DOCUMENT RESUME

ED 050 019 SP 004 802

AUTHOR Throne, John M.
TITLE Inappropriateness of Inferential and Insufficiency of Descriptive Statistics in Educational Evaluation: The Problem and a Solution.
PUB DATE Feb 71
NOTE 17p.; Paper presented at annual meeting, AERA, New York, 1971

EDRS PRICE EDRS Price MF-$0.65 HC-$3.29
DESCRIPTORS Behavioral Objectives, Behavior Change, Evaluation, Evaluation Criteria, *Evaluation Methods, *Operant Conditioning, Predictive Validity, Research Methodology, *Statistical Analysis, Statistical Data, *Teacher Evaluation, Teaching, *Validity

ABSTRACT

This paper argues that to be appropriate the evaluation of teaching must occur under circumstances entirely free of the limitations which inferential statistics necessarily impose on teaching. Regardless of whether the statistical, design, and treatment assumptions required for the valid use of inferential statistics in education are met, inferential statistical analysis is still functionally inappropriate. Descriptive statistical analysis, often recommended as an alternative, is also insufficient for evaluating teaching effectiveness. Interpretations or predictions based on descriptive or inferential statistical findings are based on presumed relationships between phenomenal variables which the statistical findings apparently--but only apparently--reflect. It is suggested that a viable replacement consists of functional analysis of behavior strategy based on operant conditioning. This method is precisely tailored for the moment-to-moment manipulation required by educational practice. (RT)

INAPPROPRIATENESS OF INFERENTIAL AND INSUFFICIENCY OF DESCRIPTIVE STATISTICS IN EDUCATIONAL EVALUATION: THE PROBLEM AND A SOLUTION[1]

John M. Throne, Ph.D.

University of Kansas

[1] To be presented to the meeting of the American Educational Research Association, New York City, February, 1971.

Insufficiency of Criticisms of Inferential Statistics

In reviewing studies which fail to yield positive findings of scholastic functionality in achievement (e.g., Coleman, et al., 1966), it is common for critics (e.g., Guba and Clark, 1967, pp. 104-107) to focus on the unfeasibility or impracticality of meeting the statistical, design, and treatment assumptions required for the valid use of inferential statistics in education. Guba and Clark identify, under requisite statistical assumptions: normality of distribution, randomness of sampling, and effective equality of treatments (additivity). They identify, as an essential design assumption for insuring internal validity: comparability of experimental and control groups; and as essential assumptions for insuring external validity: random selection of subjects from the population in question and their random assignment to experimental and control groups, and insulation of the subjects so selected and assigned from reactive or interactive effects extraneous to the independent variables under study. Essential treatment assumptions identified include: a priori treatment explication, "non-contamination" (non-confounding) of the treatment by extraneous independent variables, treatment invariance throughout the experiment, identicality of treatment application by all experimenters, and elimination of competing treatments.

Because these assumptions, at least in the aggregate, may seldom if ever be met in education, Guba and Clark regard the use of inferential statistics in evaluating teaching as inappropriate. But Guba and Clark do not go far enough. Inferential statistics are inappropriate in evaluating teaching under any circumstances. Education demands a system of evaluation indistinguishable from teaching; an epistemological framework within whose bounds may be found a single set of methods for determining the effectiveness of educational practices in both of the senses, teaching and evaluation, that the term, determination, implies (Throne, 1970a). Conceptual distinctions between teaching and evaluation have tended to obscure their operational commonalities. Through operations called teaching (or training, etc.) an educator may determine, in the sense of produce, the level of achievement a child attains. The educator may also determine, in the sense of measure, that selfsame level of achievement. However, if measurement is undertaken through an active process of response manipulation to criterion, rather than (as in the case of inferential statistics) a passive one of comparison of obtained results against a theoretical expectancy (i.e., the null hypothesis), all operational distinctions between teaching and evaluation may be dissolved.

