JOURNAL OF APPLIED BEHAVIOR ANALYSIS 1977, 10, 167-172 NUMBER 1 (SPRING) 1977

"PERHAPS IT WOULD BE BETTER NOT TO KNOW EVERYTHING"[1]

DONALD M. BAER

UNIVERSITY OF KANSAS

The advent of statistical methods for evaluating the data of individual-subject designs invites a comparison of the usual research tactics of the group-design paradigm and the individual-subject-design paradigm. That comparison can hinge on the concept of assigning probabilities of Type 1 and Type 2 errors. Individual-subject designs are usually interpreted with implicit, very low probabilities of Type 1 errors, and correspondingly high probabilities of Type 2 errors. Group designs are usually interpreted with explicit, moderately low probabilities of Type 1 errors, and therefore with not such high probabilities of Type 2 errors as in the other paradigm. This difference may seem to be a minor one, considered in terms of centiles on a probability scale. However, when it is interpreted in terms of the substantive kinds of results likely to be produced by each paradigm, it appears that the individual-subject-design paradigm is more likely to contribute to the development of a technology of behavior, and it is suggested that this orientation should not be abandoned.

DESCRIPTORS: individual-subject design, group design, Type 1 error, Type 2 error, inferential statistics

If behavior might be different under a condition known as "A" than it is under a condition known as "B", and if it were important to find out whether that possibility was an actuality, then two basic paradigms would be available for its examination.

A number of subjects might be recruited and divided at random into two equal groups. One of these groups would be exposed to the "A" condition, and its behavior noted; the other would be exposed to the "B" condition, and its behavior similarly noted. The mean behavior of those exposed to "A" could be compared to the mean behavior of those exposed to "B". A difference in those means might be interesting.

Alternatively, a single subject might be recruited and exposed to the "A" condition for some time to behave repeatedly under its influence. Then, the "A" condition would be replaced by the "B" condition, and the ongoing behavior would continue to be monitored as before. An alternation of "A" and "B" conditions would continue, and the repeated patterns of behavior seen under the repeated "A's" would be compared to the repeated patterns of behavior seen under the repeated "B's". A consistent difference in those arrays might be interesting.

Unfortunately, in either paradigm, there are conditions under which the difference would not be interesting. Primarily, those are the conditions under which suspicion arises that the difference is a result only of chance, rather than of the intrinsic difference between "A" and "B". This suspicion is profound in behavioral scientists. It arises from a past history of discovery that behavior is subject to control by many, many variables other than "A" or "B". Any prudent inductive organism, confronted with the

[1] This comment has been attributed to Oedipus Rex, shortly after his successful investigation of a public-health problem in Thebes. The article following the comment is based largely on a symposium report presented at the meeting of the Association for the Advancement of Behavior Therapy, San Francisco, December, 1975. Reprints may be obtained from the author, Department of Human Development, University of Kansas, Lawrence, Kansas 66045.

fact that a subject matter may be affected by many, many known variables, will leap reasonably to the conclusion that the same subject matter may well be affected by even more unknown variables. Consequently, any difference in behavior will always be subject to interpretation as a product of some currently unknowable fluctuation in those unknown variables.[2] The Greeks referred to unknown variables as Barbarians; scientists refer to them as Chance.

In the individual-subject paradigm, a judicious defence against chance is available. The total array of behaviors under all the "A's" and all the "B's" is examined. Because each "A" and each "B" has yielded repeated displays of the behavior, and because the "A's" and "B's" themselves have been presented repeatedly, a potential consistency of pattern is available for inspection. If behavior repeated under the repeated "A's" is repetitively different from behavior repeated under the repeated "B's", the scientist will conclude that such consistency cannot be a product of chance. After all, it has been repeated quite repetitively. This conclusion obviously has a quantitative base, but the base is never made precise. However, a subtle recourse to the other paradigm is applied: groups of scientists are recruited to examine the same total display, and when their mean reaction is that this consistency cannot be a product of chance, then the conclusion is proclaimed a scientifically sound one. The Greeks referred to this as Democracy; scientists refer to it as Editorial Review.

