Peer Commentry on 'Application of Single Subject Randomization Designs to Communicative Disorders Research' by Susan Rvachew

Random Assignment and Randomization Tests

The reader of Rvachew's article may well ask, "Is this article about randomized single subject designs or about randomization tests?" The principal focus, as the title indicates, is on the designs, not the tests. And that is as it should be in single subject research. Too frequently we hear that an experimenter should keep in mind the statistical test to be applied, and design the experiment accordingly; otherwise, it is contended, it may not be possible to carry out a test on the data. That observation is certainly important when one is restricted to standard statistical procedures, but it is not relevant when a statistical test can be made to order. A randomization test can always be developed to provide a statistical test of treatment effect, provided some type of random assignment, no matter how unusual, has been performed. That is true of multi-subject experiments for any type of random assignment of subjects to treatments, and it is true of single subject experiments for any type of random assignment of treatment times to treatments. Thus, an experiment may be designed with the random assignment that is most suitable without fear that the results will not be analyzable.

The flexibility that randomization tests provide the user of a randomized single subject design is not the only reason for using randomization tests to analyze the data. Since, in a single subject experiment, there is no random sampling, the statistical test must be one that is valid in the absence of random sampling and one that can utilize a probability distribution of a test statistic based on random assignment alone, namely, a randomization test. The importance of random assignment to internal validity of multi-subject experiments is widely recognized, and Rvachew has stressed its importance in single subject experimentation as well. And when one introduces control over extraneous variables in a single subject experiment by random assignment of treatment times to treatments, the randomizing of the extraneous variables (such as fatigue or boredom) provides the basis for statistical control over those variables by means of a randomization test.

E.S. Edgington
Department of Psychology
University of Calgary

* * *

Susan Rvachew has provided an excellent overview of the application of randomization tests to a variety of clinical situations that might otherwise prove statistically intractable. Researchers and clinicians should find this exposition very helpful. There are two general areas I would like to comment on. The first concerns the logic of statistical testing in general and the alternative interpretation for different types of tests. The second involves the extreme flexibility of randomization tests and their general compatibility with other clinical testing procedures.

Rvachew's suggestion that, "Clinical and statistical significance are two quite different ways of judging experimental outcomes," is somewhat over simplified. The central issue in the assessment of effects of treatment on behavior is, "Are they large enough to be important?" The definition of importance will vary from study to study. In some cases, the existence of subtle but reliable differences in listeners' behavior (e.g., a one percent difference in response latency for two different types of stimuli) could be of great theoretical interest. In other cases, for example, clinical applications, a highly reliable difference may be so small as to be of no practical importance.

However, reliability (or replicability) is the *sine qua non* of any kind of experimental research. No matter how large an effect appears to be for a given experiment, if there is large natural variability in the target behavior in question, a large value on any given occasion may simply represent "background noise." It is here that statistics has its most important use in scientific inference: We compare the results of a particular experiment with a family of possible results from a null hypothesis and assess how the observed results compare to typical results expected from that null model.

For parametric tests in fully randomized designs, the concern is with the natural variability expected on random sampling from specified populations of subjects. For some single subject studies, we might be concerned with comparison of expected results from replication of the same experiment on the same subject a number of times: How large is the advantage of treatment versus control trials compared to the variability of this advantage on repeated application and withdrawal of a treatment (e.g., in the ABAB experiments of Kearns, 1986).

In the case of randomization tests, the relevant population is apparently quite different. It is roughly the population of measures of a summary statistic (e.g., average number of articulation errors) with exactly the same raw responses occurring at exactly the same times, but with random association of treatments to treatment times. As Edgington notes: "The null hypothesis for a randomization test is that the obtained measurement for each experimental unit...will be the same under one assignment to treatments as under any alternative assignment that could have resulted from the random assignment procedure" (1987:2).

Despite the different "noise" populations of relevance, it is remarkable that randomization tests often lead to the same conclusions as parametric tests based on random sampling when the relevant assumptions about population parameters are met. Indeed, randomization test techniques have been employed by Kempthorne (as noted by Edgington, 1987:20) to examine the robustness of parametric tests under specific violations of parametric assumptions. However, despite the equivalence of significance levels in special cases, it is important to understand the fundamentally different logic that applies to randomization tests. In particular, as Edgington emphasizes, it is important that we are not *statistically* justified in drawing any conclusions about subject populations beyond those individuals in our study. (Nor for that matter, are inferences about replication of the same experiment on the same subject at other times statistically justified since there is no pretense of random sampling from a large set of replications of the same experiment on the same subject.) Rather, we must appeal to "non-statistical" arguments for generalization.

