The Reliability of Reliability: The Generality and Correlates of Intra-Individual Consistency in Responses to Structured Personality Inventories

Lewis R. Goldberg
University of Oregon and the
Institute for the Measurement of Personality

When a personality inventory is administered to an individual on two occasions (under similar instructions and testing conditions), some items may not elicit the same response on both occasions; and the individual's scale scores may differ between the two administrations. Traditionally, such intra-individual score variability has been averaged across a sample of individuals to furnish one type of estimate of the reliability of the scale. Analogously, the intra-individual variability in the responses to a single item has been averaged across a sample of individuals in order to provide an estimate of the stability (and, indirectly perhaps, the ambiguity) of the item. Personality scales (and items) have traditionally been compared on the basis of these reliability estimates, unreliability being equated with measurement error.

Such standard psychometric procedures for error estimation rest on the implicit assumption that intra-individual response variability to the same set of items on two occasions arises primarily from characteristics of the items themselves, and not from any particular attributes of the sample of individuals responding to them. That is, the data from subjects who manifest a large amount of variability are combined with that from more consistent responders to obtain an average reliability estimate. This estimate is then ascribed to the scale or the item. Traditionally, Scale *A*'s "reliability," derived from one sample of subjects, has been compared with a corresponding estimate for Scale *B*, derived from another sample; conclusions have been drawn relative to the differential "reliability" of the two scales. Such a procedure might be justified if either (1) there are no differences between individuals in their *general* likelihood of eliciting consistent vs. inconsistent responses or (2) the two different samples were matched, either explicitly or by random sampling.

The aim of the present investigation was to discover whether hypothesis (1) is reasonable or, on the other hand, whether intra-individual consistency in responses to structured personality inventories could be usefully conceptualized as a personality trait (i.e., as a reliable and general response disposition). The answer to such a question, about *any* possible dimension of individual differences, involves the sequential analysis of a series of questions:

1. Are measures of the possible trait reliable?
   a. How homogeneous are the measures?
   b. How stable are the measures over time?
2. Do measures of the trait appear to possess construct validity?

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 2, No. 2 Spring 1978 pp. 269-291

   a. Do they show convergent validity across diverse methods?
   b. Do they show divergent validity from other dimensions?
3. What are the psychometric correlates of these measures?
4. Is knowledge of an individual's status on the trait useful in prediction to non-test criteria?
   a. Do these measures function as traditional predictor or suppressor variables?
   b. Do these measures function as moderator variables?

The remainder of this report will be devoted to answering these questions for the possible trait of intra-individual response consistency, providing a summary of a much more extensive and detailed exploration of these questions (Goldberg & Jones, 1969).

Procedure

Experimental Design

Three widely employed personality inventories—the California Psychological Inventory (CPI), the Edwards Personal Preference

Table 1
Experimental Design

Order of Inventory Administrations

Week	Inventory	Administration
1	CPI	I
2	EPPS	I
3	MMPI	I
4	-	-
5	CPI	II
6	EPPS	II
7	MMPI	II
8-10	Other Inventories (single administration)	

Number of Identical Items in Each of the Test-Retest Administration Pairs

	Complete Inventories		225-Item Inventories		
	CPI-I	MMPI-I	EPPS-I	CPI_r-I*	$MMPI_r$-I*
CPI-II	480[a]	167	0	225	0
MMPI-II	167	566[b]	0	0	225
EPPS-II	0	0	225[c]	0	0
CPI_r-II*	225	0	0	225	0
$MMPI_r$-II*	0	225	0	0	225

* Inventories reduced to 225 items, first by the elimination of all duplicated and overlapping items, then by the elimination of other items on a random basis.

a Includes 12 items duplicated within each administration.
b Includes 16 items duplicated within each administration.
c Includes 15 items duplicated within each administration.

Schedule (EPPS), and the Minnesota Multiphasic Personality Inventory (MMPI)—were each administered on two separate occasions. Of these three inventories, two are designed to assess traits within the normal range of personality functioning (CPI and EPPS), and one is designed to measure aspects of psychopathology (MMPI). Two of the inventories employ a true-false response format (CPI and MMPI), while the third employs a forced-choice format (EPPS). Consequently, consistency measures from inventories differing in both content and response format can be compared.

The order of inventory administrations is presented in Table 1. All inventories were administered in class, one each week, to 95 male and 108 female University of Oregon undergraduate students enrolled in an introductory psychology course. As Table 1 indicates, the two administrations of the same inventory were separated by four weeks, during which time the two other inventories were administered. Therefore, it is unlikely that the subjects could recall their earlier responses when they were responding on a second occasion. Moreover, the instructions emphasized that the students should respond to each item "as you feel today, regardless of how you may have answered in the past." The students were told that the testing procedures were an integral part of the course and that the inventories had been chosen as examples of psychometric instruments which they would later study. Feedback on their test performance was promised, and the students were excused from all other experimental requirements typically associated with such courses. An evaluation questionnaire administered at the end of the course indicated that satisfaction with the testing procedures (and the course as a whole) was quite high.

Measures of Response Variability vs. Consistency

Fifteen different item subsets were used to obtain measures of response variability vs. consistency. For each item set, a subject's response consistency score was calculated as the percentage of items in the set to which he/she responded in a consistent fashion on both administrations[1]. The 15 sets of items can be logically grouped into four categories (see Table 1).

Complete item pools. The percentages of unchanged responses among the 480 CPI items (CPI-I vs. CPI-II); the 225 EPPS items (EPPS-I vs. EPPS-II); and the 566 MMPI items (MMPI-I vs. MMPI-II) constitute the first three consistency measures.

Reduced item pools. Since the three complete item pools differed in length, the CPI and MMPI were each reduced to 225 items to match the EPPS. Items were culled from the CPI and the MMPI by (1) eliminating from the MMPI the 167 items common to the CPI; (2) eliminating from both the CPI and MMPI the second administration of their duplicated items; and (3) randomly eliminating items until each of the reduced CPI and MMPI pools contained 225 items. Thus, the percentage of unchanged responses to the reduced CPI (CPI_r-I vs. CPI_r-II) and the reduced MMPI ($MMPI_r$-I vs. $MMPI_r$-II) can be directly compared with the EPPS (EPPS-I vs. EPPS-II) to assess the convergence of consistency scores for three distinct sets of 225 items, with no item overlap between sets.

Duplicated items. Each of the three complete inventories contained a small set of items duplicated within each testing session. The percentage of unchanged responses was computed for the 12 duplicated CPI items on the first

[1]The answer sheets used by all subjects permitted only a dichotomous response ("True" vs. "False" for the CPI and the MMPI and "A" vs. "B" for the EPPS). Since no intermediate response option was provided and subjects were requested to respond to every item (even if they were in doubt about the appropriate response), well over 99% of all responses could be easily coded into one of the two response alternatives. The extremely rare instances of response omission or double responding were arbitrarily coded "True" (CPI and MMPI) or "A" (EPPS). Consequently, an "inconsistent response" should be understood as reflecting a change from "True" to "False" (or "A" to "B"), or vice versa.

