqrt{\chi^2}$, ϕ', C, H, and λ (see Hays, 1963, pp. 578–614). The six different validity coefficients turned out to correlate almost perfectly with one another (e.g., the intercorrelations among the first five ranged from .93 to 1.00, median = .96). Since the correlations of the six indexes with other variables were all virtually identical, the results for only one of them (χ^2) are reported here.

Table 3 presents the correlations across items between validity coefficients and face validity scores. Note that the correlations are, in general, positive and that they are rather high within a few of the item subpools that were presumably more homogeneous with respect to content. Inspection of the scatter plots which are summarized by these correlations revealed that items of low face validity generally had low validity coefficients, while items of high face validity had validities that were distributed over the entire range of the distribution (e.g., some presumably relevant items actually were valid discriminators, while others were not).

TABLE 3
Correlations between Face Validity Scores and Item Validity Coefficients

Criterion	Achievement pool[a]	Affiliation pool[a]	Kuder pool[a]	Total SRT[b]
GPA	.49**	.03	.07	.24**
ACH	.04	.23*	.28*	.19**
SOR	.53**	−.02	.15	.13*
YLD	−.06	−.20	.16	.06
SOC	.38**	.26*	.01	.26**
DOM	−.02	.49**	−.20	.24**
M	.23**	.13*	.08	.19**

[a] $n = 60$.
[b] $n = 180$.
* $p < .05$, one-tailed test.
** $p < .01$, one-tailed test.

Examination of items that received high face validity scores for a particular criterion revealed subtle content differences between items which may have mediated their relative discriminating power. This is illustrated by the pairs of items shown in Table 4. The items within each pair are matched with respect to face validity for one of the six criteria, but for each of these matched pairs, one item was a valid discriminator while the other was not. Items were paired so as to be as closely related as possible in terms of the underlying behavior to which they referred. Thus, pair a_1 and a_2 both refer to the importance of achievement in the classroom, though they differ in that one item emphasizes examination performance, while the other is less specific. On the other hand, pair b_1 and b_2 both deal with the effect of achievement upon one's state of mind, though they also differ in a subtle way.

One possible explanation for the differences in validity among items of equally high face validity is that the judges did not know the response variance of each

item and thus could not take this parameter into account when making their ratings. Thus, in Table 4, items e_1 and e_2 both had high face validity for the criterion of sociability, presumably because both dealt with friendship. However, virtually all subjects strongly agreed to item e_2, whereas there was a relatively uniform distribution of responses to item e_1. To evaluate the possibility that item variance might moderate the relationship between face validity and empirical validity coefficients, an index of response variance was computed for each item. The squared discrepancy between the frequency of responses falling in each response category and $N/4$ (the frequency that would have been obtained if all responses were equally divided among the four possible categories) was summed across all four response categories. Scores on this index were then reflected. Therefore, the highest response variance scores were elicited from items with a uniform response distribution (e.g., .25, .25, .25, .25) and the lowest response variance scores were elicited from items to which all responses were given to the same response option (e.g., 1.00, .00, .00, .00).[4]

TABLE 4
Matched Items of High Face Validity Differing Markedly in Actual Validity

Item	Criterion	Validity	
		Face[a]	Actual
a_1 It is important for me to be among the best in the classroom.	GPA	84	High
a_2 It is important for me to do well on exams.	GPA	84	Low
b_1 I enjoy relaxing only after completion of work well done.	GPA	55	High
b_2 I feel that my future peace depends upon my accomplishment.	GPA	62	Low
c_1 I would rather be on social probation than on academic probation.	GPA	75	High
c_2 I would rather have fun than be an outstanding student in college.	GPA	68	Low
d_1 I like doing things my way, disregarding what others think.	YLD	39	High
d_2 It is important for me to feel free to do and say what I want.	YLD	39	Low
e_1 I like to make as many friends as I can.	SOC	95	High
e_2 It is important for me to have close friendships.	SOC	79	Low

[a] Percentages.

TABLE 5
Correlations among Actual Validity, Face Validity, and Response Variance for 180 SRT Items

Correlation	Criterion						*M*
	GPA	ACH	SOR	YLD	SOC	DOM	
AV vs. Va	.11	.16*	.18**	.18**	.19**	.07	.15*
FV vs. Va	.12	.10	−.05	.03	−.03	.11	.05
AV vs. FV	.24**	.19**	.13*	.06	.26**	.24**	.19**
AV vs. FV[a]	.24**	.19**	.13*	.06	.26**	.24**	.19**

Note.—Abbreviated: AV = actual validity, FV = face validity, Va = variance.

