VOL. 56, NO. 2 MARCH, 1959

Psychological Bulletin

CONVERGENT AND DISCRIMINANT VALIDATION BY THE MULTITRAIT-MULTIMETHOD MATRIX[1]

DONALD T. CAMPBELL
Northwestern University
AND DONALD W. FISKE
University of Chicago

In the cumulative experience with measures of individual differences over the past 50 years, tests have been accepted as valid or discarded as invalid by research experiences of many sorts. The criteria suggested in this paper are all to be found in such cumulative evaluations, as well as in the recent discussions of validity. These criteria are clarified and implemented when considered jointly in the context of a multitrait-multimethod matrix. Aspects of the validational process receiving particular emphasis are these:

1. Validation is typically *convergent*, a confirmation by independent measurement procedures. Independence of methods is a common denominator among the major types of validity (excepting content validity) insofar as they are to be distinguished from reliability.

2. For the justification of novel trait measures, for the validation of test interpretation, or for the establishment of construct validity, *discriminant* validation as well as convergent validation is required. Tests can be invalidated by too high correlations with other tests from which they were intended to differ.

3. Each test or task employed for measurement purposes is a *trait-method unit*, a union of a particular trait content with measurement procedures not specific to that content. The systematic variance among test scores can be due to responses to the measurement features as well as responses to the trait content.

4. In order to examine discriminant validity, and in order to estimate the relative contributions of trait and method variance, *more than one trait* as well as *more than one method* must be employed in the validation process. In many instances it will be convenient to achieve this through a multitrait-multimethod matrix. Such a matrix presents all of the intercorrelations resulting when each of several traits is measured by each of several methods.

To illustrate the suggested validational process, a synthetic example is

[1] The new data analyses reported in this paper were supported by funds from the Graduate School of Northwestern University and by the Department of Psychology of the University of Chicago. We are also indebted to numerous colleagues for their thoughtful criticisms and encouragement of an earlier draft of this paper, especially Benjamin S. Bloom, R. Darrell Bock, Desmond S. Cartwright, Loren J. Chapman, Lee J. Cronbach, Carl P. Duncan, Lyle V. Jones, Joe Kamiya, Wilbur L. Layton, Jane Loevinger, Paul E. Meehl, Marshall H. Segall, Thornton B. Roby, Robert C. Tryon, Michael Wertheimer, and Robert F. Winch.

TABLE 1

A SYNTHETIC MULTITRAIT-MULTIMETHOD MATRIX

		Method 1			Method 2			Method 3		
	Traits	A_1	B_1	C_1	A_2	B_2	C_2	A_3	B_3	C_3
Method 1	A_1	(.89)								
	B_1	.51	(.89)							
	C_1	.38	.37	(.76)						
Method 2	A_2	.57	.22	.09	(.93)					
	B_2	.22	.57	.10	.68	(.94)				
	C_2	.11	.11	.46	.59	.58	(.84)			
Method 3	A_3	.56	.22	.11	.67	.42	.33	(.94)		
	B_3	.23	.58	.12	.43	.66	.34	.67	(.92)	
	C_3	.11	.11	.45	.34	.32	.58	.58	.60	(.85)

Note.—The validity diagonals are the three sets of italicized values. The reliability diagonals are the three sets of values in parentheses. Each heterotrait-monomethod triangle is enclosed by a solid line. Each heterotrait-heteromethod triangle is enclosed by a broken line.

presented in Table 1. This illustration involves three different traits, each measured by three methods, generating nine separate variables. It will be convenient to have labels for various regions of the matrix, and such have been provided in Table 1. The reliabilities will be spoken of in terms of three *reliability diagonals*, one for each method. The reliabilities could also be designated as the monotrait-monomethod values. Adjacent to each reliability diagonal is the *heterotrait-monomethod* triangle. The reliability diagonal and the adjacent heterotrait-monomethod triangle make up a *monomethod block*. A *heteromethod block* is made up of a *validity* diagonal (which could also be designated as monotrait-heteromethod values) and the two *heterotrait-heteromethod* triangles lying on each side of it. Note that these two heterotrait-heteromethod triangles are not identical.

In terms of this diagram, four aspects bear upon the question of validity. In the first place, the entries in the validity diagonal should be significantly different from zero and sufficiently large to encourage further examination of validity. This requirement is evidence of convergent validity. Second, a validity diagonal value should be higher than the values lying in its column and row in the heterotrait-heteromethod triangles. That is, a validity value for a variable should be higher than the correlations obtained between that variable and any other variable having neither trait nor method in common. This requirement may seem so minimal and so obvious as to not need stating, yet an inspection of the literature shows that it is frequently not met,

and may not be met even when the validity coefficients are of substantial size. In Table 1, all of the validity values meet this requirement. A third common-sense desideratum is that a variable correlate higher with an independent effort to measure the same trait than with measures designed to get at different traits which happen to employ the same method. For a given variable, this involves comparing its values in the validity diagonals with its values in the heterotrait-monomethod triangles. For variables A_1, B_1, and C_1, this requirement is met to some degree. For the other variables, A_2, A_3 etc., it is not met and this is probably typical of the usual case in individual differences research, as will be discussed in what follows. A fourth desideratum is that the same pattern of trait interrelationship be shown in all of the heterotrait triangles of both the monomethod and heteromethod blocks. The hypothetical data in Table 1 meet this requirement to a very marked degree, in spite of the different general levels of correlation involved in the several heterotrait triangles. The last three criteria provide evidence for discriminant validity.

Before examining the multitrait-multimethod matrices available in the literature, some explication and justification of this complex of requirements seems in order.

Convergence of independent methods: the distinction between reliability and validity. Both reliability and validity concepts require that agreement between measures be demonstrated. A common denominator which most validity concepts share in contradistinction to reliability is that this agreement represent the convergence of independent approaches. The concept of independence is indicated by such phrases as "external variable," "criterion performance," "behavioral criterion" (American Psychological Association, 1954, pp. 13–15) used in connection with concurrent and predictive validity. For construct validity it has been stated thus: "Numerous successful predictions dealing with phenotypically diverse 'criteria' give greater weight to the claim of construct validity than do . . . predictions involving very similar behavior" (Cronbach & Meehl, 1955, p. 295). The importance of independence recurs in most discussions of proof. For example, Ayer, discussing a historian's belief about a past event, says "if these sources are numerous and independent, and if they agree with one another, he will be reasonably confident that their account of the matter is correct" (Ayer, 1954, p. 39). In discussing the manner in which abstract scientific concepts are tied to operations, Feigl speaks of their being "fixed" by "triangulation in logical space" (Feigl, 1958, p. 401).

Independence is, of course, a matter of degree, and in this sense, reliability and validity can be seen as regions on a continuum. (Cf. Thurstone, 1937, pp. 102–103.) Reliability is the agreement between two efforts to measure the same trait through maximally similar methods. Validity is represented in the agreement between two attempts to measure the same trait through maximally different methods. A split-half reliability is a little more like a validity coefficient than is an immediate test-retest reliability, for the items are not quite identical. A correlation between dissimilar subtests is probably a reliability measure, but is still closer to the region called validity.

Some evaluation of validity can take place even if the two methods

are not entirely independent. In Table 1, for example, it is possible that Methods 1 and 2 are not entirely independent. If underlying Traits A and B are entirely independent, then the .10 minimum correlation in the heterotrait-heteromethod triangles may reflect method covariance. What if the overlap of method variance were higher? All correlations in the heteromethod block would then be elevated, including the validity diagonal. The heteromethod block involving Methods 2 and 3 in Table 1 illustrates this. The degree of elevation of the validity diagonal above the heterotrait-heteromethod triangles remains comparable and relative validity can still be evaluated. The interpretation of the validity diagonal in an absolute fashion requires the fortunate coincidence of both an independence of traits and an independence of methods, represented by zero values in the heterotrait-heteromethod triangles. But zero values could also occur through a combination of negative correlation between traits and positive correlation between methods, or the reverse. In practice, perhaps all that can be hoped for is evidence for relative validity, that is, for common variance specific to a trait, above and beyond shared method variance.