In the other paradigm, a differently judicious defence against chance is available. A catechism is recited, much as follows:

What is desired is knowledge about the population of differences between "A" and "B". In particular, it is important to know whether the mean difference in the population is zero.

Unfortunately, the population of differences between "A" and "B" is too large to be available: all that is available are samples of "A-B" differences, and samples vary in their resemblance to the population from which they are drawn.

However, if the samples are drawn randomly from the population, then the likelihood that they resemble or deviate from the population to any specified degree is computable through the application of the Laws of Probability.

Unfortunately, this is true if and only if the samples are random samples from the population.[3] It is possible to sample randomly from a population, but it is extremely difficult to do so. Of course, it is always possible to consider a given sample as a random one from *some* population, but it is extremely difficult to specify from which pop-

[2] Or even some known ones unfortunately not under current experimental control.

[3] The ability to make an "if and only if" statement is both rare and wonderful in science. This one rests on an argument put forth repeatedly by statisticians. Pearson (1900), Walker and Lev (1953, pp. 10-12), and Wallis and Roberts (1956, p. 116) provide three examples of this argument.

ulation it is a random sample.

However, it is always possible to divide a sample randomly into two groups.

Unfortunately, dividing a sample randomly into two groups is not the same as randomly sampling the population of "A-B" differences.

However, it is always possible to ignore this deficiency and apply the Laws of Probability anyway. Doing so will yield an apparent probability of the sample difference having been drawn from the population, given any assumption about the population from which it was drawn.

As an act of conservatism, it may be assumed that the population of "A-B" differences has a mean of zero—in other words, that the difference between "A" and "B" is not a functional one for the behavior under study. Then, application of the Laws of Probability will produce an apparent estimate of the probability of a sample difference this large, if in fact there ordinarily would not be a difference at all. If that probability is low enough, conservatism will be abandoned; it will be concluded that the sample result is too unlikely an event, under the assumption of no functional differences between "A" and "B". Then, the assumption itself will be doubted, and will be set aside in favor of the conclusion that after all, "A" and "B" do affect this behavior differently. Usually, this will be done if the probability is as low as 0.05.

But it will always be remembered that this conclusion could be an error, and in fact, that proceeding in this manner will virtually guarantee that some 5% of all such scientific conclusions are errors. Presumably to honor this institutionalized error rate, and also because it will need to be referred to very often, it is given a title: these errors are called Type 1 errors. Of all the errors that there are to be made, these have primacy, because they are almost certain to be made. The Greeks referred to this as *hubris;* scientists refer to it as Inferential Statistics.

It should be realized immediately that Type 1 errors are not unique to the group paradigm. They occur in identical form in the individual-subject paradigm as well, for both paradigms need to defend against chance. But in the group paradigm, because of the decision to apply the laws of probability (justifiably or not), Type 1 error probabilities are computable; in the individual-subject paradigm, they are not.[4] In that paradigm, they are merely worrisome, and the basic defence against them is to consider the total array of data points available within the design with a fairly skeptical eye. That is, a difference has to be seen to be affirmed. A comparison of differences treated by skeptical examinations by eye, and by computation of the probability that they arose by chance from a zero-difference population, suggests strongly that much smaller and less consistent differences can be validated by computation than by inspection. That is, in the individual-subject paradigm, the probability of Type 1 errors is not known with any precision, but is clearly much smaller than 0.05.

