Chapter 5 from Foundations of the theory of prediction. The Dorsey Press: Homewood, Ill. 1966.

Chapter 5: Validity

Generically, the notion of "validity" has to do with the adequacy with which a test (i.e., a predictor) does, in fact, test what it is supposed to be testing; and the reader who sensibly, if naively, reasons that the correlation between test and criterion should amply suffice to describe this state of affairs has reckoned without the conceptual fecundity of test theorists. Actually, the literature abounds with phrases in which some adjective has been prefixed to "validity" for reference to some particular aspect or problem in the interpretation of test scores. Those prefix-validities which are currently most prominent in the discourse of persons who talk about such things are "empirical validity," "predictive validity," "concurrent validity," "face validity," "content validity," and "construct validity." None of these expressions has ever been defined with sufficient precision to bring joy to the beleaguered heart of a logician, and their grammatical form misleadingly suggests that they refer to various kinds of validity. Actually, what is signalized by the various prefixes "empirical," "predictive," etc., is not a plurality of validity modalities, but certain methodological distinctions among the criteria for which our predictors are supposedly testing and our grounds for believing that they are, in fact, testing successfully.

Empirical Validity

Let us define the expression, *the objective validity of variable* X *as a test for variable* Y *in population P*, or simply "the (unqualified) validity of X" when Y and P are contextually understood, to mean the linear correlation between X and Y in P.¹ It is important to be clear that a test has as many different objective validities as there are different criteria with which it can be correlated and different populations in which it and the criterion are jointly distributed. In a far from trivial sense, every variable is a "test" of every other variable—with, however, an objective validity of zero or near-zero in most cases.² Now, what is primarily to be understood by the *empirical validity* of X as a test for Y is simply X's objective validity for Y in the relevant population. However, the adjective "empirical" also carries strong overtones to the effect that by means of a sufficient number of observations on joint values of X and Y we have actually been able to learn what the approximate numerical value of this correlation is, or at least that such observations *could* successfully be carried out. Concommitently, it is intimated that what X is claimed to be a test of is something observable which has been clearly and quantitatively defined independently of the test.

¹Strictly speaking, we should refer to "validity" so defined as *linear validity*, thereby making explicit that what is meant is the accuracy of the predictor under linear estimation, even though its predictive success might well be greater than this were curvilinear prediction policies considered.

²The fact that a test has many validities of varying degrees of accuracy lies behind a useful metaphor advanced by Cronbach (1960, pp. 602 ff.) in which he speaks of a test's *bandwidth* and *fidelity*. By "bandwidth" Cronbach means the range of different criteria which can be estimated with at least moderate accuracy from the test, while "fidelity" refers to the test's validity for a particular criterion, presumably the one with. which the test correlates best. He observes that high fidelity can usually be obtained (if at all) only at the expense of bandwidth, whereas a test which has only moderate correlation with any given criterion may nonetheless justify its existence by being a modestly useful predictor of a great many variables in which its user is interested.

Suppose, for example, that some eminent hirsutologist proposes that a good test of baldness is the amount of light reflected by a person's head under certain standard conditions of illumination. And suppose too that this contention is hotly contested by other authorities. How is this vital issue to be settled? The scientific procedure would be to ascertain joint scores on the Head-reflection and Degree-of-baldness variables within a suitably large sample of the relevant population and compute the correlation between thetn. However, the intrepid researcher who actually attempted to carry this project through would quickly realize that it cannot be done not because of practical difficulties such as professional incompetence or trouble in acquiring a research grant, but for the elemental reason that nobody knows just what a "degree of baldness" is. Although generations of men have with varied emotions studied the increasing visibility of their scalps, not one seems to have had the courage to spell out the concept of "baldness" to the point where values on a scale of this variable hav any clear meaning. Consequently, a test for "baldness" in the everyday sense of this term cannot, even in principle, have any empirical validity simply because the intended criterion is too vaguely conceived. On the other hand, our researcher can readily determine the empirical validity of Head-reflection as a test of some observable criterion which is stipulated to be what he henceforth elects to mean by "degree of baldness" —e.g., the average number of hairs per square centimeter on the portion of a person's scalp above a plane passing through the eyebrows and auditory canals, or the measured resistance encountered by a comb dragged over the head in a standard manner, or the average baldness rating given to the person by three selected judges on a seven-point scale, or the proportion of test lice which upon release on the person's head are still found in residence a week later, ete. Such replacements of obscure ideas from everyday language with more technically workable concepts are routine in any scientific investigation, and are known as operational definitions. When one speaks of the "empirical validity" of a test X when what X is supposed to be a test of is described in everyday vague terms, it is implied that an operationally defined criterion has been selected. Observe, however, that the em pirical validity of X will depend upon which of the variously available operational definitions has been chosen. Thus it is most unlikely that the correlations of Head-reflection with Hair-density, Comb-resistance, Baldnessrating, and Lice-retention, respectively, would turn out to be numerically identical. It would be a sign of misunderstanding in this case to ask which correlation is the correct empirical validity of Head-reflection as a test of baldness, for none of them are. Rather, the test has a separate validity for each different proposed criterion, and it is merely a historical accident, so to speak, that these various criteria should have a common conceptual ancestor. It should also be noted that the test itself may be taken as an operational definition of what it purports to test for-e.g., we may let Head-reflection be our objective standard for "Degree-of-baldness." However, it is then no longer appropriate to speak of the test as having *empirical* validity. The correlation between Head-reflection and Head-reflection is obviously high, but the relation is logical, not empirical. This is why it was stated above that the notion of "empirical validity" carries with it the idea that the criterion has been defined independently of the test.

If the sensitive or knowledgeable reader wonders if these rather casual remarks about operational definitions and the determination of objective validity are an altogether adequate confrontation of the methodological issues involved, his suspicions are only too well justified. A satisfactory account of what our concepts mean, and how we are able to acquire knowledge about the external realities to which they supposedly refer, is perhaps the most profoundly difficult problem which has ever confronted human reason, and continues to challenge not merely test theory but the entire domain of science and philosophy. Our discussion of "validity" has so far just begun to work its shoulders under a burden which will weigh down with increasing oppressiveness as we proceed.

Predictive Validity and Concurrent Validity

As explained in Chapter 1, the term "prediction" has been used throughout this book to mean any inference about the still unknown details of some event, regardless of where that event occurs in time. However, a somewhat more common usage of the term is to construe "prediction" as an inference about the *future*. It is this temporal orientation which is reflected in the distinction between *predictive validity* and *concurrent validity*. What is implied by saying that a test has "predictive" validity is that the test scores can with some useful degree of objective validity be used to estimate a future criterion, whereas "concurrent" validity pertains to the test's correlation with a contemporaneous criterion. Thus a child's score on a scale of Authoritarianattitudes might have concurrent validity as an index of how sternly he is disciplined at home, and predictive validity for assessing how permissive he will be toward his own children twenty years from now. Similarly, a baseball team's winl loss record at midseason has predictive validity for the team's final standing at the end of the season, and concurrent validity as a clue to the size of its manager's ulcer. Of course, it is barbarous to describe this difference as predictive versus concurrent *validity*, for what is at issue is not the test's validity as such, but its time relation to the criterion. Considering also the danger of confusion between the temporal and atemporal senses of "prediction," it would seem desirable to retire the expressions "predictive validity" and "concurrent validity" in favor of prognostic utility and diagnostic utility, respectively.

