AN ANALYSIS OF CONSTRUCT VALIDITY

A Plan B Paper
Submitted to Professor Paul E. Meehl
in Psychology
(Psy. 167, Measurement of Opinions and Attitudes
Psy. 212, Research Problems)

by

Edwin L. Crosby

In partial fulfillment of the requirements for
the degree of Master of Arts, Plan B

September, 1965

Approved: Paul E Meehl
For 6 credits

CONTENTS

INTRODUCTION

How do we know what a psychological test measures? A political scientist who uses the F-Scale of the Authoritarian Personality (Adorno, 1950)* or the measure of the value of power found in the Allport, Vernon, Lindzey Study of Values must certainly find this a salient question. As a humanist, he may wonder if such a simple test is adequate to fathom the depths of the human motive for power. As a scientist, he may be disturbed by the apparent failure to define the power motive explicitly. The former interest demands an even more complex and subtle test; the latter interest leads to parsimonious tests and operational definitions.

This conflict led me to study what psychologists refer to as the "validity" of psychological tests. This is a key problem in psychology since it raises some very basic philosophical and metholological issues. The most sophisticated approach to these issues is offered by the advocates of "construct validity". On closer examination, however, this solution seemed too behaviorist to satisfy the humanist in me, while not quite rigorous enough to satisfy the scientist.

*Citations are listed by author and date rather than by footnote. See bibliography at end of paper.

This paper is an attempt to examine the basic issues involved in the problem of validity. Section I outlines construct validity and the reasons why it has been criticized. Section II discusses the philosophical issues. The conclusions reached there are applied in Section III to a revision of construct validity.

In preparing this paper it soon became clear that it could not be written for a mixed audience of psychologists and political scientists. The problems investigated are primarily of interest to the psychologist. To have written for the political scientist as well, would have been to include terms, examples, and explanations of little or no interest to the psychologist. The result would have been a paper much too cumbersome for either audience.

Also, no further reference is made to the humanist-scientist controversy. Although these terms played a role in the initial stages of the paper, they are not easily applied with precision to the particular arguments of particular men. It seemed wise to avoid discussion of whether Jones was more a humanist or more a scientist.

The following arguments and conclusions, therefore, are aimed solely at psychologists. It is assumed that the social scientist who is interested in psychological tests will be familiar enough with the jargon and assumptions to follow what is said.

SECTION I

CONSTRUCT VALIDITY

The term 'construct validity' was introduced by a committee of the American Psychological Association (Technical Recommendations, 1954) and was elaborated on by Cronbach and Meehl (1955). Although it is a widely accepted approach to validity, much controversy has arisen over it. In this Section we shall review what construct validity is, why it was introduced, how it is performed, and what the objections to it are. We shall use Cronbach and Meehl's formulation, rather than that of the earlier article.

PART 1: THE THEORY OF CONSTRUCT VALIDITY

In the field of psychological testing a test is valid if it measures what we want it to measure. The key problem of validity arises in defining what it is that we want to measure. How do we know that test A is a valid measure of construct X? One way is to correlate test A with criterion B, which is assumed to be a measure of construct X. But how do we know that criterion B is valid? To appeal to criterion C obviously leads to an infinite regress in which every criterion must be validated against some further one.

This problem arises (with regard to psychological tests) only when we are measuring psychological constructs. If we want our test to do no more than reflect a specific behavior pattern (grades in school, number of arrests, etc.) then there is no question of whether the behavior pattern is a valid measure of something else. This type of validation is referred to as predictive or concurrent validity - how well does the test predict some future or currently existing behavior pattern?

This points to one method of avoiding the infinite regress when validating the test of a psychological construct. We can say that test A is the operational definition of construct X. This makes it a trivial question to ask if A is a valid measure of X, since A measures X by definition. A similar procedure is to define X as test B and then see to what degree A correlates with B. This makes it an empirical question as to whether test A is a valid measure of X, but we are still solving the problem of the infinite regression by means of an operational definition. The logic of this latter approach is the same as that of predictive or concurrent validity.

The other solution to the problem of infinite regress is found in construct validity. "Construct validation is involved whenever a test is to be interpreted as a measure of some attribute or quality which is not 'operationally defined'" (Cronbach and Meehl, p. 175). The dissatisfaction with operational definitions (aside from their apparent arbitrariness) seems to rest on the assumption that they are phenotypical, while psychological constructs are genotypical. "The psychologist interested in construct validity for clinical devices is concerned with making an estimate of a hypothetical internal process, factor, system, structure or state and cannot expect to find a clear unitary behavior criterion" (Cronbach and Meehl, p. 180-81, quoting the *Technical Recommendations*).

Cronbach and Meehl then go on to say what they mean by a construct. "A construct is some postulated attribute of people assumed to be reflected in test performance" (p. 178). "Scientifically speaking, to 'make clear what something is' means to set forth the laws in which it occurs. We shall refer to the interlocking system of laws which constitute a theory as a nomological network. ••• A necessary condition for a construct to be scientifically admissable is that it occur in a nomological net, at least *some* of whose laws involve observables" (p. 187). Without the nomological network, "we would not be able to say anything intelligible about our constructs" (p. 192).

One characteristic of a construct is that it "goes beyond the data". In the most elementary case this means that our construct is simply ~~o~~an inductive summary: it "purports to characterize the behavior facets which belong to an observable but as yet only partially sampled cluster" (p. 189). "If we simply list the tests or traits which have been shown to be saturated with the 'factor' or which belong to the cluster, no construct is employed" (p. 190).

This approach to the meaning of psychological constructs avoids the infinite regress involved in the search for criteria. The criteria against which we validate our test are found in the relationships which the construct has in the no~~l~~mological network. When our test of a construct enters the laws hypothesized by the nomological network, then we say we have a valid test. Since we assume in science that we have never discovered all the empirical relationships into which a concept may enter, the process of test validation is never complete. We can only say that a test is more valid than others which have been proposed.

In avoiding the infinite regress of criteria, construct validity also avoids the oversimplifications of operational definitions. We realize that our constructs are genotypic; they are only partially manifested in any particular overt behavior pattern. Also we aim at general terms, rather than the specific terms of operational definitions. We leave the meaning of our terms open to further knowledge. We are able to refine our terms. Finally, our concepts go beyond the data; they serve the heuristic end of suggesting new relationships to be tested.

PART 2: THE METHODS OF CONSTRUCT VALIDITY

For the purposes of analysis, the methods of construct validity can be divided into two general catagories - intertest validity and intratest validity.

A. Intertest Validity. To perform intertest validity we correlate our test as a whole with other tests or behavior patterns which are linked to the construct by the nomological network. The "multitrait-multimethod matrix" proposed by Campbell and Fiske (1959) is the most elaborate example of this approach. Since a construct is genotypic, we must be sure that any test of it is a function of the construct and not of the particular type of behavior used to measure it. Thus Campbell and Fiske propose the use of several different tests each of which measures several different constructs. For example, we may have tests A, B, and C each of which measures constructs X, Y and Z.

A test is valid to the degree that the measure of one construct has a low correlation with the other constructs measured by the same test and a high correlation with measures of the same construct by different tests. Thus X_A (construct X as measured by test A) should correlate more highly with X_B and X_C than with Y_A or Z_A. The more divergent the methods are, the more sure we are of the results (eg, paranoia as measured by the MMPI, the Rorschach, behavior in a small group setting, and clinical interviews).

How far construct validity may be carried using the intertest method is not made clear by Cronbach and Meehl. The multitrait-multimethod approach is really no more than the correlation of several measures of the same construct with each other. But the nomological net does more than relate constructs to behavior patterns in which they are manifested. It also relates constructs to each other. This would seem to mean that we can validate a test of construct A by its ability to predict a certain configuration of constructs B, C, . . . N which are hypothesized by the nomological net to correlate with construct A.

This extended version of intertest validity is often hard to recognize. Cronbach and Meehl give an example in which a test for anxiety is validated against ratings of tenseness and against intellectual inefficiency induced by painful electric shock. Are tenseness and intellectual inefficiency behavioral manifestations of anxiety or are they separate constructs? Cronbach and Meehl provide no answer to this, perhaps because they feel it is not an important question.

The only answer they give is an indirect one. They point out that if we fail to discover the expected relationship, the fault may lie with test A, test B or with the way the construct or the nomological net has been established. We have no way of knowing where the fault lies unless there is a body of empirical evidence supporting one or two of the elements and not supporting the others. This can be taken to mean that we can use any part of the nomological net for construct validity as long as it is supported by empirical evidence. Thus in some cases the extended version of intertest validity would seem to be permissable.

B. Intratest Validity. This deals with the relationship between the various elements or items which make up a test. In constructing a test our problem is to make inferences about a single trait from a set of responses all of which are multiply determined. Although these responses may appear on the surface to be dissimilar, they must all have some element in common since they are supposed to be reflecting the same trait.

Those items which do not have this common element do not reflect the trait. To include them is to make the test to some degree invalid. The attempt to be sure that the items have this common element is what Peak (1953) refers to as the attempt to establish functional unities. In her review of the ways to do this she considers item analysis, the rational ordering of items, factor analysis and dynamic unities.

There is often a strong realist tinge to this intratest approach to validity. This is brought out by Loevinger (1957) where she speaks of the structural component of validity. The items of the test should be structured so as to reflect the structure of the construct itself (p. 661). We are interested in the validity of a test as a measure of a trait which existed prior to and independently of the psychologist's act of measuring (p. 642). Under structural validity she discusses much of what Peak included under functional unities.

Finally there is an emphasis on the theoretical derivation of items. This is found in Travers (1951) as well as in Peak and Loevinger (who discusses it under the heading of substantive validity). It is advocated most strongly by Jessor and Hammond (1957); the theory or properties of the trait should be used in the selection of items. This is merely the use of the nomological net to relate the items of the test to the construct, instead of the test as a whole to the construct and to the other tests. If we do not understand why our items reflect the trait, we cannot be sure that our test is a valid measure of the trait. Thus the purely empirical selection of items (eg, they are selected because they correlate with some criterion) or the intuitional selection of items (eg, items selected on the basis of clinical judgment as in the Taylor Manifest Anxiety Scale) is inadequate.

PART 3: CRITICISMS OF CONSTRUCT VALIDITY

The objections to construct validity come mainly from those who favor an operationist approach. This is seen most clearly in articles by Bechtoldt (1959) and Ebel (1961). There are three key points which they make.

A. The construct validity approach fails to define its concepts adequately. Cronbach and Meehl say that the meaning of a psychological construct lies in the laws of the nomological network into which it enters. But this does not give us a definition of the construct itself. Empirical concepts must be defined in terms of observables; in using a concept, we must know what observables it makes reference to. This is called the meaning of a concept. This is a matter of convention and not of empirically established laws.

Once we have given our concept an adequate definition, then we try to see how it is related to other concepts. The empirical laws into which a concept enters are called its significance. Cronbach and Meehl, in saying that the meaning of a construct lies in the laws into which it enters, have ignored the meaning of a concept (in Bechtoldt's sense) in favor of its significance. But this can only lead to confusion, since we cannot know what the laws are about unless the concepts involved in them are independently defined. A concept can have meaning without significance, but not the reverse. The requirement of a definition in terms of observables is essentially the same thing as a demand for operational definitions.

B. If we agree that constructs must be defined in terms of observables, then the need for construct validity is removed. As we noted above (p. 2), since definition is a matter of convention, it is meaningless to ask if an operational definition of a construct is valid. The question of validity arises only where we seek an alternative test for a construct which has already been operationally defined. Such an alternative test would probably be sought only on the grounds of economy; it is easier to score, can be given in less time, etc. The validity of such a test is ascertained by correlating it with the operational definition.

Since there is no other way in which a construct can be given meaning, there is no other way to validate a test. The attempt to devise constructs which make more sense out of the nomological net is really a search for significance. This is different from the search for validity. In the latter case we have a precise idea of what we are looking for, while in the former case we do not. If a proposed definition fails to achieve the expected significance, then the fault may be with the new test, with the nomological net, or with the tests of other constructs against which we correlate our test.

C. To mistake the search for significance for the search for validity leads to some potential errors. Construct validity is liable to the confusions of a vague realism. In attempting to make our test correspond with "what the trait really is", we lose track of the empirical underpinnings of psychology. Instead of showing explicitly that one definition is more significant than another, we seek out the test which fits our implicit notion (often intuitive or mentalist) of what the construct "really is".

Another potential error of construct validity lies in the use of response inferred traits without adequate emphasis on the dangers of circularity involved. For example, an investigator may define an over-compensated inferiority complex in terms of aggressive and dogmatic behavior. Given an instance of such behavior, the investigator may then turn around and explain it by the "fact" that the person is over-compensating for an inferiority complex. Such an error is too easy to make when it is assumed that there is some "real" trait which is being "indirectly" measured by test performance.

PART 4: CONCLUSIONS

It is the thesis of this paper that any consideration of validity leads to a discussion of some of the core problems faced by psychology. Any particular approach to validity represents, at the least, certain positions on the nature of construct definition, the nature of causality, and the use of mental data.

In Section II we shall discuss concept formation and definition. The differences between Cronbach and Meehl on one hand and Bechtoldt and Ebel on the other are really reflections of the philosophical differences between Herbert Feigl, Carl G. Hempel, and Rudolf Carnap (whom we shall call the Liberal Empiricists) and Gustav Bergmann and May Brodbeck (whom we shall call the Logical Empiricists). Campbell (1960) has noted that these differences exist, but claims that they are of little importance to the practicing psychologist. Contrary to this outlook, an attempt will be made to show that these differences are important. Although it is possible to achieve a "working solution" to the problems raised, this will rule out the terminology of construct validity.

The working solution of Section II will be applied in Section III to the problem of validity. We shall show why revisions must be made in the construct validity approach. In rejecting the terminology of construct validity, we still retain some of its methods. One of the values of the revised terminology is that it throws new light on some of the operations which Loevinger and Jessor and Hammond advocated.

SECTION II

DEFINING EMPIRICAL CONCEPTS

Our purpose in this Section is to investigate the differences between the Liberal and Logical Empiricists over the nature of concept definition. An attempt will be made to show that although there is disagreement over the logic of concept formation and definition, a "working solution" is possible with regard to how the operation is actually performed. This Section is divided up into three parts: concepts defined in terms of observables; concepts introduced as theoretical terms; and the existential nature of psychological concepts.

PART 1: CONCEPTS DEFINED IN TERMS OF OBSERVABLES

In stating the two opposing positions we shall use the work of Bergmann (1951, 1953) and Hempel (1952) since the former was cited by Bechtoldt and the latter by Cronbach and Meehl as providing the philosophical basis for their respective approaches to validity. This will be followed by a consideration of the differences between the philosophical positions together with an attempt to resolve them.

A. Bergmann's Logical Empiricism. In this Sub-Part we present a synopsis of those parts of *The Logic of Psychological Concepts* (1951) which deal with definitions in terms of observables. The basic point he makes is that a concept is meaningful if it can be ultimately defined in terms of the basic undefined or primitive concepts of our language. These latter can only be known by experience; they are immediately observable.

For a concept to be meaningful it must be linked to these undefined characters by means of a chain of grammatically correct definitions in which (a) the first contains, except for the concept it defines, only undefined concepts; (b) each contains, except for the concept it defines and for undefined concepts, only concepts defined in the proceeding members of the chain; (c) the last one is the definition of our concept.