Discriminant validation. While the usual reason for the judgment of invalidity is low correlations in the validity diagonal (e.g., the Downey Will-Temperament Test [Symonds, 1931, p. 337ff]) tests have also been invalidated because of too high correlations with other tests purporting to measure different things. The classic case of the social intelligence tests is a case in point. (See below and also [Strang, 1930; R. Thorndike, 1936].) Such invalidation occurs when values in the heterotrait-heteromethod triangles are as high as those in the validity diagonal, or even where within a monomethod block, the heterotrait values are as high as the reliabilities. Loevinger, Gleser, and DuBois (1953) have emphasized this requirement in the development of maximally discriminating subtests.

When a dimension of personality is hypothesized, when a construct is proposed, the proponent invariably has in mind distinctions between the new dimension and other constructs already in use. One cannot define without implying distinctions, and the verification of these distinctions is an important part of the validational process. In discussions of construct validity, it has been expressed in such terms as "from this point of view, a low correlation with athletic ability may be just as important and encouraging as a high correlation with reading comprehension" (APA, 1954, p. 17).

The test as a trait-method unit. In any given psychological measuring device, there are certain features or stimuli introduced specifically to represent the trait that it is intended to measure. There are other features which are characteristic of the method being employed, features which could also be present in efforts to measure other quite different traits. The test, or rating scale, or other device, almost inevitably elicits systematic variance in response due to both groups of features. To the extent that irrelevant method variance contributes to the scores obtained, these scores are invalid.

This source of invalidity was first noted in the "halo effects" found in ratings (Thorndike, 1920). Studies of individual differences among laboratory animals resulted in the recognition of "apparatus factors," usually more dominant than psychologi-

cal process factors (Tryon, 1942). For paper-and-pencil tests, methods variance has been noted under such terms as "test-form factors" (Vernon: 1957, 1958) and "response sets" (Cronbach: 1946, 1950; Lorge, 1937). Cronbach has stated the point particularly clearly: "The assumption is generally made . . . that what the test measures is determined by the content of the items. Yet the final score . . . is a composite of effects resulting from the content of the item and effects resulting from the form of the item used" (Cronbach, 1946, p. 475). "Response sets always lower the logical validity of a test. . . . Response sets interfere with inferences from test data" (p. 484).

While E. L. Thorndike (1920) was willing to allege the presence of halo effects by comparing the high obtained correlations with common sense notions of what they ought to be (e.g., it was unreasonable that a teacher's intelligence and voice quality should correlate .63) and while much of the evidence of response set variance is of the same order, the clear-cut demonstration of the presence of method variance requires both several traits and several methods. Otherwise, high correlations between tests might be explained as due either to basic trait similarity or to shared method variance. In the multitrait-multimethod matrix, the presence of method variance is indicated by the difference in level of correlation between the parallel values of the monomethod block and the heteromethod blocks, assuming comparable reliabilities among all tests. Thus the contribution of method variance in Test A_1 of Table 1 is indicated by the elevation of $r_{A_1B_1}$ above $r_{A_1B_2}$, i.e., the difference between .51 and .22, etc.

The distinction between trait and method is of course relative to the test constructor's intent. What is an unwanted response set for one tester may be a trait for another who wishes to measure acquiescence, willingness to take an extreme stand, or tendency to attribute socially desirable attributes to oneself (Cronbach: 1946, 1950; Edwards, 1957; Lorge, 1937).

Multitrait-Multimethod Matrices in the Literature

Multitrait-multimethod matrices are rare in the test and measurement literature. Most frequent are two types of fragment: two methods and one trait (single isolated values from the validity diagonal, perhaps accompanied by a reliability or two), and heterotrait-monomethod triangles. Either type of fragment is apt to disguise the inadequacy of our present measurement efforts, particularly in failing to call attention to the preponderant strength of methods variance. The evidence of test validity to be presented here is probably poorer than most psychologists would have expected.

One of the earliest matrices of this kind was provided by Kelley and Krey in 1934. Peer judgments by students provided one method, scores on a word-association test the other. Table 2 presents the data for the four most valid traits of the eight he employed. The picture is one of strong method factors, particularly among the peer ratings, and almost total invalidity. For only one of the eight measures, School Drive, is the value in the validity diagonal (.16!) higher than all of the heterotrait-heteromethod values. The absence of discriminant validity is further indicated by the tendency of the values in the monomethod triangles to approximate the reliabilities.

An early illustration from the ani-

TABLE 2

PERSONALITY TRAITS OF SCHOOL CHILDREN FROM KELLEY'S STUDY
(N=311)

		Peer Ratings				*Association Test*			
		A_1	B_1	C_1	D_1	A_2	B_2	C_2	D_2
Peer Ratings									
Courtesy	A_1	(.82)							
Honesty	B_1	.74	(.80)						
Poise	C_1	.63	.65	(.74)					
School Drive	D_1	.76	.78	.65	(.89)				
Association Test									
Courtesy	A_2	*.13*	.14	.10	.14	(.28)			
Honesty	B_2	.06	*.12*	.16	.08	.27	(.38)		
Poise	C_2	.01	.08	*.10*	.02	.19	.37	(.42)	
School Drive	D_2	.12	.15	.14	*.16*	.27	.32	.18	(.36)

mal literature comes from Anderson's (1937) study of drives. Table 3 presents a sample of his data. Once again, the highest correlations are found among different constructs from the same method, showing the dominance of apparatus or method factors so typical of the whole field of individual differences. The validity diagonal for hunger is higher than the heteroconstruct-heteromethod values. The diagonal value for sex has not been *italicized* as a validity coefficient since the obstruction box measure was pre-sex-opportunity, the activity wheel post-opportunity. Note that the high general level of heterotrait-heteromethod values could be due either to correlation of methods variance between the two methods, or to correlated trait variance. On a priori grounds, however, the methods would seem about as independent as one would be likely to achieve. The predominance of an apparatus factor for the activity wheel is evident from the fact that the correlation between hunger and thirst

TABLE 3

MEASURES OF DRIVES FROM ANDERSON'S DATA
(N=50)

		Obstruction Box			Activity Wheel		
		A_1	B_1	C_1	A_2	B_2	C_2
Obstruction Box							
Hunger	A_1	(.58)					
Thirst	B_1	.54	( )				
Sex	C_1	.46	.70	( )			
Activity Wheel							
Hunger	A_2	*.48*	.31	.37	(.83)		
Thirst	B_2	.35	*.33*	.43	.87	(.92)	
Post Sex	C_2	.31	.37	.44	.69	.78	( )

Note.—Empty parentheses appear in this and subsequent tables where no appropriate reliability estimates are reported in the original paper.

TABLE 4

SOCIAL INTELLIGENCE AND MENTAL ALERTNESS SUBTEST INTERCORRELATIONS FROM THORNDIKE'S DATA
(N=750)

		Memory		Comprehension		Vocabulary	
		A_1	B_1	A_2	B_2	A_3	B_3
Memory							
Social Intelligence (Memory for Names & Faces)	A_1	( )					
Mental Alertness (Learning Ability)	B_1	.31	( )				
Comprehension							
Social Intelligence (Sense of Humor)	A_2	*.30*	.31	( )			
Mental Alertness (Comprehension)	B_2	.29	*.38*	.48	( )		
Vocabulary							
Social Intelligence (Recog. of Mental State)	A_3	*.23*	.35	*.31*	.35	( )	
Mental Alertness (Vocabulary)	B_3	.30	*.58*	.40	*.48*	.47	( )

(.87) is of the same magnitude as their test-retest reliabilities (.83 and .92 respectively).