Face Validity

Suppose that you are one of the anonymous seers who construct the "personality quizzes" which have become a stock feature in Sunday supplements and popular magazines,³ and that for your next creation you have decided to whip up a test entitled "How strong is your sex drive?" for a men's magazine. If you put together a set of items of the sort

- "My favorite form of literature is (a) poetry, (b) western novels, (c) comic books, (d) pornography."
- "The trouble with women's bathing suits, nowadays, is that they are too (a) immodest, (b) expensive, (c) itchy, (d) opaque."

in which answers such as (d) in the examples given are scored as indicative of a strong sex drive, your test has been constructed to have high *face validity*. That is, a test is said to have "face validity" if intuitively it *looks* as though it should measure what it purports to. Conversely, if your test for sex drive consisted of items pertaining to the subjects' preferences among automobile styles and brands of cigarettes, most of your readers might well fail to perceive that their answers had any relevance to their sexual motivation, in which case your test would be lacking in face validity for them. Another example: A physician's query, "How do you feel?" has considerably more face validity as a test of his patient's physical health than does a white blood cell count or urinalysis.

³To the immense relief of all persons interested in psychology as a serious intellectual discipline, these seem to be waning in popularity in favor of articles on sex and space exploration.

It must be understood, however, that the "face validity" of a test has nothing to do with the objective relation between the test and what it measures. It merely reflects a psychological attitude toward the test by persons who contemplate it. (Usage of the expression, "face validity," is unclear as to whether the test's face validity should be said to vary from person to person according to how much credence that particular person places in the test, or whether it is to be understood as an average credibility within some relevant population of contemplators.) Clearly, this need have little bearing on the correlation between test and criterion. Thus it might be argued (how correctly need not concern us) from a psychoanalytic view of human motivation that persons who give conspicuously lustful responses to the questions in your sex-drive test are actually compensating for subconscious doubts about their sexual adequacy, whereas the sophisticated advertising behind competitive styles of automobiles and cigarettes have invested them with such potent symbolic values that preferences among them may richly betray a person's deepest motives, especially sexual ones. In general, human resources in honest ignorance, gullibility, self-deception, and plain stupidity are too fulsomely abundant to treat the intuitive credibility of a test as any useful sign of its objective validity. Thus with one exception, face validity has little importance for the theory of prediction.

The exception to the general irrelevance of face validity lies in the area of psychological tests on which the subject's interpretation of the test's significance may influence his performance and hence affect the test's objective validity. For example, most tests of abilities and aptitudes present the subject with a set of problems on which the quality of the subject's performance is a useful index of his ability only if he is sufficiently interested in doing well. Consequently, if the subject does not credit the test with rele ance to the purpose for which he is taking it, his effort may be inadequate and his performance hence misleading. Conversely, it is frequently important that face validity be held to a minimum on personality tests (real ones, that is), since the usefulness of an item for the intended purpose of the test may be weakened if the subject's response is biased by the way he would *like* to be viewed by himself or others. Thus on a test for honesty used by an employer to screen prospective personnel, it is perhaps questionable whether all applicants with felonious intent would answer truthfully to

"I frequently feel an urge to steal things. (True or False?)"

whereas a more subtle item such as

"A person is likely to be taken advantage of unless he is careful. (True or False?)"

might correlate reasonably well with an operational criterion of dishonesty even while affording the subject little clue as to how his response will be interpreted.⁴

Interpretive Validity

There is still another facet to the distinction between apparent and objective validity which has been oddly neglected in the test-theory literature. In keeping with the customary phrase form for validity concepts we may call this aspect of test appearances *interpretive validity*, though

⁴Of course, there is nothing objectionable in a subject's being able to interpret a question and bias his answer accordingly if the manner in which he does or does not dissemble is what is diagnostically significant about the question. In fact, one widely used personality inventory includes a scale for detecting lack of candor, in which the liar hopefully gives himself away by describing himself as a model of shining virtue on items which confess to certain near-universal human frailties.

a more accurate wording would be "validity of the test interpretation." The intent here is to contemplate the degree to which the conclusions a test user draws from the test scores are in fact correct. This is not at all the same thing as objective validity, for the information which is objectively contained in the test data in the form of reduced contingent criterion variance need not be properly extracted by the test user. That is, what a test interpreter *thinks* a given test score implies is not necessarily the same as what it *does* imply.

Suppose, for example, that scores on test **X** are converted into estimates of a criterion variable **Y** by a linear prediction equation $\hat{\mathbf{Y}} = a\mathbf{X} + b$ in which constants *a* and *b* are not necessarily optimal. The standard error of this policy is

(5.1)
$$\operatorname{SE}_{\mathbf{Y},\mathbf{X}}[\hat{\mathbf{Y}} = a\mathbf{X} + b] = \sqrt{(\mathbf{M}_{\mathbf{Y}} - a\mathbf{M}_{\mathbf{X}} - b)^2 + \sigma_{\mathbf{Y}}^2 + a^2\sigma_{\mathbf{X}}^2 - 2a\sigma_{\mathbf{X}}\sigma_{\mathbf{Y}}r_{\mathbf{X}\mathbf{Y}}}$$

(from (4.50), (4.51-M), and (4.51-V)). We know that optimal choice of *a* and *b* can reduce this standard error to a minimum of $\sigma_{\mathbf{Y}} \sqrt{1 - r_{\mathbf{XY}}^2}$. However, even when it has been possible to obtain a sampling approximation to the joint distribution of **X** and **Y** in the relevant population this still yields only an approximation to the optimal *a* and *b* while in testing practice, objective validity data are usually (1) nonexistent, (2) taken from populations other than the one of immediate concern, or (3) ignored. Hence the interpretive validity of a linear prediction policy will always be poorer than is indicated by the test's objective linear validity, perhaps considerably so. This observation also applies, of course, to prediction policies of any form if we measure objective validity by the curvilinear correlation coefficient, and in fact one especially important way in which interpretations can go astray is through linear inferences from a test whose relation to the criterion is markedly curvilinear.

To appreciate the dangers which lurk in test usage when interpretive validity diverges sufficiently from objective validity, observe from (5.1) that in order for the standard error of prediction policy $\hat{\mathbf{Y}} = a\mathbf{X} + b$ to be no greater than the standard deviation of \mathbf{Y} , it is necessary to have

$$a^{2} \leq 2a \left(\frac{\sigma_{\mathbf{Y}}}{\sigma_{\mathbf{X}}} \mathbf{r}_{\mathbf{X}\mathbf{Y}} \right) - \left(\frac{\mathbf{M}_{\mathbf{Y}} - a\mathbf{M}_{\mathbf{X}} - b}{\sigma_{\mathbf{X}}} \right)^{2},$$

which in turn requires both that a have the same sign as r_{XY} and that

(5.2)
$$|a| \le 2 \left| \frac{\sigma_{\mathbf{Y}}}{\sigma_{\mathbf{X}}} \mathbf{r}_{\mathbf{X}\mathbf{Y}} \right| \qquad (\mathrm{SE}_{\mathbf{Y},\mathbf{X}} [\hat{\mathbf{Y}} = a\mathbf{X} + b] \le \sigma_{\mathbf{Y}})$$