Any concept which is meaningful acquires its meaning in the way outlined here except for structure words ('all', 'is', 'and', etc. - these are not considered to be concepts) and certain theoretical terms which will be discussed later. One common class of definitions covered by the scheme in the above paragraph is what Bergmann calls definitions in use. These are characterized by the fact that they define one sentence in terms of another sentence rather than one concept in terms of another.

The form of one of these definitions is '$L =_{Df} R$' where 'L' is the left side where the new term is conventionally placed; '$=_{Df}$' is 'means by verbal agreement the same as';

and 'R' is the old, or defining, term. R is often a compound sentence in the form 'if R_1, then R_2'. The definition of an electric field in terms of the behavior of an electroscope is an example of a definition in use. L reads, 'This place is in an electric field'; R_1 reads 'An electroscope is at the place (of which L asserts that it is in an electric field)'; and R_2 describes the behavior of the electroscope.

After outlining this, Bergmann turns to operational definitions. They add nothing to what has been said above about the meaning of concepts. They do serve a purpose, however, in reminding us that sometimes operations are necessary in order to realize a definition. Thus, in the above case if we want to ascertain the truth or falsehood of L, we must help nature along by realizing manipulatively the condition which R_1 states.

This is the Logical Empiricist view on how concepts should be defined; it is the basis for Bechtoldt's criticisms of construct validity. Let us now turn to the Liberal Empiricist view.

B. Hempel's Liberal Empiricism. In section 6 of Fundamentals of Concept Formation in Empirical Science, Hempel criticizes the "narrower thesis of empiricism". This holds that any scientific statement can be transformed by means of definitions into equivalent statements couched exclusively in observational terms. Hempel then sets out to show that this formulation cannot adequately deal with dispositional terms.

He starts off with an example on magnetism, putting his definition in the form of a contextual definition (which seems to be identical with Bergmann's "definitions in use"). "x is magnetic at t $=_{Df}$ if, at t, a small iron object is close to x, then it moves toward x." It can be seen that this definition is in the same form as 'L $=_{Df}$ R', where R is 'if R_1, then R_2'.

But Hempel rejects this definition because L "would be satisfied by an object x not only if x was actually magnetic at time t but also if x was not magnetic but no small iron object happened to be near x at t" (p. 25). This is because in logic 'if ..., then ...' is synonymous with ' either not ..., or also ...'.

A way to get around this problem is to formulate the definition as follows: "x is magnetic at t $=_{Df}$ if, at t, a small iron object should be close to x, then that object would move toward x". This formulation in terms of a counterfactual conditional is also rejected since it raises too many unsolved problems.

The solution which Hempel then adopts is to use Carnap's (1936-37) reduction sentences, the simplest form of which is $P_1x \supset (Qx \equiv P_2x)$. This reads: "If an object x has characteristic P_1 . . ., then the attribute Q is to be assigned to x if and only if x shows the characteristic . . .P_2". If we put the example of magnetism in this form it reads: "If a small iron object is close to x at t, then x is magnetic at t if and only if that object moves toward x at t".

This form does not run into the difficulties encountered with the first example. But it does mean that reduction sentences cannot be eliminated in favor of primitives (except in the case where P_1x is analytic) since "it provides no interpretation for a sentence such as 'object x is now magnetic,' but there is no iron whatever in its vicinity'".

This indeterminacy in the meaning of a concept introduced by a reduction sentence can be made smaller by introducing additional reduction sentences which specify different test conditions (eg, relating magnetism to the electric current in a loop of wire). Such additional reduction sentences can be introduced only when there is empirical evidence showing their equivalence to the original reduction sentence. But even if additional reduction sentences are added, a concept introduced by a reduction sentence will always have an "openness of meaning".

This completes the bare outline of the two basic approaches to the definition of concepts. Let us now turn to the differences between them. How great are they? Do they make any difference for the practicing psychologist?

C. Material Implication. This is the technical term for the logical deficiency shown by explicit definitions: if we cannot perform the test for a disposition, then we are led to assert its presence (more technically, a false antecedent in the definiens leads to the truth of the conditional). This point was first made by Carnap (1936-37) and was covered above (p. 15) in our review of Hempel.

Bergmann (1957, p. 61) notes this point and says that many consider it to be obvious and of little importance.

For those who want to eliminate it, the logical 'if . . ., then . .' can be removed and some non-logical term put in its place. Bergmann clearly shows that he considers this difficulty to be an inadequate reason for rejecting explicit definitions.

Pap (1953) reaches the same conclusion. Since Pap strongly advocates the elimination of explicit definitions, it is worthy to note that he does not reject them due to this logical difficulty which they encounter. Madden (1961, p. 400) goes even further, claiming that reduction sentences run into the same problem. In spite of the extended meaning of concepts introduced by reduction sentences, there will be cases when the test conditions for the concept are not present. Thus we cannot say that a lump of sugar is soluble until it has been dropped into some liquid. Reduction sentences, relying on an ideal language, run into the same problems with counterfactuals and lawfulness that explicit definitions do.

Therefore, in deciding between explicit definitions and reduction sentences, there seems to be a consensus that this problem over their logical adequacy should be ignored. It should give the practicing psychologist no trouble.

D. Openness of Meaning. The Liberal Empiricists have often claimed that the openness of meaning of reduction sentences is a distinct advantage over the specific meaning of explicit definitions. As our knowledge about a concept grows, this can be incorporated into it through additional reduction sentences. Second, a concept often means more than just the single sentence which introduces it in the case of explicit definition.

Bergmann gets around the first criticism (operational definitions cannot be revised) in the usual way of operationists: we can always introduce new terms into our science. Since the new term is often very similar to an old term (even to the point of using the same word), this in effect allows us to revise our terms to accord with our increased knowledge. Even Hempel (1958, p. 69) does not reject this approach: what problems it raises are problems of terminology but not of methodology.

The Logical Empiricist answer to the second criticism lies in the difference they make between meaning and significance. The meaning of a defined concept resides entirely in the right side of its definition. But for a concept to be useful, it must enter into empirical laws. This aspect of a concept is its significance. A concept can be meaningful (ie, explicitly defined in terms of observables) and yet fail to have any significance (ie, have no empirical relationships with other concepts).

This difference between meaning and significance is one of the key points in the Logical Empiricist attack on reduction sentences. Conversely, it is criticized by the Liberal Empiricists. But there is no controversy over the range of material which the two approaches can encompass. Any additional meaning given to a concept by further reduction sentences can be included by the Logical Empiricists as part of the significance of the concept.

E. Problems of Explicit Definitions. If explicit definitions do not fall victim to the logic of their formulation, the possibility of revision, or the range of material they cover, what disadvantages do they show?

(i) A fairly minor objection is that, logically, we can not test for the presence of a concept except through observing the conditions which define it. This is because an explicit definition states both a necessary and sufficient condition for the use of a concept. This means that if concept X is defined as A, then we can never say that X is present unless A is. Therefore, test B is not a sufficient condition for the presence of X. This logical difficulty can be solved empirically if we can show that B is a sufficient condition of A. Evidence for such a relationship is of course never conclusive, but this is merely an instance of the inductive nature of science. But since A is always present when B is, test B will only be used if it is more economical to use than A.

(ii) A much more cogent criticism involves the case where we seem to have the same concept, yet it is measured in at least two different ways. One instance of this occurs when certain parameters of a concept are beyond the range of ordinary measurement. Thus we speak of the temperature of the sun and the temperature of a room, even though the former is defined in terms of different operations than the latter.

This raises problems for the Logical Empiricist. Let us say we define temperature in terms of the fluctuations of mercury in a glass tube. It is meaningful to talk about the temperature of the sun only in the sense that it is theoretically possible to put a thermometer near the sun. But it is wrong to say anything about the temperature of the sun if we infer it only from the intensity of its head radiation. This is because the whole meaning of the term 'temperature' lies in the behavior of the thermometer.

This may seem to be the same problem that we faced above: can we ever assert concept X on the grounds of test B rather than on the grounds of the defining observables A? But in (i) above, B was an empirically sufficient condition of A. Whenever we asserted the presence of X on the grounds of B, we could always point to A and say "this is what we mean". But the sun is too hot for a thermometer to be useful. In only certain ranges is the intensity of the radiation a sufficient condition for variations in the thermometer. When we extrapolate this relationship to very high temperatures, we can no longer point to the thermometer as our meaning.

The Logical Empiricist may try to avoid this problem by, eg, redefining temperature in terms of intensity of head-radiation. This new definition of temperature will be accepted if - (1) it enters into substantially the same laws as the old concept, thus justifying the use of the same word; (2) it is more significant than the old concept, ie, it enters into additional laws that the old concept did not enter into; and/or it accounts for more variance in the original laws than the old concept.

But what if this new definition adds significance in some cases (high temperatures) but does not include all that the mercury thermometer included (low ranges)? Assume furthermore that the non-overlapping ranges of the two definitions enter into the same laws, so that the significance of temperature$_1$ (low temperatures defined by thermometer) is the same as the significance of temperature$_2$ (high temperature defined by heat-radiation). This leads us to think that in every way except explicit definition we are dealing with the same concept.

The Logical Empiricist in this situation could do three things: search for a definition which covers the whole range; speak of temperature$_1$ and temperature$_2$; or forget the logical requirements of his philosophy and use the different operations to refer to the same concept. The last alternative is surely the one which the practicing physicist would adopt, although all three solutions are perhaps workable as long as he did not waste too much time on the first. But this example shows that the Logical Empiricist can maintain his demands for explicit definition only at the price of considerable awkwardness.

(iii) The above problem was raised by Pap (1953). He also discusses a further problem which is related to it. Suppose that we define concept X in terms of A and that this enters into lawful relationship with B, C, D, and E. Imagine that we discover the presence of B, C, D, and E, but fail to discover A. According to the Logical Empiricists we would have to say that X was not present. But in many cases the Logical Empiricist will admit that which aspect is chosen as meaning and which as significance is arbitrary. Thus it seems absurd to accept or reject the presence of a concept on the grounds of one piece of evidence, regardless of how many bits of equally reliable evidence speak against this.

For example, if we have the concept of 'mass', we may decide to define it in terms of the ratio of the accelerations produced in two interacting particles. If we then discover many other "tests" for mass which all indicate its presence in a particular case, we would be foolish to reject this because we failed to discover the "defining property".

This problem rests on the fact that one property is designated as the definition and the rest as significance. Thus one property is singled out to have an analytic relation with X while all the rest are related synthetically. Pap hopes to avoid the artificiality of this approach by calling all properties reduction sentences. But this runs into problems over the analytic-synthetic distinction.

It is clear that the definition of a word, (<u>ie</u>, the introduction of a word), is a matter of convention and thus analytic. But in the case of a reduction sentence, where does this definition lie? If we designate any one of the reduction sentences as the definition, then we are in fact using explicit definitions and are in the same position as the Logical Empiricist. We cannot get around this by designating all of them as the definition; if we did, then the absence of any one of the properties would mean that the concept was not present.

Pap gets around this by saying that the traditional analytic-synthetic distinction is inadequate. Each reduction sentence serves both functions at the same time: it serves both to introduce the concept and to give it empirical meaning. "We have seen that a systemic reduction sentence has factual content inasmuch as it is open to correction through the other members of the system, but at the same time it is a (partial) meaning-rule" (1953, p. 22). Each reduction sentence indicates a concept to a certain (undetermined) degree of probability. Thus we know properties which are sufficient conditions for the presence of a concept, but not those which are both necessary and sufficient.

The Logical Empiricists do not attempt to rework their formulations to avoid the above problem. Instead, they attack the adequacy of reduction sentences; or they try to show that the Liberal Empiricist criticism is not as damaging as it appears. Let us therefore turn to the weaknesses of the Liberal Empiricist approach.

F. Problems of Reduction Sentences. The points made under this heading will be somewhat vague with regard to which formulation of reduction sentences is being criticized. Although in large part due to my own inadequacies, this is also due to the fact that neither Brodbeck (1954, 1958, 1963) nor Bergmann (1953, 1957, 1961) make explicit reference to which Liberal Empiricist they are criticizing.

(1) The reduction sentence approach fails to define its concept. We do not know what we are talking about. This is due to a failure to distinguish between meaning and significance. A concept is significant to the degree that it enters into empirical laws. But an empirical law can exist only

where both terms can be idependently identified. They must have some meaning apart from the laws which they enter into, otherwise the laws are tautologies. Yet it is just this independent meaning which the reduction sentence approach fails to give the concept it introduces. This is because its advocates insist that the relationship between the concept and the reduction sentence is empirical.

At the worst, this lack of independent meaning leads to a holistic approach to meaning in which all truths become analytic (Bergmann, 1957, p. 64; Brodbeck, 1963, p. 65). This criticism of the reduction sentence approach goes wide of the mark, however, since it ignores the insistance by Liberal Empiricists that the empirical relationships between reduction sentences be demonstrated.

Nevertheless, the Liberal Empiricist must in some way deal with this difference between meaning and significance. The most obvious way is to designate one of the reduction sentences as the analytic relationship which introduces the concept and to designate all the rest as synthetic. But we already noted (p. 22 above) that this is just the same as the explicit definition approach and thus avoids none of its difficulties.

Pap's solution seems to be the only way out of this. Each reduction sentence serves both to introduce the concept and to add to its empirical relationships. This approach requires that the analytic-synthetic distinction be rejected. Whether this can be done is a much discussed question which goes far beyond our interests and my abilities. But there is another way to approach Pap's suggestion. Let us review his reasons for making a reduction sentence both analytic and synthetic.

The problem was this: if X is explicitly defined as A and enters into lawful relationships with B, C, D, and E, then if we find B, C, D, and E, but fail to find A, do we say that X is absent (thus rejecting laws B, C, D, and E) or do we say it is present (thus rejecting our definition)? The former is obviously counter to the ways of science, while the latter undermines the whole reason for using explicit definitions.

But there is a way for the Logical Empiricist to slip through the horns of this dilemma. This rests on the frequency with which we find B, C, D, and E but fail to find A. If it occurs very rarely and in a random fashion, then we can dismiss it as the result of errors in measurement. If it occurs more frequently, then we will surely redefine X in terms of B, assuming that B is a more reliable condition of C, D, and E than A. And the redefinition of terms is an integral part of the Logical Empiricist approach.

This of course assumes that B is more "regular" than A. But Pap's argument derived much of its force from the assumption of such regularity. The dilemma: "Would we call an apple-tasting lemon, a lemon?" is a dilemma just because the taste of lemons is so distinctive that it could be used as a defining property. The dilemma: "Would we call a person anxious if he did not have diarrhea?" is not a dilemma because no one defines anxiety solely in terms of diarrhea. Whenever we discover that one aspect of a concept is not as regular as another, we can always redefine the concept in terms of the more regular aspect. In this way, Pap's dilemma ceases to be one.

In eliminating Pap's dilemma, we also eliminate the need (at least in this case) for rejecting the distinction between analytic and synthetic. But in order to deal with the criticisms of the Logical Empiricists that reduction sentences lead to a holistic approach to meaning, one of the reduction sentences must be designated as the definition of the concept. This does not, however, lead to a complete acceptance of the explicit definition approach. It is possible to designate any one of the reduction sentences as analytic, depending on the problem one is faced with, and to designate some other reduction sentence as analytic in another situation. Still, this is a step in the direction of explicit definitions.