The second issue I want to address is the flexibility of randomization tests. As Rvachew notes, Edgington (1987: 11) emphasizes not only are randomization tests non-parametric in the usual sense of avoiding distributional assumptions, but they also do not require the random sampling procedures, which are required for valid inference by standard parametric procedures. True random sampling (e.g., by lottery) from a well-defined population rarely occurs in either clinical or university research settings. (Note that such a lottery is totally impossible for producing "random samples" from a putative population of replications of the same experiment on the same subject.) As Edgington notes, arguments that our real samples "resemble" such true random samples are not *logically* relevant to the validity of purely statistical inference. At best, we are left with another type of "non-statistical generalization" viz., that our samples are "random like" and that, if our procedures are valid for true random samples, they will be "valid-like" for our "random-like" sample. That the validity of randomization tests, within their own framework, is not dependent on such assumptions is for Edgington their most attractive feature.

However, the most important practical consequence of the non-dependence of randomization tests on random sampling is the unique way in which they are able to "neutralize" the effects of extraneous variables. The "test for treatment interventions" described by Rvachew is most noteworthy for clinical applications. This method really opens new doors. In particular, it can be applied in a variety of settings with relatively minimal disruption of normal training or treatment procedures. By following a regimen of random onset of a new treatment during the course a standard treatment program (i.e., while other proven treatment is being administered independently), the differential effects of the new method can be fairly assessed. Provided precautions are taken to ensure the administration of the new treatment at randomly selected times (but see Edgington, 1987, p 14-15 for a reminder of some things that can go wrong with an experiment, independent of randomization), the potential confounding by improvement due to the standard treatment is neutralized. Of course, as Rvachew points out, tests in such an environment may not be as sensitive (powerful) as in a more controlled situation where extraneous variables are minimized. However, they still provide valid type I error rates under conditions on minimal disruption of normal treatment and are likely to be applicable to a wider variety of clinical settings.

Finally, it should be reemphasized that other single subject designs can easily be modified to incorporate random assignment methods. Indeed, as Rvachew notes, "a randomization test can be applied to any conceivable single-subject experiment in which there is a random assignment of treatment times to treatments." In particular, appropriate randomization tests could be applied to the multiple baseline designs described by Kearns (1986: 208 ff) with minimal change of experimental methods: namely, by random assignment of training onset times for the target behaviors.

Terrance M. Nearey
Department of Linguistics
University of Alberta

* * *

Response to Commentaries by Drs. Edgington and Nearey

Both Edgington and Nearey have expanded on some of the theoretical issues raised in the paper. I agree with Nearey's comments regarding clinical significance and reliability in experimental research. My point is simply that an evaluation of experimental outcomes requires two separate judgements: It is necessary to judge both the clinical importance of the treatment effect and the internal validity of the experiment as a whole. I believe that these two issues are often confused. For example, McReynolds and Kearns (1983) state that "many applied researchers are only concerned with large, clinically significant changes and smaller changes revealed through statistical analysis may be viewed as unimportant" (p. 127). While it is true that statistical analysis may detect small effects that are difficult to discern from visual analysis alone, the purpose of statistical analysis is not to detect small changes per se. Rather, statistical analysis helps to establish the internal validity of an experiment. Conversely, effect size helps in making a judgement about clinical significance, but tells us little about the internal validity of the experiment.

Traditionally, the internal validity of single-subject experiments has been established through non-statistical means. As Edgington points out, a randomization test is the only statistical procedure which can be applied validly to single-subject data. I do not believe that there are any sound arguments for rejecting the use of random assignment and statistical analysis via randomization test as a method for establishing the internal validity of a single-subject experiment. In fact, it could be argued that the assumptions underlying single-subject randomization designs are more tenable than those underlying traditional single-subject designs. In particular, it is not necessary to assume that "performance under baseline conditions predicts future performance if the treatment were not introduced" (McReynolds & Thompson, 1986, p. 198).

Both Nearey and Edgington comment on the flexibility that randomization tests provide the single-subject researcher. However, it is true that practical considerations will likely limit the number of randomization designs that are used. As Nearey notes, the "test for treatment intervention" appears to have the most potential for communicative disorders research because it fits well with normal clinical practice (i.e., a period of diagnosis or observation, followed by a period of treatment with continued monitoring of performance). In addition, this design avoids the problems associated with repeated applications and withdrawals of the treatment (e.g., carryover effects).

The only problem with this design is that it provides less power than the other designs, given the same number of treatment sessions. In order to increase power it is necessary to: (1) schedule a relatively large number of treatment times, or (2) repeat the experiment with a number of subjects, and then combine probability values across subjects.

* * *