Now, $(\sigma_{\mathbf{Y}}/\sigma_{\mathbf{X}})\mathbf{r}_{\mathbf{XY}}$ is the regreession weight of \mathbf{X} for prediction of \mathbf{Y} and is the value of a when $SE_{\mathbf{Y},\mathbf{X}}[\hat{\mathbf{Y}} = a\mathbf{X} + b]$ is minimal. Hence if the interpretive weight given to the test scores—i.e., the coefficient of \mathbf{X} in $\hat{\mathbf{Y}} = a\mathbf{X} + b$ —exceeds its optimal magnitude by more than a factor of 2 (or worse, if it has the wrong sign), the resulting criterion estimates will be less accurate, on the whole, than if scores on \mathbf{X} are disregarded and the criterion's mean is taken as the estimate for all subjects. What gives this conclusion especial poignancy is that a great many test users disregard regression principles when interpreting test scores and naively take a subject's rank on the test as his estimated rank on the criterion. This amounts to estimating \mathbf{Y} by the policy $\hat{\mathbf{Z}}_{\mathbf{Y}} = \mathbf{Z}_{\mathbf{X}}$, which may be read into (5.1) and (5.2) by putting $a = 1, b = 0, \sigma_{\mathbf{Y}} = \sigma_{\mathbf{X}} = 1$, and $M_{\mathbf{Y}} = M_{\mathbf{X}} = 0$. On the other hand, predicting that each subject is average on the criterion is to adopt the policy $\hat{\mathbf{Z}}_{\mathbf{Y}} = \mathbf{M}_{\mathbf{Z}_{\mathbf{Y}}}$ whose standard error is $\sigma_{\mathbf{Z}_{\mathbf{Y}}}$ and by (5.2) is superior to policy $\hat{\mathbf{Z}}_{\mathbf{Y}} = \mathbf{Z}_{\mathbf{X}}$ unless $r_{\mathbf{X}} \ge .50$. Thus *if a test's objective validity for a given criterion is less than*

.50, treating unregressed standardized test scores as estimates of the similarly standardized criterion scores results in greater predictive inaccuracy than simply taking the criterion's mean for all estimates. (Of course if the use being made of the test is not a point estimate—e.g., if its application is to select subjects which surpass a certain criterion level—then this conclusion does not apply.)

We see, therefore, that while objectively a test can be no worse than irrelevant, the validity or rather, invalidity-of an inappropriate interpretation can make the test results not merely useless but actually detrimental. When the test interpretation consists in point estimates of the criterion, as assumed in the preceding paragraph, the test's objective validity is an upper bound on its interpretive validity if the latter is defined for this case as $\sqrt{1 - (SE_{Y,X}/\sigma_Y)^2}$. (The imaginary values assumed by this measure when $SE_{Y,X} > \sigma_Y$ seem wryly appropriate.) Point estimates are not the only way to interpret test scores, however, and there is an important sense in which interpretive validity can in principle approach perfection even though objective validity remains modest. Specifically, if we are content to forego categorical conclusions in favor of *probability* judgments in our interpretation of test scores, then it is possible for the degree of likelihood we attach to an inference about the criterion to be in essential agreement with the criterion's objective uncertainty given the test data. For example, suppose that the joint distribution of test X and criterion Y in the population under concern is normal, and that our mode of test interpretation is to make interval estimates of Y. If, given a subject's test score, we infer simply that his Z-score on Y lies in the interval $r_{XY}Z_{XY} \pm d$ for some given half-interval width d, then if P_d is the proportion of a normal distribution which lies within d/k_{XY} sigma units of the mean, there is a probability of $1 - P_d$ that our inference is wrong. But if our *conclusion* from the test score is that Z_Y has a $100 \times P_d$ % chance of falling within the interval $r_{XY}Z_{XY} \pm d$, then there is no way in which the test interpretation involves error and hence no imperfection of interpretive validity. In practice, of course, an interval estimate accompanied by a certain degree of subjective configence will not necessarily have been assigned the *correct* uncertainty, so this more sophisticated sort of interpretive validity can also fall considerably short of ideal, even to the point of confidence judgments which are more unrealistic than the uncertainty which would be felt about the criterion were there no test scores at hand to misinterpret. We may conclude, then, that whatever objective validity a test may have for a given criterion, the validity (i.e., accuracy) of a particular *interpretation* to which it is subjected may range anywhere from unblemished veridicality to oaf-handed abuses in which it were better that the test not be consulted at all.

The Definition of the Criterion: A Darkening Mystery

So far, our account of prefix-validities has dealt with relatively simple issues, but now the story grows more somber. With the sole exception of a moment's hesitation over the nature of "baldness" in the discussion of empirical validity, we have spoken throughout this book as though we were dealing with well-defined variables—variables, that is, whose various values have all been conceptually identified, so that we can name a particular value and know at least what it would *mean* to assert that some individual has this score, even though there might be practical difficulties in actually determining whether the assertion is true or not. (For example, we know perfectly well what it would mean to say that the earth is 7,895 miles in diameter at the poles. It means that the smallest number of mile-long measuring sticks laid end to end and passing through intervening obstacles as necessary that would span the distance from one

pole to the other is 7,895, though of course it would prove awkward to prove or disprove this claim by direct methods.) However, a little critical reflection on many of the illustrative test criteria cited previously reveals how idealistic this supposition actually is. We casually spoke, for example, of tests for disciplinary sternness, sex drive, physical health, and honesty. But really, now, have we any idea of what it is that we would be testing for? What are the different values assumed by the Sternness-of-discipline or Degree-of-honesty variables, and how, even in principle, could we determine which particular value holds for a given individual? What would it be like to have a Sex-drive score of 26 on some scale—any scale—of this variable? And if Jimmy Jones's best friend just gave him a bloody nose, how many points does this subtract from his index of Physical-health?

When confronted with vagueness of this sort in our prior Degree-of-baldness example, it was argued that since the ordinary-language concept of "baldness" has no definite meaning, we are free to replace this notion with any more precise, operational definition which serves our need, or at least with any that captures the spirit of the original. But unfortunately, despite the philosophical hopes of an earlier era (circa 1920-50), explicit operational definitions based on what is directly knowable are of but limited service (though where appropriate, they remain indispensible) in making commonsense ideas scientifically respectable. In the special case of "baldness," the ordinary intent of this notion makes apparent that if we had a complete description of the distribution of hair on a person's head, no conceivable additional data could further assist us in deciding how bald that person is;⁵ hence all that is lacking for a precise concept is simply for us to make up our minds about what abstraction from hair distribution is to count as "baldness."⁶ However, what set of observations would suffice as an abstraction basis for a clarified concept of, say, "honesty"? Presumably, our decisions about how honest a person is are based on the way he behaves when confronted with certain opportunities for self-benefit or self-protection at the expense of ethical ideals, such as finding a well-stocked billfold with the owner's name in it, or being in control of the ballot box after a close election, or being asked whether the rumor about his behavior at last night's party is really true. But for any list of honesty-testing situations we might compile, no matter how extensive, we could always conceive of still other situations in which a person's behavior, were we to know it, would further modify or confirm our judgment about his honesty. Even more awkwardly, what are we to say about the differential honesties of persons at times when they are not actually in situations which allow their honesty to be manifested. We would normally hold, say, that John Teach Smith and his wife Angelica continue to be a black-hearted scoundrel and a paragon of virtue, respectively, even while they are both sleeping peacefully in bed. Whatever we intend to denote by "honesty," it must be something other than an abstraction from the observable behaviors and circumstances from which we infer it. But if we cannot explicitly *define* honesty by an operational definition over observable attributes, but can only say that honesty, whatever it may be, is something whose symptoms are certain aspects of a person's behavior in appropriate circumstances, how are we to justify ever taking these behaviors in these circumstances as symptoms of honesty in the first place? Obviously we cannot establish their diagnostic validity empirically, for we have no knowledge about the criterion in this case apart from that which the predictors themselves afford. And it may as well be stated frankly here and now that no satis-

⁵We ignore the possibility that the circumstances under which a person loses his hair—e. g., having his head shaved—is considered relevant to whether or not he is "bald" at a given moment.