[a] Partial correlation; variance partialed out.

* $p < .05$, one-tailed test.

** $p < .01$, one-tailed test.

Examination of the variance indexes for the items in Table 4 revealed that for every pair (e.g., a_1 and a_2) the item with the higher validity also had the more uniform distribution of responses. This seems to document the effect of subtle differences in content upon item variance, an effect which may have produced, at least in part, the substantial differences in item validity. However, when item variance was correlated with item validity on the one hand and with face validity on the other, the resulting correlations (reported in Table 5) were so low that partialing out the effect of item variance did not change the correlations between face validity and empirical validity.

To provide another view of the effect of item response variance as a moderator of the relationship between face validity and

[4] The computing formula used to calculate item response variance was

$$\sigma^2 = -\sum_{i=1}^{4} [n_i - (N/4)]^2$$

where σ^2 = item response variance, N = the total number of subjects, and n_i = the frequency of cases falling in the ith response category ($\sum_{i=1}^{4} \times n_i = N$). See Walker and Lev (1953, p. 28).

predictive validity, the 180 SRT items were ranked on response variance; six variance-homogeneous subpools of 30 SRT items each were formed, and the correlations between face validity and empirical validity were computed within each subpool. The results are presented in Table 6. Note that while the 30 items of highest response variance (the top row in Table 6) had the highest correlations ($\bar{r} = .32$) and the 30 items of lowest response variance (the bottom row in Table 6) had the lowest correlations ($\bar{r} = .06$), the moderating effect of response variance was neither strong nor linear. For the trait of dominance, in fact, the highest correlation ($r = .35$) occurred within a subpool of items with a rather unimodal response distribution (i.e., low response variance). In general, the results appear to indicate that item response variance is, at best, a weak moderator of the relationship between face validity and predictive validity for items of this type.

TABLE 6
Correlations between Actual Validity and Face Validity for Six SRT Subpools of Differing Response Variance

Variance items	GPA	ACH	SOR	YLD	SOC	DOM	*M*
Highest	.44*	.26	.49*	−.01	.51*	.21	.32*
High	.03	.37*	.21	−.12	.07	.27	.14
High–medium	.26	.07	.19	−.10	.27	.34	.17
Low–medium	.52*	.45*	.06	−.06	.39*	.16	.25
Low	.09	−.20	.29	.34	.38*	.35	.21
Lowest	.13	.30	−.25	.19	−.05	.05	.06

Note.—All n's = 30.
* $p < .05$.

Discussion

The findings from the present study are highly congruent with those of Norman (1963a), who also used nonpathological personality characteristics as criteria. Apparently the PRT can best discriminate among such dissimilar groups as psychiatric patients and normals, though it has not been established that the PRT can make even such gross kinds of discriminations more accurately or efficiently than other inventories. Evidence that the PRT can validly predict more subtle individual differences has yet to be presented.

In the history of structured inventory measurement, some assumptions about the relationships between face and empirical validity have been implicit in each new strategy of scale construction that has been proposed (Hase & Goldberg, 1967). The earliest personality inventories (composed of rationally developed scales) were predicated on the assumption that face validity and empirical validity were nearly isomorphic—that self-statements reflect honest appraisals of actual behavioral tendencies (Buchwald, 1961). Over the years, this assumption began to be questioned, at first by critics proposing to substitute projective approaches to personality assessment (e.g., Frank, 1939) and soon by the proponents of structured inventories themselves. Modifications of the isomorphic assumption gave birth to two new strategies of scale construction: the empirical group-discriminative (or "criterion group") strategy (e.g., Meehl, 1945) and the variants of internal consistency (homogeneity) maximization strategies, of which the factor analytic strategy is typical (e.g., Cattell, 1946; Eysenck, 1947). Proponents of the empirical group-discriminative strategy argued that personality theory had not reached the stage where face validity could be expected to mirror actual validity (even to the most sophisticated of test constructors) and, therefore, that only the empirically determined effectiveness of each item should legitimately influence the decision whether an item belonged in a scale. Moreover, if criterion groups of subjects who fell at the polar extremes of a trait could be located, then the empirically determined differential response frequency of the two groups to each item could provide an external (nonsubjective) index of item validity. Berg's statements of his deviation hypothesis and its corollary represent an extreme position stemming from this stream of thought.

Proponents of the factor analytic strategy of scale construction appear to have been using a more complex correspondence

assumption than either rational or group-discriminative strategists. While, on the one hand, factor analysts may have used relatively nonsubjective criteria for determining which items should be included in a factor scale ("items that hang together belong together"), nonetheless, they often invoke a near isomorphic correspondence assumption for scale labeling purposes. Even Cattell, who has proposed a factor naming system of "universal index numbers" (so as to leave to an unprejudiced posterity the task of determining the "meaning" of each factor uncovered), is not immune to peeking at item content to propose a first approximation of the trait purportedly measured by the factor scale.