Teaching versus Evaluation

Traditionally, research ground rules have demanded that the manipulative steps implied by teaching be rigorously distinguished from those of evaluation; that the differences between teaching and evaluation be scrupulously maintained. Thus, meeting the assumptions of inferential statistics is intended to make more probable the "uncontaminated" determination of the effectiveness of the independent variable or variables operative. Once an independent variable is "played," no extraneous "tampering" with the dependent-variable parameters (in terms of which the effectiveness of the independent variable or variables is to be evaluated) may be permitted. But "tampering" and "contamination" are indispensable to teaching; moment-to-moment alteration of subject responses is synonymous with teaching. It is therefore not only appropriate but necessary to evaluate the effectiveness of teaching as determined by its functional effects upon dependent-variable subject-responses; that is, as the latter are deliberately and systematically manipulated to criterion by the former. In short, the evaluation of teaching requires an epistemological framework equal to the task "of supplying evaluation data yielded through the very processes of teaching which produce responses which the evaluation data represent" (Throne, 1970d).

Insufficiency of the Inductive-Deductive Method

Research ground rules have undergone little or no change since the 17th century. They are essentially Baconian. Data (expressed numerically or not: it does not matter, simple occurrence constituting the most primitive form of quantitative event) are first induced and second deduced, or vice versa: the inductive-deductive method. In practice if not in theory, induced quantitative data are frequently treated as though they are equal to, not surrogates for, phenomena extraneous to themselves. In fact, of course, quantitative data are semantic signs representing phenomenal significates; a sign merely signifies, it does not equal, a significate. For example, a number yielded through statistical analysis is a semantic sign, the performance which the number supposedly reflects, a phenomenal significate.

Through the medium of inferential statistical analysis, it is semantic signs, not phenomenal significates, which educational evaluation traditionally induces. Of course, a phenomenal significate may also be induced (or even produced: i.e., manipulated to criterion). However, the independent variables which constitute the operations of phenomenal induction (or production) are only indirectly reflected in the semantic signs which constitute inferential-statistical outcome data. Instead of being a function of actual manipulation of subject responses by teachers, these signs are a function of theoretical (e.g., mathematical) manipulation of symbols (e.g., numbers) by evaluators (who may happen to be teachers also, but who are acting in a strictly evaluative capacity; that is, functioning within a set of operational parameters distinct from teaching). That in the teaching phases prior to evaluation actual phenomenal significates provide the factual flesh and bones over which the theoretical garment of inferential statistics is eventually draped, is not the issue. By the time the garment is completed (theoretically-treated teaching-outcome data), the actual shape of the body underneath is indiscernible; for all intents and purposes, it is irrelevant to inferential statistics' tailoring aims. It is this emphasis on semantic rather than phenomenal manipulation -- requiring an unnatural and unrealistic freeze upon teaching -- that renders evaluation of teaching based on inferential statistical procedures, inappropriate. However, it is recent methodological and epistemological developments, to be described below, which make the utilization of these procedures in evaluating teaching, unnecessary.

How May Independent-Variable Contributions be Confirmed?

In the deductive stage following either theoretical or actual induction (or production), interpretations are superimposed upon the semantic sign or phenomenal significate, as the case may be. These interpretations, while defensible in principle, are entirely gratuitous from the standpoint of the variables demonstrably responsible for the sign or significate in fact. At best, interpretations provide hypotheses about contributive factors to induced (or produced) data, hypotheses whose confirmation or rejection depends on operations additional to those which yield the data from which the interpretations rise. At worst, interpretations are invoked to account for data on the basis of the data itself (tautological reasoning), or on the basis of an unvalidable relationship between the data and a hypothetical construct lacking operational definition (ascientific reasoning). Dependent-variable phenomenal significates are a function of the independent-variable manipulations of phenomena which induce (or produce) them; they are not a function of those other variables, real or imaginary, whose contributions to the phenomenal-significate responses are known or unknown. Moreover, whatever else may be true of the contributions of these other variables (the real ones only, the contributions of imaginary variables being, of course, illusory), their nature cannot be characterized as causal.