(ii) A second criticism of reduction sentences deals with their extensiveness. Technically, to define a concept we must state its meaning; in the case of reduction sentences, this means spelling out every empirical law into which the concept enters. But when we ask what a concept means, this is obviously not the sort of answer we expect. This may be a dilemma in the logic of concept definition, but it is not a practical problem. As soon as two people know what they are talking about, they will cease to mention other reduction sentences in order to complete the definition. No confusion should arise here.

The practical problem which does arise deals with how "central" or "essential" or "natural" our definitions are.

Granted that we need not list all of the reduction sentences in defining a concept, what ones do we list? Three criteria are typically offered--we want to select those aspects which are few in number; these must clearly separate the concept from all others; and these aspects must make as much "theoretical sense" as possible (ie, they must be related to as many other concepts as possible and include a wide range of phenomena while achieving clarity and even "elegance" if possible).

Neither the use of reduction sentences nor the use of explicit definitions will guarantee that we achieve central or fruitful concepts. Explicit definitions may be precise, but this does not make them fruitful or central. They must constantly be revised so as to come closer and closer to this goal. Conversely, it is so easy to add new knowledge to a concept through reduction sentences, that this may lull us into ignoring the attempt to organize our knowledge so as to make it central and fruitful. Other reasons for the need to achieve these ends will be discussed in Sub-Part G below.

(iii) A third problem with reduction sentences arises when they do not correlate perfectly with each other. We have already noted that some measures of a concept do not completely cover the full range of its variation (eg, the temperature of a room and the temperature of the sun). Instances of this sort raise no problems as long as the measures correlate highly in the ranges where they overlap.

The trouble with reduction sentences which do not correlate perfectly is that they do not enter into other laws in the same way. This may or may not lead to difficulties. Thus, time may be measured by the rotation of the earth or by electrical oscillations in a crystal. For most purposes the differences between these two measures is so small that it makes no difference which we use. But there are some measurements made in physics and astronomy where it does make a difference. Many laws which hold very closely using time measured by crystal oscillations do not hold using time measured by the rotation of the earth.

If the concept'time' is introduced by reduction sentences which include both measurements, then confusion is apt to arise. Thus, what is meant when we say that the time elapsed was "exactly two hours, to the nearest microsecond"? This obviously depends on which measurement of time was used. The confusion, of course, can be avoided by speaking of $time_1$ and $time_2$ or by specifying which measure was used. But this is one of the things which reduction sentences were supposed to avoid: They were meant to allow us to revise our concepts, rather than having to redefine them as is necessary with explicit definitions.

The similarity between the two approaches can be seen here. Using reduction sentences, time has an extended meaning, not linited to one explicit definition. Yet in practice we must use time as defined by a particular measure if confusion is to be avoided. With explicit definitions the situation is reversed. These give a specific and limited definition to time. Yet in practice we can use a variety of measures of time by calling anything which correlates highly with our original definition, a "test" of time.

This similarity exists because we designate one reduction sentence as an explicit definition. This is the same thing we did on p. 26. There we noted that one reduction sentence must be designated an explicit definition but that we can switch the analytical designation around from one reduction sentence to another depending on the problem at hand. This works well where the reduction sentences correlate highly (it would even come close to working in the example of time above).

But when reduction sentences do not correlate highly (and in psychology they typically run from .40 to .80), we must take the final step towards explicit definitions. We must designate one particular reduction sentence as the explicit definition and stick to it until it can be shown clearly that some other reduction sentence (or combination of reduction sentences) is more significant. All the other reduction sentences are then seen as the significance of the basic definition.

Unless such steps are taken, when a person speaks of, eg, anxiety, a double confusion would arise. First, we would not know what behavioral manifestations this refers to. Second, we would not know what laws link this concept to other concepts nor to what degree the laws hold.

G. Generality. One of the aims of concept formation in the empirical sciences is to attain concepts which are both general so that a wide range of phenomena is included and accurate so that we know the precise meaning of the term.

It is often felt that explicit definitions sacrifice generality to obtain accuracy while reduction sentences do the reverse. Let us examine this belief more closely.

(i) Explicit Definitions. The first point to be made is the obvious one that a concept is not a proper name. Thus all concepts, explicitly defined or defined by reduction sentences, are general in that they refer to more than one particular object. But explicit definitions are general in more than just this limited sense. Thus Bergmann notes that operational definitions are often chains of definitions of a complex form including computations as well as manipulations. When the chain is long and many computations are involved, then some may choose to call this definition abstract (1943, p. 269).

But if a concept is stated in general terms, does this not lead to a lack of clarity or precision? The problem arises as to which phenomena are instances of a concept and which are not. The first point to be made is that no matter how explicit or non-general our definition is, there will always be instances which are not easily classified. Even with something as precise as length, we always find that we cannot go beyond a certain degree of precision. Bergmann notes: "No definition is specific in the sense that it either controls all the possibly relevant variables or that it completely specifies the manipulations themselves" (1943, p. 272).

The test of the precision of our definition lies in the laws which the concept enters into. If we define concept X in terms of o•p•q, and if concept X enters into laws L_1, L_2

and L_3, then any entity or process which displays o·p·q ought to fit laws L_1, L_2 and L_3. If they do, then we assume that our definition is precise; that any instance of concept X is "essentially the same thing".

How general a term is depends on whether we are specifying a quantity or a quality. In defining magnetism in terms of the movement of a small iron object, we are defining a disposition qualitatively. It is meaningful to discuss what is meant by "iron". Does this include steel? Does the iron need to be pure or can it be mixed with other metals? But when we want to measure the degree of magnetism it is necessary to specify exactly what type of iron we mean as well as what we mean by "small".

A general concept dealing with qualities may include instances which are phenomenally very different. Thus we might define a psychopath as a person who violates society's norms. In this case both rape and robbery are indications that a person is a psychopath. There are always questions about how much a person has to steal or rape to qualify as a violator of societal norms, but other than this, the definition is quite explicit. But it is a significant concept only if both rapists and thieves fit all the laws we have established concerning psychopaths. Since this is very unlikely, we would surely revise this definition to be more specific about the type of rapist or thief we will consider to be a psychopath.

With a quantified concept, the problem of what is an instance of it is much smaller. This is because it is usually hard to state in general terms how phenomenally different entities reflect a given degree of the concept. Even where this is possible (eg, how psychopathic a person is, is shown by the number of violations regardless of what they are) the various measures may not enter into the laws about psychopaths in the same way. Since in psychology different measures of the same general concept often do not enter the same laws in the same way, the temptation exists to sacrifice generality for precision.

In formulating concepts two opposing tendancies are always at work. We seek to state our findings in more general terms so as to apply our knowledge to new phenomena. This is part of the inductive nature of science. On the other hand, as new phenomena are included in our general terms, we must test to see if they enter into the same laws. As long as we can offer proof of this, we are safe in including general concepts under the heading of explicit definitions. And proof must be offered whenever doubts exist about any specific case. When to doubt and when to test is part of the art of science.

(ii) Reduction Sentences. It is usually felt that reduction sentences are more useful in defining general concepts than explicit definitions. While an explicit definition introduces a concept in terms of one set of criteria, reduction sentences can expand the meaning of a concept to cover a broad range of material.

But when we look more closely, this difference disappears. First, the Logical Empiricist can handle much of the "generality" added through reduction sentences by means of his notion of significance. Second, an explicit definition need not be limited to one criterion. The class approach to definition can be used in which several criteria are listed. Thus concept X may be defined in terms of o•p•q. This is not the same as the reduction sentence approach where any one of the criteria is an indication of concept X. In the case of an explicit definition all of the criteria must be present. We use the class approach only if it is shown to be more significant than any one of the criteria used by itself.

Third, reduction sentences may increase the range of concept X, but they do not make the concept itself more general. Cronbach and Meehl seem to make this point (p. 190-91) when they say that the concept 'psychopath' must be defined genotypically rather than by a mere listing of its phenotypical manifestations.

Fourth, the use of reduction sentences rather than explicit definitions often arises from the demands of a behaviorist psychology which refuses to use mental data. This argument will be discussed in Section III, but we can sketch its outlines here. Part of the observables which make up the basic givens (or knowledge by acquaintance) of our language may be classed as mental data. These, of course, are ignored by the behaviorists. But many of the genotypical concepts of psychology rely on mental data. An obvious dilemma exists if we can state our concepts in general terms only by using mental data.

One solution is to use explicit definitions and pitch them at a low level of generality. But this is inadequate: no science can be built without general terms. But reduction sentences provide a way out of this by listing all the phenotypic situations in which the genotype manifests itself. It is interesting that Cronbach and Meehl find this inadequate and insist that a term be defined genotypically in terms of the network of laws surrounding it. But they never sketch out the network itself. In their example about the psychopath they say it is there but never what it is. My guess is that they failed to do so because it involved some basically mentalist terms much as "introject", "without emotional attachment", etc. Although these have behavioral correlates, the argument will be made in Section III that they are more fruitfully viewed as based on mental data.

(iii) Causality. The final objection to the use of reduction sentences to attain generality is that this may well lead to a blurring of the "lines of causality". The problem of causality has evidently been so thoroughly worked over in psychology that probably only a non-psychologist-non-philospher dares talk in such naive terms. Nevertheless, whether one speaks of lines of causality, systematic import, or significance, there seems to be some important points to be made.

The first centers around multi-causality. We have already said that for an entity or behavior pattern to be accepted as an instance of a concept it must enter the same laws as other instances of the concept. But since most phenomena are multiply determined, any one instance of a concept will probably be the function of independent variables other than those which enter into lawful relationship with the concept as a whole. Thus if one aspect of anxiety is trembling, we must differentiate between trembling as part of anxiety and trembling as the result of a low temperature.

In order to rule out the effects of "irrelevent causes", we must show ~~t~~what they could be in the case of any one reduction sentence and show how they were controlled for. One control is to remove the conditions under which the irrelevant causes operate (eg, be sure the room is warm). Another is to specify a group of reduction sentences which together define the concept. We assume here that the "irrelevant causes" do not overlap so that this combination could only be the manifestation of the concept we are interested in. To use reduction sentences without specifying irrelevant causes and the controls used is to blur the lines of causality, to lower the significance of our concept. Correlation alone is not sufficient grounds for accepting a new reduction sentence as part of the concept.

The second point deals with how the reduction sentences themselves interrelate. We mentioned in Sub-Part F above that we must try to define our concept in the most central terms possible. This aims not only at parsimony and fruitfulness in defining concepts but also at eliminating reduction sentences which are the causal effects of more central reduction sentences. In the terminology of psychological tests, we must investigate the structure of the items on the test.

The third point is of greater importance. We must be careful not to increase the significance of a concept by including elements of other concepts in its definition. For example, we may discover that anxiety leads to intellectual inefficiency. But we must be sure that this correlation does not exist because we have defined anxiety party in terms of behavior patterns which are very similar or identical to those used in defining intellectual inefficiency. Often it may be hard to decide whether a certain phenomenon should be included as part of one concept or part of another. But in any case we should not include it in both concepts if we are going to say that one is the cause of (or is empirically related to) the other.

The above three concerns (the attempt to account for multi-causality of items, to use central rather than peripheral items, and to prevent overlap of concepts) will be referred to in the rest of the paper as "the systematic ordering of items". Explicit definitions are less likely to run into trouble on these points than reduction sentences (unless a set of items is stipulated as a whole to be the explicit definition). But if our explicit definition is limited to one coherent set of criteria, then we are led to phenomena which at least have certain characteristics in common other than mere correlation.

(iv) Conclusions of Sub-Part G. In the above discussion I have tried to show that both explicit definitions and reduction sentences encounter problems with general concepts and these problems are to some degree common to both approaches. Typically, however, the danger of explicit definitions is that they are stated at too low a level of generality in order to assure precision, or that, if fairly general, insufficient proof is offered that instances of it enter the same laws in the same way. Reduction sentences, on the other hand, run the danger of blurring the "lines of causality" so that the laws we construct with the concepts are either spuriously low or spuriously high in the accuracy of their predictions.

PART 2: THEORETICAL LANGUAGE

One of the basic tenants of Logical Positivism is that there is a very important class of concepts which is not defined in terms of observables. These are often referred to as making up a "theoretical language". Included here are logical concepts and logical systems such as mathematics. In the 1920$^{\underline{s}}$ many of the concepts of physics and chemistry were seen as part of a theoretical language. By the 1950$^{\underline{s}}$ Carnap was even claiming that most psychological concepts should be defined in this way. This shows an attempt to escape from some of the limitations of traditional empiricism.

For this reason, the advocates of construct validity may feel that Part 1 was a waste of time. Since Cronbach and Meehl say that construct validity is used where an operational definition is inadequate, it is unlikely that even Hempel's views on definitions in terms of observables would allow for its use. It is much more likely that construct validity would be called for when our constructs are stated in a theoretical language.

In considering theoretical concepts we shall review in Sub-Part A the positions of Hempel, Bergmann, and Carnap. Carnap advocates defining most psychological concepts in terms of a theoretical language, while Bergmann advocates the reverse. In Sub-Part B we shall try to resolve these differences. Finally in Sub-Part C we shall examine the surplus meaning of theoretical concepts.

A. Three Approaches.

(i) Carl G. Hempel "A theoretical system may then be conceived as an uninterpreted theory in axiomatic form which is characterized by (1) a specified set of primitive terms; these are not defined within the theory, and all other extra-logical terms of the theory are obtained from them by logical deduction" (1952, p. 33).

As such, this theoretical system makes no statements about the physical world. Empirical meaning can be given to the theory by linking some of the terms of the theory to some aspect(s) of the physical world. This may be done with the primitives of the theory, as in the case of physical geometry where we link the theoretical primitive 'point' to something in the physical world such as a pinpoint or the intersection of cross-hairs. This can also be done with the defined terms of the theoretical system, as in chemistry where the elements (as defined in terms of the primitives of atomic structure) are linked with certain gross chemical and physical characteristics of elements as found in the physical world.

This operation of giving a theoretical language empirical meaning has been discussed by Norman Campbell (1920) in terms of the hypothesis and the dictionary which interprets it and by Carnap (1956) as the theoretical language which is linked to the language of observables by correspondance rules.

The interpretation of the theoretical term in the language of observables is never a necessary and sufficient definition of the term. Theoretical terms are at best only partially assigned an empirical content. Therefore an empirical test of a theory can never prove it conclusively true or false.

This raises the question of whether or not a concept has "surplus meaning" beyond that assigned by the observables in terms of which it is defined. Some insist that concepts should not have surplus meaning since the aim of empiricism is to eliminate vagueness. Others insist that surplus meaning is necessary on heuristic grounds. This controversy will become more clearly defined in what follows.

(ii) Gustav Bergmann. The theoretical language outlined above certainly seems to have a surplus meaning beyond that assigned to it by its interpretation in observation language. What, then, are Bergmann's views on this?

Bergmann sees theory as being of two types: the empirical construct type and the axiomatic model type*. The latter achieves its meaning through the partial interpretation of a calculus. Thus in the kinetic theory of heat the laws of mechanics serve as a model for deriving laws of thermodynamics. Since the laws of mechanics are already formulated, it is

*The terms are from Spence, 1957

always possible, through further interpretation of them, to derive further laws of thermo-dynamics. In this sense he agrees that theories carry excess meaning (Bergmann, 1953).

