The Psychological Record, 1975, 25, 265-272.

ON THE ALLEGED FALSITY OF THE NULL HYPOTHESIS[1]

WILLIAM F. OAKES
Brooklyn College of the City University of New York

Consideration is given to the contention by Bakan, Meehl, Nunnally, and others that the null hypothesis in behavioral research is generally false in nature and that if the *N* is large enough, it will always be rejected. A distinction is made between self-selected-groups research designs and true experiments, and it is suggested that the null hypothesis probably *is* generally false in the case of research involving the former design, but is *not* in the case of research involving the latter. Reasons for the falsity of the null hypothesis in the one case but not in the other are suggested.

The U.S. Office of Economic Opportunity has recently reported the results of research on performance contracting. With 23,000 *S*s—13,000 experimental and 10,000 control—the null hypothesis was not rejected. The experimental *S*s, who received special instruction in reading and mathematics for 2 hours per day during the 1970-71 school year, did not differ significantly from the controls in achievement gains (American Institutes for Research, 1972, p. 5).

Such an inability to reject the null hypothesis might not be surprising to the typical classroom teacher or to most educational psychologists, but in view of the huge *N* involved, it should give pause to Bakan (1966), who contends that the null hypothesis is generally false in behavioral research, as well as to those writers such as Nunnally (1960) and Meehl (1967), who agree with that contention. They hold that if the *N* is large enough, the null is sure to be rejected in behavioral research. This paper will suggest that the Falsity contention does not hold in the case of experimental research—that the null hypothesis is *not* generally false in such research.

THE FALSITY CONTENTION

Bakan says that there are a priori reasons for believing that the null hypothesis is generally false. As he puts it:

> **The probability of rejecting the null hypothesis is a function of five factors; whether the test is one-or two-tailed, the level of significance, the standard deviation, the amount of deviation from the null hypothesis, *and the number of observations*. The choice of a one- or two-tailed test is the investigator's; the level of significance is also based on the choice of the investigator; the standard deviation is a given of the situation and is characteristically reasonably well estimated; the deviation from the null hypothesis is what is unknown; and the choice of the number of cases is [sic] in psychological work is characteristically arbitrary or expeditious.**

[1] **Requests for reprints should be sent to William Oakes, Psychology Department, Brooklyn College of the City University of New York, Brooklyn, New York 11210.**

> Should there be any deviation from the null hypothesis in the population, *no matter how small*—and we have little doubt but that such a deviation usually exists—a sufficiently large number of observations will lead to the rejection of the null hypothesis [1966, p. 426]

Bakan presents the following evidence for having little doubt but that a deviation from the null hypothesis usually exists:

> One of the common experiences of research workers is the very high frequency with which significant results are obtained with large samples. Some years ago, the author had occasion to run a number of tests of significance on a battery of tests collected on about 60,000 subjects from all over the United States. Every test came out significant. Dividing the cards by such arbitrary criteria as east versus west of the Mississippi River, Maine versus the rest of the country, North versus South, etc., all produced significant differences in means. In some instances, the differences in the sample means were quite small, but nonetheless, the *p* values were all very low [p. 425].

Nunnally (1960), who also holds that the null hypothesis is generally false in behavioral research, reported as evidence for this contention an experience involving 700 *S*s in a study of public opinion. After a factor analysis of the results, he calculated the correlation coefficients of the factors with age, sex, income, and a number of other variables: "Nearly all correlations were significant, including ones that made little sense [p. 643]."

Meehl (1967), supporting the same contention, presented evidence from a study involving "a huge sample of over 55,000 Minnesota high school seniors" in which it was found that "91% of pairwise associations among a congeries of 45 miscellaneous variables such as sex, birth order, religious preference, number of siblings, vocational choice, club membership, college choice, mother's education, dancing, interest in woodworking, liking for school, and the like" showed statistically significant relationships [p. 109].

The following report will distinguish between two basic classes of research design, the self-selected-groups (SSG) design and a true experimental design, and it will suggest that while the Falsity contention is very likely true with respect to the SSG design, it does not hold for the experimental design.