⁶In light of this consideration, it would seem that of the various "operational definitions" for baldness suggested earlier, only Hair-density is really a legitimate definition of the Degree-of-baldness variable.

factory answer to this problem will be forthcoming, or even attempted, in this book. Search for clarification and explanation of how it is that we are apparently able to devise concepts which penetrate beyond what has been observed, and succeed after a fashion in acquiring knowledge about realities inaccessible to our direct experience, is one of the most active quests of advanced contemporary research in philosophy and scientific method; and while some answers are little by little beginning to take shape, any serious discussion of the issues is grossly beyond our present scope.⁷ We can and shall, however, have a look at the manner in which test theorists, in their own humble⁸ way, have grappled with these matters as they bear upon the interpretation of test scores.

To press onward, then, the concepts of "content validity" and "construct validity" derive from considerations of how we might justify inference from test observations to criteria about which we have no knowledge except *by means of* these tests or similar observations. That is, when confronted with a problem of content or construct validity, we cannot appeal to an empirically determined correlation between test and criterion but must provide an argument for why it is reasonable to interpret the test as a measure of something else which, in general, can be observed only indirectly. In cases to which "construct validity" applies, there is an unabashed postulation of unobserved or "theoretical" variables which are assumed to be more or less casually responsible for the observed test scores. "Content validity"," on the other hand, pertains to an older, less audacious, form of construction in which the criterion does not, prima facie, differ in kind from the test itself (or at least not from the "true score" component discussed later—of the test variable), and in some cases may even be in principle directly determinable.

Content Validity

By the "composite" of a person's scores on several different variables, let us mean the merger of these scores by some fixed procedure into a single resultant score for that person. The distribution of these composite scores in a given population is then the distribution of a "composite variable" which is defined by a certain function of its constituent variables. (In physics, for example, "Density" is a composite variable whose constituent variables are Mass and Volume, since an object's density is its mass divided by its volume.) A test may then be said to have content validity when (a) the criterion variable is a composite of all the variables of a certain specified kind, which will be called the "domain" of the criterion, and (b) the test obtains a composite score (or, in the limiting case, a single score) on a samPle of the criterion's domain.⁹ For instance, how anemic a person is may be defined by the density of red cells in his blood. Direct determination of this by counting all the red cells in the subject's blood and dividing by its volume would be difficult to manage technically and would greatly inconvenience the subject. However, it is quite practical to count the red cells in a small drop of the subject's

⁷The reader who would like to acquire some familiarity with the more respectable literature on these matters may get off to an excellent start by browsing in such notable works as Feigl and Broadbeck (1953), Braithwaite (1953), Feigl, Maxwell, and Scriven (1956, 1958, 1962), Nagel (1961), Feigl and Maxwell (1961), and Pap (1962). ⁸Net reselve

⁸Not really.

⁹Content validity has often been characterized only as a test's more or less representative sampling from a domain of items, with no concern expressed by the discussant for what composite over that domain is to be considered the criterion, or even signs of recognition that the concept of "content validity" requires some such composite (see, e.g., Lennon, 1956). What such accounts construe a content-valid test to be a predictor of is never made clear.

blood and to take the computed density in this sample as an estimate of the red-cell density in the aggregate of all his drops of blood. Again: On what grounds may a list of spelling words as administered, say, by an English teacher to her students, be justified as a test of English-spelling ability-not just as a test of ability to spell the particular words on the list, mind you, but as a test of English-spelling ability without restriction? Simply that if "English-spelling ability," unrestricted, is defined as a person's total ability to spell all the words of the English language, then performance on the test exhibits at least a part of a person's ability on the total. Or suppose that for the edification of entering Freshmen, the fraternities at Brainsweat University develop a checklist of ten items of offensive professorial behavior on which to rate members of the faculty for how big an s.o.b. each one is. Clearly, the ways in which a professor can be an s.o.b. are not limited to any ten items, even if these include, say, how strict he is about cutting class, how frequently he gives written assignments and surprise quizzes, how tough it is to pass his course without studying for it, and how nasty he gets about cribbing on exams and term papers, reading newspapers or matching pennies during class lectures, and other natural exuberances of misunderstood youth. Even so, these are certainly important contributors to a professor's total s.o.b. stature, and if professor P rates substantially higher than professor U on the checklist, it is unlikely that this order is reversed on the complete criterion. In each of these three examples-blood count, spelling test, and s.o.b. checklist-it will be seen that the test variable comprises a selection from the constituents which, in aggregate, make up the criterion, so that a person's score on the test logically contributes to his score on the criterion. Such a test has "content validity" because its contents, or components, are included in the criterion's domain.

There is a great deal which can be said, mathematically and methodologically, about the subject of content validity, more than has appeared in the technical literature and certainly more than can satisfactorily be discussed here, especially since we have eschewed consideration of statistical sampling. We should, however, call attention to certain difficulties which make appraisal of a test's significance from the standpoint of content validity considerably more problematic than seems generally to be appreciated.

To begin with, content validity guarantees for a test only that its items are *logically relevant* to the criterion in virtue of actually being ingredients of the latter. By itself, this implies very little about the statistical correlation between test and criterion. It is possible for a sample to be highly representative of what it samples (e.g., blood-sampling as a test for anemia), but it can also fail abjectly to do so. For example, if a test of English-spelling ability were to contain only names of commercial products heavily advertised on TV, or technical terms of some highly specialized profession, we might well expect to find poor correlation between the test and other less biased measures of English spelling. The *face* validity of a test which samples its criterion¹⁰ can easily seduce its user into negligence about the test's more important predictive credentials. Actually, content validity is a reasonable basis on which to accredit a test only if it meets minimal standards of representative sampling from the criterion's domain.

