Cronbach's (1971, p. 447) statement that, "Because every interpretation has its own degree of validity, one can never reach the simple conclusion that a particular test 'is valid'." should be qualified in that when interpretations are made within scientific inquiry they are of one basic type. Within scientific work there is only one type of construct validity--the nature of which stems from the purpose of science.

This paper has attempted to reintroduce the problem of construct validity by showing that operational methodology does not solve the problem but actually misinterprets it. When a distinction is made between definition and indication, the problem of construct validity can be better understood and hence better dealt with. This paper attempted to take the investigation thereof more consciously in that direction.

References

Alston, W. P. Philosophy of Language. Englewood Cliffs, N. J.: Prentice-Hall, 1964.

Bechtoldt, H. P. "Construct Valdity: A Critique." American Psychologist, 14, 1959.

Cronbach, L. J. "Test Validation." In R. L. Thorndike, ed., Educational Measurement, 1971, pp. 443-507.

Cronbach, L. J. & Meehl, P. E. "Construct Validity in Psychological Tests." Psychological Bulletin, 52, 1955.

Edwards, W. H., et al. "Bayesian Statistical Inference for Psychological Research." Psychological Review, 70, 1963.

Giere, R. N. "Objective Single Case Probabilities and the Foundations of Statistics." In Proceedings of the Fourth International Congress on Logic, Methodology and Philosophy of Science, 1971.

WHAT IS THE PROBLEM OF CONSTRUCT VALIDITY?

Jerome A. Popp

Associate Professor of Education
Southern Illinois University
at Edwardsville

Presented to:

Special Interest Group--
Philosophical Studies of Education
American Educational Research
Association
Washington, D. C. April, 1975

WHAT IS THE PROBLEM OF CONSTRUCT VALIDITY?

While philosophic inquiry and scientific inquiry are distinct cognitive activities, the products of one often contribute to the products of the other. Philosophers of science and methodologically conscious scientists constitute two populations which should be able to communicate readily. Yet in educational research the lack of communication between philosophers and scientists seems marked. The gap between educational research and educational practice is often emphasized, but there is also often a gap between the methodology of educational research and the practice of educational research. The present paper is an attempt to bridge part of, or one aspect of, the gap between the two.

In the following discussion the background is developed which is requisite for the adequate formation of the problem of construct validity. Once formulated, several aspects of an adequate solution to the problem will be explored.

1. What Is a Construct?

Scientific inquiry proceeds, to put it briefly, by raising questions about the properties and their relationships of specific populations. The properties which describe the population under study are the constructs of the inquiry--the postulated attributes of a particular domain of study. One should not be distracted by the use of other names such as 'concept', 'variable', 'attribute'. Different labels do not entail different meanings. When one wishes to emphasize that properties are: (1) products of the researcher's creative efforts he calls them "constructs"; (2) postulated of

the members of a population he calls them "attributes"; (3) ontological entities which can be conceived of or entertained mentally he calls them "concepts"; (4) possessed in different quantities or qualities by different individuals he calls them "variables."

Some of these properties may be directly observable, others indirectly observable, i.e., an instrument of some type is required, and finally, some may not be observable at all. In this case the possession of the attributes by the members of the domain of study is inferred, not observed. This way of approaching the meaning of 'construct' deviates from Cronbach and Meehl (1955) who held that the attribute must be thought to be reflected in test performance. The present account does not assume in advance that given any meaningful construct it is always possible to build a test to measure it.

At this point it is important to note that a construct is not a word but is named by a word. 'Intelligence' or 'scholastic aptitude' name constructs. The construct is the meaning of the name. The question "What does 'intelligence' mean?" asks for an account of what is named by 'intelligence'; it asks for the construct involved. The constructs of science are of two basic types: continuous and categorical. The former are attributes or properties possessed to some degree by every member of the domain while the latter are possessed by only some members in dichotomous (or trichotomous, etc.) fashion, i.e., it is either present or absent in each case, or if the construct is possessed by all members, this possession is of a categorical nature. Intelligence and sex are examples of each type, respectively.

The meaning of 'intelligence' is not to be equated with the actual individuals who possess this property. This is to reject the referential theory of meaning (see Alston, 1964). Stated differently, in the language of social scientific methodology, the target population of the particular investigation is always greater than the population which is accessible (in principle, technically, or practically) to the researcher. For example, the individuals who possess the property, intelligence, make up the extension of the term 'intelligence'; and as is being claimed, the meaning of a term is not the same as its extension.