R. L. Thorndike's study (1936) of the validity of the George Washington Social Intelligence Test is the classic instance of invalidation by high correlation between traits. It involved computing all of the intercorrelations among five subscales of the Social Intelligence Test and five subscales of the George Washington Mental Alertness Test. The model of the present paper would demand that each of the traits, social intelligence and mental alertness, be measured by at least two methods. While this full symmetry was not intended in the study, it can be so interpreted without too much distortion. For both traits, there were subtests employing acquisition of knowledge during the testing period (i.e., learning or memory), tests involving comprehension of prose passages, and tests that involved a definitional activity. Table 4 shows six of Thorndike's 10 variables arranged as a multitrait-multimethod matrix. If the three subtests of the Social Intelligence Test are viewed as three methods of measuring social intelligence, then their intercorrelations (.30, .23, and .31) represent validities that are not only lower than their corresponding monomethod values, but also lower than the heterotrait-heteromethod correlations, providing a picture which totally fails to establish social intelligence as a separate dimension. The Mental Alertness validity diagonals (.38, .58, and .48) equal or exceed the monomethod values in two out of three cases, and exceed all heterotrait-heteromethod control values. These results illustrate the general conclusions reached by Thorndike in his factor analysis of the whole 10×10 matrix.

The data of Table 4 could be used to validate specific forms of cognitive functioning, as measured by the different "methods" represented by usual intelligence test content on the one hand and social content on the other. Table 5 rearranges the 15 values for this purpose. The monomethod values and the validity diagonals exchange places, while the heterotrait-heteromethod control coefficients are the same in both tables.

TABLE 5

MEMORY, COMPREHENSION, AND VOCABULARY MEASURED WITH SOCIAL AND ABSTRACT CONTENT

		Social Content			Abstract Content		
		A_1	B_1	C_1	A_2	B_2	C_2
Social Content							
Memory (Memory for Names and Faces)	A_1	( )					
Comprehension (Sense of Humor)	B_1	.30	( )				
Vocabulary (Recognition of Mental State)	C_1	.23	.31	( )			
Abstract Content							
Memory (Learning Ability)	A_2	*.31*	.31	.35	( )		
Comprehension	B_2	.29	*.48*	.35	.38	( )	
Vocabulary	C_2	.30	.40	*.47*	.58	.48	( )

As judged against these latter values, comprehension (.48) and vocabulary (.47), but not memory (.31), show some specific validity. This transmutability of the validation matrix argues for the comparisons within the heteromethod block as the most generally relevant validation data, and illustrates the potential interchangeability of trait and method components.

Some of the correlations in Chi's (1937) prodigious study of halo effect in ratings are appropriate to a multitrait-multimethod matrix in which each rater might be regarded as representing a different method. While the published report does not make these available in detail because it employs averaged values, it is apparent from a comparison of his Tables IV and VIII that the ratings generally failed to meet the requirement that ratings of the same trait by different raters should correlate higher than ratings of different traits by the same rater. Validity is shown to the extent that of the correlations in the heteromethod block, those in the validity diagonal are higher than the average heteromethod-heterotrait values.

A conspicuously unsuccessful multitrait-multimethod matrix is provided by Campbell (1953, 1956) for rating of the leadership behavior of officers by themselves and by their subordinates. Only one of 11 variables (Recognition Behavior) met the requirement of providing a validity diagonal value higher than any of the heterotrait-heteromethod values, that validity being .29. For none of the variables were the validities higher than heterotrait-monomethod values.

A study of attitudes toward authority and nonauthority figures by Burwen and Campbell (1957) contains a complex multitrait-multimethod matrix, one symmetrical excerpt from which is shown in Table 6. Method variance was strong for most of the procedures in this study. Where validity was found, it was primarily at the level of validity diagonal values higher than heterotrait-heteromethod values. As illustrated in Table 6, attitude toward father showed this kind of validity, as did attitude toward peers to a lesser degree. Attitude toward boss showed no validity. There was no evidence of a generalized attitude toward authority which would include father and boss, although such values as the

TABLE 6

ATTITUDES TOWARD FATHER, BOSS, AND PEER, AS MEASURED BY INTERVIEW AND CHECK-LIST OF DESCRIPTIVE TRAITS

		Interview			Trait Check-List		
		A_1	B_1	C_1	A_2	B_2	C_2
Interview ($N=57$)							
Father	A_1	( )					
Boss	B_2	.64	( )				
Peer	C_1	.65	.76	( )			
Trait Check-List ($N=155$)							
Father	A_2	*.40*	.08	.09	(.24)		
Boss	B_2	.19	*−.10*	−.03	.23	(.34)	
Peer	C_2	*.27*	.11	*.23*	.21	.45	(.55)

.64 correlation between father and boss as measured by interview might have seemed to confirm the hypothesis had they been encountered in isolation.

Borgatta (1954) has provided a complex multimethod study from which can be extracted Table 7, illustrating the assessment of two traits by four different methods. For all measures but one, the highest correlation is the apparatus one, i.e., with the other trait measured by the same method rather than with the same trait measured by a different method. Neither of the traits finds

TABLE 7

MULTIPLE MEASUREMENT OF TWO SOCIOMETRIC TRAITS ($N=125$)

		Sociometric				Observation			
		by Others		by Self		Group Interaction		Role Playing	
		A_1	B_1	A_2	B_2	A_3	B_3	A_4	B_4
Sociometric by Others									
Popularity	A_1	( )							
Expansiveness	B_1	.47	( )						
Sociometric by Self									
Popularity	A_2	*.19*	.18	( )					
Expansiveness	B_2	.07	*.08*	.32	( )				
Observation of Group Interaction									
Popularity	A_3	*.25*	.18	*.26*	.11	( )			
Expansiveness	B_3	.21	*.12*	.28	*.15*	.84	( )		
Observation of Role Playing									
Popularity	A_4	*.24*	.14	*.18*	.01	*.66*	.58	( )	
Expansiveness	B_4	.25	*.12*	.26	*.05*	.66	*.76*	.73	( )

any consistent validation by the requirement that the validity diagonals exceed the heterotrait-heteromethod control values. As a most minimal requirement, it might be asked if the sum of the two values in the validity diagonal exceeds the sum of the two control values, providing a comparison in which differences in reliability or communality are roughly partialled out. This condition is achieved at the purely chance level of three times in the six tetrads. This matrix provides an interesting range of methodological independence. The two "Sociometric by Others" measures, while representing the judgments of the same set of fellow participants, come from distinct tasks: Popularity is based upon each participant's expression of his own friendship preferences, while Expansiveness is based upon each participant's guesses as to the other participant's choices, from which has been computed each participant's reputation for liking lots of other persons, i.e., being "expansive." In line with this considerable independence, the evidence for a method factor is relatively low in comparison with the observational procedures. Similarly, the two "Sociometric by Self" measures represent quite separate tasks, Popularity coming from his estimates of the choices he will receive from others, Expansiveness from the number of expressions of attraction to others which he makes on the sociometric task. In contrast, the measures of Popularity and Expansiveness from the observations of group interaction and the role playing not only involve the same specific observers, but in addition the observers rated the pair of variables as a part of the same rating task in each situation. The apparent degree of method variance within each of the two observational situations, and the apparent sharing of method variance between them, is correspondingly high.

In another paper by Borgatta (1955), 12 interaction process variables were measured by quantitative observation under two conditions, and by a projective test. In this test, the stimuli were pictures of groups, for which the *S* generated a series of verbal interchanges; these were then scored in Interaction Process Analysis categories. For illustrative purposes, Table 8 presents the five traits which had the highest mean communalities in the over-all factor analysis. Between the two highly similar observational methods, validation is excellent: trait variance runs higher than method variance; validity diagonals are in general higher than heterotrait values of both the heteromethod and monomethods blocks, most unexceptionably so for Gives Opinion and Gives Orientation. The pattern of correlation among the traits is also in general confirmed.