Secondly, a fundamental but outstandingly neglected prerequisite for judging a test's content validity is specification of just what the composite criterion *is*. The composition of the test itself does little to clarify this, for there are an unlimited number of potential composite criteria whose components are simultaneously sampled by the test. Thus the spelling test given to her sixth-grade pupils by Miss Smith presumably has content validity for, among other things, (1) Sixth-grade-spelling (defined, say, by what Miss Smith's school system expects a sixth grader

¹⁰Note that while face validity and content validity are not the same, a test which has the latter is very apt to have the former as well.

to know of spelling), (2) the ability to spell all words known to Miss Smith, (3) Commandof-commercial-spelling (i.e., ability over a domain of words specified, say, by the graduation requirements of the local business school), (4) English-spelling ability (without restriction), (5) Sixth-grade-scholastic-achievement (i.e., a composite of all things which Miss Smith has tried to teach her sixth graders), and (6) General-competence, this last being a composite of how good a person is at everything. Moreover, it takes much more than a few commonplace phrases (as have been used in the present examples) to specify the domain of a composite variable with any useful precision, and definitions of composites which determine, either explicitly or by implication, exactly what variables do or do not fall within their scope virtually never occur in practice. Even in unusually simple cases such as spelling abilities, whose constituents can explicitly be enumerated by reference to extant word lists such as dictionaries, the final operational definition is seldom actually carried out. What is more likely is that the composite's domain presents near-insuperable problems of delimitation to begin with. What items, for example, are comprised by "arithmetic ability," an expression used frequently in everyday discourse to denote, presumably, a composite measure of a person's abilities at various problems in arithmetic? For even the most elementary forms of arithmetic, it is impossible to enumerate the different possible problems of that form individually - e.g., the phrase, "the ability to add n_l and n_2 ," generates descriptions of an infinitude of different specific abilities as various numbers are substituted for " n_l and n_2 " A good start toward defining the domain of "arithmetic ability" would be to list the various problem forms which are to be classed as arithmetic—e.g., problems of the forms $n_l + n_2$, $n_l \div n_2$, $n_l \times n_2 - n_3$, etc. But how does one list all the forms of arithmetic? And where does arithmetic leave off and more complex mathematics begin- is

"Find to the 20th decimal the limit of $\sum_{i=1}^{n} 1/i^2$ as *n* approaches infinity," or "What is 69.13 raised to the 172th power?" or "Find the cube root 4," a problem in *arithmetic* or not? What restrictions, if any, are to be set on the numbers on which the arithmetic forms operate-e.g., do or do not "What is the sum of π and e?" and "What is the sum of $3 + 2\sqrt{-1}$ and $17/4 - \sqrt{-1}$?" count as arithmetic instances of the form "What is the sum of n_1 and n_2 ?" And how, in defining our domain of specific items, are we to deal with different ways of presenting a problem-e.g., do "Multiply 7 by 2," "Multiply 2 by 7," "What is seven times two?" and "What is 2×7 ?" define four different specific abilities or only one? (After all, a person who is shaky on the fact that $n_l \times n_2 = n_2 \times n_l$ or on the equivalence of "7" with "seven" and " \times " with "times" might well be able to work the problem in one form but not in another.) In the case of mathematical problems, moreover (and if definition of "arithmetic ability" presents problems, think how much worse the more general composite, "mathematical ability" must be), we at least feel intuitively that we can recognize one when we see it, even if precise definitions elude us. But how often can we honestly feel even this minimal assurance about the meaning of composite concepts actually in use. Consider, for example, the important variable "Presidential-suitability," meaning a composite of a person's qualifications for the Presidency of the U.S. Certainly we periodically. judge how various men of affairs stand on this variable, often with high conviction and strident fervor, yet have we really any notion of how to circumscribe what goes into the determination of Presidential-suitability?

To compound the woes of defining a composite criterion, even if its domain has been successfully identified, a decision must still be reached about the manner in which the composite is formed. Is, e.g., a person's English-spelling ability a simple arithmetic mean of his abilities on the individual words, or do some words receive more weight than others -or should the com-

posite be defined by an even more complicated nonlinear function over the individual words it comprises? The question of weighting may seem to be of small consequence when dealing with a relatively homogeneous domain such as that of spelling problems, but it appears much less so when the domain includes different types of items. For example, should a problem in long division be worth no more in determining what is meant by "arithmetic ability" than a simple ability to add two digits? And how should a man's personal charm be weighed against his decisiveness when evaluating his Presidential-suitability? This issue of differential weighting also draws attention to uncertainties about the definition and scaling of the individual variables in the composite's domain. How, for example, do we scale the ability to spell "hippopotamus"? Is it a two-valued, or "dichotomous," variable-either the subject can spell it or he can't-or do we allow for gradations in this ability? And for that matter, what do we mean by "the ability to spell 'hippopotamus,' " anyway? We test for it by confronting a subject with the problem and observing his response; yet uttering the series of sounds, "aich-eye-pee-pee-oh-pee-oh-teeaye-em-you-ess," is not logically equivalent to having the ability to spell "hippopotamus," but is merely an imperfectly correlated symptom thereof. We shall have a bit more to say about the nature of abilities later. Here the intent is merely to call attention to the un palatable fact that over and above problems of domain and weighting, the definition of a composite variable also suffers from whatever vagueness and ontological uncertainty may adhere to any of its constituent variables. To think, then, that we likely have much idea of what we are talking about when we appraise the "content validity" of some particular test is simply open-mouthed naivete.

Because many test theorists have held composite criteria to be more respectably in keeping with the spirit of scientific empiricism than are inferred theoretical criteria, it is also worth observing that even if a composite were to be so clearly defined that the value of anyone of its constituent variables could readily be observed for a given individual, it would still be impossible, in most cases of practical interest, to determine scores on the composite variable except by inference from scores on a test which samples only a small portion of the domain. Even so simple a composite criterion as the ability to add two integers comprises an infinity of individual items, only 0% of which can actually be tested directly "insomuch as any finite number, no matter how large, is still only 0% of an infinite totality. In fact, even when the domain of the composite is finite, as in the case of English-spelling ability, an attempt actually to observe the entire domain by successive presentation of all the individual items would introduce practice effects, fatigue, etc., which should modify the character of the later items to a greater or lesser extent. (Thus a person's ability to spell "hippopotamus" now is not necessarily altogether the same as his ability to spell "hippopotamus" 30 seconds from now after having first been asked how to spell "sphygmomanometer.") It is a mistake, then, to think that composite variables defined over a domain of variables which individually raise no special problems of observation are free from the taint of nonobservability which stigmatizes the theoretical variables contemplated by "construct validity" (see below).

Suppose, then, that we have identified a set of n_C variables X_1, \ldots, X_{n_C} whose sum is taken to define a composite criterion C, i.e.,

(5.3)
$$\mathbf{C} =_{\text{def}} \mathbf{X}_1 + \ldots + \mathbf{X}_{n_c}$$
$$= \sum_{i=1}^{n_c} \mathbf{X}_i.$$

Also, let **T** be a test formed by summing \mathbf{n}_{T} items, say $\mathbf{X}_1, \ldots, \mathbf{X}_{\mathsf{n}_{\mathsf{T}}} (\mathbf{n}_{\mathsf{T}} \leq \mathbf{n}_{\mathsf{C}})$, from **C**'s domain,

i.e.,

(5.4)
$$\mathbf{T} =_{\text{def}} \mathbf{X}_1 + \ldots + \mathbf{X}_{n_T}$$
$$= \sum_{i=1}^{n_T} \mathbf{X}_i.$$

(It might seem more psychologically satisfying to define C and T as averages, rather than sums, of the items they comprise, but this would modify the present definitions only by multiplication of each by a constant, namely, $1/n_{C}$ and $1/n_{T}$, respectively, and would hence change only the units of measurement for raw scores on T and C.) Clearly test T has content validity as a measure of criterion C, but what we would now like to know is the *objective* validity of T for **C**.