2. Definition versus Indication

For any non-logical scientific term two distinct questions may be asked: (1) What is the meaning of the term? (2) Which of the individuals of the domain possess this construct, and if all individuals of the domain posses it then to what degree does each? In other words, what is the extension of this term (given either categorical or continuous variation)?

These two questions are answered in distinct ways. The first asks for the relation of the term being investigated with other terms, while the second asks for the relation of the term with the objects of study. To provide answers to these two questions the researcher is required to engage in two distinct types of activities, the processes of definition and indication, respectively. These are schematized in Figure 1.

Since words are objects themselves, a further specification of these relationships is required. Given the vocabulary of a science, definition relates only the members of that vocabulary. All non-logical words in a given definition from a science must come from the vocabulary of that

science. Indication is the process of relating a word or term from a given scientific vocabulary to some object which is not a member of that vocabulary. This allows for scientific languages which refer to languages, e.g., empirical linguistics.

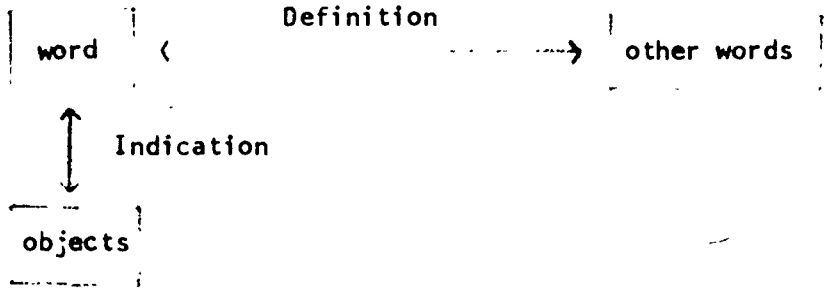

Figure 1.

Definitional inquiry may proceed in two directions (Popp, 1973). Given the usually adhered to convention of placing the definiendum (or term defined) on the left, the activity of relating a word to a previously established definiens can be called 'left-handed defining'. For example, let '(a,b)' stand for 'the greatest common factor of a and b'. The construct or notion of a greatest common factor is not at issue. We are simply looking for a symbol. Left-handed defining is a question of <u>notation, not notion</u>. On the other hand, when one seeks the definiens of a previously known or accepted term (like 'intelligence') he could be said to be engaged in right-handed defining. In this case, the description of the construct is at issue, i.e., it is a question of <u>notion, not notation.</u>

Indication can also go on in two directions: <u>denotation</u> and <u>satisfaction</u>. If a given term is correctly said of or applied to a given object the

term denotes the object, and the object satisfies the term. The relating of a word to an object often requires or involves an instrument of some sort.[1] For example, if one wanted to know whether a given rock specimen was radioactive he might make use of a Geiger counter. This instrument produces an indicator reading which allows the user to decide the amount of the property, radioactivity, present. Or, to take a more familiar example, the Scholastic Aptitude Test is an instrument which if correctly used (and the conditions for correct use are formally stated) produces an indicator reading (the SAT score) for each individual examined. Given any individual or member of the population, it is possible to produce an indicator of his scholastic aptitude. Some instruments will produce better indicators for this construct than others.

A good example of the separation of the two distinct relationships of definition and indication can be found in the chapter of P. S. Wilson's (1971) on children's interests. He discusses the problems of, one, "to know what sort of thing we are looking for" or what "sort of notion which 'interest' itself is, and, second, with questions about any special problems which may arise in connection with identifying and recognizing such feelings as these in ourselves and in others." In the language of this paper, Wilson's chapter is concerned with finding (1) the definition of 'interest' or a description of the construct, interest, and (2) indicators for interests--the symptoms one exhibits when he has an interest in something.

---

[1]What is often called 'ostensive definition' is better called 'ostensive indication'--pointing to an object while uttering the word. This activity has more value in teaching science than in the conduct of science.

Any particular indicator of a construct is in principle expendable; that is, the research can proceed without it. This is not the case for the constructs of science for these are in a sense the primary variables or essential properties of science, whereas indicators are not. Of course, indicators are variables. The number of red spots on a patient's face or the rate of a Geiger counter's clicking are variable properties; but they are secondary to the point of the research effort. Spots or rates of clicking are observed not for their own sake but because of what they indicate. They constitute evidence or data for the ascription of a theoretical construct to a particular individual within a domain of study.