Of greater interest because of the greater independence of methods are the blocks involving the projective test. Here the validity picture is much poorer. Gives Orientation comes off best, its projective test validity values of .35 and .33 being bested by only three monomethod values and by no heterotrait-heteromethod values within the projective blocks. All of the other validities are exceeded by some heterotrait-heteromethod value.

The projective test specialist may object to the implicit expectations of a one-to-one correspondence between projected action and overt action. Such expectations should not be attributed to Borgatta, and are not necessary to the method here proposed. For the simple symmetrical model of this paper, it has been as-

TABLE 8

INTERACTION PROCESS VARIABLES IN OBSERVED FREE BEHAVIOR, OBSERVED ROLE PLAYING AND A PROJECTIVE TEST
(N=125)

		Free Behavior					Role Playing					Projective Test				
		A_1	B_1	C_1	D_1	E_1	A_2	B_2	C_2	D_2	E_2	A_3	B_3	C_3	D_3	E_3
Free Behavior																
Shows solidarity	A_1	( )														
Gives suggestion	B_1	.25	( )													
Gives opinion	C_1	.13	.24	( )												
Gives orientation	D_1	−.14	.26	.52	( )											
Shows disagreement	E_1	.34	.41	.27	.02	( )										
Role Playing																
Shows solidarity	A_2	*.43*	.43	.08	.10	.29	( )									
Gives suggestion	B_2	.16	*.32*	.00	.24	.07	.37	( )								
Gives opinion	C_2	.15	.27	*.60*	.38	.12	.01	.10	( )							
Gives orientation	D_2	−.12	.24	.44	*.74*	.08	.04	.18	.40	( )						
Shows disagreement	E_2	.51	.36	.14	−.12	*.50*	.39	.27	.23	−.11	( )					
Projective Test																
Shows solidarity	A_3	*.20*	.17	.16	.12	.08	*.17*	.12	.30	.17	.22	( )				
Gives suggestion	B_3	.05	*.21*	.05	.08	.13	.10	*.19*	−.02	.06	.30	.32	( )			
Gives opinion	C_3	.31	.30	*.13*	−.02	.26	.25	.19	*.15*	−.04	.53	.31	.63	( )		
Gives orientation	D_3	−.01	.09	.30	*.35*	−.05	.03	.00	.19	*.33*	.00	.37	.29	.32	( )	
Shows disagreement	E_3	.13	.18	.10	.14	*.19*	.22	.28	.02	.04	*.23*	.27	.51	.47	.30	( )

TABLE 9

MAYO'S INTERCORRELATIONS BETWEEN OBJECTIVE AND RATING MEASURES OF INTELLIGENCE AND EFFORT

($N = 166$)

		Peer Ratings		Objective	
		A_1	B_1	A_2	B_2
Peer Rating					
Intelligence	A_1	(.85)			
Effort	B_1	.66	(.84)		
Objective Measures					
Intelligence	A_2	*.46*	.29	( )	
Effort	B_2	.46	*.40*	.10	( )

sumed that the measures are labeled in correspondence with the correlations expected, i.e., in correspondence with the traits that the tests are alleged to diagnose. Note that in Table 8, Gives Opinion is the best projective test predictor of both free behavior and role playing Shows Disagreement. Were a proper theoretical rationale available, these values might be regarded as validities.

Mayo (1956) has made an analysis of test scores and ratings of effort and intelligence, to estimate the contribution of halo (a kind of methods variance) to ratings. As Table 9 shows, the validity picture is ambiguous. The method factor or halo effect for ratings is considerable although the correlation between the two ratings (.66) is well below their reliabilities (.84 and .85). The objective measures share no appreciable apparatus overlap because they were independent operations. In spite of Mayo's argument that the ratings have some valid trait variance, the .46 heterotrait-heteromethod value seriously depreciates the otherwise impressive .46 and .40 validity values.

Cronbach (1949, p. 277) and Vernon (1957, 1958) have both discussed the multitrait-multimethod matrix shown in Table 10, based upon data originally presented by H. S. Conrad. Using an approximative technique, Vernon estimates that 61% of the systematic variance is due to a general factor, that $21\frac{1}{2}\%$ is due to the test-form factors specific to verbal or to pictorial forms of items, and that but $11\frac{1}{2}\%$ is due to the content fac-

TABLE 10

MECHANICAL AND ELECTRICAL FACTS MEASURED BY VERBAL AND PICTORIAL ITEMS

		Verbal Items		Pictorial Items	
		A_1	B_1	A_2	B_2
Verbal Items					
Mechanical Facts	A_1	(.89)			
Electrical Facts	B_1	.63	(.71)		
Pictorial Items					
Mechanical Facts	A_2	*.61*	.45	(.82)	
Electrical Facts	B_2	.49	*.51*	.64	(.67)

tors specific to electrical or to mechanical contents. Note that for the purposes of estimating validity, the interpretation of the general factor, which he estimates from the .49 and .45 heterotrait-heteromethod values, is equivocal. It could represent desired competence variance, representing components common to both electrical and mechanical skills—perhaps resulting from general industrial shop experience, common ability components, overlapping learning situations, and the like. On the other hand, this general factor could represent overlapping method factors, and be due to the presence in both tests of multiple choice item format, IBM answer sheets, or the heterogeneity of the *S*s in conscientiousness, test-taking motivation, and test-taking sophistication. Until methods that are still more different and traits that are still more independent are introduced into the validation matrix, this general factor remains uninterpretable. From this standpoint it can be seen that $21\frac{1}{2}\%$ is a very minimal estimate of the total test-form variance in the tests, as it represents only test-form components specific to the verbal or the pictorial items, i.e., test-form components which the two forms do *not* share. Similarly, and more hopefully, the $11\frac{1}{2}\%$ content variance is a very minimal estimate of the total true trait variance of the tests, representing only the true trait variance which electrical and mechanical knowledge do *not* share.

Carroll (1952) has provided data on the Guilford-Martin Inventory of Factors STDCR and related ratings which can be rearranged into the matrix of Table 11. (Variable R has been inverted to reduce the number of negative correlations.) Two of the methods, Self Ratings and Inventory scores, can be seen as sharing method variance, and thus as having an inflated validity diagonal. The more independent heteromethod blocks involving Peer Ratings show some evidence of discriminant and convergent validity, with validity diagonals averaging .33 (Inventory × Peer Ratings) and .39 (Self Ratings × Peer Ratings) against heterotrait-heteromethod control values averaging .14 and .16. While not intrinsically impressive, this picture is nonetheless better than most of the validity matrices here assembled. Note that the Self Ratings show slightly higher validity diagonal elevations than do the Inventory scores, in spite of the much greater length and undoubtedly higher reliability of the latter. In addition, a method factor seems almost totally lacking for the Self Ratings, while strongly present for the Inventory, so that the Self Ratings come off much the best if true trait variance is expressed as a proportion of total reliable variance (as Vernon [1958] suggests). The method factor in the STDCR Inventory is undoubtedly enhanced by scoring the same item in several scales, thus contributing correlated error variance, which could be reduced without loss of reliability by the simple expedient of adding more equivalent items and scoring each item in only one scale. It should be noted that Carroll makes explicit use of the comparison of the validity diagonal with the heterotrait-heteromethod values as a validity indicator.