To keep the analysis as simple as possible, we assume that all variables in the domain of C have unit variance and that the interitem correlations are all equal-i.e.,

(5.5-ij
$$(\mathbf{i}, \mathbf{j} = 1, \dots, \mathbf{n_c})$$
) $\operatorname{Cov}(\mathbf{X}_{\mathbf{i}}, \mathbf{X}_{\mathbf{j}}) = \begin{cases} 1 & \text{if } \mathbf{i} = \mathbf{j} \\ \\ r & \text{if } \mathbf{i} \neq \mathbf{j} \end{cases}$ (assumed)

The notation " $\mathbf{i}, \mathbf{j} = 1, \dots, \mathbf{n}_{C}$ " means that **X** the indices **i** and **j** run over all combinations of an assignment to i from 1 to nc with an assignment to j from 1 to nc, so formula (5.5-ij) actually represents $\mathbf{n}_{\mathbf{C}}^2$ different equations. For $\mathbf{i} = \mathbf{j}$, assumption (5.5- $\mathbf{i}\mathbf{j}$) yields $Cov(\mathbf{X}_{\mathbf{i}}, \mathbf{X}_{\mathbf{i}}) = Var(\mathbf{X}_{\mathbf{i}}) = Var(\mathbf{X}_{\mathbf{i}})$ $\sigma_{\mathbf{X}_{i}}^{2} = 1$, while for two different items \mathbf{X}_{i} and \mathbf{X}_{j} , $r_{\mathbf{X}_{i}\mathbf{X}_{j}} = \text{Cov}(\mathbf{X}_{i}, \mathbf{X}_{j})/(\sigma_{\mathbf{X}_{i}}\sigma_{\mathbf{X}_{j}}) = r/(1 \times 1) = r$. It follows immediately from (4.40)¹¹ and assumptions (5.5) that

(5.6/5.5¹²)

$$Var(\mathbf{T}) = \mathbf{n}_{\mathsf{T}}(\mathbf{n}_{\mathsf{T}} - 1)r$$

$$= \mathbf{n}_{\mathsf{T}}(\mathbf{n}_{\mathsf{T}}r + 1 - r)$$
(5.7/5.5)

$$Var(\mathbf{C}) = \mathbf{n}_{\mathsf{C}}(\mathbf{n}_{\mathsf{C}} - 1)r$$

$$= \mathbf{n}_{\mathsf{C}}(\mathbf{n}_{\mathsf{C}}r + 1 - r)$$

while using $(4.39)^{13}$,

(5.8/5.5)

$$\operatorname{Cov}(\mathbf{T}, \mathbf{C}) = \operatorname{Cov}\left[\mathbf{T}, \mathbf{T} + \sum_{i=n_{T}+1}^{n_{C}} \mathbf{X}_{i}\right]$$

$$= \operatorname{Var}(\mathbf{T}) + \operatorname{Cov}\left(\sum_{i=1}^{n_{T}} \mathbf{X}_{i}, \sum_{i=n_{T}+1}^{n_{C}} \mathbf{X}_{i}\right)$$

$$= \operatorname{Var}(\mathbf{T}) + \mathbf{n}_{\mathsf{T}}(\mathbf{n}_{\mathsf{C}} - \mathbf{n}_{\mathsf{T}})r$$

$$= \mathbf{n}_{\mathsf{T}}(\mathbf{n}_{\mathsf{C}}r + 1 - r).$$

 11 equation 4.40.

¹²To continue our practice of parenthetical annotation of restrictive conditions which hold on stated formulas, we shall list the index number of prior equations which assert these restrictions when such equations have been given.

¹³equation 4.39

Hence,

(5.9/5.5)
$$r_{TC}^{2} = \frac{\text{Cov}(T, C)^{2}}{\text{Var}(T) \cdot \text{Var}(C)}$$
$$= \frac{\mathbf{n}_{T}(\mathbf{n}_{C}r + 1 - r)}{\mathbf{n}_{C}(\mathbf{n}_{T}r + 1 - r)}$$

where we show r_{TC}^2 rather than r_{TC} to save the nuisance of a square root sign.

Formula (5.9) is not very communicative as it stands, but when properly caressed, it confides interesting secrets. One is that if = 0

(5.10/5.5;
$$r = 0$$
) $r_{TC}^2 = \frac{n_T}{n_C}$

which says that if the items in the domain of a composite C all have zero intercorrelations, the validity of test T for criterion C is the square root of the proportion of items composing C which are sampled by T. If the variables in C's domain are unrelated to one another, therefore, T will have negligible validity for C unless the test contains a sizable proportion of the criterion's domain.

For most composite concepts of practical interest, however, the domain is much, much larger than the number of items on any test of manageable length. To see what happens in this case, observe that (5.9) can be put into form

(5.11/5.5)
$$\mathbf{r}_{\mathsf{TC}}^2 = \frac{\mathbf{n}_{\mathsf{T}}r + \epsilon}{\mathbf{n}_{\mathsf{T}}}(r+1-r)$$

where

$$\epsilon =_{\rm def} \frac{\mathbf{n}_{\mathsf{T}}}{\mathbf{n}_{\mathsf{C}}}(1-r),$$

and is insignificant if the proportion of **C**'s domain included in **T** is sufficiently small. That is, assuming *r* to be nonnegative,¹⁴

(5.12/5.5;
$$r \ge 0$$
) $r_{TC}^2 = \frac{\mathbf{n}_T r}{\mathbf{n}_T r + 1 - r}$

in which the inequality approaches an identity as the sampling proportion $\mathbf{n}_T/\mathbf{n}_C$ approaches zero. Hence if r > 0, the validity of **T** for **C** rapidly approaches unity as the number, \mathbf{n}_T , of test items increases, even when **T** is relatively low. This fact can be appreciated most readily by computing the number of test items required to attain a fixed level of validity. Solving (5.11) for \mathbf{n}_T shows that

(5.13/5.5;
$$r \ge 0$$
) $\mathbf{n}_{\mathsf{T}} \le \left(\frac{\mathbf{r}_{\mathsf{TC}}^2}{1 - \mathbf{r}_{\mathsf{TC}}^2}\right) \left(\frac{1 - r}{r}\right) < \left(\frac{\mathbf{r}_{\mathsf{TC}}^2}{1 - \mathbf{r}_{\mathsf{TC}}^2}\right) \left(\frac{1}{r}\right)$

which says that the number of test items necessary to achieve a given validity is less than proportional to the reciprocal of the interitem correlation r. (From (5.13) it may be observed, e.g., that even when the interitem correlations are as small as r = .10, less than 10 test items are needed for a validity as high as $\sqrt{.5}$, or .71.) Hence so long as the variables in the domain of a composite criterion do not have essentially zero correlation with one another, it is fairly easy,

¹⁴Negative r is a pathological case which has little if any practical importance.

in this simplified case at least, for a content-valid test to yield a respectably accurate estimate of the criterion. Moreover, correlation of test T with any unsampled item $X_k(k > n_T)$ in C's domain is easily seen (the proof may be left to the reader) to be

(5.14/5.5)
$$r_{\mathsf{TX}_k}^2 = \frac{\mathbf{n}_{\mathsf{T}}r^2}{\mathbf{n}_{\mathsf{T}}r + 1 - r}$$

or from (5.14) and (5.11) when the sampling proportion is negligible,

$$(5.15/5.5, \mathbf{n}_{T}/\mathbf{n}_{C} = 0)$$
 $r_{TX_{k}} = r_{TC} \sqrt{r}$

We may also quickly see by adapting (5.11) to the limiting case where $\mathbf{n}_{T} = 1$ that the correlation between the full composite C and any item \mathbf{X}_{k} its domain is

(5.16/5.5)
$$r_{CX_k} = \sqrt{r + \frac{1-r}{n_C}}$$

or simply $r_{CX_k} = r$ when the domain is very large. Hence if test T samples a moderately large number of items from the criterion's domain (even if the sampling *proportion* is vanishingly small), it yields an almost perfect measure of the criterion, while its correlation with any single unsampled item in the criterion's domain is approximately \sqrt{r} , which is also the criterion's own correlation with this constituent.