Table 2
Means, Standard Deviations, and Split-half Reliability (Homogeneity) Coefficients for 15 Consistency Scores

Duplicated Items	Admin.	No. of Items	Mean		σ		r_{oe} [a]		r_{tt} [b]		r_{ii} [c]	
			M	F	M	F	M	F	M	F	M	F
EPPS	I	15	.80	.79	.17	.10	.06	.08	.11	.15	.01	.01
CPI	I	12	.89	.90	.11	.10	.34	.18	.51	.30	.05	.04
MMPI	I	16	.94	.95	.08	.06	.43	.00	.60	.00	.09	.00
EPPS	II	15	.83	.80	.11	.11	.21	.13	.35	.23	.04	.02
CPI	II	12	.92	.93	.08	.08	.07	.10	.13	.18	.01	.02
MMPI	II	16	.95	.95	.07	.08	.38	.45	.55	.62	.07	.09
Common Items												
CPI-MMPI	I-I	167	.85	.86	.05	.04	.49	.49	.66	.66	.01	.01
CPI-MMPI	I-II	167	.85	.85	.05	.04	.56	.49	.72	.66	.02	.01
CPI-MMPI	II-I	167	.89	.89	.05	.04	.64	.58	.78	.73	.02	.02
CPI-MMPI	II-II	167	.90	.90	.04	.04	.66	.59	.80	.74	.02	.02
Inventories												
EPPS	I-II	225	.76	.77	.05	.05	.53	.53	.69	.69	.01	.01
CPI_r	I-II	225	.85	.85	.04	.03	.52	.41	.68	.58	.01	.01
$MMPI_r$	I-II	225	.86	.87	.04	.04	.65	.62	.79	.77	.02	.01
CPI	I-II	480	.85	.85	.04	.03	.77	.57	.87	.73	.01	.01
MMPI	I-II	566	.87	.87	.04	.03	.84	.78	.91	.88	.02	.01

Note:--N = 93 males (M) and 108 females (F).
[a] Split-half correlations, odd vs. even items.
[b] r_{oe} corrected by the Spearman-Brown formula.
[c] Estimated reliability of a single item, by reversal of the Spearman-Brown formula.

(CPI_d-I) and the second (CPI_d-II) administration for the 15 EPPS duplicated items ($EPPS_d$-I and $EPPS_d$-II) and the 16 MMPI duplicated items ($MMPI_d$-I and $MMPI_d$-II). Since each inventory was administered on two occasions, there were four administrations of each of these duplicated item sets.

Common items. As Table 1 indicates, there were 167 identical items common to the CPI and the MMPI (see Goldberg & Rust, 1964); and since both inventories were administered on two occasions, there were four administrations of this common item pool. Four inventory-occasion pairings were employed: (1) CPI-I vs. MMPI-I; (2) CPI-I vs. MMPI-II; (3) CPI-II vs. MMPI-I; and (4) CPI-II vs. MMPI-II.

The analyses of these 15 measures will be reported in a sequential order as they bear on the general hypothesis that intra-individual response variability vs. consistency can be conceptualized as a personality trait.

Results

How Homogeneous are Measures of Response Consistency?

Table 2 presents the means, standard deviations, and three types of split-half reliability coefficients for the 15 measures of response consistency. The upper section of the table lists the values for the short (12- to 16-item) duplicated item pools, the middle section for various parings of the 167 CPI-MMPI common items, and the bottom section for the repeated administration of the reduced (225-item) and the complete item pools.

As Table 2 indicates, the findings from the male and female samples were generally quite similar. The average subject in both samples responded consistently to about 80% of the EPPS duplicated items, to about 90% of the CPI duplicated items, and to about 95% of the MMPI duplicated items. The standard deviations of these measures were each about 10%, indicating considerable skewness in the distribution of consistency scores within these short item pools. For the longer item pools, the average subject responded consistently to about 75% of the EPPS items and to between 85% and 90% of the CPI and MMPI items. The standard deviations of these longer measures were about 5% for the EPPS and about 3% to 4% for the CPI and the MMPI.[2]

While the various homogeneity values for the short duplicated item pools were generally low, the comparable values for the longer item pools were quite sizable. The corrected odd-even reliability coefficients based on the 167 CPI-MMPI common items ranged from .66 to .80, those based on the 225-item pools ranged from .58 to .79, and those based on the two complete CPI and MMPI pools from .73 to .91. The average homogeneity values (r_{rr}) for consistency scores from the complete 225-item EPPS pool (.70), the complete 480-item CPI pool (.80), and the complete 566-item MMPI pool (.90) were at least as high as the corresponding values for the regular scales scored from these inventories.

To determine whether or not the homogeneity estimates might have been produced by a few careless responders (whose extreme scores could have spuriously inflated the overall reliability coefficients), all consistency score distributions (and all bivariate scatterplots) were examined visually. For the longer item pools (167 to 566 items), all of the consistency score distributions from the sample of 108 females were smooth, unimodal, and symmetric; all scatterplots appeared to be bivariate normal. The corresponding univariate distributions from the sample of 95 males showed two deviant male subjects; one displayed relatively low consistency on the CPI

[2]As Table 2 indicates, there was a general tendency for responses to be slightly more consistent upon later pairs of test-retest administrations than upon earlier ones. That is, the mean consistency of response to the second administration of the duplicated items was slightly higher than that to the first, and the responses to the second administrations of the common CPI-MMPI items were slightly more consistent than those given to the first administration. For previous evidence of this same effect, see Howard (1964).

Table 3
The Correlations among the 15 Consistency Scores

	Complete Inventories		225-Item Inventories			167 CPI-MMPI Common Items				Duplicated Items: Administration I			Duplicated Items: Administration II		
	CPI	MMPI	EPPS	CPI_r	$MMPI_r$	I-I	I-II	II-I	II-II	CPI	MMPI	EPPS	CPI	MMPI	EPPS
CPI		[.57]	.35	(.93)	.48	[.68]	[.64]	[.62[	[.55]	[.30]	-.10	.09	[.36]	.33	.19
MMPI	[.67]		.47	.56	(.86)	[.68]	[.65]	[.76]	[.78]	.11	[.12]	.13	.17	[.47]	.13
EPPS	.33	.45		.32	.46	.30	.26	.39	.48	.03	.09	[.20]	.11	.32	[.45]
CPI_r	(.87)	.58	.28		.49	.67	.61	.58	.50	.24	-.07	.07	.36	.30	.18
$MMPI_r$	.51	(.87)	.37	.40		.55	.53	.56	.60	.06	.08	.08	.19	.41	.13
I-I	[.77]	[.67]	.37	.65	.53		[.84]	[.68]	(.54)	.12	.01	.08	.18	.18	.02
I-II	[.83]	[.70]	.42	.69	.50	[.80]		(.59)	[.60]	.17	-.02	.12	.12	.25	.07
II-I	[.70]	[.80]	.43	.60	.62	[.64]	(.65)		[.77]	.13	.07	.21	.17	.29	.04
II-II	[.68]	[.76]	.39	.62	.55	(.57)	[.71]	[.74]		.16	-.03	.21	.20	.40	.20
CPI-I	[.34]	.35	.06	.29	.24	.30	.28	.27	.26		-.07	-.05	(.21)	.02	.00
MMPI-I	.37	[.43]	.15	.30	.32	.27	.26	.40	.44	.25		-.15	-.04	(.03)	-.09
EPPS-I	.17	.22	[.35]	.10	.13	.20	.16	.21	.14	.12	.17		.08	.02	(.33)
CPI-II	[.27]	.21	.09	.32	.18	.21	.26	.29	.31	(.27)	.09	-.04		.01	.03
MMPI-II	.18	[.32]	-.01	.19	.21	.14	.20	.20	.32	.27	(.19)	-.13	.24		.16
EPPS-II	.14	.20	[.39]	.07	.19	.23	.22	.20	.22	.12	.16	(.08)	.22	.05	

Note:--Correlations for the male sample (N = 93) are listed above the main diagonal; those for the female sample (N = 108) are listed below the diagonal. Figures in parentheses are part-whole correlations. Figures in brackets, while not part-whole correlations, nonetheless have some common elements spuriously influencing the correlation. Circled values are test-retest reliability coefficients from items administered on four occasions. Correlations above .16 and .23 are significantly different from zero at probabilities of .05 and .01, respectively.

and the EPPS, the other on the MMPI and the EPPS. Consequently, these two subjects were removed from the male sample, and all analyses were recomputed on the remaining 93 males. While the analyses showed no substantial changes between the complete (N = 95) and reduced (N = 93) male samples, the values from the latter (smaller) sample have been tabled throughout this paper in order to rule out any possibility of spuriously increased correlations due to extreme scores. Consequently, it now appears safe to conclude that intra-individual variability vs. consistency can be measured with adequate reliability (homogeneity) when such measures are based upon responses to item pools of substantial size.