Ratings in the Assessment Study of Clinical Psychologists

The illustrations of multitrait-multimethod matrices presented so far give a rather sorry picture of the validity of the measures of individual differences involved. The typical case shows an excessive amount of

TABLE 11

GUILFORD-MARTIN FACTORS STDCR AND RELATED RATINGS

(N=110)

	Inventory					Self Ratings					Peer Ratings				
	S	T	D	C	-R	S	T	D	C	-R	S	T	D	C	-R
Inventory															
S	(.92)														
T	.27	(.89)													
D	.62	.57	(.91)												
C	.36	.47	.90	(.91)											
-R	.69	.32	.28	−.06	(.89)										
Self Ratings															
S	*.57*	.11	.19	−.01	.53	( )									
T	.28	*.65*	.42	.26	.37	.26	( )								
D	.44	.25	*.53*	.45	.29	.31	.32	( )							
C	.31	.20	.54	*.52*	.13	.11	.21	.47	( )						
-R	.15	.30	.12	.04	*.34*	.10	.12	.04	.06	( )					
Peer Ratings															
S	*.37*	.08	.10	−.01	.38	*.42*	.02	.08	.08	.31	(.81)				
T	.23	*.32*	.15	.04	.40	.20	*.39*	.40	.21	.31	.37	(.66)			
D	.31	.11	*.27*	.24	.25	.17	.09	*.29*	.27	.30	.49	.38	(.73)		
C	.08	.15	.20	*.26*	−.05	.01	.06	.14	*.30*	.07	.19	.16	.40	(.75)	
-R	.21	.20	−.03	−.16	*.45*	.28	.17	.08	.01	*.56*	.55	.56	.34	−.07	(.76)

method variance, which usually exceeds the amount of trait variance. This picture is certainly not as a result of a deliberate effort to select shockingly bad examples: these are ones we have encountered without attempting an exhaustive coverage of the literature. The several unpublished studies of which we are aware show the same picture. If they seem more disappointing than the general run of validity data reported in the journals, this impression may very well be because the portrait of validity provided by isolated values plucked from the validity diagonal is deceptive, and uninterpretable in isolation from the total matrix. Yet it is clear that few of the classic examples of successful measurement of individual differences are involved, and that in many of the instances, the quality of the data might have been such as to magnify apparatus factors, etc. A more nearly ideal set of personality data upon which to illustrate the method was therefore sought in the multiple application of a set of rating scales in the assessment study of clinical psychologists (Kelly & Fiske, 1951).

In that study, "Rating Scale A" contained 22 traits referring to "behavior which can be directly observed on the surface." In using this scale the raters were instructed to "disregard any inferences about underlying dynamics or causes" (p. 207). The *S*s, first-year clinical psychology students, rated themselves and also their three teammates with whom they had participated in the various assessment procedures and with whom they had lived for six days. The median of the three teammates' ratings was used for the Teammate score. The *S*s were also rated on these 22 traits by the assessment staff. Our analysis uses the Final Pooled ratings, which were agreed upon by three staff members after discussion and review of the enormous amount of data and the many other ratings on each *S*. Unfortunately for our purposes, the staff members saw the ratings by Self and Teammates before making theirs, although presumably they were little influenced by these data because they had so much other evidence available to them. (See Kelly & Fiske, 1951, especially p. 64.) The Self and Teammate ratings represent entirely separate "methods" and can be given the major emphasis in evaluating the data to be presented.

In a previous analysis of these data (Fiske, 1949), each of the three heterotrait-monomethod triangles was computed and factored. To provide a multitrait-multimethod matrix, the 1452 heteromethod correlations have been computed especially for this report.[2] The full 66×66 matrix with its 2145 coefficients is obviously too large for presentation here, but will be used in analyses that follow. To provide an illustrative sample, Table 12 presents the interrelationships among five variables, selecting the one best representing each of the five recurrent factors discovered in Fiske's (1949) previous analysis of the monomethod matrices. (These were chosen without regard to their validity as indicated in the heteromethod blocks. Assertive—No. 3 reflected—was selected to represent Recurrent Factor 5 because Talkative had also a high

[2] We are indebted to E. Lowell Kelly for furnishing the V.A. assessment date to us, and to Hugh Lane for producing the matrix of intercorrelations.

In the original report the correlations were based upon 128 men. The present analyses were based on only 124 of these cases because of clerical errors. This reduction in *N* leads to some very minor discrepancies between these values and those previously reported.

TABLE 12

RATINGS FROM ASSESSMENT STUDY OF CLINICAL PSYCHOLOGISTS
($N = 124$)

		Staff Ratings					Teammate Ratings					Self Ratings				
		A_1	B_1	C_1	D_1	E_1	A_2	B_2	C_2	D_2	E_2	A_3	B_3	C_3	D_3	E_3
Staff Ratings																
Assertive	A_1	(.89)														
Cheerful	B_1	.37	(.85)													
Serious	C_1	−.24	−.14	(.81)												
Unshakable Poise	D_1	.25	.46	.08	(.84)											
Broad Interests	E_1	.35	.19	.09	.31	(.92)										
Teammate Ratings																
Assertive	A_2	*.71*	.35	−.18	.26	.41	(.82)									
Cheerful	B_2	.39	*.53*	−.15	.38	.29	.37	(.76)								
Serious	C_2	−.27	−.31	*.43*	−.06	.03	−.15	−.19	(.70)							
Unshakable Poise	D_2	.03	−.05	.03	*.20*	.07	.11	.23	.19	(.74)						
Broad Interests	E_2	.19	.05	.04	.29	*.47*	.33	.22	.19	.29	(.76)					
Self Ratings																
Assertive	A_3	*.48*	.31	−.22	.19	.12	*.46*	.36	−.15	.12	.23	( )				
Cheerful	B_3	.17	*.42*	−.10	.10	−.03	.09	*.24*	−.25	−.11	−.03	.23	( )			
Serious	C_3	−.04	−.13	*.22*	−.13	−.05	−.04	−.11	*.31*	.06	.06	−.05	−.12	( )		
Unshakable Poise	D_3	.13	.27	−.03	*.22*	−.04	.10	.15	.00	*.14*	−.03	.16	.26	.11	( )	
Broad Interests	E_3	.37	.15	−.22	.09	*.26*	.27	.12	−.07	.05	*.35*	.21	.15	.17	.31	( )

loading on the first recurrent factor.)

The picture presented in Table 12 is, we believe, typical of the best validity in personality trait ratings that psychology has to offer at the present time. It is comforting to note that the picture is better than most of those previously examined. Note that the validities for Assertive exceed heterotrait values of both the monomethod and heteromethod triangles. Cheerful, Broad Interests, and Serious have validities exceeding the heterotrait-heteromethod values with two exceptions. Only for Unshakable Poise does the evidence of validity seem trivial. The elevation of the reliabilities above the heterotrait-monomethod triangles is further evidence for discriminant validity.

A comparison of Table 12 with the full matrix shows that the procedure of having but one variable to represent each factor has enhanced the appearance of validity, although not necessarily in a misleading fashion. Where several variables are all highly loaded on the same factor, their "true" level of intercorrelation is high. Under these conditions, sampling errors can depress validity diagonal values and enhance others to produce occasional exceptions to the validity picture, both in the heterotrait-monomethod matrix and in the heteromethod-heterotrait triangles. In this instance, with an N of 124, the sampling error is appreciable, and may thus be expected to exaggerate the degree of invalidity.

Within the monomethod sections, errors of measurement will be correlated, raising the general level of values found, while within the heteromethods block, measurement errors are independent, and tend to lower the values both along the validity diagonal and in the heterotrait triangles. These effects, which may also be stated in terms of method factors or shared confounded irrelevancies, operate strongly in these data, as probably in all data involving ratings. In such cases, where several variables represent each factor, none of the variables consistently meets the criterion that validity values exceed the corresponding values in the monomethod triangles, when the full matrix is examined.

To summarize the validation picture with respect to comparisons of validity values with other heteromethod values in each block, Table 13 has been prepared. For each trait and for each of the three heteromethod blocks, it presents the value of the validity diagonal, the highest heterotrait value involving that trait, and the number out of the 42 such heterotrait values which exceed the validity diagonal in magnitude. (The number 42 comes from the grouping of the 21 other column values and the 21 other row values for the column and row intersecting at the given diagonal value.)