Then, one might reasonably ask why Type 1 errors should be tolerated at an 0.05 probability level in the group paradigm, when they might well be reduced by relying on visual inspection rather than computation, as in the individual-subject paradigm. The answer, of course, is that there is a second type of error, fittingly known as a Type 2 error, that may be committed in

[4] Maybe. See, for example, Jones, Vaught, and Weinrott, 1977.

either paradigm, when Type 1 errors are avoided. In fact, when worrying over "A-B" differences, there are *always* two errors available: we may conclude from our sample that there is a difference, when in the population there is none (Type 1); or we may conclude from our sample that there is no difference, when in the population there is indeed one (Type 2). Furthermore, an inevitable arithmetic relates these two types of errors. That arithmetic rates a chapter in the usual textbook; here, let it suffice to remember that whenever we decrease the probability of one type of error, necessarily we increase the probability of the other. Individual-subject-design practitioners, operating at very, very low probabilities of Type 1 errors, consequently operate at very high probabilities of Type 2 errors. Group-design practitioners, able to operate at higher levels of Type 1 error probability like 0.05 (a practice difficult to match by visual inspection unaided by computation) thereby are also able to operate at somewhat lower probabilities of Type 2 errors.

In one sense, then, the advocates of each paradigm are really very similar in their scientific research practices. Both are aptly enough described by the model of inferential statistics: they are all Type 1 and Type 2 error-avoiders. It is simply that one (the individual-subject-design practitioner) usually does not calculate the probabilities of Type 1 errors, and as a result is forced to estimate them by visual inspection of the total array of data available. Consequently, the individual-subject-design practitioner makes very few Type 1 errors and very many Type 2 errors. The group-design practitioner, typically committed to calculating Type 1 error probabilities and choosing to operate at an 0.05 probability level of making them, thereby makes somewhat more Type 1 errors than does the individual-subject buff, but at the same time makes considerably fewer Type 2 errors than that imprecise person. Thus, what has sometimes been seen as a most profound difference in scientific practice, in this interpretation dissolves into a mere difference of opinion about where to set two parameters: the probabilities of Type 1 and Type 2 errors.

On the other hand, that very same difference could also be described in terms that suggest a more profound difference than a few centiles on a probability scale. This is the difference that emerges if we ask what kind of errors Type 1 and Type 2 errors are, in substantive terms.

To make a Type 1 error is to affirm that a certain variable is a functional one, when in fact it is not. Scientists who commit relatively many Type 1 errors are bound to memorize very long lists of variables that are supposed to affect diverse behaviors, some predictable proportion of which are not variables at all. By contrast, scientists who commit very few Type 1 errors have relatively short lists of variables to remember. Furthermore, and much more important, it is usually only the very robust, uniformly effective variables that will make their list. Those who will risk Type 1 errors more often will uncover a host of weak, occasional, or otherwise highly specialized variables. Unquestionably, they will know more, although some of that more is wrong, and much of it is tricky.

To make a Type 2 error is to deny that a certain variable is a functional one, when in fact it is. Thus, those who keep their probability of Type 2 errors low do not often reject an actually functional variable, relative to those whose Type 2 error probability is higher. Again, unquestionably, the practitioner with the lower probability of Type 2 errors will know more; but again, the nature of that more is seen often in its weakness, inconsistency of function, or its tight specialization.

Thus, to sum up, movement of a few centiles on a probability scale when establishing the acceptable risks of Type 1 and Type 2 errors, can alter rather importantly the character of what will be learned. Individual-subject-design practitioners, operating without calculation of the pertinent probabilities, necessarily fall into very low probabilities of Type 1 errors and very high probabilities of Type 2 errors, relative to their group-paradigm colleagues. As a result, they learn about fewer variables, but these variables are typically more powerful, general, dependable, and—very important—sometimes actionable. These are exactly the variables on which a technology of behavior might be built.[5] Thus, it is no coincidence that the individual-subject-design practitioners proved to be foremost in the development of behavioral technologies—considering the methods under which they usually operated, there was little else they could discover. Furthermore, they were comfortable in their method: their interest, when technological, was to solve social and personal problems. If a problem has been solved, you can *see* that; if you must test for statistical significance, you do not have a solution. This, after all, was the major conclusion to be made, not whether an experimental effect had been uncovered.