In the empirical construct type of theory the meaning of the concepts rests in the definitions which link them to observables. Thus in the electromagnetic theory of light the constructs in question -- currents, electrical and optical fields, etc. -- have real referents rather than the calculational referents of an axiomatic model type theory (1943, p. 283). But empirical construct theories are theories only in the sense that they are general or abstract. The relm of empirical constructs can be seen as hierarchical: there is no point in the hierarchy where we can draw a line and say that all constructs beyond this point are theoretical (1943, p. 271).

This presentation shows that Bergmann accepts the use of a theoretical language which is similar to that proposed by Hempel. The difference lies in how many theories are encompassed by the axiomatic model type. Bergmann feels that no theories of this type exist in psychology. This is not because the nature of the discipline rules them out, but merely because psychology is not far enough advanced to make use of them. Or, perhaps more accurately, physiology is not yet far enough advanced to serve as a model for psychology.

Even the hypothetico-deductive approach of Hull is seen to be of the empirical construct type (Bergmann and Spence, 1941). This was true of Hull's Mathematico-deductive Theory of Rote Learning (1940) and was reaffirmed by Spence (1957) with regard to Hull's more ambitious, if less rigorous, schemes (1943, 1951). Brodbeck (1958) even goes a step further: she says that there is no counterpart to the axiomatic model type theory in any of the sciences dealing with macroscopic, observable variables, whether these be in the social or the physical sciences.

One final note on Bergmann's views is necessary. There is one sense in which he agrees that empirical construct theories have surplus meaning. This is the result of the inductive nature of science. First, there are some cases in which the right side of a definition includes "one or several generalities and negative existential statements" even after all defined terms have been eliminated from R. If this happens, the truth of R cannot be conclusively established by any finite set of observations, and thus neither can the truth of L.

In a later article (1953), Bergmann elaborates on this point. He notes that in an empirical construct theory there are two types of laws: elementary laws which decide the basic relationship and composition laws which apply the elementary laws to complex situations. The use of the composition laws allows us to generate many specific laws about specific situations which have not yet been formulated. Thus Bergmann concludes: "as soon as these intervening variables of behavior theory are put into use, they acquire automatically excess meaning" (p. 446).

From this outline of Bergmann's views on theory, it can be seen that he considers the definitions in terms of observables outlined in Part 1 as an adequate basis for any theoretical terms in use in psychology.

(iii) Rudolf Carnap. This presentation of Carnap's views on a theoretical language is taken from The Methodological Character of Theoretical Concepts (1956). He starts off by outlining the characteristics of an observation language. This is similar to what Hempel called the narrower interpretation of empiricism: it is required that all terms of this language be definable in terms of the observable primitive descriptive terms rather than merely being reducible to them by conditional definitions.

There is also a theoretical language which is very similar to the theoretical language outlined by Hempel. This is linked to the observation language by means of correspondance rules. What is different about Carnap's approach is that he introduces dispositional concepts as a class of terms occupying an intermediate position between the observation and the theoretical languages.

Dispositional concepts are introduced when we say that a thing has the disposition to react to S by R. The introduction of the first dispositional concept must be such that both S and R are expressable in terms of the observation language. Further dispositional concepts may be described in terms of previously defined dispositional concepts. (How this is done is left open, but Carnap lists as one possibility the use of reduction sentences.)

So far, Carnap's scheme resembles Hempel's fairly closely, except that Carnap sees dispositions as based on the observation language rather than being part of it. It also seems that Carnap's theoretical language is similar in form to Bergmann's axiomatic model theories: they both are calculi which are only partially interpreted in terms of observables. (They may not be similar in substance, since Bergmann seems to feel that the calculus must be borrowed from some other more completely formulated discipline, while Carnap does not.)

The difference between Carnap and Bergmann is clear, however, when we consider the range of concepts they include in the theoretical language (or axiomatic model theory). Carnap states that most of the theoretical terms of science, including psychology, are best reconstructed as terms in the theoretical language rather than as dispositional concepts. Bergmann, as we have already noted, feels that none of the terms of psychology can be stated in an axiomatic model theory. The consideration of this difference is the substance of Part B.

B. Dispositions versus Theoretical Language. Before comparing Carnap's and Bergmann's views on this matter, let us first look at the differences Carnap sees between the theoretical language and dispositional concepts.

The first difference is that a disposition is completely interpreted in observation language ("the specified relation between S and R constitutes the whole meaning of the term"), while a theoretical concept is never completely interpreted in observation language. In stating this difference, Carnap notes that formerly (1936-7) he attempted to allow for the openness of scientific terms by admitting the addition of further reduction sentences to dispositional terms. Now, however, he finds it better to represent this openness by couching scientific terms in a theoretical language.

The second difference is that "the negative result of a test for a disposition must be taken as conclusive proof that the disposition is not present" (p. 68). This is not the case for theoretical concepts. To see the significance of this, let us look at one of Carnap's examples.

Assume that the term "an IQ higher than 130" is a disposition to react to a test S by a certain kind of response R. This means that if a person reacts to S with R then we can say he has an IQ higher than 130. But if he does not react with R then we must say that he does not have an IQ higher than 130 no matter what other evidence we may have that his IQ is indeed over 130. This is true even if we later learn that when he took the test he was very depressed, but that this did not show in his behavior at the time of the test.

It might be thought that this difficulty could be avoided by saying that at the time of the test there must be no sign of a disturbed emotional state. But this condition was fulfilled. If we go further and say that any future indication of a disturbed emotional state invalidates the test, then the test procedure is made useless since it is not finished until the subject's death.

These difficulties are not encountered if we designate "an IQ higher than 130" as a theoretical term. The same test procedure with S and R may be accepted, but its specification is no longer an operational definition of the term. This means that the result of no single test, or any number of tests, is ever conclusive evidence for the presence of the disposition (in theoretical language) in the person. This non-conclusive nature of a test accords with the way tests are generally used by psychologists.

Before comparing this to Bergmann's position, we must first note Carnap's reasons for assigning concepts to the theoretical rather than the dispositional language. It might be thought that there are some basic linguistic properties of scientific concepts, other than those just cited, which influenced the choise. But if there are, Carnap does not mention them. In fact he makes a point of saying that the form of reconstruction of a term is not uniquely determined by the accepted formulations of science. For instance, temperature can be interpreted in either language (p. 66). This means that if we can show that the disadvantages of dispositional terms, listed above, do not necessarily hold, then there is no reason why our terms cannot be formulated in the dispositional language.

Let us see how Bergmann could deal with Carnap's position. There are three points to be covered.

(i) Bergmann's views on the first difference between dispositional and theoretical concepts mentioned by Carnap (ie, the latter is open in meaning while the former is not) have already been discussed in Part 1 above. There we said that Bergmann can allow for openness of meaning through his notion of significance. Also, by replacing old terms with new ones, concepts can in effect be revised. This means that Carnap's first objection to dispositional concepts, as he has formulated them, does not apply to dispositional terms as formulated by Bergmann in the language of observables.

(ii) Carnap's second difference (ie, tests must be accepted as conclusive evidence for the presence or absence of an entity if it is defined in dispositional terms) can also be handled by Bergmann. There are several points to be made.

Let us assume we have a test which is the operational definition of a dispositional concept. (This need not be the case: (a) the disposition could be defined in more general terms of which the test is an instance; (b) the test is used because it correlates highly with the (operational) definition of the disposition.) In this case the test is conclusive evidence for the presence or absence of the disposition, with one proviso: the test must be performed correctly.

This point seems to have been ignored by Carnap. And yet "errors of measurement" is a meaningful concept in Bergmann's terminology. An "objective observer" could tell us that we forgot to read the directions on the intelligence test to the subject. In other cases we may not know whether we have performed the operations correctly, but the possibility of error is recognized and this means that no test is conclusive evidence for the presence or absence of a construct.

(iii) The other aspect of Carnap's second objection to dispositional concepts (see pp. 44-45 above) was that they are unable to deal with the effects of some unaccounted for factor (eg, when depression affected the IQ). There are several reasons why such a factor would be unaccounted for.

(a) No test for this factor exists. This is obviously not what Carnap intends. If there is no test for the factor then we cannot say what the factor is nor when it operates. To say that our test of intelligence might not have really been measuring intelligence due to such a factor is no different than invoking gremlins or deamons which bias our results.

(b) A test for the factor exists, but it did not turn up the factor until well after the original IQ test was completed. Here the obvious question to ask is why the test was not given at the same time as the IQ test. If there are good reasons why this cannot be given together with the IQ test, then we can include this in our instructions for scoring the IQ test; we will then not complete the score of the original test until the later one is completed. But if there are no good reasons why the tests cannot be given together, then any failure to do so is an error in measurement.

(c) A test exists for this "disturbing factor" but failed to note its presence when given with the IQ test. It was not until later that evidence (including a retake of the test) showed that the test of the disturbing factor was in error when given the first time. This might be due to errors in measurement, in which case it is just an instance of what we discussed above. But if we rule out errors in measurement, then we are faced with the alternative that our test which measures the factor was unreliable.

Although the term reliability raises problems, it still can be dealt with in Bergmann's terminology. We saw on p. 19 above that we may define construct X as observable A and then use test B as a measure of X if it correlates highly with A. The validity of B refers to the degree it correlates with A over the long run. The variations in the correlation between A and B which occur when the correlation is repeated are called the reliability of B as a test of X.

The problem which Carnap raises, however, is whether the definition of a concept can be called reliable or not. The answer rests on whether or not we are able to assign some sort of meaning to concept X apart from its definition in terms of A. This raises the problem of existential reference which will be discussed in Part 3 below. But for the present we shall limit ourselves to the discussion of some of the reasons why the Logical Empiricists seem to have trouble with reliability.

In the first place, there are some response-inferred constructs where it is clearly meaningless to talk about the reliability of the definition. Let us say we have defined anxiety as the score on the GSR, the degree of muscle tension, the frequency of diarrhea, and the dilation of the pupils. Since this is our definition of anxiety, we have no criteria against which to correlate it to test for its reliability. On what grounds then could we call the definition unreliable?

The appeal to other signs of anxiety is meaningless. If there are any which we find important, then they should be included in our definition. This is the same problem we faced above pp. 29-33 when we decided that Pap's dilemma about the apple-tasting lemon was not really a dilemma.

We might then say, as Carnap did with intelligence, that some other factor biased the results of our test. There is a further answer to this besides the ones we gave above. Let us say that our measure of anxiety is performed by a handsome, personable, older man who is relaxed, sympathetic, and has little sexual interest in young women. We may discover that he consistently reports lower anxiety for young women than a dynamic, intellectual, young female research assistant. Surely one of the two measures is unreliable.

The answer to this dilemma is that anxiety fluctuates. The measures performed by the man were not unreliable; rather the women actually were less anxious with him than they were with the research assistant. We may want to eliminate the effects of the test situation on the construct being measured so that we can get at the degree the construct is present under "ordinary" conditions. But this is irrelevant to the fact that anxiety fluctuates with the nature of the social environment.

Even if our critic grants that this answer is correct, he may push on to other traits. What about intelligence? Surely this does not fluctuate from day to day and from environment to environment. The answer to this depends on how we define intelligence. We may define it operationally in terms of what the Stanford-Binet measures. In this case intelligence may well fluctuate. The person who takes the test when he is tired probably does less well than when he is not.

But this means simply that he is less intelligent when he is tired than when he is not. This, however, clashes with the ordinary notion of intelligence. Why?

It olashes with common sense because we usually think of intelligence as an underlying trait or faculty which is stable in a way in which anxiety is not. Such a notion is useful because it allows us to predict a person's performance on a variety of intellectual tasks. But what indications, other than common sense, do we have that such a trait exists independent of measures of it and that it is stable? I can think of none. Certainly we do not ascertain our intelligence by introspection in the way we might examine our pain or our memory of yesterday's lunch.

But we can avoid the narrowness of operational definitions by stating definitions in general terms. We can say that intelligence is the average ability with which a person solves an intellectual problem. In this case the Stanford-Binet is only one task out of many. This definition of course runs into the problem of all general definitions: do all instances of it fit the same laws? And how wide a range of material need we examine before establishing the average?

The drawbacks of such a broad definition will probably lead us to compromise and define intelligence in terms of several tests which have been given more than once. In this case we can say that any one test is reliable to the degree that it approximates the average score. It is in some such way as this that the Logical Empiricist deals with reliability. Any method which talks about reliability in terms of the deviations of a test from the assumed stability of a trait lead to confusion unless the assumed stability is given some empirical referent.

This means that the problems of dispositional concepts claimed by Carnap can be avoided by a person adopting the position of Bergmann. We have shown how (i) openness of meaning, (ii) errors of measurement, and (iii) reliability are all adequately handled by concepts of the observation language. But this only partially solves our problems with a theoretical language. Although we have shown that Bergmann's language is adequate to deal with Carnap's theoretical concepts, we have not shown that Carnap's theoretical language cannot be used. What is the meaning of terms defined in this language? In particular, do they have any surplus meaning?

C. Surplus Meaning and Carnap's Theoretical Language. The criterion of empirical meaningfulness for terms in the theoretical language proposed by Carnap is quite complex. The outline of this criterion is that any term M of the theoretical language is empirically meaningful only if it makes a difference in the prediction of an observable event. This means that if we have a set of postulates, T, in the theoretical language and a set of correspondance rules, C, then M is empirically meaningful only if we can infer some sentence, S_o, in the observation language which could not have been inferred from T and C alone. (S_o can be inferred either deductively or probabilisticly.)

This outline leaves out many of the refinements which

Carnap introduces. Nevertheless, it serves to show that the theoretical language has a meaning which is "open", but not a meaning which is unspecified. If the theoretical language is elaborate, there will be many postulates which are not empirically meaningful in the way we have outlined above. In this case these terms are not part of the meaning of the concepts we use unless they are specifically linked to the observation language and shown to make a difference there.

As an example of this let us look at an article by L. W. Beck (1950). He discusses the kinetic theory of gases as an example of the value of a theoretical language. From Boyle's and Charles's laws about gasses it is impossible to deduce Graham's law about the rate of diffusion of a gas through a porous membrane. But it is possible to deduce this from the kinetic theory of gasses.

In order to explain Boyle's and Charles's laws, the kinetic theory of gasses assumes that a gas is made up of a great number of tiny "balls" or molecules which are perfectly elastic, move at a high velocity, have contact times which are short compared to their time of free movement, etc. To explain Boyle's and Charles's laws it is not necessary to assume that the molecules have a mass. If this additional assumption is made, however, then we are able to deduce Graham's law (with the aid of some other postulates).

If we look at mass as the term M in the theoretical language, then we can say that it did not have empirical meaningfulness until Graham's law was deduced. Before that time it was an untested part of our theoretical language. It could serve as a hunch to be tested, and indeed, did so. But anyone who claimed that part of the empirical meaning of a molecule was that it had mass was in error until the concept of mass was shown to make a difference in the case of Graham's law.

Thus the surplus meaning of theoretical terms is something which serves as a hunch. To use it for anything other than this leads to the use of terms whose meaning is vague and unspecified. The empirical meaning of concepts in a theoretical language at any one time is clearly specified by those postulates whose empirical relevance has been demonstrated.