On the requirement that the validity diagonal exceed all others in its heteromethod block, none of the traits has a completely perfect record, although some come close. Assertive has only one trivial exception in the Teammate-Self block. Talkative has almost as good a record, as does Imaginative. Serious has but two inconsequential exceptions and Interest in Women three. These traits stand out as highly valid in both self-description and reputation. Note that the actual validity coefficients of these four traits range from but .22 to .82, or, if we concentrate on the Teammate-Self block as most certainly representing independent methods, from but .31 to .46. While these are the best traits, it seems that most of the traits have far above

TABLE 13

VALIDITIES OF TRAITS IN THE ASSESSMENT STUDY OF CLINICAL PSYCHOLOGISTS, AS JUDGED BY THE HETEROMETHOD COMPARISONS

	Staff-Teammate			*Staff-Self*			*Teammate-Self*		
	Val.	Highest Het.	No. Higher	Val.	Highest Het.	No. Higher	Val.	Highest Het.	No. Higher
1. Obstructiveness*	.30	.34	2	.16	.27	9	.19	.24	1
2. Unpredictable	.34	.26	0	.18	.24	3	.05	.19	29
3. Assertive*	.71	.65	0	.48	.45	0	.46	.48	1
4. Cheerful*	.53	.60	2	.42	.40	0	.24	.38	5
5. Serious*	.43	.35	0	.22	.27	2	.31	.24	0
6. Cool, Aloof	.49	.48	0	.20	.46	10	.02	.34	36
7. Unshakable Poise	.20	.40	16	.22	.27	4	.14	.19	10
8. Broad Interests*	.47	.46	0	.26	.37	6	.35	.32	0
9. Trustful	.26	.34	5	.08	.25	19	.11	.17	9
10. Self-centered	.30	.34	2	.17	.27	6	−.07	.19	36
11. Talkative*	.82	.65	0	.47	.45	0	.43	.48	1
12. Adventurous	.45	.60	6	.28	.30	2	.16	.36	14
13. Socially Awkward	.45	.37	0	.06	.21	28	.04	.16	30
14. Adaptable*	.44	.40	0	.18	.23	10	.17	.29	8
15. Self-sufficient*	.32	.33	1	.13	.18	5	.18	.15	0
16. Worrying, Anxious*	.41	.37	0	.23	.33	5	.15	.16	1
17. Conscientious	.26	.33	4	.11	.32	19	.21	.23	2
18. Imaginative*	.43	.46	1	.32	.31	0	.36	.32	0
19. Interest in Women*	.42	.43	2	.55	.38	0	.37	.40	1
20. Secretive, Reserved*	.40	.58	5	.38	.40	2	.32	.35	3
21. Independent Minded	.39	.42	2	.08	.25	19	.21	.30	3
22. Emotional Expression*	.62	.63	1	.31	.46	5	.19	.34	10

Note.—Val. = value in validity diagonal; Highest Het. = highest heterotrait value; No. Higher = number of heterotrait values exceeding the validity diagonal.

* Trait names which have validities in all three heteromethod blocks significantly greater than the heterotrait-heteromethod values at the .001 level.

chance validity. All those having 10 or fewer exceptions have a degree of validity significant at the .001 level as crudely estimated by a one-tailed sign test.[3] All but one of the variables meet this level for the Staff-Teammate block, all but four for the Staff-Self block, all but five for the most independent block, Teammate-Self. The exceptions to significant validity are not parallel from column to column, however, and only 13 of 22 variables have .001 significant validity in all three blocks. These are indicated by an asterisk in Table 13.

This highly significant general level of validity must not obscure the meaningful problem created by the occasional exceptions, even for the best variables. The excellent traits of Assertive and Talkative provide a case in point. In terms of Fiske's original analysis, both have high loadings on the recurrent factor "Confident self-expression" (repre-

[3] If we take the validity value as fixed (ignoring its sampling fluctuations), then we can determine whether the number of values larger than it in its row and column is less than expected on the null hypothesis that half the values would be above it. This procedure requires the assumption that the position (above or below the validity value) of any one of these comparison values is independent of the position of each of the others, a dubious assumption when common methods and trait variance are present.

sented by Assertive in Table 12). Talkative also had high loadings on the recurrent factor of Social Adaptability (represented by Cheerful in Table 12). We would expect, therefore, both high correlation between them and significant discrimination as well. And even at the common sense level, most psychologists would expect fellow psychologists to discriminate validly between assertiveness (nonsubmissiveness) and talkativeness. Yet in the Teammate-Self block, Assertive rated by self correlates .48 with Talkative by teammates, higher than either of their validities in this block, .43 and .46.

In terms of the average values of the validities and the frequency of exceptions, there is a distinct trend for the Staff-Teammate block to show the greatest agreement. This can be attributed to several factors. Both represent ratings from the external point of view. Both are averaged over three judges, minimizing individual biases and undoubtedly increasing reliabilities. Moreover, the Teammate ratings were available to the Staff in making their ratings. Another effect contributing to the less adequate convergence and discrimination of Self ratings was a response set toward the favorable pole which greatly reduced the range of these measures (Fiske, 1949, p. 342). Inspection of the details of the instances of invalidity summarized in Table 13 shows that in most instances the effect is attributable to the high specificity and low communality for the self-rating trait. In these instances, the column and row intersecting at the low validity diagonal are asymmetrical as far as general level of correlation is concerned, a fact covered over by the condensation provided in Table 13.

The personality psychologist is initially predisposed to reinterpret self-ratings, to treat them as symptoms rather than to interpret them literally. Thus, we were alert to instances in which the self ratings were not literally interpretable, yet nonetheless had a diagnostic significance when properly "translated." By and large, the instances of invalidity of self-descriptions found in this assessment study are not of this type, but rather are to be explained in terms of an absence of communality for one of the variables involved. In general, where these self descriptions are interpretable at all, they are as literally interpretable as are teammate descriptions. Such a finding may, of course, reflect a substantial degree of insight on the part of these *S*s.

The general success in discriminant validation coupled with the parallel factor patterns found in Fiske's earlier analysis of the three intramethod matrices seemed to justify an inspection of the factor pattern validity in this instance. One possible procedure would be to do a single analysis of the whole 66×66 matrix. Other approaches focused upon separate factoring of heteromethods blocks, matrix by matrix, could also be suggested. Not only would such methods be extremely tedious, but in addition they would leave undetermined the precise comparison of factor-pattern similarity. Correlating factor loadings over the population of variables was employed for this purpose by Fiske (1949) but while this provided for the identification of recurrent factors, no single over-all index of factor pattern similarity was generated. Since our immediate interest was in confirming a pattern of interrelationships, rather than in describing it, an efficient short cut was available: namely to test the similarity of the sets of heter-

otrait values by correlation coefficients in which each entry represented the size values of the given heterotrait coefficients in two different matrices. For the full matrix, such correlations would be based upon the N of the $22 \times 21/2$ or 231 specific heterotrait combinations. Correlations were computed between the Teammate and Self monomethods matrices, selected as maximally independent. (The values to follow were computed from the original correlation matrix and are somewhat higher than that which would be obtained from a reflected matrix.) The similarity between the two monomethods matrices was .84, corroborating the factor-pattern similarity between these matrices described more fully by Fiske in his parallel factor analyses of them. To carry this mode of analysis into the heteromethod block, this block was treated as though divided into two by the validity diagonal, the above diagonal values and the below diagonal representing the maximally independent validation of the heterotrait correlation pattern. These two correlated .63, a value which, while lower, shows an impressive degree of confirmation. There remains the question as to whether this pattern upon which the two heteromethod-heterotrait triangles agree is the same one found in common between the two monomethod triangles. The intra-Teammate matrix correlated with the two heteromethod triangles .71 and .71. The intra-Self matrix correlated with the two .57 and .63. In general, then, there is evidence for validity of the intertrait relationship pattern.