The problem of construct validity can now be stated. An instrument (assuming it is correctly used) will produce an indicator for each of the individuals examined. If the property is categorical or qualitative, then the problem of construct validity is the question of whether the indicators produced accurately classify the individuals being measured. If, on the other hand, the variance is continuous or quantitative, then the question becomes: do the individuals being measured possess the property to the degree indicated? In each case, is the instrument measuring what it is supposed to measure, i.e., is the instrument producing valid indicators for the construct? Any instrument which, when used as specified in the situations specified, measures a property accurately is said to possess construct validity with respect to a given construct. Thus, construct validity is a property of instruments. It is the ability of the instrument, to put it differently, to produce true indicator laws, e.g., the barometer reading and the storm. Or, breeding confusion, construct validity is the strength of

the relation between secondary and primary properties.

3. Operational Definitions

Some, such as Bechtoldt (1959), have concluded that the notion of construct validity creates unnecessary confusion within the social sciences. It is claimed that "operational methodology" or more particularly operational definitions are sufficient for the establishment of scientific constructs. In the following discussion, it will be shown that Bechtoldt's conclusion is logically and methodologically premature, and that it is the operational definition which is at the base of much unnecessary confusion within social science.

In an earlier paper (Popp, 1969), it was argued that the reliance of educational research upon operational definition is a significant block to the development of educational theory. To establish this methodological conclusion it is necessary to analyze the function of such moves in research. Giving an operational definition in a research report consists of defining a construct (like intelligence) in terms of a test score (e.g., a CTMM score). Such definitions are operational in that the construct under study is "defined" via certain operations required to measure the construct. In this way the research and report are thought to be rendered more objective. But why is it that some operations are acceptable and some are not? What if a researcher defined intelligence as the score of a test for manual dexterity? Why would this operational definition be rejected and others

accepted? Simply because some operational definitions are more justified than others. One can distinguish good operational definitions from those which are not so good. But again, it is asked, what makes one operational definition better than others? Some operational definitions, of say 'intelligence', better correspond to the underlying understandings or background knowledge of the construct being measured. Operational definitions and operational defining are theory-laden. One's antecedent knowledge reveals to him what operations to perform in order to measure a certain property.

The problem with operational methodology is that it cannot set out clearly the constructs of the science, for as we have just seen, it presupposes some notion of the constructs measured but never makes these presuppositions explicit. Stated differently, if the point of science is to set out the properties and relationships of a domain of study, then more is required than operational definition and hypothesis construction, for operational definitions presuppose an understanding of the nature of the constructs measured in the experiment, and related by the hypothesis. Where do these understandings come from? How do they emerge within scientific inquiry?

Operational methodology muddles the notions of definition and indication. Definition is the process by which one makes explicit the constructs of the inquiry--developing constructs is a part of theorizing and done via definition. Indication does not set out constructs but selects instances of them. It is now obvious that if one accepts the definition-indication distinction, operational definitions are not definitions but indications. More precisely, what is usually called 'operational definition' is better called 'indicator specification'. That is, any published report should specify the

instruments used to collect the data. This is the function of operational definitions. Operational methodology is a part of verificational methodology, whereas definition is a part of methodology for theorizing--theory construction methods. Thus, operational definitions presuppose theory and do not produce it.

The importance of this discussion is that operational definitions depend upon valid instruments, when instruments are mentioned in the operational definition. The reason why we would admit the CTMM as a measure of intelligence and reject the test of manual dexterity is that the CTMM has greater construct validity. The operational definition which refers to the use of an instrument is only as justified as the validity of that instrument. Construct validity, thus, occupies a central position within verificational methodology.

4. Achieving Construct Validity

Constructing valid measures consists of bringing an instrument into a certain relationship with a construct. This process may go on in two directions: (1) assuming the construct is defined properly one may create or modify instruments, or (2) one may retain the present structure of the instrument and modify the construct so that the indications of the instrument better agrees with the extension of the construct's name. The former process may be called instrument adjustment, and the latter construct redefinition.