PART 3: EXISTENTIAL REFERENCE

The final difference between the Logical and Liberal Empiricists which we shall consider is over the existance of psychological constructs. Feigl (1950) claims that the difference between his approach and that of the "radical empiricist" is that he assigns factual reference to indirectly observable concepts while the radical empiricist deals with them in terms of "epistemic reduction".

Bergmann (1951) specifically took issue with Feigl on this point. He feels that Feigl's insistance on the reality of an electric field is misleading. This is because a philosophical realist considers the "real" chair to be just as unobservable as the "real" electric field. Therefore we only create confusion if we say that theoretical terms mean more than their definitions and that this meaning lies in their existence.

This reply to Feigl is unfortunate. Feigl explicitly says that "real" chairs are observable while electric fields are not. But when we look at what he does mean by existential reference, we see that this is an issue of concern to philosophers rather than psychologists. The realism he proposes is a semantic realism. The main claim of this is that it can give an adequate account of the operations of designation and reference, while the "radical empiricist" cannot.

This can be seen in cases where the evidence for an hypothesis is not logically equivalent to the hypothesis itself. Thus the evidence for the presence of an unobservable entity is located at a different point in space and time from the unobservable it makes reference to.

This seems to be a valid argument, but I cannot see that it makes much difference in our use of concepts. It means that using Feigl's approach we say that an electric field is some entity which we define as X. Bergmann, on the other hand, would probably say that the electric field is X. The logical difference between these two formulations makes an empirical difference if we assign empirical meaning to the phrase "some entity" which goes beyond the definition of the field itself. But Feigl does not do this, as a close reading of his article shows. To use the title of Sellar's article (1948) from which Feigl has drawn heavily, his realism is "a new way in words" rather than a realism which postulates entities with some sort of vague implications beyond their definitions.

There are two other aspects of existential reference which Fiegl mentions. First, the "radical empiricist" is unable to account for inaccuracies in measurement. Second, the meaning of the term is restricted to known methods of testing. The first point was covered above pp. 46-51 where we showed that Logical Empiricists can handle inaccuracies in measurement. The Logical Empiricists can deal with the second objection by the revision of their concepts to make them more significant (pp. 17-18).

Thus the differences between Feigl's position and Bergmann's Logical Empiricism are differences in the formulation of language rather than differences which are important for the user of the language. Feigl himself points this out when he quotes C. I. Lewis to the effect that a critical realist and a critical idealist should have no substantial quarrel with each other.

PART 4: SUMMARY

Let us now review the problems of the Logical and Liberal Empiricist approaches to the definition of concepts. What differences, if any, do these two approaches make?

A. Material Implication. It was decided that this is a logical problem of no interest to the practicing psychologist.

B. Range and Openness of Meaning. Explicit definitions can cover the same range of phenomena as reduction sentences. This is done in many ways. The range of knowledge we have about a concept beyond its definition can be included as its significance. We can construct empirical tests to substitute as indicators of the presence of a concept. In the case where we have different measures for different parameters of the concept, we can designate one as the definition and assume that if the other correlates with it in the ranges where they overlap, that it will measure the concept in ranges where they do not. This does not solve the logical problem, but as long as both measures enter the same laws in the same way, no confusion should arise.

Finally, explicit definitions can be general. This means that we can accomodate all the aspects of a concept under explicit definitions that are included under reduction sentences. The generality of explicit definitions plus the fact that they can be revised means that they also have the same openness of meaning that reduction sentences do. This removes one of the reasons Carnap gives for introducing psychological concepts in the theoretical language.

At times, when there is no aspect of a concept that is central or peripheral, then explicit definitions may often seem arbitrary in singling out one factor rather than another. But other than this, the user of a language should find no differences between reduction sentences and explicit definitions with regard to their range and openness of meaning.

C. Surplus Meaning. A concept may have surplus meaning in several ways:

(i) In defining the concept the basic definition may be universal in form and may thus never be completely tested for.

(ii) If we have an elementary and a composition law, then this applies to new (and sometimes phenomenally different) entities or processes besidesthose found in the elementary law. Here too we have surplus meaning.

(iii) In a theoretical language, all the elements may not yet have been given empirical meaning. This is similar to the instances of a general term which have not yet been tested to see if they fit the same laws as the rest of the concept. In both cases we have a surplus meaning beyond what has been empirically verified. This includes what Bergmann has called the axiomatic model type theory.

(iv) The surplus meaning of a concept can also be said to lie in its existential reference. But we have seen that no empirical meaning is given to this existing entity, apart from the observables which define it. This is a very limited type of surplus meaning, introduced to deal with linguistic problems.

The main difference which seems to arise between the Liberal and Logical Empiricists is that the former may be able to assign more surplus meaning to their concepts through a theoretical language than the latter can assign through the use of general concepts. But it seems questionable whether we can go very far in psychology in spelling out in a theoretical language relationships a concept has with other concepts. Although the requirements of Carnap's theoretical language are more liberal than those for Bergmann's axiomatic model theory, it remains for Carnap to give examples in psychology of this sort.

What is more to the point is that all of this surplus meaning serves a heuristic purpose rather than expanding the empirical meaning of the term. Even if Carnap could give examples of a theoretical language, Bergmann could deal with these by calling them hunches about the significance of the concept. The heuristic advantages of Carnap's approach are probably balanced by the dangers of assuming some aspect of the theoretical language to have empirical meaning before adequate evidence has been presented. Thus the differences between the Logical and Liberal Empiricists over surplus meaning are not of any great practical importance.

D. Confusions of Meaning. This is the first and major difference between explicit definitions and reduction sentences. Unless one (or one set) of the reduction sentences is designated as the definition of the concept, then confusions arise if the reduction sentences do not correlate highly. First, we do not know what behavioral manifestations to expect. Second, we do not know exactly what laws the concept will enter into nor to what degree it will do so. These ambiguities can only be avoided by designating one of the reduction sentences as the definition. This modification in effect adopts the major characteristic of the explicit definition approach. This is acceptable as long as caution is exercised to avoid the difficulties of explicit definitions pointed out by the Liberal Empiricists. Much of Section II, in fact, dealt with how these difficulties could be avoided.

E. Systematic Ordering of Items. We must attempt to establish the causal relationships into which the elements of a concept enter. What external variables must be controlled if an element is to be seen as part of the concept? Are we sure that some of the elements are not caused by other elements included in the concept? Are we sure that separate concepts do not have overlapping elements? Although these stipulations apply to both explicit definitions and reduction sentences, the multiplicity of the latter in defining some psychological concepts together with the use of what are essentially correlational techniques, means that reduction sentences run a greater risk of error here (see pp. 34-37).

F. Conclusions. The conclusion of this very long section is that we have provided a working solution of the problems raised by the differences between the Liberal and Logical Empiricists. As long as one keeps in mind the desirable aspects of the openness and range of meaning, surplus meaning, and systematic import while avoiding confusions of meaning, it matters little which approach one says he is using. In effect, the revisions of the two approaches noted here mean that there is no longer any clear dividing line between them.

Any philosopher having the perseverence to read this paper would probably be either amused or upset. In the first place my modification of reduction sentences turns them into explicit definitions. This is justified, however, by the elimination of those aspects of explicit definitions which they found objectionable. Second, and more seriously, there are surely many logical and philosophical points which I have left out or about which I am mistaken. But since this paper is aimed at the practicing psychologist, any philosophical mistakes I have made are of importance only if they upset the working solution which has been presented.

SECTION III

THE VALIDITY OF PSYCHOLOGICAL TESTS

In this section we shall apply the working solution of Section II to the problem of validity. The aim is to show that the terminology of construct validity does not fit the working solution. It must be revised if confusions are to be avoided. In revising it, some of the methods of construct validity will be retained while others will be modified or rejected.

PART 1: VALIDITY AND SIGNIFICANCE*

It will be remembered that the reason for undertaking the long discussions on concept definition in Section II were to provide a basis for evaluating the criticisms of construct validity (Section I Part 3). The gist of these criticisms was that all concepts must be defined in terms of observables. Any test for the concept must be validated against these observables. It is meaningless to ask whether the definition itself is valid. Construct validity is really no

*In the following discussion we shall use the words 'meaning' and 'significance' in the way the Logical Empiricists have defined them. (See pp. 18-20; Bergmann, 1951, p. 94; Bechtoldt, p. 623).

more than the search to give our concepts significance.

If construct validity is to defend itself against this criticism, it must show that concepts are not defined in this way. The two appeals advocates of construct validity might make are either to reduction sentences or to some sort of theoretical language.

A. Reduction Sentences. The argument might be made that a concept may be defined in terms of reduction sentences. In this case, any test of the concept must correlate with all of the reduction sentences together rather than with just one criterion which has been designated the operational definition.

But we noted in Section II Part 1 F that the use of reduction sentences is possible only when they all correlate highly. Since this is not the case in psychology, it is necessary to designate one of the reduction sentences (or a combination of them) as the explicit definition. Thus the attempt to use reduction sentences does not lead to a refutation of the criticisms of construct validity. This point is easier to accept in light of the many refinements of explicit definitions covered in Section II. The most important refinement was the need to constantly redefine our concepts so as to make them more significant.

Since this search for significance is meant to replace what Cronbach and Meehl mean by construct validity, what are the differences between significance and validity, as used by the Logical Empiricists? The validity of a test is the degree to which a test correlates with the observables which define the concept we are interested in. In correlating two we must note what other factors have been controlled for; it is unlikely in the social sciences that the test is a function of only the observables which define our concept.

It is purely for conventional reasons that we do not call a test significant. If the test is a valid measure of a concept which enters into lawful relations with other concepts, then the test of the concept must enter into these laws too. It is therefore significant. But since we are interested in the test only as a more economical way to measure the concept, the significance of the latter is all we are concerned with. It only leads to confusion to speak of the significance of the test, especially if it is not completely valid.

Another point will show the closeness in meaning between validity and significance: they both refer to the same empirical relationship. Thus, if test B correlates highly with observables A which define concept X, then we say that B is a valid test of X. But the significance of X lies in its empirical relationships with other observables. Thus B is part of the significance of X.

Why differentiate between significance and validity if they are just different ways of looking at the same relationship? The reason is that in the case of validity we have a clear idea, stated in observables, of what we are looking for. In the case of significance we do not. In the above example, if B shows no correlation with A, then we know it is not a valid test of X. But this failure of A and B to correlate does not allow us to say that X has no significance. There is always the possibility that A will correlate with some other observable than B, thus making X significant.

Even this difference can be whittled down by showing that in some cases of significance we do, in fact, know what we are looking for. Imagine that concept X, defined as A, is significant due to its lawful relationships with C, D, and E. If we seek to improve the significance of X by redefining it as A_1, our primary aim will be to improve the accuracy of the laws with C, D, and E. But what is the difference between revising A so that it correlates more highly with C, D, and E and revising B so that it correlates more highly with A? In both cases we have specific criteria we are aiming at.

But even under these conditions there are still two differences. First, in validating B against A we seek a correlation of 1.00, while in examining the significance of A with regard to C, D, and E, we probably do not, since C, D, and E may be functions of other variables besides A. Second, in validity we correlate the test to a specific definition in terms of observables; in significance, there are an infinite number of possible observables with which our original concept may correlate.

Two questions now arise. First, is this difference between validity and significance really very important? Second, if it is, why should we not say that this is the difference between criterion validity on one hand and construct validity on the other? Are the Logical Empiricists really doing anything more than substituting the word 'significance' for 'construct validity'? The answer is that the difference between significance and validity does make a difference in actual psychological practice; and that construct validity should be eliminated from our terminology since it causes confusion by including both validity and significance under the same heading. But we must wait until the end of Section III before the reasons for this conclusion are complete.

B. Theoretical Language. It might seem that a theoretical language would solve the problems of construct validity. A theoretical language states terms and defines the relationships between them. In formulating empirical concepts our aim is to approximate the meaning of the concept of the theoretical language. The advocate of construct validity could claim that in this case we have a clear idea of what our concept should be. Here is an instance of defining a concept where it is meaningful to ask whether it is valid.

But such an assertion does not hold up under closer examination. It is clear that the first correspondance rule is applied to a theoretical term as a matter of convention. There is no empirical meaning yet assigned to the theoretical language and so it is impossible to speak of either validity or significance since both of these are empirical relationships.

The advocate of a theoretical language might agree to this and yet insist that the assigning of further correspondance rules to other theoretical terms is an empirical matter, and in fact is a case of validity. Thus our theoretical language may say that A is lawfully related to X•Y•Z. If we give A empirical meaning by defining it in terms of A_e (empirical A), then we must define X_e such that A_e is lawfully related to X_e. If it is not, then this shows that X_e is an invalid operationalization of X.

This argument does not hold, however. We must remember that validity is possible only when a concept is clearly defined in terms of observables. This is obviously not the case with X, which is only defined in terms of its relationships in the theoretical language. Thus the assigning of a correspondance rule in this case is a matter of significance and not of validity. We seek to define our terms so that they enter the significant relationships which are hypothesized by the theoretical language.

Again the advocate of a theoretical language might object. The above conclusion, that the assigning of correspondance rules is a matter of significance, rests on the assumption that the theoretical language serves only as a set of unverified hypotheses. But in some cases these hypothetical relationships may have been partially verified empirically. Let us imagine in our above example that X, Y, and Z all have characteristics o and p such that we call them instances of W (which is a general term defined as o and p). Thus A is lawfully related to W. If, further, we find that Y_e and Z_e are lawfully related to A_e, then we can say that we have given empirical weight to W and thus to X and that the correspondance of X_e to X is now a matter of validity.

But this argument derives its force by ignoring the role of o and p. If the relationship of X_e to X is one of validity due to the empirical meaning we have assigned to X, then just what is this empirical meaning? If the empirical meaning is that X is related to A, then this is the same as the previous case, which we decided was significance. The only other empirical meaning must come from o and p. If we have given them an empirical definition through a correspondance rule, then we have the general term we defined as o_e and p_e. In this case we must choose X_e so that it has characteristics o_e and p_e. But this is not a matter of either validity or significance, but of finding an instance of an existing definition. If X_e is not lawfully related to A_e this is a matter of significance, meaning that we must either redefine o_e and p_e, W, or A_e or else reject the theoretical language. This position is the same as that taken in Section II Part 1 G with regard to general concepts.

C. Conclusions. The above analysis has shown that neither reduction sentences nor a theoretical language give the advocates of construct validity a way to avoid the criticisms of Section I, Part 3. We are concerned with validity when we are constructing an alternate test for a concept already defined in terms of observables. We aim at complete correlation between the two. Significance is the concern with the lawful relation of one concept to another. There is no definite set of concepts with which any particular concept should be significant. Also, the relationship will often be in terms of correlations well below unity.

PART 2: INTERVENING VARIABLES*

Although Part 1 above covers the differences between validity and significance, a special problem arises when we deal with internal processes of an organism. This is the problem of defining these internal processes in terms of observables. If we cannot do this, then we cannot raise the question of whether our tests of these concepts are valid. This leaves us with a dilemma. Either we say that these internal processes <u>are</u> the observables in terms of which they are defined, which leads to a contradiction in terms (how can an internal process be a directly observable process?); or we say that we can know these processes through their effects, which is what Cronbach and Meehl said and which leads to a blurring of the validity-significance distinction.

This problem is of special concern to behaviorist psychologists. For those outside this tradition (and even for some within it) there are two ways to define internal processes in terms of observables.