How Stable are Measures of Response Consistency?

To obtain test-retest reliability estimates for consistency scores, it is necessary to administer the same item pool on at least four occasions and then to correlate consistency scores obtained on one pair of administrations with those obtained on another pair. Since the 167 CPI-MMPI common items, as well as the shorter duplicated item pools within each of the three inventories, were all administered on four occasions, it was possible to obtain some estimates of the stability of these consistency scores. The circled values in Table 3 present the evidence regarding this second aspect of reliability. Note that the test-retest reliabilities, like the homogeneity coefficients, were quite low for the three short duplicated item pools (.03 to .33), while they were of at least moderate size for the 167 CPI-MMPI common items (.54 to .65). That is, the test-retest reliability values were of approximately the same order as the homogeneity values previously reported. Thus, while the test-retest stability of consistency scores from the three complete inventories were not available in this study, it seems reasonable to conclude from the corresponding values for the 167 common items that such reliabilities would be substantial. Even for the 167-item pool, the obtained values were roughly comparable to those previously reported for many regular inventory scales.

Do Measures of Response Consistency Show Convergent Validity?

To demonstrate that the consistency measures, though reliable, are not simply test-specific, it becomes necessary to discover whether measures derived from different item pools display convergent validity. Table 3 presents the evidence bearing on this question, namely, the correlations among the 15 measures of intra-individual consistency. As might be expected, the short (and unreliable) duplicated item pools generally showed low correlations with all other measures. However, all of the consistency measures based upon the longer item pools showed statistically significant positive correlations, generally of substantial size. For example, the consistency measures from the complete inventories had an average correlation of about .50, ranging from .34 (CPI vs. EPPS) to .62 (CPI vs. MMPI).

The corresponding correlations among measures from the 225-item pools averaged around .40 and ranged from a low of .30 (CPI vs. EPPS) to a high of .45 (CPI vs. MMPI). Again, all bivariate scatterplots were examined visually; as would be expected from knowledge of the (essentially normal) univariate score distributions, all scatterplots produced by measures from the longer item pools appeared bivariate normal. Consequently, it can be concluded that for consistency measures based upon longer item pools, at least some modest degree of convergent validity has been demonstrated.

What are the Psychometric Correlates of Response Consistency?

To obtain a global index of response consistency, each subject's consistency proportions on the three non-overlapping 225-item pools (CPI,-

I vs. CPI$_r$-II, EPPS-I vs. EPPS-II, and MMPI$_r$-I vs. MMPI$_r$-II) were summed; the resulting composite score (Σ) was used as an overall measure of intra-individual consistency. Scores on this composite, as well as on each of the five 225-item and longer consistency measures, were then correlated with scores on scales from the CPI, EPPS, MMPI, and other inventories. These correlations are available in Goldberg and Jones (1969). Table 4 summarizes some of the CPI and MMPI correlates of the three full-length consistency measures plus the composite consistency index (Σ). Since the correlational pattern differed somewhat between the two samples, the correlations are listed separately for males and females. The values presented in Table 4 are based upon scores from the first administration of the inventories; correlates of scores from the second administration were quite similar.

MMPI and CPI correlates. Perhaps the most striking feature of Table 4 is the fact that the CPI and MMPI measures of consistency correlated quite substantially with a number of CPI and MMPI scales, while the corresponding correlations for the EPPS consistency measure were markedly lower. Since the same effect was found for the 225-item reduced CPI and MMPI consistency scales, this finding can be attributed neither to the greater length and reliability of the CPI and MMPI consistency measures nor to the fact that these two measures share some common items. Moreover, few EPPS scales were significantly correlated with the consistency scales in either sample; and none were significantly associated in both samples.[3]

The CPI and MMPI consistency measures, as well as the composite global consistency index (Σ), correlated most highly with those scales marking the first CPI factor, especially the scales constructed to measure "achievement potential and intellectual efficiency" (Gough, 1957). The composite consistency index correlated .33 and .30 in the two samples with Nichols and Schnell's (1963) first factor marker *(VO)*. In the male sample, all consistency measures correlated significantly with Gough's Achievement via Independence scale *(Ai)*. In the female sample, all of the consistency measures correlated significantly with Gough's Intellectual Efficiency Scale *(Ie)* and Hase and Goldberg's (1967) rationally constructed Psychological-mindedness scale *(Psy)*. In general, response consistency was negatively related to the number of "True" responses on the CPI and positively related to Gough's Tolerance *(To)* and Sense of Well-Being *(Wb)* scales.

The bottom half of Table 4 lists some of the MMPI correlates of response consistency. As with the CPI, consistency was negatively correlated with the number of "True" responses (as well as with the number of "Deviant True" responses) and with various measures of acquiescence response style, including Hanley's *At* (Wiggins, 1962), Fricke's (1957) *B*, and Jackson and Messick's (1961) *Dy-3* scales. Consistency turned out to be positively related to the MMPI *K* scale and negatively related to the Schizophrenia *(Sc)* and Psychasthenia *(Pt)* scales, Welsh's (1956) Factor *A* scale, Block's (1965) Neurotic Undercontrol *(Nu)* scale, and Meehl and Hathaway's (1946) scale of General Personality Variance *(G)*.

Correlations with response style measures. Since response consistency appears to be negatively related to endorsing MMPI and CPI items, it is important to discover whether this relationship would extend to other types of inventories. Fortunately, five inventories of rather diverse content were administered to the subjects in this study; these inventories included a number of putative measures of acquiescence and deviance response styles. The correlations between the consistency indices and seven of these stylistic measures are presented in Table 5. Included in Table 5 are the correlates of scores on:

[3]The highest EPPS correlations were found in the female sample, negative with Abasement and Autonomy and positive with Affiliation and Achievement. In the male sample, Introception was positively related and Aggression was negatively related to consistency.

Table 4
Some CPI and MMPI Correlates of Response Consistency

	CPI-I vs. CPI-II		EPPS-I vs. EPPS-II		MMPI-I vs. MMPI-II		$\bar{\Sigma}$	
CPI Scale	M	F	M	F	M	F	M	F
Ie	.40**	.49**	.19	.20*	.39**	.49**	.37**	.44**
Ai	.45*	.37**	.30**	.11	.45**	.33**	.47**	.26**
Total True	-.32**	-.37**	-.12	-.22*	-.41**	-.42**	-.35**	-.38**
To	.33**	.48**	.11	.10	.40**	.43**	.33**	.37**
Psy	.32**	.48**	.09	.27**	.36**	.43**	.31**	.43**
Wb	.25*	.37**	.09	.12	.35**	.43**	.27**	.36**
VO	.32**	.41**	.13	.04	.43**	.37**	.33**	.30**
MMPI Scale								
Total True	-.38**	-.36**	-.21*	-.15	-.49**	-.49**	-.42**	-.36**
Deviant True	-.29**	-.41**	-.13	-.14	-.53**	-.58**	-.36**	-.42**
B	-.36**	-.31**	-.21*	-.12	-.48**	-.40**	-.43**	-.33**
At	-.20*	-.37**	-.05	-.27**	-.37**	-.46**	-.25*	-.44**
Dy-3	-.42**	-.27**	-.26**	-.08	-.49**	-.39**	-.44**	-.24*
K	.25*	.33**	.07	.15	.47**	.46**	.32**	.35**
Sc	-.14	-.35**	-.10	-.07	-.44**	-.54**	-.25*	-.38**
Pt	-.13	-.35**	-.12	-.07	-.40**	-.51**	-.24*	-.34**
A	-.16	-.34**	-.08	-.11	-.42**	-.50**	-.23*	-.35**
Nu	-.23	-.31**	-.15	-.11	-.48**	-.41**	-.34**	-.30**
G	-.20*	-.29**	-.09	-.15	-.47**	-.50**	-.28**	-.36**