Neither the empiricists nor the factor analysts have spelled out very clearly their expectations regarding the relationship between face validity and actual validity. For example, the MMPI item pool was initially assembled explicitly with "content" considerations in mind. Hathaway and McKinley (1940) state:

> The individual items were formulated partly on the basis of previous clinical experience. Mainly, however, the items were supplied from several psychiatric examination direction forms, from various textbooks of psychiatry, from certain of the directions for case taking in medicine and neurology, and from the earlier published scales of personal and social attitudes [p. 249].

Implicit in this statement is the assumption that such items were more likely to differentiate psychiatric patients from normals than were miscellaneous assortments of verbal items (or nonverbal PRT-type items). A similar implicit assumption seems to have guided Gough in writing the initial items for the CPI, as well as Cattell in initially selecting items to begin the factorial development of the Sixteen Personality Factor Questionnaire. The results of the present study indicate that these choices were probably wise ones.

However, for many kinds of personality scales (e.g., those to be used in selection situations) it may be necessary to begin scale development using less efficient item pools in order to ultimately develop more valid scales (see Campbell, 1950; Loevinger, 1955). That is, the use of face valid item pools could serve to purchase validity (among relatively honest subjects) at the price of increasing the transparency (and the fakability) of the resulting scale. While Berg's corollary aims investigators towards nonverbal questionnaire items of the sort found in the PRT and the Welsh Figure Preference Test, the results of both the present investigation and that of Norman (1963a) indicate the folly of following this road. Consequently, future personality scale developers must either develop more sensitive procedures for handling individual differences in impression management (e.g., Norman, 1963b) or they must find items which are valid under relatively diverse conditions (see Fiske & Butler, 1963). For this latter purpose, the search could lead away from subjective questionnaire items of all sorts and, as Cattell has repeatedly suggested, towards the development of "objective" or "maximum performance" tests of personality traits.

REFERENCES

Adams, H. E. Statistical rigidity in schizophrenic and normal groups measured with auditory and visual stimuli. *Psychological Reports,* 1960, **7,** 119–122.

Barnes, E. H. The relationship of biased test responses to psychopathology. *Journal of Abnormal and Social Psychology,* 1955, **51,** 286–290.

Berg, I. A. Response bias and personality: The deviation hypothesis. *Journal of Psychology,* 1955, **40,** 60–71.

Berg, I. A. Deviant responses and deviant people: The formulation of the deviation hypothesis. *Journal of Counseling Psychology,* 1957, **4,** 154–161.

Berg, I. A. The unimportance of test item content. In B. M. Bass & I. A. Berg (Eds.), *Objective approaches to personality assessment.* New York: Van Nostrand, 1959. Pp. 83–99.

Berg, I. A. Measuring deviant behavior by means of deviant response sets. In I. A. Berg & B. M. Bass (Eds.), *Conformity and deviation.* New York: Harper, 1961. Pp. 328–379.

Boozer, D. G. Response sets as indicators of senescence and of psychopathology in old age. Unpublished doctoral dissertation, Louisiana State University, 1961.

Brozek, J., & Erickson, N. K. Item analysis of the psychoneurotic scales of the Minnesota Multiphasic Personality Inventory in experimental semistarvation. *Journal of Consulting Psychology,* 1948, **12,** 403–411.

Buchwald, A. M. Verbal utterances as data. In

H. Feigl & G. Maxwell (Eds.), *Current issues in the philosophy of science: Symposium of scientists and philosophers.* (A.A.A.S. Section on History and Philosophy of Science; Proceedings of Section L, 1959) New York: Holt, Rinehart & Winston, 1961. Pp. 461–472.

CAMPBELL, D. T. The indirect assessment of social attitudes. *Psychological Bulletin,* 1950, **47,** 15–38.

CATTELL, R. B. *The description and measurement of personality.* New York: World Book, 1946.

CIEUTAT, L. G. Deviant responses as a function of mental deficiency. Unpublished doctoral dissertation, Louisiana State University, 1960.

DUFF, F. L. Item subtlety in personality inventory scales. *Journal of Consulting Psychology,* 1965, **29,** 565–570.

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

ENGEN, E. P. Response set of pulmonary tuberculosis patients. Unpublished doctoral dissertation, Louisiana State University, 1959.

EYSENCK, H. J. *Dimensions of personality.* London: Kegan Paul, 1947.