The first is to use physiological concepts. The logic of this approach is interesting since it allows us to define processes, which we often cannot observe, in terms of an observation language. Let us use the obvious example of hunger, defined as contractions of the stomach. Contractions of the

*In this paper 'intervening variable' is the generic term referring to all internal processes of an organism. But when referring to MacCorquodale and Meehl's (1948) distinction between 'intervening variables' and 'hypothetical constructs', we shall follow Meissner (1960) and call the former 'IV' and the latter 'HC'.

stomach is obviously a concept which is stated in terms of observables. The problem arises because we cannot observe the actual contractions themselves. The solution is to place a balloon in the stomach with a tube leading out and fill the balloon with a body-temperature liquid. The fluctuations of the liquid are taken to indicate the contractions of the stomach.

Here it is meaningful to ask if our measure is valid, since our original concept is clearly defined in the language of observables. But we cannot perform validity using the usual correlational techniques. Therefore we must infer the cause of the fluctuations of the liquid using laws which have been independently established. Thus we know the empirical laws that when the volume of a container grows smaller it will expell any fluid in it, temperature and pressure remaining constant. From this we infer that the liquid expelled was the result of stomach contractions. This is not a sound inferrence logically, but it is typical of scientific reasoning. As long as we can show that "other things were equal" (eg, it was not an excess flow of gastric juice which caused the volume of the baloon to decrease) this is an acceptable method.

Thus we maintain the difference between significance and validity. It is possible for our measure of hunger, defined as stomach contractions, to be completely valid and yet for the stomach contractions themselves to have no relationship to "eating activity". In this case we would have a valid measure of a concept which has no significance (at least with regard to our hypotheses).

But some may object that this contradicts the conclusion reached in the section on theoretical concepts. There it was decided that whether or not the empirical concept X_e is an adequate operationalization of X is a matter of significance in spite of the fact that we assumed that X enters a lawful relationship with A. Is not X's lawful relationship to A_e the same as the lawful relationship between the stomach contractions and the liquid coming out of the tube? In the case of both X and stomach contractions we are dealing with something which is not observable.

But on closer examination, the analogy does not hold. In the case of X we have a concept defined only in relation to other concepts of a theoretical language; X_e is meant to give empirical meaning to it. In the case of stomach contractions we have something which is defined in the language of observables but which cannot for technical reasons be observed; the liquid coming out of the tube is not meant to be an operational definition of stomach contractions, but rather an empirical correlate of them.

Second, the empirical laws do not serve a parallel purpose. We said that X_e should be related to A_e because X_e is an instance of W_e which is lawfully related to A_e. Here our empirical law connects X_e to A_e rather than X_e to X. But in the case of stomach contractions we have a law relating the liquid in the tube to the contractions themselves. This is not relating the liquid in the tube to some other observable (as in X_e to A_e) but relating it to our unobservable stomach contractions (as in X_e to X).

Thus we can use independently validated laws to ascertain the validity of an observable phenomenon as a measure of an unobservable one. This means that when we are dealing with physiological processes within the organism we can ask meaningfully whether our measure of them is valid, but that when we are dealing with an operational definition of a theoretical concept, we cannot.

The other way to define internal processes in terms of observables is by making reference to mental data. In this case a person's introspective report of his internal states is taken as a basic or undefined term in our language. We know it by acquaintance rather than definition. Such an approach means that we validate a test of a concept against a person's observations of the concept within himself. But since this approach raises many thorny issues of its own, we shall not discuss it until the next section.

Many psychologists refuse to use either of these two approaches. They may feel that physiological data is sufficient to deal with something like hunger in terms of stomach contractions or anxiety in terms of hippuric acid, but that physiology is not yet far enough advanced to deal with many important psychological processes. On the other hand, if they are in the behaviorist tradition, they will surely reject the use of mental data.

The rejection of these two approaches is of no concern to psychologists who see no need for dealing with processes within the organism. Thus, Skinner defines such concepts as 'drive' purely in observable terms. This is an aspect of overt behavior and nothing more. In many cases this might be satisfactory. A good case could be made for treating such concepts as intelligence or anxiety in this way: they are dispositions and should be defined explicitly in the same way that we define dispositions like magnetism.

The drawback of this approach is two-fold. First, concepts defined in terms of responses of the organism cannot be used to explain these responses. This point was made by Bechtoldt in discussing response inferred constructs.(see p. 10 above). The second, related point is that often there is reason to believe that something happens in the organism between stimulus and response and we want to get at these processes directly and examine their variations.

For example, we may want to use the concept 'drive'. We may decide to give meaning to this concept by means of two overt physiological measurements: GSR and pulse rate. But this raises problems: is this a definition or a test of drive? Let us say that this is a definition of drive. This means that drive <u>is</u> these observables.

What, it may be asked, is the difference between this and the non-intervening-variable approach? It is different in intent, in that we see it as an intervening variable: something between the stimulus and the response. In this way it may be used to explain the response, unlike response inferred constructs. But such an approach runs into the problem of causality. It is hard to see how something like high GSR can lead to something else like "higher levels of performance in aversive forms of conditioning" (Spence, 1958). Another way of looking at this is that such a definition of drive is in terms of its peripheral rather than its central aspects.

If we accept this criticism, then we must say that drive is reflected in GSR and pulse rate. This approach is not uncommon. Even Tolman seemed to feel that 'demand' was a function of food deprivation; the food deprivation is thus an indication of demand. This is true even if we attempt to make demand an IV by saying that it has no physiological referent and that we know nothing about demand other than what we state in the definition. Also Malmo, who advocates physiological measures of drive, speaks of these as concomitants of drive rather than the drive itself (1958, p. 234).*

Such an approach runs into new difficulties. If drive is something which is reflected in GSR and pulse rate, or which is measured by them, then how do we know this is a valid measure? The answer is that we do not, since it is not defined in terms of observables (other than our measure of it). This can of course be avoided by going back to saying that drive is GSR plus pulse rate, but then we are back with the problems we wanted to avoid.

But there is still one way in which empirical meaning could be assigned to an unobservable concept like drive. We

*This view may seem to be an adoption of Feigl's suggestions about existential reference. In a way it is, in that it avoids the radical empiricism of saying that drive is GSR. But Feigl would insist on existential reference even in the case of dispositions like magnetism, while this approach would not necessarily do so.

can say that we know when drive is present by observing the variations in dependent variables. These variations are assumed to be functions of our unobservable construct. But this "assigning of meaning" can not be a definition of drive or we are back with the first position of response defined constructs which have no explanatory power. Thus, the empirical knowledge that is given to drive is its significance. Moreover, the temptation becomes very strong to validate our measure of drive against the concepts which make up the significance of drive. In doing so, we are of course performing construct validity.

The Logical Empiricists are faced here with a type of concept which they want to use and which they cannot define in terms of observables. They can avoid this problem only by dropping the use of intervening variables or by so defining them that they sacrifice much of their theoretical significance. But if they refuse to do this, they are faced by the problems of construct validity. It will now be argued that the advantages of such undefined, unobservable concepts are outweighed by their disadvantages.

First, the claim that such concepts are undefined is not insignificant. What is something like drive if we cannot know anything about it other than its significance? To deal with such a concept is certainly contrary to the intent of Empiricist philosophy, both Logical and Liberal. Feigl's realism was a semantic realism and nothing more. Pap's use of reduction sentences which ignored the analytic-synthetic distinction was successful only where there was a very high correlation between them. Carnap's theoretical language was not intended as an explanatory system apart from the empirical meaning assigned to it.

It might then be claimed that drive is a theoretical concept similar to those found in Bergmann's axiomatic-model type theories. But surely psychology is not so advanced as to make use of the abstract calculi which such an approach seems to require. Furthermore, in such an approach it is seldom that we attempt to measure the variations of the theoretical entities directly. We do not try to measure the speed of one molecule in a gas directly; rather we infer the effects of such an entity, given certain unobservable properties. But in the case of drive we attempt to measure its variations directly. And here we run into the great question of what it is that we are measuring. Such measures were attempted in physics only long after the theory had shown its great usefulness. In psychology we seem to reverse the process: if only we could perform the measure, then surely it would be useful.

Turning to less theoretical considerations, it will be noted that there is no effective difference between GSR and pulse rate as a definition of drive and GSR and pulse rate as a test of drive. In both cases we have the same concept. In both cases when we attempt to improve the concept, we aim at higher correlations with its hypothesized consequents (such as higher levels of performance in aversive forms of conditioning). In the former case it is called significance, in the latter, validity plus significance (or construct validity). But there is no difference between these two operations.

Parsimony would therefore require the dropping of the term validity.

The only appeal left is that of the heuristic properties of concepts like drive. They bring together and give meaning to a large number of observables which are phenomenally quite diverse. But this can be accomplished by explicit definitions which are general. These can single out certain properties of observables and thus bring together many phenomena which are otherwise quite different. Such an approach is preferrable to the positing of some unobservable concept with no explicit definition. As we noted in Section II Part 1 G on general concepts, it is better to approach these explicitly so that we know what is to be included and what not, rather than tying many entities together under the guise of reduction sentences merely because they correlate.

Therefore it is concluded that in dealing with intervening variables, these should be explicitly defined in terms of observables, rather than implying that the definition is a test or measure of some internal process. This means that not only do we eliminate the use of HCs but also many of the IVs like Tolman's demand. If this leaves us in the position of trying to explain one set of responses in terms of another set, this is all the more reason to turn to mental and internal physiological data. Feigl's (1955) criticism of the "psychology of the empty organism" should lead not to unobservable and undefined theoretical concepts like drive but to the use of mental and physiological concepts.*

*Some have interpreted HCs as the call for a physiological

PART 3: VERBAL BEHAVIOR

It is clear by now that this paper advocates the definition of psychological concepts in terms of observables. But what if one of the defining characteristics of a trait is the answer to a question? Is verbal behavior to be regarded as the same as other behavior?

Some think it is. "The behavior scientist ... accepts verbal response as just one more form of behavior and he proposes to use this type of data in exactly the same manner as he does other types of behavior variables. Thus he attempts to discover laws relating verbal responses to environmental events of the past or present, and he seeks to find what relations they have to other types of response variables" (Spence, 1948, p. 574). For convenience we shall call this the behavioral approach to language.

The other position is that language makes reference to something, that it is symbolic. Feigl (1958, pp. 417-18) points out that discourses involving aboutness (ie, intentional or referential terms) cannot be reduced to the language of behavioral or neurophysiological description. The relation of designation is not an empirical one, but a construct of semantical discourse. This is the referential approach to language.

Let us look at an example to see what the difference

psychology. But physiological data was only part of the surplus meaning of a behaviorally defined concept. What is advocated here is the explicit definition of psychological concepts in terms of internal physiological data.

between these two views is. Assume that the answer "yes" to the question "Do you have diarrhea more often than twice a month?" is part of the behavior which makes up our definition of anxiety. If we adopt the behavioral approach, then we must say that the utterance of the word "yes" is part of the behavior making up anxiety; it is not the diarrhea itself. We accept this response as part of the definition of anxiety because it correlates with the other defining properties and because its inclusion makes our definition of anxiety more significant.

Put in such simple terms as these, it is hard to imagine anyone advocating the behavioral approach. But there are other reasons, besides its apparent foolishness, which rule against its use. If we adhere to it strictly, then we must use correlational techniques rather than a dictionary to discover what a word means (to ask what a word means is already a more liberal view than the hard-line approach of the preceding paragraph). This leads to awkwardness. Either we must perform new correlations for every new verbal stimulus and response, or we must establish classes of concepts which are similar. What is this similarity, though? If someone answers our question with "guess" rather than "yes", are these effectively the same response since they sound the same?

Another problem lies in the formulation of a general definition of anxiety. It is conceivable that the other aspects of anxiety such as high GSR, trembling, tenseness, high pulse rate could be grouped together under a general concept. Being physiological, it would seem that diarrhea would fit in this definition. But it is hard to see how the word "yes" could possibly be included.

Along the same lines, the stimuli which lead to the rest of the defining properties are different from the stimulus (ie, the question about diarrhea) which leads to the "yes". Thus it must be asked whether the inclusion of "yes" is not done to the detriment of the systematic import (in this case functional unity) of our concept.

The above objections should certainly cast some doubt on the correctness and usefulness of the behavioral approach to language. The referential approach fares better. Using this we would consider the diarrhea as a defining property of anxiety rather than the answer "yes". But since it is mainly a philosophical viewpoint, the referential approach alone is not sufficient for empirical purposes. We must show in any particular case that the answer "yes" is a function of the referent alone and not of something else. We cannot see the significance of the answer to a question solely with the use of a dictionary.

Thus we must try to rule out any confusions as to what the question makes reference to. We must attempt to eliminate the effects of response set, lying, defense mechanisms, etc. The attempt to correct for these is the attempt to clarify the lines of causality. We want to make sure that the referent alone is what influences the answer. Thus, in a way, the referent approach uses verbal behavior as a test of a trait which has already been defined in terms of observables.

There is a third way to use verbal behavior which falls between the above two approaches. This attempts to use the answers to questions as signs or tests of traits, but does so while ignoring the referent of the question. For example, the answer "false" to the statement "It takes a lot of argument to convince some people of the truth" is a sign of hysteria in the MMPI. This also rejects the behavioral approach in that it does not say that this answer to this question is part of the definition of hysteria. When used in this way, we shall call the verbal behavior a sign of a concept.

This is an acceptable approach as long as certain precautions are taken. First we must have an explicit definition of the trait of which this question and answer is a sign. This is not always the case. Sometimes the original criteria against which a test is validated are ignored and the test itself is taken to be the basic definition of the concept. Often when this is done, no attempt is made to show that the test is more significant than the original criteria. Even if it were, we would not want to use the test as the definition due to the problems of the behavioral approach listed above.

The other way in which the original criteria are ignored is when it is claimed that the test is a measure of some concept which is unobservable and thus has no explicit definition. This approach is exactly the same as that of Part 2 above where the unobservable concept 'drive' was postulated. We reject such a use of verbal behavior as signs of unobservables for the same reasons that we rejected concepts like drive which were known only through their significance but where basic definition was lacking.

But even when we have explicitly defined criteria against which to correlate our verbal signs, this approach must be used with caution. It is much harder to show here that our sign is mainly a function of the given criterion and not of some correlate of the criterion. This is much easier to show in the referential approach than in the sign approach. Also this approach used by itself is atheoretical in that it does not try to show why the verbal behavior is a sign of some trait. In the case of the referential approach we know why this is related to the criterion. If the sign approach is to be used, we ought to devote a fair amount of effort to discovering its relationship to its criterion.*

Finally, the sign approach runs the risk that some people may answer the question in terms of its referent rather than reading their own emotions, or whatever, into it. This might well be the case with those high in intelligence, compulsiveness, or defensiveness. We must show that a question and answer are either a sign or a referent for all types of people or else some sort of configural scoring must be used <u>on the item level</u> in order to take this into account.

Two final points must be made. The above emphasis on the referent approach makes no assertions about how a language is formed. It may be that linguistic rules evolve behavioristically (<u>eg</u>, the work of G. H. Mead), but this does

*Mellenbruch (1962) proposes the use of hypnosis as one way to do this.

not mean that verbal behavior fails to make reference to other things. Feigl can still maintain that the relation of designation is not an empirical one.