* $p \leq .05$
** $p \leq .01$

1. Bass's (1956) Social Acquiescence (BSA) scale, comprised of 56 maxims or aphorisms such as "Love is the greatest of the Arts," " 'Tis vain to quarrel with our destiny," and "Still water runs deep." Each is presented with "Agree," "Uncertain," and "Disagree" response options; the acquiescence score is the number of "Agree" responses.
2. Rust and Davie's (1961) Reported Behavior Inventory (RBI), which includes 320 activities (e.g., Have you ever: "donated blood," "given money to charity," "been in an automobile accident," "had a theatre date," "gotten drunk," "shot dice," "acted in a play," "owned stocks or bonds," "signed a petition." The total number of "YES" responses was utilized as the RBI score.
3. Welsh's (1959) Figure Preference Test (WFPT; 400 line-drawings to which the subject responded either "Like" or "Dislike"). Three scales were scored: *(CF)* Welsh's 38-item Conformance scale, composed of items of greatest response agreement among both artists and non-artists (13 keyed "Like" and 25 keyed "Dislike"); *(X)* a measure of "Deviant Like" preferences analogous to the MMPI "Deviant True" scale (based on 200 items which more than 50% of Welsh's normal sample marked "Dislike"); and *(0)* a similarly constructed measure of "Deviant Dislike" preference, analogous to the MMPI "Deviant False" scale. The latter two scales were constructed by Kliewer (1962), who also constructed male and female deviant preference scales for the Perceptual Reaction Test.
4. The Perceptual Reaction Test (PRT; Berg,

Table 5
Other Psychometric Correlates of Response Consistency

Scales	CPI-I vs. CPI-II		EPPS-I vs. EPPS-II		MMPI-I vs. MMPI-II		$\overline{\Sigma}$	
	M	F	M	F	M	F	M	F
BSA[a]	-.22*	-.29**	-.19	-.06	-.23*	-.23*	-.25*	-.21*
RBI[b]	-.29**	-.05	-.05	-.12	-.26**	-.03	-.23*	-.12
CF[c]	.14	.16	.00	.10	.01	.11	.05	.14
$\overline{X}$[d]	-.23*	-.25**	.14	-.10	-.07	-.18	-.03	-.20*
$\overline{O}$[e]	.03	-.06	-.12	-.03	-.06	.00	-.08	-.04
PRT[f]	-.01	-.04	.11	.03	.01	.00	.04	-.04
ARS[g]	-.36	-.38**	-.10	-.12	-.33**	-.36**	-.28**	-.31**

* $p \leq .05$
** $p \leq .01$
[a] Bass (1956) Social Acquiescence Scale.
[b] Rust and Davie (1961) Reported Behavior Inventory: Number of activities experienced (Kliewer, 1962).
[c] Conformance scale from the Welsh (1959) Figure Preference Test.
[d] Deviant "Like" scale from the Welsh Figure Preference Test (Kliewer, 1962).
[e] Deviant "Dislike" scale from the Welsh Figure Preference Test (Kliewer, 1962).
[f] Deviance scale (Kliewer, 1962) from the Perceptual Reaction Test (Berg, Hunt, & Barnes, 1949).
[g] Couch and Keniston (1960) Agreement Response Scale.

Hunt, & Barnes, 1949). The PRT consists of 60 line-drawings presented with four response options: "Like much," "Like slightly," "Dislike slightly," and "Dislike much." The PRT deviancy scales include those response options chosen by less than 10% of Barnes's (1955) male (and female) normal samples.

5. Couch and Keniston's (1960) Agreement Response Scale (ARS). The ARS consisted of 20 items, each presented with seven response options (ranging from strongly disagree to strongly agree).

If response consistency is generally related to response "acquiescence" and "deviance" across diverse content, then the consistency indices should be *positively* correlated with *CF* and *negatively* correlated with the other six measures listed in Table 5. Of the 27 non-zero entries for males in Table 5, 22 (81%) were in this direction; of the 26 non-zero entries for females, 25 (96%) were in the same direction. And while most of the correlations between response consistency and WFPT and PRT scales (all involving "Like-Dislike" judgments of line-drawings) did not differ significantly from zero, the BSA and the ARS scales in both samples and the RBI in the male sample manifested significant negative correlations with the CPI and MMPI response consistency indices.

Summary. The overall evidence regarding the construct validity of the indices of response

consistency is hardly clear-cut. On the one hand, all three major consistency measures showed reasonable convergent validity (Table 3) as well as similar reliability characteristics (Table 2); on the other hand, the EPPS consistency measure showed a distinctively lower pattern of psychometric correlates than did the consistency indices derived from the CPI and the MMPI. Moreover, while the male and female samples produced quite similar findings in the reliability analyses (Tables 2 and 3), there were considerable differences between the two samples in their patterns of psychometric correlates (Tables 4 and 5). Consequently, the overall construct validity of the putative consistency trait must still remain in doubt. One possible means of resolving this doubt may be to examine the correlates of response consistency with measures other than those from self-report inventories and in samples other than the two utilized in previous analyses.

What are the Non-test Correlates of Response Consistency?

In order to discover whether consistency might function as a predictor of non-test criteria of the sort which self-report inventories are traditionally constructed to assess, the sample of 152 women from Hase and Goldberg (1967) was reanalyzed. Response consistency scores, based on two administrations of the CPI (with a four-week test-retest interval), were correlated with each of 13 criterion indices (see Hase & Goldberg [1967] for a description of the 13 criteria) as well as with the CPI scale scores. The results are summarized in Table 6. The left half of Table 6 lists the CPI correlates of response consistency in the new sample; the corresponding correlations for the original female sample are listed in parentheses. In both of these female samples, Gough's Intellectual Efficiency *(Ie)* scale was the highest consistency correlate (r= .49 and .42; $p <$ .01). The Spearman rank-order correlation between the values in the two samples for the 18 standard CPI scales was .73; the corresponding correlation for the original male vs. female samples was .80.

The right half of Table 6 lists the correlations between response consistency and the non-test criteria. Note that while consistency was not significantly related to most of these diverse criterion indices, there was a significant positive correlation between consistency and "responsibility" as rated by peers (r = .22; $p < .01$) and also with peer-rated "psychological-mindedness" (r = .18; $p < .05$). Since in our society responsible individuals are expected to be reasonably consistent, the former relationship adds an increment of construct validity to both of these measures.

Finally, one might hypothesize that consistent individuals should manifest a less ambiguous social demeanor and consequently should be less equivocally viewed by others than less consistent individuals (see Peterson, 1965). To test this hypothesis, response consistency was correlated with the *variance* of each of the peer ratings taken separately, as well as with a composite index of rater variance. The results of these analyses are also summarized in Table 6. As hypothesized, all of the consistency vs. variance correlations were negative (response consistency relating to lower peer rating variance), though none of the correlations differed significantly from zero. Consequently, with the possible exception of peer-rated "responsibility" (and "psychological-mindedness"), response consistency does not appear promising as a *direct* predictor of at least this class of non-test criteria.

Does Response Consistency Function as a Suppressor Variable?