The most critical simplifying restriction in the preceding analysis is the assumption that all intercorrelations among items in the composite's domain are the same. Even in the more general case, however, these results still hold approximately, with r replaced by the average interitem correlation, so long as the items included on the test are a suitably representative sample of the total domain. We may conclude, then, that a necessary and prima facie sufficient condition for a composite concept defined over a large number of individual variables to be potentially useful is for the average correlation among the latter to be positive. The practical payoff in this case is that any decent estimator of the composite—in particular, a well-constructed test with content validity for this criterion-will also usefully (just how usefully depending on the correlation coherence of the composite's domain) predict various specific items included in the composite as need for such predictions arises. And is this not, in fact, what we expect from our composite concepts in everyday life? When we assess a person's "English-spelling ability" or "arithmetic ability" by observing his performances on a sample of problems, we certainly feel, justifiedly or not, that we have also learned something about how well he can do on still other spelling or arithmetic problems. And when we judge a candidate's "Presidential-suitability" on the basis of presently observed constituents thereof (past governmental experience, political affiliations, charisma, etc.), we are above all hopefully forecasting his probable success at dealing with the various national and international crises which will grace the next term of office. Analysis of composite concepts in everyday use makes unmistakeably clear that a nearuniversal implicit presupposition of these is that the individual elements of the composite's domain are at least moderately valid estimators of one another.

In summary of this rather extended critique of "content validity," then, we have the following conclusions: (1) While no attempt was made to quantify the notion, a test has "content validity" for any criterion conceived as acomposite-i.e., an average or other summary measureover a group of variables of which the items composing the test are a sample. (2) Composite criteria of the sort envisioned in discussions of content validity are practically never defined with greater precision than the commonsense intuitions of a bright ten-year-old child. (3) For virtually all composite criteria of interest, the domain is far too large to permit simultaneous observation of a person's scores on all the variables included therein; hence scores on the criterion cannot be determined except by inference from scores on a content-valid test of it. (4) In general, it is not only useless but misleading to introduce a composite criterion unless we have reason to believe that the variables participating in the composite are positively intercorrelated. This last conclusion, however, has a further important implication. If the variables in the domain are too numerous to observe individually, on what grounds *could* we reasonably believe them to be positively intercorrelated? Only, it would seem, if we surmise *that they have some feature in common which would induce this correlation*.¹⁵ If so, then it may be suspected that a test which is alleged to assess merely a composite over observables is in actuality being surreptitiously construed to measure a *theoretical* variable hypothesized to unify the composite's domain. Hence consideration of content validity ineluctably feeds into the problems of *construct* validity.

Construct Validity

It was previously claimed that a great many concepts of science and everyday life refer to entities which cannot themselves be perceived, even in part, but which we are somehow able to learn about by inference from their effects on the things which we can observe. For example, consider once again the attribute of being able to spell "hippopotamus." As pointed out earlier, the *ability* to do this is not the same thing as the actual *doing* of it, for either can occur without the other. Thus the reader is perfectly capable, presumably, of spelling "hippopotamus" i.e., this ability is an attribute which he has right at this very moment—even though he is not now actually vocalizing or writing this series of letters. Conversely, a person whose literary talents are less advanced than the reader's might recite the sequence "h-i-p-p-o-p-o-t-a-m-u-s" by reading the letters one by one from a prompt card his remedial spelling teacher has placed before him and still remain bereft, even during the act of vocalization, of what is normally meant by "the ability to spell 'hippopotamus'." And yet, it is only through observations of what a person *does* when confronted with the task of spelling this word that we come to know his ability to spell it. The epistemological relevance of the act to the ability is that Degree-ofability-to-spell-"hippopotamus" is a theoretical variable postulated specifically to account for the observed fact that given equal opportunity and incentive to spell "hippopotamus," people characteristically differ from one another in the way they respond. When one person does consistently better than another on a problem where the difference cannot be ascribed to a systematic bias in the conditions under which the problem is presented, we must assume - or least we inevitably do assume - that this is due to some underlying difference in the persons themselves which persists (though not necessarily without hope of modification) even at times when they are not actually doing the problem. Performance scores obtained under the standard test-circumstances are then a measure of the ability variable postulated to account for them. Such a test is said to have *construct validity* because, while test and criterion are logically distinct from one another, if we did not assume the test to have validity for the criterion, we

¹⁵If the variables included in a good-sized sample of the domain are found to be significantly correlated, then it seems reasonable to infer that this intercorrelation undoubtedly pervades the entire domain. But a good case can be made for the position that it is precisely the likelihood, given such an observation, that one or more unifying factors underlie the domain which gives the statistical inference its tenability.

would have no good reason for assuming the criterion to exist at all. That is, the observed test-circumstance \rightarrow test-result connection is the evidential basis (or at least part of it) for the concept of the criterion in the first place.

By definition, then, a test has "construct validity" when it is construed to be a test of some variable hypothesized to explain a given body of data of which the test behavior is itself a paradigm instance. It is only recently that this theory-dependent characteristic of tests has been made explicit,¹⁶ for it derives from an interpretation of scientific concepts which was vehemently rejected by the positivistic views of scientific knowledge that were ascendent during the first half of this century.¹⁷ It is now coming generally to be agreed,¹⁸, however, that much of scientific and commonsense belief pertains to entities - not just "convenient fictions," but genuine existents - which we learn about only indirectly through their effects on what can be observed. The detailed nature of the process by which we acquire such knowledge is as yet understood but crudely.¹⁹ Even so, one facet which seems clear is that, however justified, an especially potent source of inference to underlying determinants is a pattern of observed interrelationships. The postulation of "abilities" to help explain the effect of problem presentations on a person's manifest behavior is one illustration of this. Actually, abilities are only one instance of a vast array of low-grade theoretical attributes technically known as "dispositions" and identified by means' of (not identified with) a person's or object's re4ction to particular test conditions. Thus "fragility" refers to whatever state of a chunk of matter is responsible for its tendency to shatter when subjected to sudden stress, while "combustibility" accounts for why some but not all objects burst into flame when heated.²⁰ A somewhat different case of inference to theoretical entities arises when an array of observation variables manifests a provocative pattern of empirical correlations. This situation has been the inspiration for much of the advanced research in psychometrics and test theory during the past several decades, though the mathematical developments thereof have reached a much higher level of sophistication than have their methodological foundations. Of these developments, the most notable has been factor analysis.