Second, this discussion of verbal behavior takes no stand on the use of mental data. It is possible to insist on the referential nature of language while at the same time insisting that mental data has no place in a scientific psychology.

The conclusions of this section are: (1) Traits should never be defined in terms of verbal behavior; they must instead be defined in terms of the actual behavior itself. (2) Tests may be devised using verbal behavior as long as it can be shown that the verbal behavior is an accurate reflection of the referent, or that it is an accurate sign of the trait. Language is a tool to get at the behavior we are interested in rather than being itself the center of our concern (except, of course, when studying language).

PART 4: MENTAL DATA

So far in this paper we have assumed a working consensus existed between the Liberal and Logical Empiricists. But in what follows we shall depart from the views of the Logical Empiricists by advocating the use of mental data. It is my belief that such an approach is completely within an Empiricist philosophy: it is part of our knowledge by acquaintance (Russell, 1912 p. 49). Not only was Hume a phenomenalist (Feigl, 1958, p. 371), but Carnap (1956, pp. 70-71), Feigl (1958, p. 399) and Pap (1951) have all to a greater or lesser degree advocated the use of mental data.

There are two problems which any full discussion of mental data should discuss. First, it should be clearly stated what is meant by mental data. This could refer to (a) raw-feels: hunger, pain, fear, etc; (b) processes such as thinking, willing, intention, rationality, etc; (c) constructs which fall somewhere in between "a" and "b": the self, the unconscious, memory, etc; (d) the meaning of the situation - the interpretation a person gives to his actions and his environment. Even with such loose catagories as these, it can be seen that not all mentalists will advocate the use of all catagories. The Liberal Empiricists would surely prefer the use of the raw-feels over processes and constructs. Some British philosophers, on the other hand, have evolved a completely different view of motivation and human causality by concentrating on processes to the neglect of raw-feels (eg, Peters, 1958; Anscombe, 1957).

Second, any thorough discussion of mental data should consider the mind-body problem. What do we refer to when we speak of mental data? How is this related to the physical world? As Feigl (1958, p. 372) puts it: "What are the logical relations of the raw-feel-talk (phenomenal terms, if not phenomenal language) to the terms and statements in the language of behavior (or neurophysiology)?" It is hard to imagine anyone doing sophisticated work with mental data until he has made clear just what he is talking about.

Having stressed the importance of these two points, I must now admit that neither will be covered in what follows.

Both of these problems are so complex that I could not possibly do justice to them. Therefore, what follows will be limited to a short discussion aimed at showing that as a start we should accept or reject mental data on pragmatic rather than on methodological or philosophical grounds.

Hardly anyone has claimed that statements about mental data (at least with regard to raw-feels) are meaningless. Watson was one of the few who has done so, and even then only in his more extreme moments. Most behaviorists admit the existence of the mental, but at the same time rule out its use in the science of psychology.

Spence (1957) rejects the use of mental data on the grounds that it does not provide the high degree of intersubjective consistency of observation which is possible with regard to physical data. But there are obviously other criteria for a science besides intersubjective reliability. These are not considered by Spence, however.

The most obvious of these is significance. It is an empirical question as to whether concepts are more significant when based on mental rather than physical data. For example, we could define anxiety in three different ways: by the subjective report of a person on how afraid he is; by some physiological characteristic such as the amount of hippuric acid in the blood; and by behavioral manifestations such as tenseness, trembling, intellectual inefficiency, etc. We may then try out these measures first for their reliability (to the degree we are sure we want anxiety to be a stable concept; see pp. 48-51 above) and second for their significance -- the degree it relates to other concepts and is able to predict behavior. We will then use that concept which is most reliable and most significant, be it mental, physiological, or behavioral.

Spence's rejection of mental data on the grounds of one methodological criterion is unfortunate. This is doubly so because it could be argued that the main reason for intersubjective reliability is that this increases the significance of a concept. The emphasis placed on significance here is really an emphasis on the use of constructs which work well; and this, of course, is a pragmatic outlook.

Another argument against the use of mental data is that it indeed has no significance. Thus Bergmann (1951) tries to show that anything which can be said about mental data must be said in terms of its physiological correlates. Therefore all significant relationships in psychology can be stated in terms of physical data; mental data is useless, other than to provide us with hunches which must be verified in terms of physical data.

This view runs into difficulty because there is one thing which we can say about a mental state beyond its physiological correlates. This is the statement: "I feel it" or "There it is". These are "pointing statements", the basis of all knowledge by acquaintance. If we were to limit our statements about mental data to what we can infer from another person's behavior (disregarding his verbal statements since this brings up the problem of reference), then indeed we could say nothing about his mental states other than in terms of physical data.

But such an approach would not be fruitful without being guided by the mental data we are trying to get at. And mental data as "hunches" is not enough. An example will show why.

In defining anxiety we may be faced with two definitions: anxiety_1 is the trembling of hands, dilated pupils, and a high GSR; anxiety_2 is frequent diarrhea, intellectual inefficiency, and tenseness. If we assume that anxiety is a purely behavioral term, then we must choose between these two definitions on the grounds of significance. If we constructed anxiety_1 because we had a hunch based on mental data, that this comes close to being "what anxiety really is", still we do not accept it because it seems to approximate our hunch. Instead we will accept it only because it is more significant than anxiety_2 (we would probably consider reliability and functional unity as well).

On the other hand we may choose to see anxiety as a mental state. In this case anxiety_2 and anxiety_1 would be used as tests which should correlate with our basic notion of anxiety. Thus we would not accept one over the other on the grounds of which is most significant; rather we would accept one over the other by the degree to which it correlates with the mental state, as independently observed. Thus if mental data serves only as a hunch we do not validate our tests against it; if it serves as the observables to which our concepts refer, then any behavioral test must be validated against it.

Having clarified this, let us turn to the mental data itself. The stumbling block of the mentalist approach is of course the independent observation of mental states. The only person who can do this is the subject himself. And how do we know that what subject A calls fear is the same as what B calls fear? B was taught to use 'fear' to apply to an internal state (eg, B was on the edge of a cliff, was seen to tremble, and was presented with the word 'fear'). But how do we know that fear does not apply to these physical circumstances (ie, the trembling of his hands when placed in dangerous situations) rather than to the mental state itself? And if it does refer to the mental state, how do we know that it is similar to A's?

The answer is that we do not know it is similar to A's, any more than we know that the green or the rectangle which A sees is similar to the green or the rectangle which B sees when viewing the same object from the same place. But B does know that 'fear' is meant to refer to some emotion. And surely B is able to differentiate between his introspective awareness of an emotion and his awareness of its external correlates, at least in the sense that he uses different methods of observation. Thus if B feels some emotion while being presented the word 'fear', he will associate that emotion with 'fear'. And as long as this emotion is introspectively different from other emotions, then he has a mental referent for the word 'fear'.

Thus our independent measure of a mental state is the subject's introspection, the results of which are conveyed to us verbally. The problems of validity connected with this are very great. Not only must we ascertain the validity of the introspection, but we must also be sure that the verbal report is accurate. It is permissable to treat the verbal report as a sign of the mental state so long as we can validate this against a referential verbal report of the introspection.)

But how can we validate a person's introspection of his own mental states when it is just this introspection which is to serve as our independent measure of the mental state? This seems very similar to the unobservable drive of Part 2 above which we decided to reject because of its inadequacies. But in spite of similarities, it is still possible to validate introspective observations of mental states. This is done by accumulating knowledge about the factors which lead to inaccurate observation. This includes defense mechanisms, lying, etc. We also investigate ways in which we can recognize the operation of these factors: we attempt to establish rapport, we note pauses in the conversation, etc. These are of course the methods which psychiatrists and clinical psychologists have developed; many of these are still in the stage of an art rather than a science.

Our knowledge of these factors is built up by our own introspection and by seeing whether a subject will modify his introspective reports in the face of suggestions that biasing factors are at work. Pap (1951) has shown that the inductive method by which we make inferrences from our own mental states to those of others can be performed within the accepted cannons of proper scientific method. The validation of introspection is difficult and the results often tenuous. Spence's objections about low intersubjective reliability are not to be lightly dismissed.

The conclusions of this section are as follows: (a) It seems that the arguments against the use of mental data are philosophical or methodological. Not only can these arguments be questioned on their own grounds, but they ignore the vital criteria of significance (in its pragmatic sense). (b) The use of mental data is not outside the Empiricist tradition. Although very difficult, it is possible to validate our measures of it. (c) In any full examination of mental data both the mind-body problem and types of mental data should be considered.

One final comment is in order. It is my guess that construct validity was acceptable to many people because it allowed the implicit use of mental data. This comes through the failure of construct validity to demand the definition of concepts in terms of observables, the vagueness of surplus meaning associated with this approach, and the imprecise nature of the nomological network. A good deal of this vagueness is surely the fault of the user. Nevertheless, construct validity probably appeals to the clinically oriented who seek a methodology which allows them to use mental data without the appearance that they are actually doing so.

Such a result was surely not the intention of Cronbach and Meehl. Therefore the explicit use of mental data advocated here is certainly preferrable to any implicit use allowed by a loose interpretation of the proceedures of construct validity.

PART 5: THE METHODS OF CONSTRUCT VALIDITY

The theoretical parts of this paper are completed. It only remains to be shown that these do make a difference in the actual performance of test validation. In what follows, we shall review the methods of construct validity listed in Section I Part 2, showing what revisions, if any, are necessary.

A. Intertest Validity. It will be remembered that this operation was divided into two catagories. In the first was included the correlation of tests all measuring the same trait. This was exemplified in the multimethod-multitrait matrix of Campbell and Fiske. In the second was included the correlation of a test with constructs which are hypothesized to be related to the construct which our test is supposed to measure.

Looking at Campbell and Fiske's method in the light of the conclusions of this paper, there are several reasons for rejecting it. The first is that it fails to give a clear definition of the construct. We are given several different tests of a construct, which correlate at well below unity, with the implication that all of these together are measures of (but do not define) the construct. But this is either the reduction sentence approach (which we rejected unless its elements correlate at close to unity) or the approach in which an undefined construct is known only through its significance (see pp. 74-76 for the reasons this was rejected).

The only way to avoid this difficulty is to say that all the tests together make up an explicit definition of the construct. Campbell and Fiske imply something like this when they say that their approach is one of multi-operationalism. But if they carry out the implications of this, then all of the tests would have to be used together whenever we wish to test for the construct. Also, if this is multi-operationalism, then there is no question as to whether any one of the tests is valid, since they all are part of the definition.

Finally, Campbell and Fiske fail to mention the importance of significance in deciding whether or not to include a test as part of their multi-operational definition. The fact of correlation is not enough. It must be shown that by adding another element to our construct, we have increased its significance in relation to other constructs.

For these reasons we reject Campbell and Fiske's methods as they stated them. But with modifications, they can still be used. The attempt to use different methods to define the same trait can be seen as the attempt to rule out irrelevant factors which may lower the significance of a construct. As such it is part of intratest validity and will be discussed below.

Let us now turn to the second type of intertest validity. Here we validate our test not against the construct which it is supposed to measure, but against other constructs which are related to it by the nomological network. But this is necessary only where we are unable to define the construct in terms of observables. This, of course, is the type of construct we rejected in Part 2 above. Therefore this approach to validity is also ruled out.

But the method itself is not rejected. The relationship of a construct to others is, of course, what makes up the significance of the construct. It is a vital element in the search for more useful constructs. By realizing that the relationship between constructs is one of significance, we are no longer forced to say that the test and the theory are validated simultaneously. There is no question about whether the observables which define a construct are valid. Any substitute test of this construct is validated against the observables and not against the significant relationships into which the construct enters.

On page 64 we asked why we cannot use the word 'significance' and the word 'construct validity' interchangeably. The answer can now be given:

a) As a practical matter, to speak of construct validity allows the mistake of confusing validity and significance. This in turn is a necessary distinction because:

i) In the case of validity we know what our criterion is and we seek a correlation of unity. In the case of significance we have only hunches as to where the relationships should be and the correlations are expected to be much less than unity.

ii) This leads to a difference in research strategy. If we seek a valid test for an already defined construct, ~~then we seek a valid test for an already defined construct,~~ then we do so for purposes of economy. If we fail to find a valid test this is of little concern scientifically, since we can still use our original definition. But a failure to find significance will lead us to reject our construct and search for new ones.

b) Construct validity condones the failure to define constructs, while an emphasis on significance does not.

i) The use of reduction sentences which do not correlate highly is ruled out by the demand for explicit definitions but not by construct validity.

ii) The failure to specify the nature of the nomological net means that construct validity leaves itself open to much confusion. We have seen that neither a theoretical language, surplus meaning, nor existential reference provide the basis for validation. This is not made clear in the terminology of construct validity. Moreover, an unspecified nomological network may in effect be the implicit use of mental data.

iii) The demand for explicit definition of concepts led to the rejection of intervening variables which were known only through their significance but which were undefined in terms of observables.

iv) The two requirements of explicit definitions and systematic ordering of items lead to the rejection of verbal behavior as a basic definition of a trait. But in the construct validity approach, it is by no means clear that this is the case. Often verbal behavior is used without specifying what it refers to or what it is a sign of.

c) Construct validity failed to differentiate between cases where a test for an intervening variable was to be judged by its significance and cases where it was to be judged by its validity. Behavioral concepts like drive fall in the former catagory while tests of mental and physiological concepts fall in the latter. Thus in no cases are a test and a theory validated simultaneously.

B. Intratest Validity. This is the concern with the relationship of the various elements of a trait to each other. As outlined by Peak, Loevinger, and Jessor and Hammond, this refers to the relationship between items on a verbal test. It will be seen that such a view of intratest validity is too narrow.

Our first concern is whether we are dealing with a definition of a construct or with a test of a construct which is defined in terms of some other set of observables. All three of the above articles are vague on this point. They imply that the tests are measures of some construct, but they also assume that this construct is not explicitly defined. In the case of Loevinger I have the feeling she had some sort of implicit mentalism in mind, while Jessor and Hammond seemed partial to undefined behavioral constructs.

In order to fit intratest validity into our scheme, let us imagine that our test is a definition of the construct. In the first place, this rules out all of the examples used in the three articles since they all make use of verbal items (see Part 3 above). Instead we must list a set of non-verbal behavior patterns which we feel make up a construct. In doing this we will probably not be looking for phenomenal similarities but for some other factor which seems to tie the behavior patterns together. This is what Peak called the search for functional unity.

At the most basic level this means no more than that the behavior patterns occur together. In this case items will be included in the definition to the degree they correlate with each other. But the number of items we include in a definition should be limited. Parsimony requires that we use very few defining traits and add others only if they increase the significance of the construct. All other correlates of the trait should be included as part of this significance.

One of the reasons for including many items on the test was that a person was assumed to manifest more of the trait when he manifested more of the items. But this must be shown to be a significant assumption. This is done by comparing it to other methods of measuring the degree of the trait present such as Guttman's rational ordering of items or Likert's method of noting the degree to which an item is present. That method is adopted which leads to the most significant relationships between our construct and others.

But the notion of functional unity usually extends beyond the mere desire to find behavior patterns which occur together. Usually we assumd that many diverse phenomena cohere because of their causal relationships - in spite of their diversity, they are all the result of one type of cause or they lead to particular types of effects. This means that we must rule out "irrelevant causes" of items (see pp. 34-36). Only in this way can we be sure that the behavior patterns make up the construct we desire rather than being the result of "accidental" conditions.