Discussion

Relation to construct validity. While the validational criteria presented are explicit or implicit in the discussions of construct validity (Cronbach & Meehl, 1955; APA, 1954), this paper is primarily concerned with the adequacy of tests as measures of a construct rather than with the adequacy of a construct as determined by the confirmation of theoretically predicted associations with measures of other constructs. We believe that before one can test the relationships between a specific trait and other traits, one must have some confidence in one's measures of that trait. Such confidence can be supported by evidence of convergent and discriminant validation. Stated in different words, any conceptual formulation of trait will usually include implicitly the proposition that this trait is a response tendency which can be observed under more than one experimental condition and that this trait can be meaningfully differentiated from other traits. The testing of these two propositions must be prior to the testing of other propositions to prevent the acceptance of erroneous conclusions. For example, a conceptual framework might postulate a large correlation between Traits A and B and no correlation between Traits A and C. If the experimenter then measures A and B by one method (e.g., questionnaire) and C by another method (such as the measurement of overt behavior in a situation test), his findings may be consistent with his hypotheses solely as a function of method variance common to his measures of A and B but not to C.

The requirements of this paper are intended to be as appropriate to the relatively atheoretical efforts typical of the tests and measurements field as to more theoretical efforts. This emphasis on validational criteria appropriate to our present atheoretical level of test construction is not at all

incompatible with a recognition of the desirability of increasing the extent to which all aspects of a test and the testing situation are determined by explicit theoretical considerations, as Jessor and Hammond have advocated (Jessor & Hammond, 1957).

Relation to operationalism. Underwood (1957, p. 54) in his effective presentation of the operationalist point of view shows a realistic awareness of the amorphous type of theory with which most psychologists work. He contrasts a psychologist's "literary" conception with the latter's operational definition as represented by his test or other measuring instrument. He recognizes the importance of the literary definition in communicating and generating science. He cautions that the operational definition "may not at all measure the process he wishes to measure; it may measure something quite different" (1957, p. 55). He does not, however, indicate how one would know when one was thus mistaken.

The requirements of the present paper may be seen as an extension of the kind of operationalism Underwood has expressed. The test constructor is asked to generate from his literary conception or private construct not one operational embodiment, but two or more, each as different in research vehicle as possible. Furthermore, he is asked to make explicit the distinction between his new variable and other variables, distinctions which are almost certainly implied in his literary definition. In his very first validational efforts, before he ever rushes into print, he is asked to apply the several methods and several traits jointly. His literary definition, his conception, is now best represented in what his independent measures of the trait hold *distinctively* in common. The multitrait-multimethod matrix is, we believe, an important practical first step in avoiding "the danger . . . that the investigator will fall into the trap of thinking that because he went from an artistic or literary conception . . . to the construction of items for a scale to measure it, he has validated his artistic conception" (Underwood, 1957, p. 55). In contrast with the *single operationalism* now dominant in psychology, we are advocating a *multiple operationalism*, a *convergent operationalism* (Garner, 1954; Garner, Hake, & Eriksen, 1956), a *methodological triangulation* (Campbell: 1953, 1956), an *operational delineation* (Campbell, 1954), a *convergent validation.*

Underwood's presentation and that of this paper as a whole imply moving from concept to operation, a sequence that is frequent in science, and perhaps typical. The same point can be made, however, in inspecting a transition from operation to construct. For any body of data taken from a single operation, there is a subinfinity of interpretations possible; a subinfinity of concepts, or combinations of concepts, that it could represent. Any single operation, as representative of concepts, is equivocal. In an analogous fashion, when we view the Ames distorted room from a fixed point and through a single eye, the data of the retinal pattern are equivocal, in that a subinfinity of hexahedrons could generate the same pattern. The addition of a second viewpoint, as through binocular parallax, greatly reduces this equivocality, greatly limits the constructs that could jointly account for both sets of data. In Garner's (1954) study, the fractionation measures from a single method were equivocal—they could have been a function of the stimulus distance being fractionated, or they

could have been a function of the comparison stimuli used in the judgment process. A multiple, convergent operationalism reduced this equivocality, showing the latter conceptualization to be the appropriate one, and revealing a preponderance of methods variance. Similarly for learning studies: in identifying constructs with the response data from animals in a specific operational setup there is equivocality which can operationally be reduced by introducing transposition tests, different operations so designed as to put to comparison the rival conceptualizations (Campbell, 1954).

Garner's convergent operationalism and our insistence on more than one method for measuring each concept depart from Bridgman's early position that "if we have more than one set of operations, we have more than one concept, and strictly there should be a separate name to correspond to each different set of operations" (Bridgman, 1927, p. 10). At the current stage of psychological progress, the crucial requirement is the demonstration of some convergence, not complete congruence, between two distinct sets of operations. With only one method, one has no way of distinguishing trait variance from unwanted method variance. When psychological measurement and conceptualization become better developed, it may well be appropriate to differentiate conceptually between Trait-Method Unit A_1 and Trait-Method Unit A_2, in which Trait A is measured by different methods. More likely, what we have called method variance will be specified theoretically in terms of a set of constructs. (This has in effect been illustrated in the discussion above in which it was noted that the response set variance might be viewed as trait variance, and in the rearrangement of the social intelligence matrices of Tables 4 and 5.) It will then be recognized that measurement procedures usually involve several theoretical constructs in joint application. Using obtained measurements to estimate values for a single construct under this condition still requires comparison of complex measures varying in their trait composition, in something like a multitrait-multimethod matrix. Mill's joint method of similarities and differences still epitomizes much about the effective experimental clarification of concepts.

The evaluation of a multitrait-multimethod matrix. The evaluation of the correlation matrix formed by intercorrelating several trait-method units must take into consideration the many factors which are known to affect the magnitude of correlations. A value in the validity diagonal must be assessed in the light of the reliabilities of the two measures involved: e.g., a low reliability for Test A_2 might exaggerate the apparent method variance in Test A_1. Again, the whole approach assumes adequate sampling of individuals: the curtailment of the sample with respect to one or more traits will depress the reliability coefficients and intercorrelations involving these traits. While restrictions of range over all traits produces serious difficulties in the interpretation of a multitrait-multimethod matrix and should be avoided whenever possible, the presence of different degrees of restriction on different traits is the more serious hazard to meaningful interpretation.

Various statistical treatments for multitrait-multimethod matrices might be developed. We have considered rough tests for the elevation

of a value in the validity diagonal above the comparison values in its row and column. Correlations between the columns for variables measuring the same trait, variance analyses, and factor analyses have been proposed to us. However, the development of such statistical methods is beyond the scope of this paper. We believe that such summary statistics are neither necessary nor appropriate at this time. Psychologists today should be concerned not with evaluating tests as if the tests were fixed and definitive, but rather with developing better tests. We believe that a careful examination of a multitrait-multimethod matrix will indicate to the experimenter what his next steps should be: it will indicate which methods should be discarded or replaced, which concepts need sharper delineation, and which concepts are poorly measured because of excessive or confounding method variance. Validity judgments based on such a matrix must take into account the stage of development of the constructs, the postulated relationships among them, the level of technical refinement of the methods, the relative independence of the methods, and any pertinent characteristics of the sample of *S*s. We are proposing that the validational process be viewed as an aspect of an ongoing program for improving measuring procedures and that the "validity coefficients" obtained at any one stage in the process be interpreted in terms of gains over preceding stages and as indicators of where further effort is needed.