These two avenues to the achievement of construct validity require further consideration. How does one build an instrument to measure a given construct which has been previously defined? If valid measures of this

construct exist, one could modify the items of the instrument in question in such a way as to result in a test whose scores have negligible reliable variance with the scores of the other valid measures. But what if no valid measures have been developed? How does one build a test to measure a previously defined construct within a given domain of inquiry?

The definition of the construct should contribute some understanding of what is to be measured. Even the researcher mentioned above who attempts to modify items to produce a measure which correlates with existing valid measures will, unless he is a hack, utilize his understandings of the construct in the modification of his test items. Knowledge of the phenomena being studied can be used in both the construction of instruments, and the validation thereof. Once a test has been written which is thought to be valid, and this process often requires great creativity even though the construct may be clearly embedded in a theory, the test is subject to evaluation--its validity is verified. This testing procedure also requires creative efforts.

If a given population is known to possess a given property, then this population can be used in the validation process. The test is administered to this known population. If the actual results of this administration correspond to the expected results, then one can conclude that the measure is valid for this population. The question of to how large a population one can infer which will allow this instrument to produce valid indicators is the question of inductive inference and will not be considered directly in this paper. If the results of the test administration are negative, then the instrument must be reconstructed.

Before considering other methods, one may ask how it is that the researcher knows that the given population possesses the construct in a given degree? Would this not presuppose some other measurement device? A negative answer is possible if the members of the population sampled can, by direct observation, be examined for certain propertie[illegible] the theory relates to the construct being studied. In this way [illegible] cases can be set up, and null hypothesis procedures invoked. (This point will be returned to below.) It should be noted that if one builds a test to produce an indicator which correlates well with a given cluster of observable properties, he does not encounter directly the problem of construct validity, but rather the problem of predictive validity, as some would call it. Construct validity studies are concerned with an explicitly definable construct.

A third possibility for establishing construct validity is via experiment. How the members of a given population respond to a given treatment subsequent to exposure to the measurement device (whose construct validity is being questioned) can be taken as evidence for inferring construct validity, if there are established hypotheses as to how the population which possesses the construct will respond to the treatment.

Fourthly, it seems that the possibility must be left open that construct validity can be established by theory independently of separate empirical evidence. If there were known relationships between the construct in question and several other constructs, then it may be possible to infer the characteristics of the measurement device. This, of course, depends upon having an explicit definition of the construct to be measured.

It is possible that researchers may be reluctant to modify further a given instrument even though the instrument does not produce valid indicators. .terberg (1965) has noted, one need not continue to reconstruct the instrument, for the desired validity can be achieved by redefining the construct. Suppose that an instrument is valid for only some regions of the domain under study. By narrowing the meaning of the variable, an adequate measure thereof is produced.

When it is said that "Intelligence is what an intelligence test measures." an absurdity results if this expression is interpreted as asserting a definition. If, on the other hand, what is meant is that the researcher has greater confidence in his ability to measure the construct than his ability to define it, then his expression is not functioning as a circular definition. It is possible for researchers to hold a greater confidence in the indication relation than in the definition relation, for a particular construct. However, it should be remembered that measurement is a tool of science, not its goal--such expressions as that mentioned above do nothing for explicating the definition relationship.

The modification of a scientific concept should take into account more than construct validity; that is, the achievement of construct validity is only one value served in the process of redefinition. Constructs and the hypotheses which relate them constitute explanatory systems; and any change in one construct has consequences throughout the complement of the system. Modification of a construct should be done in light of the results of hypothesis testing and construct validity studies. A given scientific

definition is not testable via observation but this does not mean that such constructions cannot be evaluated in terms of the experience of those engaged in the inquiry. A fuller development of this methodology takes the discussion into the problems of theory construction in science--an area of inquiry beyond present purposes.

5. Degree of Construct Validity

Construct validity is usually thought to be possessed by all instruments but in different degrees. Some are quite valid and some are quite invalid. Moreover, it seems reasonable to view the degree of construct validity as a probability; that is, the construct validity of a test or instrument is the probability that the indicators it produces are valid indicators (for a specific population). This interpretation raises the question of which view of probability should be utilized; that is, is the usual frequency view better in this regard than the propensity of subjectivist views? (Discussion of logical probabilities is omitted for it is being assumed that construct validity is an empirical property of instruments.) Classical statistics is based upon the frequency interpretation of probability, and it seems that researchers who utilize these procedures would view construct validity as a frequency--the limiting relative frequency of adequate indications in an infinite sequence of indications. The frequency view encounters various problems, one of which is important in this context: it does not allow for the determination of the probabilities in single cases. As Giere (1971) put it, "What good is it to know that a certain method will give a correct estimate 95% of the time on the average? The problem is to know what to think about the particular case. . .". For

example, if an individual gets a certain score on some test and it is known that the instrument has a construct validity rating of p, all that should be said is that in cases like this one, the long run frequency of successful measurement will be p. Single cases are thus indeterminant.