One way to do this is to use many items selected by item analysis. The assumption is that any one "irrelevant factor" will effect only a few of the items.

But this is not an adequate approach to functional unity. First, the number of items does not guarantee that irrelevant causes are removed (eg, response set can effect an indefinite number of items). Second, this ignores parsimony and systematic import (in the sense of central rather than peripheral items). Third, item analysis by itself does not rule out a high correlation between constructs A and B due to an overlap in the items which make up both traits. For these reasons it would seem best to use other methods than mere item analysis. This accords with the conclusions of Peak and Loevinger.

There is much more to be said on how items should be interrelated. But this is not primarily a paper on methods. Our interest is in showing that the methods of intratest validity are necessary operations for concept formation but that they are not related to the validity of a test. Their purpose is to lead psychologists to form more significant concepts rather than to aid in designing tests of concepts which have already been explicitly defined. It is hoped that the emphasis here on explicit definition, significance, systematic import, causality, and parsimony will make these methods more palitable to Logical Empiricists than appeals to realism, undefined genotypes, or response inferred constructs. It is in this new terminology that such methods as configural scoring and factor analysis must be considered.

It should now be clear that Campbell and Fiske's emphasis on diverse measures of the same trait is one aspect of a revised use of the methods of intratest validity.

C. The above discussion centered on the behavior patterns which define a construct. If we want to build a test for such a construct, then we accept it not on the grounds of significance and systematic ordering of items but simply on the grounds of validity and economy. From the test we should ideally be able to say anything about the construct which we can say from the defining characteristics themselves.

But it is hard to imagine that such tests would be frequently used. The bulk of "mental tests" were proposed as measures of constructs which were not defined in terms of observables. By ruling out implicit mentalism and undefined behavioral constructs we have removed the major need for these tests.

Although the above discussion of methods ruled out much of the construct validity approach, it also places limits on what is acceptable in terms of operational definitions.

As an example of the type of operational definition which has been ruled out, let us look at the much criticized Taylor Manifest Anxiety Scale. In a later article on the A-Scale, Taylor says that "'manifest anxiety' has been defined operationally only in terms of test scores" (1956, p. 304). Such an operational definition is acceptable only if:

a) it has been shown that this new construct is more significant than former definitions of anxiety (or definitions of drive, if this is what Taylor intended);

b) it has been shown that irrelevant causes have been ruled out;

c) the items of this construct have been shown not to overlap with the items of related constructs;

d) the use of verbal behavior has been eliminated;

e) an effort has been made to achieve parsimony;

f) it has been shown that the use of many items is a better way to measure the degree of presence of the trait than Guttman's or Likert's approach.

Of course it is impossible to demonstrate all of this in the initial article. But when we look at the literature on the A-Scale it seems that many of these requirements are still unfulfilled. None of the Logical Empiricists including Bechtoldt and Ebel devote much time to the elimination of such abuses of operationalism. If Taylor fails to meet many of these criteria, she is not alone in her faults.

D. One final point must be made. This concerns Jessor and Hammond's suggestion that the theory of properties of the construct should determine the nature of the test. But what does it mean to derive the items of a test from the theory about the construct? We saw in Section II Part 2 that a theoretical language can do no more than serve as a hunch which guides us. The assigning of correspondence rules to a theoretical language is not an empirical matter.

But Jessor and Hammond do not seem to have hunches in mind. The core of their article deals with the Taylor Manifest Anxiety Scale as a measure of Hull's concept of drive. The implication is that the items of this test should be derived from the nomological net surrounding this concept. But since this is one of the undefined intervening variables, there is no way to derive items from the construct itself. The only other thing to do is to relate the items of the test to other constructs hypothesized to be related to it. But this requirement fits nicely into what we have already said about intratest validity - this is the search for systematic ordering of items and significance rather than the search for validity.

There are two other ways in which elements of a test may be derived from theory. One is when the construct is stated in general terms. In this case we may try to find elements of behavior which are instances of the general term. The only drawback to this in constructing a test is that all instances of a general term need not manifest themselves in the same individual at the same time. The problem of the degree to which the trait is present must still be investigated empirically, as we noted above.

The other type of derivation occurs when we use laws to relate unobservable physiological or mental data to their behavioral manifestations. This is not solely a matter of logical deduction as Jessor and Hammond imply (p. 163). Instead we assume that laws which have been independently validated will hold in cases where we can observe only one of the variables of the relationship. This is the case whether we are inferring stomach contractions from the fluctuations of fluid in a tube or inferring the operation of defense mechanisms from a long pause in the subject's conversation.

Thus the derivation of items from theory may be one of two operations. In the case of mental and physiological data this derivation is done via an empirical law. Whether it leads to a correct inference is a matter of validity. But in other cases the derivation is no more than the search for significance and systematic ordering of items which has already been outlined.

PART 6: CONCLUSIONS

The conclusion of this paper is that we must drop the terminology of construct validity and use instead the terms validity and significance. The reasons for this were listed on pp. 92-94.

The use of the terms validity and significance is based upon the difference which Logical Empiricists make between meaning and significance. In Section II it was shown that this terminology can be used if we are careful to avoid shortcomings which the Liberal Empiricists have pointed out. It was further shown in Section III Part I that the differences between significance and validity are not as great or as clear-cut as Ebel and Bechtoldt imply.

The working solution of Section II leads to other important conclusions. It shows that surplus meaning, existential reference, and a theoretical language provide no basis for the performance of construct validity. The consensus also allowed some conclusions about the type of psychological construct which is acceptable. Undefined intervening variables like drive are ruled out, while indirectly testable concepts in terms of mental or physiological data are accepted.

The adoption of these suggestions led to a revision of most of the methods of construct validity (Part 5). Both intertest and intratest validity were shown to be a demand for significant concepts with a systematic ordering of items. Since the multitrait-multimethod matrix of Campbell and Fiske fails to give an explicit definition of a concept, it is not to be used as a method of intertest validity but as a method for the systematic ordering of the elements of a concept.

Furthermore, the use of verbal behavior as the explicit definition of a psychological construct was ruled out (*eg*, Taylor's operational definition of manifest anxiety in terms of the A scale). Since verbal behavior is to be used in a test only if we have an explicit definition of the construct to be measured, the validity of such tests as the MMPI must be questioned. Finally, Jessor and Hammond's demand for the theoretical derivation of the items of a test was shown to be an empirically meaningful method only where explicit definitions of the construct were available.

Therefore, it can be seen that our "working solution" of certain philosophical problems leads directly to the revision of several well-known techniques of validity. This in turn should lead us to question anew the validity and significance of current psychological tests. Under the guise of construct validity such tests as the MMPI could avoid the question of validity as we have defined it. Conversely, an oversimplified type of operational definition allows such tests as Taylor's A scale to avoid questions of significance and systematic ordering of items.

Being neither a psychologist nor a philosopher, I shudder to think of the holes a sophisticated reader will be able to find in my arguments. If my rather shaky rejection of undefined intervening variables is not accepted, then the whole question of how to measure them arises, with some sort of construct validity as the only solution.

With such a dour thought as this in mind, it seems that this paper still could make one contribution. It has hopefully been shown that any solution to the problem of validity in psychology rests on a solution of the more basic problems of psychology. Most important of these are the problems of concept definition, causality, and the use of mental data. Any solution arises out of a combination of the substantive requirements of psychology and the methodological and logical requirements of philosophy.

BIBLIOGRAPHY

Adorno, T. W. The Authoritarian Personality, New York: Harpers, 1950.

Allport, Gordon W.; Vernon, Philip E.; and Lindzey, Gardner. Study of Values, Revised Edition, Boston: Houghton Mifflin Co.

Anscombe, G.E.M. Intention. Oxford, Blackwell Press, 1957.

Bechtoldt, Harold P. "Construct Validity: A Critique". American Psychologist, 1959, v. 14 pp. 619-629.

Beck, Lewis White. "Constructions and Inferred Entities," 1950. Reprinted in Feigl and Brodbeck (see Feigl) pp. 368-381.

Bergmann, Gustav. "Outline of an Empiricist Philosophy of Physics," 1943. Reprinted in Feigl and Brodbeck (see Feigl) pp. 262-287.

Bergmann, Gustav. "The Logic of Psychological Concepts," Philosophy of Science, 1951, v. 18, pp. 93-110.

Bergmann, Gustav. "Theoretical Psychology," Annual Review of Psychology, 1953, v. 4, pp. 435-458.

Bergmann, Gustav. Philosophy of Science, Univ. of Madison, Wisconsin Press, 1957.

Bergmann, Gustav. "Physics and Ontology," Philosophy of Science, 1961, v. 28, pp. 1-14.

Bergmann, Gustav and Spence, Kenneth W. "Operationism and Theory in Psychology," Psychological Review, 1941, v. 46, pp. 1-14.

Brodbeck, May. "On the Philosophy of the Social Sciences," Philosophy of Science, 1954, v. 21, pp. 140-156.

Brodbeck, May. "Models, Meaning and Theories," in L. Gross (ed.) Symposium on Sociological Theory, Evanston: Row Peterson, 1958.

Brodbeck, May. "Logic and Scientific Method in Research in Teaching" in N. L. Gage Handbook of Research on Teaching, Chicago: Rand McNally, 1963, pp. 44-93.

Campbell, Donald T. "Recommendations of APA Test Standards Regarding Construct, Trait, or Discriminant Validity," American Psychologist, 1960, v. 15, pp. 546-553.

Campbell, Donald T. and Fiske, D. W. "Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix," Psychological Bulletin, 1959, v. 56, pp. 81-105.

Campbell, Norman R. "The Structure of Theories," from Physics: The Elements, 1920 Reprinted in Feigl and Brodbeck (see Feigl) pp. 288-308.

Carnap, Rudolf. "Testability and Meaning," Philosophy of Science, 1936, v. 3, pp. 420-468; 1937 v. 4, pp. 1-40.

Carnap, Rudolf. "The Methodological Character of Theoretical Concepts," In H. Feigl and M. Scriven (Eds.) Minnesota Studies in the Philosophy of Science, v. 1, Minneapolis: Univer. Minnesota Press, 1956, pp. 38-76.

Cronbach, L. J. and Meehl, P. E. "Construct Validity in Psychological Tests," 1955, Reprinted in Feigl and Scriven (see citation directly above) pp. 174-204.

Ebel, Robert L. "Must All Test be Valid?" American Psychologist, 1961, v. 16, pp. 640-647.

Feigl, Herbert. "Existential Hypotheses," Philosophy of Science, 1950, v. 17, pp. 35-62.

Feigl, Herbert, "Functionalism, Psychological Theory, and the Uniting Sciences: Some Discussion Remarks," Psychological Review, 1955, v. 62, pp. 232-235.

Feigl, Herbert. "The Mental and the Physical" in H. Feigl, M. Scriven, and G. Maxwell (eds.) Minnesota Studies in the Philosophy of Science, v. 2, Minneapolis: Univer. Minnesota Press, 1958, pp. 370-497.

Feigl, Herbert and Brodbeck, May. Readings in the Philosophy of Science, New York: Appleton - Century - Crofts, 1953.

Hempel, Carl G. Fundamentals of Concept Formation in Empirical Science, v. 2, Number 7 of The International Encyclopedia of Unified Science, University of Chicago Press, 1952.

Hempel, Carl G. "The Theoretician's Dilemma: A Study in the Logic of Theory Construction" in H. Feigl, M. Scriven, and G. Maxwell (eds.) Minnesota Studies in the Philosophy Science, v. 2, Minneapolis Univer. Minnesota Press, 1958, pp. 37-98.

Russell, Bertrand. The Problems of Philosophy, 1912, Reprinted 1959: Galaxy paperback; New York: Oxford Univer. Press.

Sellars, Wilfrid. "Realism and the New Way of Words," Philosophical and Phenomenological Research, 1948, v. 8, pp. 601-634.

Spence, Kenneth W. "The Postulates and Methods of 'Behaviorism'," 1948, Reprinted in Feigl and Brodbeck (see Feigl) pp. 571-585.

Spence, Kenneth W. "The Empirical Basis and Theoretical Structure of Psychology," Philosophy of Science, 1957, v. 24, pp. 97-108.

Spence, Kenneth W. "A Theory of Emotionally Based Drive," American Psychologist, 1958, v. 13, pp. 131-141.

Taylor, Janet A. "The Relationship of Anxiety to the Conditioned Eyelid Response," Journal of Experimental Psychology, 1951, v. 41, pp. 81-92.

Taylor, Janet A. "A Personality Scale of Manifest Anxiety," Journal of Abnormal and Social Psychology, 1953, v. 48, pp. 285-290.

Taylor, Janet A. "Drive Theory and Manifest Anxiety," Psychological Bulletin, 1956, v. 53, pp. 303-320.

"Technical Recommendations for Psychological Tests and Diagnostic Techniques" Psychological Bulletin (Supplement) 1954 v. 51 pp. 1-38.

Travers, Robert M. W. "Rational Hypotheses in the Construction of Tests," Educational and Psychological Measurement, 1951, v. 11, pp. 128-137.

Hull, Clark L. et al. Mathematico-Deductive Theory of Rote Learning, New Haven: Yale Univer. Press, 1940.

Hull, Clark L. Principles of Behavior, New York: Appleton-Century, 1943.

Hull, Clark L. Essentials of Behavior, New Haven: Yale University Press, 1951.

Jessor, Richard and Hammond, Kenneth R. "Construct Validity and the Taylor Anxiety Scale," Psychological Bulletin, 1957, v. 54, pp. 161-170.

Loevinger, Jane. "Objective Tests as Instruments of Psychological Theory," Psychological Reports, 1957, v. 3, pp. 635-694.

MacCorquodale, Kenndth and Meehl, Paul E. "Hypothetical Constructs and Intervening Variables," Psychological Review, 1948, v. 55, pp. 95-107.

Madden, Edward H. "Definition and Reduction," Philosophy of Science, 1961, v. 28, pp. 390-405.

Malmo, Robert B. "Measurement of Drive: An Unsolved Problem in Psychology," in M. R. Jones (ed.) Nebraska Symposium on Motivation, Lincoln: Univer. Nebraska Press, v. 6, 1958, pp. 229-264.

Meissner, W. W. "Intervening Constructs-Dimensions of Controversy," Psychological Review, 1960, v. 67, pp. 51-72.

Mellenbruch, P. L. "The Validity of a Personality Inventory Tested by Hypnosis," American Journal of Clinical Hypnosis, 1962, v. 5 (2) pp. 111-114.

Pap, Arthur. "Other Minds and the Principle of Verifiability," Revue Internationale de Philosophie, 1951, v. 5, pp. 280-303.

Pap, Arthur. "Reduction-Sentences and Open Concepts" Methodos, 1953, v. 5, pp. 3-30.

Peak, Helen. "Problems of Objective Observation" in L. Festinger and D. Katz Research Methods in the Behavioral Sciences, New York: Dryden Press, 1953, pp. 243-299.

Peters, R. S. The Concept of Motivation, London: Routledge and Kegan Paul, 1958.

Rommetweit, Ragnar. "Model Construction in Psychology: A Defense of 'Surplus Meaning' of Psychological Concepts," Acta Psychologica, 1955, v. 11, pp. 335-345.