While consistency may not be a potent predictor of most non-test criteria, inconsistency might serve to attenuate predictor-criterion relationships; and the addition of an index of response consistency to a battery of other measures might increase prediction by the suppression of this extraneous variance. To test this hypothesis, the

Table 6
Correlates of Response Consistency (CPI-I vs. CPI-II) in a New Sample (N = 152 Females): CPI Scale Scores and Criterion Indices from Hase and Goldberg (1967)

CPI Scales			Criterion Indices	
Do	.35**	(.25**)	**Mean Peer Ratings**	
Cs	.26**	(.30**)	Dominance	.01
Sy	.31**	(.16)	Sociability	.15
Sp	.20*	(.15)	Responsibility	.22**
Sa	.27**	(.15)	Psychological-mindedness	.18*
Wb	.39**	(.37**)	Femininity	.15
			How Well Known	.13
Re	.39**	(.43**)	**Other Social Criteria**	
So	.23**	(.23*)	Number of Dates per Month	-.02
Sc	.21**	(.29**)	Sorority vs. Independent	-.01
To	.34**	(.48**)	Yielding (Conformity)	.00
Gi	.24**	(.18)		
Cm	.25**	(.20*)	**Academic Criteria**	
			College Dropout	.05
Ac	.37**	(.35**)	College Major	-.05
Ai	.30**	(.37**)	Grade Point Average	.12
Ie	.42**	(.49**)	Under- vs. Over-Achievement	-.13
Py	.15	(.31**)	**Variance of Peer Ratings**	
Fx	.02	(.10)	Dominance	-.13
Fe	-.01	(-.04)	Sociability	-.05
			Responsibility	-.01
VO	.33**	(.41**)	Psychological-mindedness	-.06
PO	.23**	(.17)	Femininity	-.15
			Composite Variance	-.13

* $p \leq .05$
** $p < .01$
Note:--Corresponding CPI scale correlations from the original study are listed in parentheses.

consistency index was included as a potential suppressor variable in a series of linear regression analyses along with each of four different sets of five CPI scales. Each of these four sets of six variables was then used to predict, in turn, each of the 13 criteria listed in Table 6. While the addition of the response consistency measure slightly improved the prediction of peer-rated "responsibility" and college GPA within some scale sets, in each of these analyses consistency was functioning as a direct predictor in the regression equation. In *none* of the 52 multiple

regression analyses did response consistency function as a significant suppressor variable.

Does Response Consistency Function as a Moderator Variable?

While response consistency may not markedly improve the prediction of non-test criteria within the standard multiple regression model, it might function as a moderator (Saunders, 1956) of the relationships between test scores and criteria. That is, it might be argued that if response inconsistency reflects test "error," then the test scores from subjects with less error (consistent responders) should correlate more highly with various non-test criteria than the test scores from less consistent responders. In the extreme case, this prediction seems straightforward:

1. If some subjects were to respond on a completely random basis, they should manifest chance levels of response consistency (i.e., very low consistency) and their test scores should show no significant relationships to any other measures.
2. If a total group of subjects were to include a subgroup of random responders, then the test-criterion correlations for the total group should be attenuated due to the introduction of these random scores.
3. If the total group were to be split into subgroups on the basis of an index of response consistency, then the low consistency subgroup (which would include all of the random responders) should manifest lower test score vs. criterion correlations than the more consistently responding subgroups.

In this case, then, the introduction of response consistency as a moderator should serve to produce subgroups of subjects who differ markedly in their general predictability (Ghiselli, 1956, 1960, 1963), with the more consistent subjects being more predictable than the less consistent ones.

To test this hypothesis, a number of moderator analyses were carried out using the sample and the 13 criteria described by Hase and Goldberg (1967). First, the total sample of 152 females was divided on the basis of their CPI consistency scores into (1) a subsample of highly consistent responders ($N = 33$), (2) an average group ($N = 88$), and (3) a subsample of low consistency responders ($N = 31$). The single CPI scale with the highest overall validity for each criterion for the total sample ($N = 152$) was then correlated with all of the criterion scores within each of the three subsamples. The results of these analyses are summarized in Table 7. (For a description of the 10 CPI scales, see Hase and Goldberg [1967]). The average of these correlations within each subsample across all 13 criteria are listed at the bottom of Table 7, along with the average of the 130 correlations between all 10 of the predictor scales and all 13 of the criteria, again computed separately within each subsample.

Note that the moderator effect, though weak, is generally in the opposite direction from that hypothesized. For the two extreme (high vs. low consistency) subsamples, 9 of the 13 pairs of correlations were higher in the low consistency subsample, 1 pair was the same, and 3 were higher among the more consistent subjects. The overall averages reflected this same paradoxical finding: The average correlations across 13 criteria (and across all 130 correlations) were higher for the low consistency than for the high consistency responders, which is the exact opposite of the initial prediction. And while most of the high vs. low sample correlations were not significantly different from each other, one of them (yielding in an experimental conformity situation, as predicted by the rationally constructed Conformity scale) approached statistical significance—though in the opposite direction from the original prediction.

Faced with these surprising findings, a second series of analyses was carried out to extend their generality. The total sample of 152 females was split into two equal subsamples ($N = 76$) on the

Table 7
Response Consistency (CPI-I vs. CPI-II) as a Moderator Variable: The Prediction of 13 Criterion Indices from Hase and Goldberg (1967)

		Response Consistency Sub-Groups		
Criterion	Predictor	High (N = 33)	Medium (N = 88)	Low (N = 31)
cDOM	Dom	.46	.43	.46
cSOC	fSu	.64	.39	.66
cRES	Res	.30	.43	.37
cPSY	So	.27	.23	.31
cFEM	f0r	.29	.32	.55
cHWK	Soc	.46	.40	.33
cDAT	nPl	.46	.37	.49
cSOR	fSu	.34	.41	.44
cYLD	Con	.14	.23	.56
cCDO	So	.25	.10	.05
cMAJ	f0r	.31	.15	.43
cGPA	Ach	.33	.30	.38
cACH	fEx	.27	.16	.08
Average r (13)		.35	.30	.39
Average r (130)		.19	.16	.22

basis of their response consistency scores. Within each of the two subsamples (high vs. low consistency), a multiple regression analysis was carried out using each criterion in turn and five Rational CPI scales *(Dom, Soc, Res, Ach,* and *Con).* Of the 13 pairs of multiple regression coefficients generated in these analyses, 10 were higher within the low consistency sample, two were approximately the same, and one was higher within the high consistency responders. Moreover, once again the "Yielding" criterion turned out to be the most highly moderated one; low consistency subjects were again the most predictable.

Consequently and contrary to the original hypothesis, there was a slight tendency for highly consistent responders to be less predictable than more inconsistent ones. One possible explanation for this finding stems from the correlates of response consistency (see Table 4). Since consistency was highly associated with high scores on scales marking the CPI first factor and the MMPI second factor, the moderating effects of response consistency may simply stem from the correlation of consistency with these other variables.

As a preliminary check on this hypothesis, scores on the Nichols and Schnell (1963) *VO* scale were used to subdivide the sample into two halves; and the identical multiple regression analyses were repeated within the high *VO* and low *VO* subsamples. As predicted, the low *VO* subjects were slightly more predictable than the high scoring subsample. While this analysis can provide no definitive conclusions, it does suggest that the moderating effects of response consistency might simply arise as an artifact of the correlation between consistency and various scale

scores. Therefore, it was relevant to examine the correlates of response consistency on the first administration of the inventory.

How Well Can Consistency be Predicted from First Administration Responses?

At least three previous investigators have attempted to predict intra-individual variability vs. consistency on one inventory, the MMPI, from the responses to the first administration of the items. The correlations between the consistency indices and six MMPI scales developed in these three studies are presented in Table 8.

Schofield (1950) compared the responses of 24 female neurotic patients who had completed the MMPI before and after hospital treatment with the responses of 42 normal females who also took the inventory on two occasions. The proportion of subjects who changed their responses to each item over the two administrations was compared between each group.