Some practitioners may object to this view on the grounds that many decisions made in, say, the schools are in terms of single cases. For example, a student's advisor may, on the basis of several instrument readings, advise one thing over another. Advisors may find it too weak and uninforming to say that in cases like this, p percent of the population in the long run will actually have the property indicated. Practitioners will desire a solution to the single case problem. The propensity interpretation of probability offers such a solution. In the context of instrumentation, this view would amount to the claim that the construct validity of a given instrument is the propensity (disposition or tendency) of that instrument, under certain standard conditions of usage and with a specific population of individuals, to produce indications of a property when that property is actually present (or present in the degree indicated) in the individual being measured. Thus, the degree of construct validity of a given instrument is seen as a property of a chance set-up. It is the propensity of the measurement set-up or system to produce valid indicators. More precisely, the degree of construct validity is the propensity of a chance set-up, including an instrument, standard conditions for its use, and a specific population, to produce valid indicators for a given construct in that population.

The propensity interpretation of probability converts probability into a theoretical property (Giere, 1971). In the present context, this interpretation renders construct validity a theoretical property of measurement set-ups. The modification of some elements of the chance set-up will produce different probabilities. Thus, construct validity as a property of a measurement system can be seen as related to other properties of the system via theoretical principles. Moreover, while construct validity can be improved by simple trial and error, it is more adequately seen as the goal of a special type of theorizing. More specifically, the instrument builder can be seen as one who theorizes about construct validity. Such theorizing is then subjected to empirical verification.

This consequence of the propensity view of construct validity seems to me to be especially important for it clearly brings test construction under the influence of theory. It allows for the theoretical explanation of why a particular device possesses a high degree of construct validity and why another may not. This is precisely the point that the operationists never seemed to have fully grasped. Measurement procedures do not define scientific constructs, but rather, are formulated in light of already defined constructs.

Before considering a more personalistic view of construct validity, one final point should be clarified. One may ask, what really is the difference between propensities and frequencies? Is it not the case that a propensity is determined by conducting a series of random experiments upon the chance set-up and then examining the relative frequency of success? In a sense this is true; as Giere (1971) claimed, "relative frequencies may provide

evidence for propensity hypotheses." But this should be sharply distinguished from the notion that definitionally there is no difference in these two views of probability.

Subjective or personalistic Bayesians view probabilities as degrees of belief. A random experiment produces data which modifies these beliefs, for rational believers. One may be asked what he thinks is the ratio of red balls to blue balls in a urn of 100 balls. If he believes that the probability of drawing a red ball is p, then after he has drawn n balls (with or without replacement) his belief can be updated as it were in light of these results. The personalistic Bayesians have a precise description of these procedures. (See Edwards, et al, 1963.)

One problem with this view is the establishment of the prior or beginning probabilities. Mathematically, the prior beliefs dominate the posterior or resultant beliefs in that any given posterior probability on a given set of data can be obtained if the appropriate prior probabilities is held. This is not just a matter of changing your prior views after the evidence is obtained. The prior probability, which is highly arbitrary, seems to unduly control our conclusions. This is not intended to be a convincing or adequate critique of the subjectivist view. But it does have consequences for the notion of construct validity. When an individual obtains a certain test score our belief that he actually possesses the property being measured will of course depend upon our belief in the validity of the score. Construct validity will be a posterior probability for the subjective Bayesian. This probability is arrived at by establishing one's prior belief about the construct validity of the set-up, observing the test results of

various individuals, and then processing these via Bayes' theorem in order to obtain a posterior belief about the validity of the measurement set-up. Thus, a student and his advisor may rationally disagree over the meaning of his test scores. The objectivity of the meaning of such scores vanishes. In scientific contexts, different scientists may rationally disagree on the warrant of the conclusions of empirical research studies on the basis of their beliefs about the instruments used therein.