Causal contributions to dependent-variable phenomenal responses can neither be induced through theoretical manipulation of semantic, including mathematical, signs; nor deduced through ratiocinated processes which utilize semantic signs. They must be produced through the introduction or withdrawal (or withholding) of independent-variable phenomenal stimuli. It is only through the production of dependent-variable responses by independent-variable stimuli that the causal contributions of these stimuli may be confirmed. Moreover, hypothesizing functional relationships between independent and dependent variables is obviously needless if such relationships have been empirically produced; conversely, if functional relationships have yet to be produced, mere hypotheses that they exist, even if theoretically confirmed through inductive and/or deductive mediation, cannot be substituted for their actual production. In a word, prediction of produced responses is unnecessary, while prediction without production is insufficient. It is not being argued that independent-variable stimuli do not determine dependent-variable responses (which would be patently self-contradictory), but that confirmation of the functionality of hypothesized independent-variable-dependent-variable relationships must be demonstrated empirically through the deliberate and systematic manipulation of phenomenal significates to criterion; and that empirical demonstration makes theoretical demonstration through semantic induction and/or ratiocinative deduction, superfluous.

Production versus Prediction

It should be perfectly obvious, therefore, that if the deliberate and systematic manipulation of responses to criterion is the quintessential modus operandum of education, the predictive purposes of inferential statistics are inappropriate precisely to the extent that their assumptions have been met. If teaching implies the production rather than prediction of behavior, then, given the moment-to-moment manipulative processes intrinsic to teaching, the requirements necessary for evaluation of independent variables through inferential statistical analysis must not be met; the demands of teaching, not those of evaluation antithetical to teaching, must dictate the arrangements under which evaluation data are obtained.

Descriptive statistical analysis, often recommended as an alternative to inferential statistics (e.g., by Coats, 1970), is, in and of itself, insufficient for evaluating teaching effectiveness. Exactly as in the case of inferential statistics, the operations of descriptive statistics are undertaken independently of those of teaching. Consequently, interpretations, including predictions, based on descriptive no less than on inferential statistical findings are based on presumed relationships between phenomenal variables (stimuli and responses) which the statistical findings apparently (but only apparently) reflect. Strictly speaking, of course, the results of descriptive statistical analysis, like those of inferential statistical analysis unsupported by the satisfaction of necessary statistical, design, and treatment assumptions, cannot be validly used as a basis for prediction. Nevertheless, both kinds of statistical results are regularly used for predictive purposes by teachers and evaluators alike. Validly or invalidly, however, results induced through either descriptive or inferential statistical analysis can only lead to hypothetical deductions about phenomenal-variable relationships which, at worst, have already been produced; or, at best, still remain to be. In any event, the results of both descriptive and inferential statistical analysis do not speak unequivocally for themselves. (Thus the inevitable, "Further research is needed.")

The problem is that predictions based on semantic signs are not <u>impelled</u> by the phenomenal significates they represent. Rather, they are <u>compelled</u> by the logic of semantic analysis. In the case of descriptive statistics, predictions of future phenomenal-significate responses represent "best guesses" based on signs representing mathematical probabilities; in the case of inferential statistics, these "best guesses" are qualified by null hypothesis tests of no significant differences between obtained and theoretical scores. Qualified or not, however, "best guesses" about phenomenal-significate responses cannot be substituted for their unequivocal production. (Theoretically, deduction may precede induction or production of data. In such cases, however, confirmation of the predicted effectiveness of independent variables upon dependent-variable significates must still be either induced or produced. In practice, deduction [therefore, prediction] rarely if ever precedes data induction or production.)

To summarize: First, to be appropriate, the evaluation of teaching must occur under circumstances entirely free of the limitations which inferential statistics necessarily impose on teaching. Insulating dependent-variable phenomenal-significates from "extraneous" independent variables makes sense only if the independent-variable phenomenal-significates under investigation are conceptualized in x predictor terms for which the dependent variables, to the extent they are achieved with pre-established p probability under q conditions, serve as y criteria. (The probability and conditions must be pre-established in order to refer to the prediction of y by x; otherwise, the validity of the prediction is indeterminable and reference to its validity, meaningless.) Obviously, insulating dependent-variable phenomenal-significates is contra-indicated if their production rather than their prediction is the y-criterion goal. Second, to be sufficient, the circumstances of evaluation must include whatever manipulative procedures lead to the production of x-y (independent-variable-dependent-variable) phenomenal-significate relationships; not lead -- as in the case of descriptive statistics -- merely to their measurement according to a formal semantic-sign system.