DISTINGUISHING BETWEEN A TRUE EXPERIMENT AND THE SSG RESEARCH DESIGN

Subject characteristics are quite commonly taken as the independent variables in behavioral research, as, for example, when behavior is observed to be a function of *S*s' age, sex, socioeconomic level, ethnic identification, education, nAch level, drive level, anxiety level, psychiatric diagnosis, etc. With some subject variables, e.g., anxiety level, the *E* may have a choice. He may treat anxiety as a personality variable and choose two groups of *S*s for the levels of the independent variable on the basis of their scores on some instrument designed to assess anxiety level, such as the Taylor Manifest Anxiety Scale. Or, rather than selecting *S*s for the levels of the independent variable on the basis of their degree of pre-existing anxiety, he may instead set up conditions designed to *induce* in his *S*s different levels of the independent variable, high versus low anxiety, then observe the subject behavior that is his dependent variable under the two levels of induced anxiety. In the latter case, when *E* induces the level of anxiety in his *S*s, he is manipulating the level of the independent variable at will. He is, in effect, *assigning* his *S*s to the levels of the independent variable. It is this assignment by *E* of the levels of the independent variable to *S*s at will that qualifies the latter type of study as a member of the class of true experimental designs. On the other hand, in the former case *E* cannot *assign* high or low scores on the Manifest Anxiety Scale to his *S*s at will. Rather, it is the characteristics of *S*s themselves that determine their membership in the groups constituting the levels of the independent variable. The level of the independent variable for each *S* in this type of design is thus self-determined, i.e., determined by his own characteristics, *independent of E's assignment*—the SSG research design. A research study utilizes the true experimental design whenever *E* assigns the levels of the independent variable to *S*s at will, and it is an SSG design whenever the level of the independent variable is determined for each *S* by his own characteristics. Research in which a subject characteristic is taken as the independent variable usually involves the SSG design, as it is rarely possible for *E* to manipulate subject variables at will. He must take such variables as age, sex, race, socioeconomic status, religion, etc., as he finds them.

EVIDENCE FOR THE FALSITY CONTENTION— EXCLUSIVELY SSG

Taking Bakan's example, when he separated his *S*s into those from east versus those from west of the Mississippi River, from Maine versus the rest of the country, from the North versus South, he was using the SSG design. He was setting up those contrasted groups based on subject characteristics as the levels of his independent variable and was then testing the significance of the difference in the mean test scores (the dependent variable) of the groups thus formed. Similarly, when Meehl separated his *S*s on the basis of sex, birth order, religious preference, number of siblings, etc., and found relationships among the variables, he was also in effect setting up independent variable levels based on subject characteristics and comparing the groups on various dependent variables—he was also using the SSG design. Nunnally was also using the SSG design when he found differences in the factor scores (the dependent variable) of *S*s differing on age, sex, income, and other such variables (the independent variables), as evidenced by the correlation of the factors with such subject variables.

The evidence those writers have presented to support their contention that a large enough *N* will always result in rejection of the null can thus be seen without exception to involve the SSG research design. The contention is not supported by evidence resulting from the use of a true experimental design, i.e., research involving a manipulated independent variable, the levels of which are assigned at will to *S*s by *E*. The OEO study mentioned in the first paragraph above was a true experimental design and did not reject the null, even with an *N* of 23,000. The reason this can happen is that although the Falsity contention may well be true with respect to research involving the SSG design, it is not true for the experimental design.

WHY THE FALSITY CONTENTION HOLDS FOR THE SSG DESIGN

Why should the Falsity contention be true in the case of the SSG design? In the first place, it should always be recognized that the subject variable which is the independent variable in an SSG design is itself *caused*. That is, there is some combination of other variables which has causally determined those subject characteristics which qualify the individuals for membership in the groups constituting the levels of the independent variable. Continuing with the example of high versus low anxiety levels (TMAS scores) as the independent variable, there is for each *S* some combination of other (extraneous) variables that has determined whether he will be in the high- or low-anxiety-level group. If this is the case, then it follows that a comparison of the high versus low groups with respect to any of those variables that are involved in the determination of anxiety level for the *S*s will find the high and low groups differing with respect to that variable. That is, if the *N* is sufficiently large, the null hypothesis will be rejected for any such influential variables.

But such influential variables likely do not exert their causal influence *only* on the particular subject variable taken as the independent variable—anxiety in this example. Many other subject characteristics and aspects of behavior of the *S*s will also be influenced by this same combination of influential variables that has determined the high versus low level of anxiety. If this is true, it follows that a comparison of the two anxiety level groups on any dependent variable whose determining variables overlap with those determining anxiety level will result in the rejection of the null if the *N* is sufficiently large.