The 24 items which manifested the most significant differences between the proportions in the two groups were selected to comprise the *Pc* scale (originally called *Px*); the items were all keyed in the deviant direction (based upon the original Minnesota norms). While it is not obvious why scores on *Pc* should be particularly predictive of response consistency in the present normal samples, as Table 8 indicates, the *Pc* scale turned out to be one of the best predictors of MMPI consistency for females ($r = -.61$). Moreover, *Pc* was negatively correlated with all of the consistency measures, significantly so with the MMPI and the composite consistency indices in both samples and with the CPI consistency measure in the female sample.

Mills (1954) constructed a pair of scales to differentiate between (1) college students whose MMPI profiles changed considerably over a period of 19 months and (2) students whose profiles remained much the same over this period. The MMPI responses of 83 males and 76 females who took the inventory as part of a freshman orientation battery and then again near the end of their sophomore year were used to construct two profile stability scales *Sb-m* and *Sb-f*

Table 8
Replication of Previously Constructed Predictors of MMPI Consistency

	CPI-I vs. CPI-II		EPPS-I vs. EPPS-II		MMPI-I vs. MMPI-II		$\bar{\Sigma}$	
Scales	M	F	M	F	M	F	M	F
Pc[a]	-.09	-.37**	-.04	-.10	-.42**	-.61**	-.22*	-.43**
Sb-m[b]	.14	.22*	.00	.12	.39**	.36**	.21*	.25**
Sb-f[b]	.36**	.02	.17	-.05	.35**	.14	.36**	.05
IC[c]	-.20*	-.39**	-.13	-.09	-.52**	-.58**	-.34**	-.41**
SC[c]	-.20*	-.42**	-.15	-.12	-.52**	-.59**	-.35**	-.41**
RC[c]	-.02	.05	.00	.01	.16	.14	.07	.12

* $p \leq .05$
** $p \leq .01$
[a] Schofield (1950)
[b] Mills (1954)
[c] Pepper (1964)

(originally called *Ps-m* and *Ps-f*). While Mills's attempt to cross-validate these scales on new samples was not successful, the scales did show some slight relationship to response consistency in the present study. As Table 8 indicates, *Sb-m* was significantly correlated with the MMPI and composite consistency indices in both samples and with the CPI consistency measure in the female sample. *Sb-f*, on the other hand, was significantly correlated with the CPI, MMPI, and composite consistency indices in the male sample, while showing no significant correlations with any of the consistency measures in the female sample.

The study of most direct relevance to the present investigation was one carried out by Pepper (1964). A sample of 198 male students in an introductory psychology course was administered the MMPI on two occasions, with a three-week test-retest interval. Pepper constructed three scales to predict test-retest variability:

1. Item Change *(IC)* was developed by contrasting the responses of 60 subjects who changed the most items from test to retest with the responses of the 60 subjects displaying the least amount of item change; the 67 items with significantly different ($p < .01$) endorsement proportions between the two groups were keyed in the direction of the responses from the high change group.
2. Scale Change *(SC)* was constructed by contrasting the responses of 60 subjects whose scale scores were most variable from test to retest with the 60 subjects whose scale scores remained most stable; the 50 items with significantly different ($p < .05$) endorsement proportions between the two samples were keyed in the direction of the high change group.
3. Rank Change *(RC)* was developed by contrasting the responses of 60 subjects for whom the rank order of clinical scales was most variable from test to retest with those from the 60 subjects displaying the least change in scale ranks; 26 items which differentiated the groups at $p < .20$ were keyed in the direction of the high change group.

The male sample in the present study allows a direct cross-validation of Pepper's *IC* scale, and the correlation of −.52 displayed in Table 8 between *IC* and MMPI response consistency compares quite favorably with the value of −.60 reported by Pepper (1964) for the original derivation sample. Moreover, while *IC* was developed on a male sample, it was at least as valid a predictor of response consistency in the present female sample ($r = -.58$). Pepper's *IC* and *SC* scales produced quite similar correlations in the present study; both scales were significantly correlated with the MMPI, CPI, and composite indices of response consistency in both samples. Pepper's *RC* scale, on the other hand, showed no significant correlations with any of the consistency measures in either sample.

The findings presented in Table 8 thus demonstrate that given the reliability of the consistency indices (see Tables 2 and 3), the validity of predictions of response consistency by means of scales scored from responses to the first administration of the inventory may well approach the maximum validity possible. For example, Pepper's (1964) *IC* scale, while constructed solely on male subjects, provided quite valid predictions of response consistency in the present study for both the male and the female samples. Such findings replicate those of Chance (1955), who earlier developed a similar response consistency predictor for the Bell Adjustment Inventory.

What is the Content of the Items Which Predict Response Consistency?

To answer this question, all of the first-administration item responses were correlated with six indices of response consistency, separately within the male and the female samples. Table 9 presents the number of items from each

Table 9
The Number of Items in Each of Three Inventories Which Were Significantly Correlated with Each of Six Response Consistency Indices

	Inventory	Number Expected by Chance	Consistency Index MMPI	CPI	EPPS	CPI_r	$MMPI_r$	Σ
Males	MMPI	(28)	122	59	37	73	85	73
	CPI	(24)	85	53	42	57	81	68
	EPPS	(12)	16	11	23	7	14	17
Females	MMPI	(28)	139	88	22	79	117	90
	CPI	(24)	72	73	24	60	54	52
	EPPS	(12)	18	15	17	12	11	14
Both Samples	MMPI	(1.4)	35	13	0	14	22	11
	CPI	(1.2)	14	17	2	12	12	6
	EPPS	(.6)	0	0	0	1	0	0

Note:--The values in the table indicate the number of items from each inventory for which the correlations between responses to the first administration of the inventory and test-retest response consistency were .20 or higher ($p < .05$). The entries in the upper section of the table are based on the male sample (N = 93), and those in the middle section on the female sample (N = 108). The number of items which were significant in both samples are listed in the bottom section of the table.

of the three inventories which produced correlations of .20 or larger ($p \leqslant .05$) with each of the six consistency measures. For example, 122 MMPI items in the male sample and 139 MMPI items in the female sample were significantly related to the full-length measure of MMPI response consistency; 35 of these items replicated across both samples. Only about 28 items in each sample would be expected to correlate this highly by chance alone, and only one or two of such chance effects should replicate across the two samples.

As Table 9 indicates, for the MMPI and the CPI consistency measures and the MMPI and CPI item pools, there were considerably more significant correlations than would be expected by chance. On the other hand, the EPPS item pool did not generally produce more significant correlations than might be expected by chance; and the EPPS consistency measure produced relatively few significant correlations in all three item pools. Moreover, of the 23 EPPS items significantly associated with EPPS response consistency in the male sample and the 17 such items in the female sample, none replicated across the two samples.

Pepper (1964), in describing the content of his Item Change *(IC)* scale, noted that the items suggest "greater psychopathology in high change subjects than in low change subjects. High change subjects complain more frequently about an inability to make decisions, distractibility, self-doubt, and physical complaints and concerns" (Pepper, 1964, pp. 44-45). Those CPI

Table 10
The Content of Some Items Correlating Highly with Response Consistency in the MMPI and the CPI

	MMPI Items
32	I find it hard to keep my mind on a task or job.
146	I have the wanderlust and am never happy unless I am roaming or traveling about.
179	I am worried about sex matters.
248	Sometimes without any reason or even when things are going wrong I feel excitedly happy, "on top of the world."
303	I am so touchy on some subjects that I can't talk about them.
334	Peculiar odors come to me at times.
390	I have often felt badly over being misunderstood when trying to keep someone from making a mistake.
470	Sexual things disgust me.
	CPI Items
136	Most people make friends because friends are likely to be useful to them.
219	Most people inwardly dislike putting themselves out to help other people.
220	I feel uneasy indoors.
226	Most people worry too much about sex.
268	At times I have been very anxious to get away from my family.
456	I have more trouble concentrating than others seem to have.
458	People who seem unsure and uncertain about things make me feel uncomfortable.