Currently, we are offered methods that might allow calculation of Type 1 errors in the use of individual-subject designs (*e.g.,* Gentile, Roden, and Klein, 1972; Jones, Vaught, and Weinrott, 1977). The offer is controversial, in that the data of individual-subject designs, typically consisting of the repeated behavior of an organism under sometimes similar, sometimes different conditions, do not seem to meet the independence-of-data assumptions characterizing the data around which parametric statistical methods had originally been conceived. However that controversy might be resolved, there will remain the one just sketched: if we calculate (correctly or incorrectly) the probability of Type 1 errors in our future research, will we promote ourselves out of that exquisitely valuable small corner in which we have been trapped these past years—the discovery of nothing more than a technology of behavior?

Of course, calculation of Type 1 error probability levels does not require that we set them as high as 0.05. We may set them at 0.00001. But there is a nonscientific contingency waiting for us. Results significant at the 0.05 level usually are publishable; publication is usually reinforcing, and sometimes essential; and what most of us know is what all of us publish. Thus, considering what the behavioral technology that we already know has taught us, it seems likely that if we accept this new offer, then we will be pressed to learn more about the less basic, less general, less dependable, less consistent, and less usable aspects of behavior. That is, we may have to become scholarly.

Scholarship has not usually been considered a bad value or an undesirable outcome, and even in this field, we do not ordinarily classify it with tantrums and headbanging. True, it is occasionally linked to impotence, but only in a metaphorical way. As my mother (Baer, *personal communication*) often said, "What can it hurt?" If we continue to discriminate carefully those effects that are strong, consistent, and dependable from those that are otherwise, then nothing much can be hurt. We will, of course, need more journals, or more pages, or both; but we learned to read very, very selectively long ago,[6] so that should be survivable.

On the other hand, what if we lost our old value for only the robust variables? We might, because we fell into that value rather than adopted it systematically, mainly because we were individual-subject-design practitioners.

[5] It is sometimes suggested that a technology of behavior could also be built by packaging together many variables that have weak, occasional, or otherwise specialized functions, on the premise that the resultant package will thereby contain something for everyone, and that some of the somethings will cumulate in their effectiveness and thereby become powerful. That would be interesting to see. At present, it is still a difficult problem to package the known powerful variables well enough to be universally useful. Perhaps this latter problem deserves priority over the former.

[6] We read selectively because we are not looking for experimental effects, but for useful effects. That is, we do not need to expand the list of at-least-sometimes effective variables; rather, we need to array the already known, highly effective variables into useful programs that solve problems. The occasional complaint that our indifference to new demonstrations of experimental effects leaves us nothing exciting to do, ignores the fact that we have not yet designed the *programs* of already known effective variables that will solve drug abuse, alcoholism, birth control, *etc. etc.*

Then, conceivably the offer of statistical methods for those designs could divert us from the much needed further development of that technology we almost have in hand. If it did, we would cease to be distinctive, hopeful, or useful. The Greeks referred to that as Tragedy.

REFERENCE NOTE

Baer, I. S. *Personal communications,* 1931 —.

REFERENCES

Gentile, J. R., Roden, A. H., and Klein, R. D. An analysis-of-variance model for the intrasubject replication design. *Journal of Applied Behavior Analysis,* 1972, **5**, 193-198.

Jones, R. R., Vaught, R. S., and Weinrott, M. Time-series analysis in operant research. *Journal of Applied Behavior Analysis,* 1977, **10**, 151-166.

Pearson, K. On a criterion that a given set of deviations from the probable in the case of a correlated system of variables is such that it cannot be reasonably supposed to have arisen from random sampling. *Philosophical Magazine,* 5th series, 1900, **50**, 339-357.

Walker, H. M. and Lev, J. *Statistical inference.* New York: Holt, 1953.

Wallis, W. A. and Roberts, H. V. *Statistics: a new approach.* Glencoe, Illinois: Free Press, 1956.

Received 10 May 1976.
(*Final acceptance 29 June 1976.*)