Further, those influential variables that have determined the level of the independent variable for *S*s are likely to be found to covary in nature with other influential variables, which variables, plus others also influenced by them, will also be found to differ for the independent variable groups—if the *N* is large enough. In any SSG design it can be seen that there will exist a vast network of interrelated variables covarying with the independent variable.

Thus if one assumes a deterministic or at least a quasi-deterministic position (as any serious behavioral scientist must do), that the subject variable that is the independent variable in an SSG study is *caused,* then it must be expected that the null hypothesis is likely to be rejected for any dependent variable of interest—whenever the SSG design is used. As Meehl suggests,

> Our general background knowledge in the social sciences, or, for that matter, even 'common sense' considerations, makes . . . an exact equality of all determining variables, or a precise 'accidental' counterbalancing of them, so extremely unlikely that no psychologist or statistician would assign more than a negligibly small probability to such a state of affairs [Meehl, 1967, p. 108].

It therefore seems quite reasonable that the Falsity contention should be accepted as true whenever the SSG design is used. This more restricted proposition may be stated as follows: *With groups selected on the basis of a difference on one variable of psychological interest, the null hypothesis with regard to any other behavioral variable is probably false in the state of nature.*

WHY THE FALSITY CONTENTION DOES NOT HOLD FOR A TRUE EXPERIMENT

But what about the case of research involving a true experimental design with a manipulated independent variable? It can be seen that in order for the Falsity contention to hold, and for the probability of the null being rejected to increase with increasing *N,* it is necessary to assume some fixed, nonzero amount of difference in the dependent variable values for the independent variable groups. It is true that if there is a nonzero difference, no matter how small, in the dependent variable for the independent variable groups, the null hypothesis will be rejected if the *N* is large enough. Such a fixed difference would place a constant in the numerator of the *t*-fraction for the significance test; and since the denominator of the *t*-fraction (the measure of variability) decreases as *N* increases, the value of *t* must eventually be sufficiently large to reject the null. But such a fixed value of the numerator does not necessarily occur in the case of research involving a true experimental design.

The defining characteristic of the true experimental design is, as pointed out above, the assignment *at will* by the researcher of levels of the independent variable to *S*s. Continuing with the example of two levels of anxiety, high versus low, presumably induced in *S*s by *E*'s manipulations, in this true experiment *E* assigns his *S*s *at random* to the high versus low anxiety groups. He then compares the dependent variable values for the groups thus constituted. This random assignment of *S*s to the two groups produces groups which, before *E*'s manipulations, have a probability of differing on any variable equal to the alpha-level of the significance test used. The important point here is that this probability of the groups differing initially on any variable of interest is *independent of the size of N*. In other words, by using random assignment of *S*s to levels of the independent variable, *E* is setting up a condition in which there is *no* relationship between the basis for assignment to groups and any behavioral measure. That is the definition of a random assignment, that the assignment is independent of the characteristics of the individuals assigned. So no matter what behavioral characteristic one may choose, no matter what its base rate in the population, the probability of an individual with a high score on the behavioral characteristic being in the one group initially is equal under random assignment to his probability of being in the other. The same is also true for any individual with a low value on the behavioral characteristic of interest. With a random procedure, as such *S*s are assigned to independent variable groups, the likelihood is that an individual who scores above the mean of the population on a measure of any subject variable and who is assigned to one of the groups will be matched by an individual with a similarly high score who is assigned to the other. And those high-scoring individuals in each of the groups are likely also to be matched by equally low-scoring individuals assigned to the two groups. So the expectation is that the means of the samples drawn from that population—and note that they are drawn from the *same* population—in the two independent-variable groups will approximate equivalence on any subject variable as they balance out. If the sample size of the two groups is small, one expects greater disparity in their means on any subject variable, because of the occasional fortuitous assignment of a high- or low-scoring individual to one group without its being balanced by a similarly high- or low-scoring individual being assigned to the other. However, as the sample sizes are increased, the contribution to the mean of any such extreme score is diminished, and thus the probability of extreme deviations of the difference in the means from a value of zero *decreases with increasing N*. In short, the expected population value of a difference in means of *zero* is approached stochastically as *N* increases. This is accommodated for in the formula for the test of significance of a difference in means by having the denominator in the *t*-fraction, the measure of variability, decrease as a function of increasing *N*, so that a smaller absolute difference in the means is required for a given level of significance as *N* increases, and a larger absolute value of the difference in means is required for significance at a given level as *N* decreases. The Falsity contention that the probability of rejecting the null increases with an increase in *N* ignores the stochastic approach of the numerator of the *t*-fraction to the population value of zero, and instead assumes some fixed, nonzero difference between the sample means. It therefore does not apply in the case of a true experimental design—unless the independent variable does indeed have an effect upon the dependent variable.