Functional Analysis of Behavior

Fortunately, an approach to evaluation not only permitting but requiring manipulation of dependent-variable phenomenal significates is available in the form of the strategies and tactics of the functional analysis of behavior, based on operant conditioning, the radical behaviorist model originated by B.F. Skinner in the 1930's. (See e.g., Skinner, 1938, 1953, 1968.) Fundamentally, the functional analysis of behavior entails the continual manipulation of the consequential stimulus conditions which responses encounter, sufficient to produce dependent-variable alterations in responses to criterion. Under such a strategy, the independent variables responsible for phenomenal-significate responses need be neither induced through semantic nor deduced through ratiocinative processes. Their effectiveness is unequivocally revealed by the effects which they produce.

"In the usual case,"

> successive approximations of criterion behavior are differentially reinforced as they are emitted. Behavior is differentially reinforced when it and it alone is reinforced. Towards this end, consequences are arranged so that criterion behavioral approximations are more likely to be emitted. Consequential stimuli sufficient to evoke them are introduced on a gradual, sequential basis, until the ultimate criterion behavior is achieved; this process is called shaping. To be sure, antecedent stimuli may evoke criterion behavioral approximations reflexively, or because, in the past, stimuli like them have been followed by criterion behavioral approximations consequated -- naturally, accidentally, or deliberately -- by reinforcement. These latter, non-reflexive stimuli are called discriminative-- because, for a given subject, they evoke criterion behavioral approximations on a selective basis, i.e., according to the probability of reinforcement as previously determined. Thus, the effectiveness of a non-reflexive antecedent stimulus is due to the previous reinforcement of the effected (effectuated)

> behavior following the stimulus (that is, a member of the stimulus class). Reinforcement is empirically defined in terms of increases in the frequency, percentage, or rate with which members of the class of behavior emitted prior to the presentation or withdrawal of a stimulus or set of stimuli are emitted subsequent to the occurrence of the consequence. The functional consequence, in such a case, is designated a reinforcer. Other consequences a behavior may encounter are extinction and punishment. The former results from the withholding of reinforcement; the latter, from the presentation of an aversive stimulus, i.e., a stimulus which the subject takes steps to avoid or escape, or from the withdrawal of reinforcement. Both extinction and punishment, by definition, result in subsequent behavioral decreases of prior behavior; however, depending on which procedure is employed, the amplitude, longevity, and generalizability of these and other effects tend to be significantly diverse. (Consequences with no effects are designated neutral.) (Throne, 1970b)

Under the procedures of the functional analysis of behavior, then, the causal independent-variable stimuli of criterion dependent-variable responses may be empirically determined through the consequential manipulation of the latter by the former. Indeed, they must be; under these procedures it is unequivocally the stimuli of teaching which cause the production of responses in subjects revealed (through these responses) to have been taught.

> Under [these procedures] the evaluation of independent-variable effectiveness in teaching is a function of that selfsame manipulation of dependent-variable response effects by which the effectiveness of the independent variables of teaching is produced. Unlike the case with inferential and descriptive procedures, evaluation data is achieved in the process of, not extraneous to, teaching; the variables responsible for response effects are empirically determined by those teaching operations which constitute the independent variables employed. If teaching is successful, an evaluation of the effective independent variables is determined ipso facto by the dependent-variable response effects. If teaching fails, this is also determined by the effects (Throne, 1970d).

In a related context, the writer has asserted:

> [N]either the success of treatment nor its failure is determined independently of treatment. Positive results (of treatment) prove the presence of independent variables sufficient for success; negative results prove only that such variables are absent. The belief that the results of one set of operations called treatment need be determined by another set called diagnosis, is a fallacy (Throne, 1970a).

If "teaching" is substituted for "treatment," and "evaluation" for "diagnosis" in the above statement, it may readily be interjected here.

It is important to note that "[t]he question of which aspects of the effective independent variables are necessary"

> to evoke the response effects obtained is a separate issue. To the extent and in the forms it is desirable that this question be resolved, the procedure of choice is the differential reinforcement of criterion responses with the independent-variable aspects in question controlled (e.g., cost, efficiency, convenience, preference, safety, etc.). In other words, the question of independent-variable <u>necessity</u> is always transformed into the question of <u>sufficiency</u>. The principle of consequential determinism demands it (Throne, 1970d).