Note:--All items listed in this table correlated with within-inventory response consistency at $p \leq .01$ in both the male ($N = 93$) and female ($N = 108$) samples. True responses to these items predict response change; False responses predict response consistency.

and MMPI items which correlated significantly ($p < .01$) with response consistency in the present study for both the male and the female samples are listed in Table 10.

For all of these items, high consistency is associated with the more desirable responses. Relatively inconsistent responders present themselves as restless and moody and admit to a host of complaints reflecting mild distress and distrust. However, the content of these items does not reflect gross pathology; the items are not of the extreme type which are obvious candidates

for the MMPI, *F*, *Pa*, and *Sc* scales. Moreover, it is important to realize that the items which predict response consistency are *not* the same type of items which are themselves responded to in an inconsistent fashion. The correlations across items between item response stability (Goldberg, 1963, 1968) and consistency predictability ranged from zero to around −.25 in various item pools. These two item parameters are not generally related to each other.

Discussion

By far the most important contribution to our knowledge of response consistency has come from the work of Fiske. A major review of the literature through 1953 on intra-individual response variability was published by Fiske and Rice (1955). Later, Fiske and his students carried out a series of studies of inconsistency within diverse instruments (e.g., Fiske, 1957b; Mitra & Fiske, 1956; Osterweil & Fiske, 1956). Their results were integrated into a general model of response consistency in a theoretical paper by Fiske (1957a), who suggested that:

> Although variability scores can show appreciable intercorrelations, it must be concluded that variability is not an important *general* trait of temperament or personality, independent of the situation in which the person finds himself. The pattern and magnitude of relationships involving variability are comparable to those obtained for many familiar personality characteristics, such as rigidity and leadership. . . . Variability is primarily determined by response tendencies or sets. . . . Variability scores are sometimes associated with external and independent measures. . . . In the normal range, several studies report modest correlations between variability and personality measures (p. 332).
>
> Variability is a meaningful characteristic of test responses. Variability scores for psychological tests are homogeneous; they often show intercorrelations between similar tests; they may demonstrate consistent patterns of association with conventional test scores. There is probably no single general trait of variability. Variability tendencies are largely specific to total constellations of stimuli and conditions (p. 335).

How well do these conclusions agree with the present findings? In regard to the homogeneity of consistency measures, the evidence is now overwhelming: Indices of response consistency are reasonably homogeneous if they are are based upon item pools of considerable size. In the present study (Table 2), estimated homogeneity coefficients ranged from around .50 to .80 for item pools of 167 to 566 items. Moreover, the retest stability of such consistency measures was about at the same level as their homogeneity values (Table 3).

Evidence regarding the convergent validity of such consistency measures is less clear-cut. While Mitra and Fiske (1956) and Fiske (1957b) reported median convergent validities between .10 and .20, their consistency measures were all based on relatively few items. Convergent validities in the present study (for the longer item pools) ranged from .30 to .60 (Table 3). Such coefficients, if corrected for the attenuation resulting from the less-than-perfect reliability of response consistency, are as high as would be expected on theoretical grounds (e.g., Fiske, 1961).

On the other hand, all of this evidence is still ambiguous, since consistency is clearly related to first-administration responses and these, in turn, may be related across the inventories. Moreover, similar problems must be faced in interpreting the correlates of response consistency. Artifactual relationships between response consistency and test scores based upon the same items have long been recognized. (See Glaser [1949, 1952] for some early explanations of such problems, and Bentler [1964] for a more recent re-appraisal.)

Unfortunately, interpretive problems do not disappear when consistency scores are correlat-

ed with other-test and non-test measures; these relationships, too, may stem—in whole or in part—from their relationship with the (consistency-related) first-administration item responses. That is, since inconsistency is related to endorsing items reflecting mild anxiety and distress, the across-test correlates of response consistency *could* simply reflect anxiety-related (rather than inconsistency-related) effects. This "chicken or egg?" interpretive problem will plague all research on response consistency, until some method is discovered for untangling these confounding effects.

To illustrate this problem, one of the best known effects of secular trends (Loevinger, 1957) can be considered. A number of previous studies have demonstrated that response changes upon retesting are typically in the socially desirable direction. Anxiety and neuroticism scores decrease upon retesting, while "adjustment" scores increase (e.g., Howard, 1964; Neprash, 1936; Pintner & Forlano, 1938; Windle, 1954, 1955). However, such retest effects could be viewed as inevitable consequences of the confounding of pathology and social undesirability. For example, subjects who make many socially undesirable responses on the first administration of an inventory (perhaps because they are unhappy, anxious, or otherwise upset) tend to change more of their responses upon retest than subjects who initially do not make as many undesirable responses. To the extent that this effect operates systematically, scores on social desirability scales will correlate positively with response consistency; and mean scores will change in the socially desirable direction. Moreover, in existing item pools there is a substantial confounding of content and item phrasing (Wiggins & Goldberg, 1965).

In the case of the MMPI and the CPI, most items are phrased so that the undesirable response is also a "True" response (e.g., "I am worried about sex matters"). Consequently, subjects who, for one reason or another, answer items in the socially undesirable direction will tend to respond "True" more often than other subjects. Since such subjects tend to change more responses on retesting, "Acquiescence" and "Deviant True" scales will correlate negatively with response consistency.

In the present study, "Acquiescence" and "Deviant True" scales typically correlated negatively, while social desirability scales correlated positively, with CPI and MMPI response consistency; the great majority of these correlations were statistically significant (Goldberg & Jones, 1969). However, most of these same scales did not correlate significantly with consistency on the EPPS, suggesting that the former correlations may be an artifactual consequence of the confounding effects already discussed.

In fact, analyses of the EPPS measure of response consistency should provide the most compelling evidence regarding a putative consistency trait. Since the EPPS items are presented in forced-choice format (the two stems having been roughly equated on social desirability), response variance potentially attributable to social desirability and acquiescence response styles should not be particularly prevalent in the EPPS. Moreover, individual differences in response consistency were larger in the EPPS than in either the reduced or the complete CPI and MMPI item pools. As Table 2 indicates, the standard deviation of consistency scores on the EPPS was .05 in both samples as compared to values of .03 and .04 for the other inventories. Consequently, it is appropriate to re-examine some of the questions posed in this report, especially as they bear on the EPPS measures of response consistency.

How homogeneous are measures of response consistency? The corrected odd-even homogeneity coefficients for the EPPS consistency index averaged .69, slightly higher than the comparable value for the reduced CPI (.63) and slightly lower than that for the reduced MMPI (.78). While the latter value might be somewhat inflated due to the correlation between consistency and initial responses to the surfeit of content-homogeneous items in the MMPI item pool, the EPPS homogeneity coefficient should

not share this problem to any sizeable extent. Consequently, individual differences in EPPS response consistency *could* correlate as high as .83 with some perfectly reliable other measure.

Do measures of response consistency show convergent validity? The correlations between EPPS response consistency and the CPI and MMPI consistency measures ranged from .28 to .47 (all $p < .01$) and averaged .38. When the coefficients were corrected for the unreliability of the consistency measures, they ranged from .44 to .62 and averaged .52. These values *could* be inflated due to the correlation of consistency with first-administration item responses (responses which, in turn, might correlate across the inventories). Given the general dearth of significant relationships between EPPS items and those from the CPI and MMPI pools, however, it would appear that these measures of response consistency *do* show evidence of convergent validity.

What are the psychometric correlates of response consistency? Unlike the CPI and MMPI consistency measures, EPPS response consistency had only modest correlations with any of the 150 inventory scales scored in this study (Goldberg & Jones, 1969). Moreover, in no case did EPPS consistency correlate significantly with the same scale in both the male and the female samples.