It is clear that if there is no effect of the independent variable on the dependent variable in a true experiment, the null hypothesis will not be rejected veridically no matter how large the *N*. But is it possible for any independent variable to have no effect on the dependent variable? Of course it is. One can think of a whole host of what would be considered "trivial" variables that would have no effect on a particular dependent variable one might be interested in. The color of the *E*'s tie wouldn't affect the *S*'s critical flicker fusion frequency; the sign of the zodiac under which rats are born would not influence their rate of learning an avoidance response, etc. Those are, of course, trivial examples. But how does one know whether a variable is trivial with respect to a particular dependent variable—how does one know whether a variable actually does influence behavior? Meehl would have us believe that "it is highly unlikely that *any* discriminable stimulation which we apply to an experimental subject would exert literally *zero* effect upon any aspect of his performance [1967, p. 109]." Actually, one can find out whether a variable is causally related in a consistent manner to a behavioral dependent variable of interest only by conducting an experiment.

Meehl may be correct with respect to the individual *S*. In the example above it may be that the color of *E*'s tie, if discriminated by a particular *S*, may influence some aspect of his behavior, possibly that being observed as the dependent variable. But in order for this to result in the rejection of the null hypothesis when the dependent variable values are compared for *S*s run with one tie color versus another, the tie color variable would have to influence the critical flicker fusion frequency of many more than just that one *S*—and to influence it for *S*s *consistently in the same direction*. In other words, it would have to exert a consistent causal influence on the dependent variable over *S*s.

Probably a great many graduate students have had the experience of running an experiment and getting "almost significant" results, then increasing the *N* only to have the effect wash out as nonsignificant with the larger *N*. If the Falsity contention were true, that shouldn't happen. One should always be able to get significance by just increasing the *N*. But it doesn't hold for the experimental design, and it is clear that there are some independent variables that don't affect some dependent variables. And they are not all "trivial" variables. For example, Slamecka (1968) has been able to demonstrate the unexpected finding that in a free-recall task *S*s' ability to recall the remaining items on a list is not influenced by *E*'s supplying some of the words on the list for him—even with 29 of the 30 words being supplied. And Guttman and Kalish (1956) found that the height of the generalization gradients along a continuum of wavelengths of light was not influenced by the discriminability of the hues for their pigeons.

Researchers have a powerful tool for detecting causal relationships *and their absence* in the true experimental design. The claims of Bakan and others to the contrary notwithstanding, one should not assume that everything is related to everything else behaviorally, and that it is always necessary only to increase the *N* to show any variable as significantly related to any other. One doesn't need to assume that if only the OEO study had used 46,000 or 92,000 instead of only 23,000 *S*s, it would have shown performance contracting as significantly related to student achievement. When a true experimental design has been used, the conclusion of no relationship can be accepted, or if one is a real purist bothered by the fact that an infinite *N* has not been used, it is possible to specify, giving consideration to the power of the test of significance, just how small an effect would have had to be in order to have escaped being detected in the experiment. But it shouldn't be assumed that the null hypothesis is generally false in an experiment.

REFERENCES

AMERICAN INSTITUTES FOR RESEARCH. 1972. OEO reports performance contracting a failure. *Behavioral Sciences Newsletter for Research Planning,* 9, 4-5.

BAKAN, D. 1966. The test of significance in psychological research. *Psychological Bulletin,* 66, 423-437.

GUTTMAN, N., & KALISH, H. I. 1956. Discriminability and stimulus generalization. *Journal of Experimental Psychology,* 51, 79-88.

MEEHL, P. E. 1967. Theory testing in psychology and physics: A methodological paradox. *Philosophy of Science,* 34, 103-115.

NUNNALLY, JUM. 1960. The place of statistics in psychology. *Educational and Psychological Measurement,* 20, 641-650.

SLAMECKA, N. J. 1968. An examination of trace storage in free recall. *Journal of Experimental Psychology,* 76, 504-513.