The design of a multitrait-multimethod matrix. The several methods and traits included in a validational matrix should be selected with care. The several methods used to measure each trait should be appropriate to the trait as conceptualized. Although this view will reduce the range of suitable methods, it will rarely restrict the measurement to one operational procedure.

Wherever possible, the several methods in one matrix should be completely independent of each other: there should be no prior reason for believing that they share method variance. This requirement is necessary to permit the values in the heteromethod-heterotrait triangles to approach zero. If the nature of the traits rules out such independence of methods, efforts should be made to obtain as much diversity as possible in terms of data-sources and classification processes. Thus, the classes of stimuli *or* the background situations, the experimental contexts, should be different. Again, the persons providing the observations should have different roles *or* the procedures for scoring should be varied.

Plans for a validational matrix should take into account the difference between the interpretations regarding convergence and discrimination. It is sufficient to demonstrate convergence between two clearly distinct methods which show little overlap in the heterotrait-heteromethod triangles. While agreement between several methods is desirable, convergence between two is a satisfactory minimal requirement. Discriminative validation is not so easily achieved. Just as it is impossible to prove the null hypothesis, or that some object does not exist, so one can never establish that a trait, as measured, is differentiated from all other traits. One can only show that this measure of Trait A has little overlap with those measures of B and C, and no dependable generalization beyond B and C can be made. For example, social poise could probably

be readily discriminated from aesthetic interests, but it should also be differentiated from leadership.

Insofar as the traits are related and are expected to correlate with each other, the monomethod correlations will be substantial and heteromethod correlations between traits will also be positive. For ease of interpretation, it may be best to include in the matrix at least two traits, and preferably two sets of traits, which are postulated to be independent of each other.

In closing, a word of caution is needed. Many multitrait-multimethod matrices will show no convergent validation: no relationship may be found between two methods of measuring a trait. In this common situation, the experimenter should examine the evidence in favor of several alternative propositions: (*a*) Neither method is adequate for measuring the trait; (*b*) One of the two methods does not really measure the trait. (When the evidence indicates that a method does not measure the postulated trait, it may prove to measure some other trait. High correlations in the heterotrait-heteromethod triangles may provide hints to such possibilities.) (*c*) The trait is not a functional unity, the response tendencies involved being specific to the nontrait attributes of each test. The failure to demonstrate convergence may lead to conceptual developments rather than to the abandonment of a test.

Summary

This paper advocates a validational process utilizing a matrix of intercorrelations among tests representing at least two traits, each measured by at least two methods. Measures of the same trait should correlate higher with each other than they do with measures of different traits involving separate methods. Ideally, these validity values should also be higher than the correlations among different traits measured by the same method.

Illustrations from the literature show that these desirable conditions, as a set, are rarely met. Method or apparatus factors make very large contributions to psychological measurements.

The notions of convergence between independent measures of the same trait and discrimination between measures of different traits are compared with previously published formulations, such as construct validity and convergent operationalism. Problems in the application of this validational process are considered.

REFERENCES

American Psychological Association. Technical recommendations for psychological tests and diagnostic techniques. *Psychol. Bull., Suppl.*, 1954, **51**, Part 2, 1–38.

Anderson, E. E. Interrelationship of drives in the male albino rat. I. Intercorrelations of measures of drives. *J. comp. Psychol.*, 1937, **24**, 73–118.

Ayer, A. J. *The problem of knowledge.* New York: St Martin's Press, 1956.

Borgatta, E. F. Analysis of social interaction and sociometric perception. *Sociometry*, 1954, **17**, 7–32.

Borgatta, E. F. Analysis of social interaction: Actual, role-playing, and projective. *J. abnorm. soc. Psychol.*, 1955, **51**, 394–405.

Bridgman, P. W. *The logic of modern physics.* New York: Macmillan, 1927.

Burwen, L. S., & Campbell, D. T. The generality of attitudes toward authority and nonauthority figures. *J. abnorm. soc. Psychol.*, 1957, **54**, 24–31.

Campbell, D. T. *A study of leadership among submarine officers.* Columbus: Ohio State Univer. Res. Found., 1953.

Campbell, D. T. Operational delineation of "what is learned" via the transposition experiment. *Psychol. Rev.*, 1954, **61**, 167–174.

CAMPBELL, D. T. *Leadership and its effects upon the group.* Monogr. No. 83. Columbus: Ohio State Univer. Bur. Business Res., 1956.

CARROLL, J. B. Ratings on traits measured by a factored personality inventory. *J. abnorm. soc. Psychol.*, 1952, **47**, 626–632.

CHI, P.-L. Statistical analysis of personality rating. *J. exp. Educ.*, 1937, **5**, 229–245.

CRONBACH, L. J. Response sets and test validity. *Educ. psychol. Measmt*, 1946, **6**, 475–494.

CRONBACH, L. J. *Essentials of psychological testing.* New York: Harper, 1949.

CRONBACH, L. J. Further evidence on response sets and test design. *Educ. psychol. Measmt*, 1950, **10**, 3–31.

CRONBACH, L. J., & MEEHL, P. E. Construct validity in psychological tests. *Psychol. Bull.*, 1955, **52**, 281–302.

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

FEIGL, H. The mental and the physical. In H. Feigl, M. Scriven, & G. Maxwell (Eds.), *Minnesota studies in the philosophy of science.* Vol. II. *Concepts, theories and the mind-body problem.* Minneapolis: Univer. Minnesota Press, 1958.

FISKE, D. W. Consistency of the factorial structures of personality ratings from different sources. *J. abnorm. soc. Psychol.*, 1949, **44**, 329–344.

GARNER, W. R. Context effects and the validity of loudness scales. *J. exp. Psychol.*, 1954, **48**, 218–224.

GARNER, W. R., HAKE, H. W., & ERIKSEN, C. W. Operationism and the concept of perception. *Psychol. Rev.*, 1956, **63**, 149–159.

JESSOR, R., & HAMMOND, K. R. Construct validity and the Taylor Anxiety Scale. *Psychol. Bull.*, 1957, **54**, 161–170.

KELLEY, T. L., & KREY, A. C. *Tests and measurements in the social sciences.* New York: Scribner, 1934.

KELLY, E. L., & FISKE, D. W. *The prediction of performance in clinical psychology.* Ann Arbor: Univer. Michigan Press, 1951.

LOEVINGER, J., GLESER, G. C., & DUBOIS, P. H., Maximizing the discriminating power of a multiple-score test. *Psychometrika*, 1953, **18**, 309–317.

LORGE, I. Gen-like: Halo or reality? *Psychol. Bull.*, 1937, **34**, 545–546.

MAYO, G. D. Peer ratings and halo. *Educ. psychol. Measmt*, 1956, **16**, 317–323.

STRANG, R. Relation of social intelligence to certain other factors. *Sch. & Soc.*, 1930, **32**, 268–272.

SYMONDS, P. M. *Diagnosing personality and conduct.* New York: Appleton-Century, 1931.

THORNDIKE, E. L. A constant error in psychological ratings. *J. appl. Psychol.*, 1920, **4**, 25–29.

THORNDIKE, R. L. Factor analysis of social and abstract intelligence. *J. educ. Psychol.*, 1936, **27**, 231–233.

THURSTONE, L. L. *The reliability and validity of tests.* Ann Arbor: Edwards, 1937.

TRYON, R. C. Individual differences. In F. A. Moss (Ed.), *Comparative Psychology.* (2nd ed.) New York: Prentice-Hall, 1942. Pp. 330–365.

UNDERWOOD, B. J. *Psychological research.* New York: Appleton-Century-Crofts, 1957.

VERNON, P. E. Educational ability and psychological factors. Address given to the Joint Education-Psychology Colloquium, Univer. of Illinois, March 29, 1957.

VERNON, P. E. *Educational testing and test-form factors.* Princeton: Educational Testing Service, 1958. (Res. Bull. RB-58-3.)

Received June 19, 1958.