"Factor analysis" is a statistical technique by which to discern within the confusingly manifold interrelations of an aggregate of empirical variables the concealed presence of a smaller, more elegant, array which may (or then again, may not) explain why the observed variables have the correlational affinities they do. To return for illustration to the domain of spellng abilities, suppose that upon analysis of the scores obtained from some population of subjects on a long test of spelling words, it is found that quality of performance on each individual word of the test has a correlation of .25 with performance on every other. Our first inferential move is to treat a subject's performance on each word as a measure (though not necessarily a perfect one)

¹⁶The term "construct validity" first made its appearance on pp. 14ff. of the American Psychological Association's 1954 manual, Technical recommendations for psychological tests and diagnostic techniques (APA, 1954). Its most definitive exposition to date has been made by Cronbach and Meehl (1955).

¹⁷For a somewhat gamey but stimulating taste of this older outlook as applied to test theory, see Bechtoldt (1959)

¹⁸See numerous articles collected in the references cited in fn.??, p.???.

¹⁹For the beginnings of a theory of the formal patterns by which theoretical constructs inductively emerge from empirical observations, see Rozeboom (1961).

 $^{^{20}}$ For a time, it was hoped by philosophers of science that dispositional concepts could be defined as abbreviations for "If ..., then ..." statements. Careful logical analysis has shown this to be generally untenable, but while it now appears inescapable that dispositional concepts presuppose theoretical states in one fashion or another, the analysis is somewhat more complicated than indicated here.

of his *ability* to spell that word. More than this, however, the pervasive intercorrelations among the individual spelling abilities (i.e., the ability to spell "hippopotamus," the ability to spell "hypotenuse," the ability to spell "hypothesis," etc.) urges a further inference to something that these specific abilities have in common. In pursuit of this speculation, let us see what follows from the hypothesis that in the population tested,

$$\mathbf{Z}_{i} = .5\mathbf{F} + \mathbf{e}_{i \bullet F}$$

for each spelling word i on the test, where Z_i is the Z-scale for the Performance-on-word-i variable, F is the "common factor" with unit variance hypothesized to underlie performances on the various words, and $e_{i\bullet F}$ is the residual component of Z_i uncorrelated with F.²¹ If we also assume that $Cov(e_{i\bullet F}, e_{j\bullet F}) = 0$ for any two different test words i and j—i.e., that there is no linear correlation among testword performances in this population not attributable to factor F—then the factorial hypothesis entails that

$$r_{Z_iZ_j} = \text{Cov}(Z_iZ_j) = \text{Cov}(.5\mathbf{F} + \mathbf{e}_{i \circ \mathsf{F}}, .5\mathbf{F} + \mathbf{e}_{j \circ \mathsf{F}})$$
$$= .25\text{Var}(\mathbf{F} + \text{Cov}(\mathbf{e}_{i \circ \mathsf{F}}, \mathbf{e}_{j \circ \mathsf{F}}))$$
$$= .25$$

We can hence explain (correctly or incorrectly as the case may be) the observed pattern of intercorrelations in this population by postulating a single theoretical variable which, like the individual spelling abilities, cannot be observed directly, but which exposes its existence by inducing a correlation among spelling performances on different words and which, moreover, is highly correlated (as shown by arguments developed later) with a subject's total score on the spelling test. What more can be learned about the nature of this construct variable depends on further correlational research. If spelling abilities do not correlate appreciably with other psychological traits, we might call F something like "general spelling ability" and let it go at that. On the other hand, if item performances on spelling tests, vocabulary tests, tests of grammar and literary style, etc., all show the same intercorrelations, we would begin to think more broadly of F as a "primary verbal factor," or, if the correlation pattern were to extend to all mental abilities, "general intelligence."²²

Of course this example, with its assumption of precisely the same correlation for each pair of spelling items, is somewhat idealized. In practice, we would expect a more complex pattern. We might find, for instance (though this is still unrealistically simple), that spdling words separate into several groups—say nouns vs. verbs vs. adjectives, etc.—such that performances on words within each group correlate highly with one another while words selected from different groups show zero correlation, so that knowing how well a person did on a list of nouns would be useful for predicting how well he can spell other nouns but would reveal nothing about his skill at verbs or adjectives. In this case we could not reasonably postulate a common "general spelling ability" factor (though we could still define "general spelling ability" as a composite of all specific spelling abilities); instead, the evidence would support the existence of several independent spelling factors, one for each intracorrelated group. Still another possibility is that the specific word abilities might all show some degree of positive intercorrelation, yet form clusters such that the correlations within each cluster are appreciably higher than the correlations between items from different clusters—e.g., performances on nouns correlate .80 with

²¹I.e. $e_{i \in F}$ is the "linear residual", or error score on Y under linear estimation of Y given X.

²²For an important recent discussion of correlational patterns and factor identity, see Campbell and Fiske (1959).

one another, performances on verbs have an intercorrelation of .90, but performance on a noun correlates only .20 with performance on a verb? This, in turn, would suggest a general factor common to all spelling abilities as well as several additional more specific group factors.

We shall turn to the mathematical details of factor analysis in a moment. The present point is that for better or worse, modern test theory has evolved powerful analytic techniques with which to prune a thicket of empirical correlations down to an inner structure of hypothetical sources, and which simultaneously establish the construct validities of the tests from which these theoretical variables are inferred by deriving an estimate of the correlation between each test and theoretical construct. Admittedly, factor analysis is by no means so sure a route to hidden truth as the present remarks might seem to imply, for as will be discussed in Chapter 6, a given correlational pattern will support a plurality of alternative factorial hypotheses and convincing procedures for adjudicating among these are still lacking. What is to be noted here is that the notion of "construct validity" is far from the appeal to magical spirits or search for metaphysical absolutes that some writers with commendably sceptical temperament have taken it to be. It merely makes explicit—and hence amenable to clarification and correction—certain primordial inductive extrusions from past experience that shape our thinking irrespective of whatever conscious assent we give to them.

References

- APA. (1954). *Technical recommendations for psychological tests and diagnostic techniques*. Washington, D.C.: American Psychological Association.
- Bechtoldt, H. P. (1959). Construct validity: A critique. American Psychologist, 14, 619-629.
- Braithwaite, R. B. (1953). Scientific explanation. Cambridge: Cambridge University Press.
- Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56, 81–105.
- Cronbach, L. J. (1960). *Essentials of psychological testing* (2nd edition ed.). New York: Harper.
- Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*, 281–302.
- Feigl, H., & Broadbeck, M. (Eds.). (1953). *Readings in the philosophy of science*. New York: Appleton-Century-Crofts.
- Feigl, H., & Maxwell, G. (Eds.). (1961). *Current issues in the philosophy of science*. New York: Holt, Rinehart & Winston.
- Feigl, H., Maxwell, G., & Scriven, M. (Eds.). (1956, 1958, 1962). *Minnesota studies in the philosophy of science* (Vols. 1–3). Minneapolis: University of Minnesota Press.
- Lennon, R. T. (1956). Assumptions underlying the use of content validity. *Educational and Psychological Measurement*, *16*, 294–304.
- Nagel, E. (1961). The structure of science. New York: New York: Harcourt, Brace & World.
- Pap, A. (1962). An introduction to the philosophy of science. Glencoe, Ill.: The Free Press.
- Rozeboom, W. W. (1961). Ontological induction and the logical typology of scientific variables. *Philosophy of Science*, 28, 337-377.