Principle of Consequential Determinism

The principle of consequential determinism (Throne, 1970a) declares: "Behavior is a function of its consequences" (Skinner, 1938). This principle is the basis of the operant conditioning model, from which the strategies and tactics of the functional analysis of behavior have been derived. In its succinct form above, the principle is shorthand for the full expression:

> A representative sample of a class of behavior is more likely to occur if another representative of the behavioral class has been reinforced, as evidenced by an increase in the frequency, rate, or percentage with which class representatives occur subsequent to consequation (Throne, 1970c).

In other words, only as a behavioral class has been demonstrated to be functionally related to reinforcing consequences, may the consequences be identified as determinative. (Moreover, if demonstration of a functional relationship between a behavioral class and a class of consequences has been provided, the latter must be identified as determinative.)

"It follows that if behavior is a function of its consequences,"

> the presentation or withdrawal (or withholding) of whichever consequences demonstrably determine the fulfillment of criterion behavior may be made contingent on occurrences of that behavior or successive approximations of it (Throne, 1970a).

The principle of consequential determinism is the key, therefore, to the epistemological problem posed in the second paragraph; through its application the requisite dissolution of all methodological distinctions between teaching and evaluation may be achieved. If the dependent-variable response effects of teaching are determined (produced) by consequential, independent-variable stimuli presented or withdrawn (or withheld) on a contingent basis, then the effectiveness of the consequential, independent variables is simultaneously determined (measured) by those selfsame dependent-variable response effects. Evaluation serves the end of determining stimulus-variable effectiveness actually, in terms of phenomenal-significate response effects deliberately and systematically manipulated to criterion; instead of theoretically, in terms of manipulated semantic signs. "Teaching and evaluation thus coalesce" (Throne, 1970d). "Insofar as"

> an educator succeeds in introducing or withdrawing (or withholding) consequential stimulus variables able to determine (produce) dependent-variable responses to criterion, he determines (measures) by the outcome the effectiveness of the independent variables employed. ... Since the effectiveness of teaching may be unequivocally produced through the processes of teaching themselves, its mere measurement through inferential statistics is not only inappropriate; and through descriptive statistics, not only insufficient. In both cases, it is unnecessary (Throne, 1970d).

Indeed, it might even be asserted that the possibility of evaluating teaching effectiveness at the level of phenomenal-significate manipulation, makes evaluation through inferential or descriptive statistical analysis (and ratiocination) at the level of manipulation of semantic-signs, absurd.

References

Coats, W. A case against the normal use of inferential statistical models in educational research. Educational Researcher, 1970, XXI, 6-7.

Coleman, J.S., Campbell, E.Q., Hobson, C.J., McPartland, J., Mood, A.M., Weinfeld, F.D., and York, R.L. Equality of educational opportunity. Washington, D.C.: U.S. Government Printing Office, 1966.

Guba, E.G. and Clark, D.L. A proposal for the National Institute for the Study of Educational Change. NISEC, Indiana University, Bloomington, Indiana, 1967.

Skinner, B.F. Behavior of organisms. New York: Appleton-Century-Crofts, 1938.

Skinner, B.F. Science and human behavior. New York: Macmillan, 1953.

Skinner, B.F. The technology of teaching. New York: Appleton-Century-Crofts, 1968.

Throne, J.M. A radical behaviorist approach to diagnosis in mental retardation. Mental Retardation, 1970, 8(3), 2-5. (a)[2]

Throne, J.M. The psychometric assessment of intelligence: Towards what end? Paper presented to the joint meeting of the American Academy on Mental Retardation and the American Association on Mental Deficiency, Washington, May, 1970. Submitted to Journal of Special Education. (b)

---

[2] A corrected copy of this paper may be obtained from J.M. Throne, Center for Mental Retardation and Human Development, University of Kansas, Lawrence, Kansas 66044.

Throne, J.M. Logical tautologies, construct validity, and motivation. Paper presented to the meeting of the Kansas Psychological Association, Topeka, April, 1970. Submitted to Psychological Review. (c)

Throne, J.M. Statistical analysis in the evaluation of teaching: Is it necessary? Unpublished paper, 1970. Submitted to Educational Researcher. (d)