6. Construct Valdity and Usefulness

Many of the instruments developed in education are used in both social scientific research and practical decision-making. Construct vaiidity questions can thus be a source of problems for scientific conclusions and conclusions about proper educational practice, which of course has great moral significance. Adequate measurement capabilities will have utility or usefulness for both theory and practice. We can distinguish theoretical usefulness and practical usefulness. In other words, a given instrument with construct validity p for a specific population may produce indicators which may be judged by the educational researcher and the educational practitioner to have different significance; for example, the construct validity of a test may be seen by teachers as high enough to base practical decisions upon its indicators, whereas the researcher might judge it as much too low for a given experiment.

The upshot of this is that judgments of the adequacy of the degree of construct validity are not of one type, and should be made in light of the purposes of the inquiry. This repeats the point in "Significance and Utility" (Popp, 1971) about levels of statistical significance. Levels

adequate for an educational practice situation may be higher or lower than adequate levels in research. If construct validity studies utilize null-hypothesis procedures, then the level of rejection becomes focal. A conclusion that "There is no significant difference between those individuals who have the property in question (or have it in the specified degree) and those individuals who were selected by the instrument as having the property in question (or have it in the specified degree)." may allow for different judgments of the usefulness of the instruments by educational practitioners and researchers.

Thus, there seems to be no basic disagreement with Cronbach (1971, p. 447) when he writes, "one validates, not a test, but an interpretation of data arising from a specified procedure." Different purposes of interpretation may produce different conclusions about validity. Put differently, the purposes of the inquiry will determine, on the evidence available, whether any given so-called indicator law is acceptable. We can, therefore, set up a pair of questions to parallel those of Maxwell (1972, p. 138).

(i) What criteria ought to govern our choice of a theory [or instrument] from two or more theories [or instruments] if our concern is with scientific growth?

(ii) What criteria ought to govern our choice of a theory [or instrument] from two or more rival theories [or instruments], if our concern is with the trustworthiness of the theory [or instrument], for purposes of technological application?

Theoretical and practical inquiries will require different criteria deriving from different values. These considerations led to the conclusion that

Cronbach's (1971, p. 447) statement that, "Because every interpretation has its own degree of validity, one can never reach the simple conclusion that a particular test 'is valid'." should be qualified in that when interpretations are made within scientific inquiry they are of one basic type. Within scientific work there is only one type of construct validity--the nature of which stems from the purpose of science.

This paper has attempted to reintroduce the problem of construct validity by showing that operational methodology does not solve the problem but actually misinterprets it. When a distinction is made between definition and indication, the problem of construct validity can be better understood and hence better dealt with. This paper attempted to take the investigation thereof more consciously in that direction.

References

Alston, W. P. Philosophy of Language. Englewood Cliffs, N. J.: Prentice-Hall, 1964.

Bechtoldt, H. P. "Construct Valdity: A Critique." American Psychologist, 14, 1959.

Cronbach, L. J. "Test Validation." In R. L. Thorndike, ed., Educational Measurement, 1971, pp. 443-507.

Cronbach, L. J. & Meehl, P. E. "Construct Validity in Psychological Tests." Psychological Bulletin, 52, 1955.

Edwards, W. H., et al. "Bayesian Statistical Inference for Psychological Research." Psychological Review, 70, 1963.

Giere, R. N. "Objective Single Case Probabilities and the Foundations of Statistics." In Proceedings of the Fourth International Congress on Logic, Methodology and Philosophy of Science, 1971.

Maxwell, N. "A Critique of Popper's Views on Scientific Method." Philosophy of Science, 39, 1972.

Popp, J. A. "Operational Definitions in Educational Inquiries." In D. Arnstine, ed., Proceedings of the Philosophy of Education Society, 1969.

Popp, J. A. "Significance and Utility: A Study in the Inductive Problems of Educational Research." In R. Heslep, ed., Proceedings of the Philosophy of Education Society, 1971.

Popp, J. A. "Definition, Logical Analysis, and Educational Theory." In B. Crittenden, ed., Proceedings of the Philosophy of Education Society, 1973.

Wilson, P. S. Interest and Discipline in Education. London: Routledge & Kegan Paul, 1971.

Zetterberg, H. On Theory and Verification in Sociology. Totowa, N. J.: Bedminster, 1965.