FISKE, D. W., & BUTLER, J. M. The experimental conditions for measuring individual differences. *Educational and Psychological Measurement,* 1963, **23,** 249–266.

FRANK, L. K. Projective methods for the study of personality. *Journal of Psychology,* 1939, **8,** 389–413.

FRICKE, B. G. Subtle and obvious test items and response set. *Journal of Consulting Psychology,* 1957, **21,** 250–252.

GOLDBERG, L. R. Test-retest item statistics for the Oregon Instructional Preference Inventory. *Oregon Research Institute Research Monograph,* 1963, **3**(Whole No. 4).

GOLDBERG, L. R., & RORER, L. G. The use of two different response modes and repeated testings to predict social conformity. *Journal of Personality and Social Psychology,* 1966, **3,** 28–37.

GOUGH, H. G. Some common misconceptions about neuroticism. *Journal of Consulting Psychology,* 1954, **18,** 287–292.

HARRIS, J. L. Deviant response frequency in relation to severity of schizophrenic reaction. Unpublished master's thesis, Louisiana State University, 1958.

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

HASTORF, A. H., & PIPER, G. W. A note on the effect of explicit instructions on prestige suggestion. *Journal of Social Psychology,* 1951, **33,** 289–293.

HATHAWAY, S. R., & MCKINLEY, J. C. A multiphasic personality schedule (Minnesota): I. Construction of the schedule. *Journal of Psychology,* 1940, **10,** 249–254.

HAWKINS, W. A. Deviant responses, response variability, and paired associate learning. Unpublished doctoral dissertation, Louisiana State University, 1960.

HAYS, W. L. *Statistics for psychologists.* New York: Holt, Rinehart & Winston, 1963.

HESTERLY, S. O. Deviant response patterns as a function of chronological age. Unpublished doctoral dissertation, Louisiana State University, 1960.

HESTERLY, S. O., & BERG, I. A. Deviant responses as indicators of immaturity and schizophrenia. *Journal of Consulting Psychology,* 1958, **22,** 389–393.

HOUSE, C. W. Response bias as a measure of emotional disturbance in children. Unpublished doctoral dissertation, Louisiana State University, 1960.

JACKSON, D. N. Desirability judgments as a method of personality assessment. *Educational and Psychological Measurement,* 1964, **24,** 223–238.

KIMBER, J. A. M. The insight of college students into the items on a personality test. *Educational and Psychological Measurement,* 1947, **7,** 411–420.

LOEVINGER, J. Some principles of personality measurement. *Educational and Psychological Measurement,* 1955, **15,** 3–17.

LOEVINGER, J. Objective tests as instruments of psychological theory. *Psychological Reports,* 1957, **3**(Monogr. Suppl. 9).

MCCALL, R. J. Face validity in the *D* scale of the MMPI. *Journal of Clinical Psychology,* 1958, **15,** 77–80.

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

MEHLMAN, B., & RAND, M. E. Face validity of the MMPI. *Journal of General Psychology,* 1960, **63,** 171–178.

NORMAN, W. T. Relative importance of test item content. *Journal of Consulting Psychology,* 1963, **27,** 166–174. (a)

NORMAN, W. T. Personality measurement, faking, and detection: An assessment method for use in personnel selection. *Journal of Applied Psychology,* 1963, **47,** 225–241. (b)

ROITZSCH, J. C., & BERG, I. A. Deviant responses as indicators of immaturity and neuroticism. *Journal of Clinical Psychology,* 1959, **15,** 417–419.

SECHREST, L. B., & JACKSON, D. N. Deviant response tendencies: Their measurement and interpretation. *Educational and Psychological Measurement,* 1963, **23,** 33–53.

SEEMAN, W. "Subtlety" in structured personality tests. *Journal of Consulting Psychology,* 1952, **16,** 278–283.

SHIMAN, E. S. The comparative validity of different strategies of predicting college achievement. Unpublished master's thesis, University of Oregon, 1966.

STONE, L. A. Subtle and obvious response on the MMPI. *Psychological Reports,* 1964, **15,** 721–722.

STONE, L. A. Subtle and obvious response on the MMPI as a function of acquiescence response style. *Psychological Reports*, 1965, **16**, 803–804.

WALKER, H. M., & LEV, J. *Statistical inference*. New York: Holt, Rinehart & Winston, 1953.

WHYTE, W. H., JR. *The organization man*. New York: Doubleday, 1957.

WIENER, D. N. Subtle and obvious keys for the MMPI. *Journal of Consulting Psychology*, 1948, **12**, 164–170.

(Received December 27, 1966)