How well can consistency be predicted from the first-administration responses? Previous attempts to predict response consistency from initial item responses have focused on either the MMPI (Mills, 1954; Pepper, 1964; Schofield, 1950) or the Bell Adjustment Inventory (Chance, 1955). While a scale previously developed to predict MMPI response consistency cross-validated at a very high level in the present study, none of the scales developed in this project to predict EPPS consistency cross-validated on a sample of the other sex (Goldberg & Jones, 1969).

What is the content of the items which predict response consistency? Unfortunately, of 23 EPPS items significantly related to EPPS response consistency in the male sample and 17 such items in the female sample, none replicated across the two samples. Moreover, none of the items from the MMPI and only two from the CPI correlated significantly with EPPS response consistency in both samples. The content of these two CPI items ("I looked up to my father as an ideal man" and "As a child I used to be able to go to my parents with my problems"—answered more often False by consistent than by less consistent subjects) contrasts with that of the items which predict CPI and MMPI response consistency, the latter items reflecting mild anxiety, restlessness, and distrust.

Summary

The evidence regarding a trait of response consistency is equivocal. While consistency measures can be constructed which possess substantial homogeneity and a significant degree of convergent validity, such measures inevitably are confounded by variance attributable to the particular set of stimuli (items) used to elicit the responses, a problem shared with all other putative response sets and styles. (See Block [1965] and Rorer [1965].) The unconfounding of consistency effects from those of other sources of variance remains an important problem for future experimental investigations.

References

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

Bass, B. M. Development and evaluation of a scale for measuring social acquiescence. *Journal of Abnormal and Social Psychology,* 1956, *53,* 296–299.

Bentler, P. M. *Response variability: Fact or artifact?* Unpublished doctoral dissertation, Stanford University, 1964.

Berg, I. A., Hunt, W. A., & Barnes, E. H. *The Perceptual Reaction Test.* Evanston, IL: I. A. Berg, 1949.

Block, J. *The challenge of response sets.* New York: Appleton-Century-Crofts, 1965.

Chance, J. E. Prediction of changes in a personality inventory on retesting. *Psychological Reports,* 1955, *1,* 383–387.

Couch, A., & Keniston, K. Yeasayers and naysayers: Agreeing response set as a personality variable. *Journal of Abnormal and Social Psychology,* 1960, *60,* 151–174.

Fiske, D. W. The constraints on intra-individual variability in test responses. *Educational and Psychological Measurement,* 1957, *17,* 317–337. (a)

Fiske, D. W. An intensive study of variability scores. *Educational and Psychological Measurement,* 1957, *17,* 453–465. (b)

Fiske, D. W. The inherent variability of behavior. In D. W. Fiske & S. R. Maddi (Eds.), *Functions of varied experience.* Homewood, IL: Dorsey, 1961, 326–354.

Fiske, D. W., & Rice, L. Intra-individual response variability. *Psychological Bulletin,* 1955, *52,* 217–250.

Fricke, B. G. A response bias (B) scale for the MMPI. *Journal of Counseling Psychology,* 1957, *4,* 149–153.

Ghiselli, E. E. Differentiation of individuals in terms of their predictability. *Journal of Applied Psychology,* 1956, *40,* 374–377.

Ghiselli, E. E. The prediction of predictability. *Educational and Psychological Measurement,* 1960, *20,* 3–8.

Ghiselli, E. E. Moderating effects and differential reliability and validity. *Journal of Applied Psychology,* 1963, *47,* 81–86.

Glaser, R. A methodological analysis of the inconsistency of responses to test items. *Educational and Psychological Measurement,* 1949, *9,* 727–739.

Glaser, R. The reliability of inconsistency. *Educational and Psychological Measurement,* 1952, *12,* 60–64.

Goldberg, L. R. A model of item ambiguity in personality assessment. *Educational and Psychological Measurement,* 1963, *23,* 467–492.

Goldberg, L. R. The interrelationships among item characteristics in an adjective check list: The convergence of different indices of item ambiguity. *Educational and Psychological Measurement,* 1968, *28,* 273–296.

Goldberg, L. R., & Jones, R. R. The reliability of reliability: The generality and correlates of intra-individual consistency in responses to structured personality inventories. *Oregon Research Institute Research Monograph,* 1969, *9*(2).

Goldberg, L. R., & Rust, R. M. Intra-individual variability in the MMPI-CPI common item pool. *British Journal of Social and Clinical Psychology,* 1964, *3,* 145–147.

Gough, H. G. *California Psychological Inventory Manual.* Palo Alto, CA: Consulting Psychologists Press, 1957.

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

Howard, K. I. Differentiation of individuals as a function of repeated testing. *Educational and Psychological Measurement,* 1964, *24,* 875–894.

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

Kliewer, V. D. *Multiple stylistic effects and self-report personality assessment.* Unpublished doctoral dissertation, University of Oregon, 1962.

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

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

Mills, W. W. *MMPI profile pattern and scale stability throughout four years of college attendance.* Unpublished doctoral dissertation, University of Minnesota, 1954.

Mitra, S. K., & Fiske, D. W. Intra-individual variability as related to test score and item. *Educational and Psychological Measurement,* 1956, *16,* 3–12.

Neprash, J. A. The reliability of questions in the Thurstone Personality Schedule. *Journal of Social Psychology,* 1936, *7,* 239–244.

Nichols, R. C., & Schnell, R. R. Factor scales for the California Psychological Inventory. *Journal of Consulting Psychology,* 1963, *27,* 228–235.

Osterweil, J., & Fiske, D. W. Intra-individual variability in sentence completion responses. *Journal of Abnormal and Social Psychology,* 1956, *52,* 195–199.

Pepper, L. J. The *MMPI: Initial test predictors of retest changes.* Unpublished doctoral dissertation, University of North Carolina, 1964.

Petersen, P. G. *Reliability of judgments of personality as a function of subjects and traits being judged.* Unpublished doctoral dissertation, University of California, 1965.

Pintner, R., & Forlano, G. Four retests of a personality inventory. *Journal of Educational Psychology,* 1938, *29,* 93–100.

Rorer, L. G. The great response-style myth. *Psychological Bulletin,* 1965, *63,* 129–150.

Rust, R. M., & Davie, J. S. The personal problems of college students. *Mental Hygiene,* 1961, *45,* 247–257.

Saunders, D. R. Moderator variables in prediction. *Educational and Psychological Measurement,* 1956, *16,* 209–222.

Schofield, W. Changes in responses to the Minnesota Multiphasic Inventory following certain therapies. *Psychological Monographs,* 1950, *64* (5, Whole No. 311).

Welsh, G. S. Factor dimensions A and R. In G. S. Welsh & W. G. Dahlstrom (Eds.), *Basic readings on the MMPI in psychology and medicine.* Minneapolis: University of Minnesota Press, 1956, 264–281.

Welsh, G. S. *Preliminary manual for the Welsh Figure Preference Test: Research edition.* Palo Alto, CA: Consulting Psychologists Press, 1959.

Wiggins, J. S. Strategic, method, and stylistic variance in the MMPI. *Psychological Bulletin,* 1962, *59,* 224–242.

Wiggins, J. S., & Goldberg, L. R. Interrelationships among MMPI item characteristics. *Educational and Psychological Measurement,* 1965, *25,* 381–397.

Windle, C. Test-retest effect on personality questionnaires. *Educational and Psychological Measurement,* 1954, *14,* 617–633.

Windle, C. Further studies of test-retest effect on personality questionnaires. *Educational and Psychological Measurement,* 1955, *15,* 246–253.

Acknowledgment

This study was supported by Grants MH 12972 and MH 10822 from the National Institute of Mental Health, U.S. Public Health Service.

Author's Address

Lewis R. Goldberg, Institute for the Measurement of Personality, 1201 Oak Street, Eugene, OR 97401.