

Proposed Solutions for Nonrandomness in Educational Research

Thomas O. Maguire

W. Todd Rogers

university of alberta

Experimental, quasi-experimental, and survey research in education is often carried out using accessible subjects. Such circumstances violate the assumptions of most traditional statistical inference procedures. This paper reviews and illustrates two computer-intensive procedures appropriate in such circumstances: randomization tests clarify treatment effects, and the bootstrap technique estimates the uncertainty of generalizations to populations. These techniques provide guidance in nonrandomized studies by approximating to a best-case scenario of statistical uncertainty.

En éducation, des recherches expérimentales et quasi-expérimentales ainsi que des sondages sont souvent effectués avec les sujets qu'on a sous la main; il s'agit là de conditions contraires aux postulats de la plupart des méthodes d'inférence statistique traditionnelles. Dans cet article, les auteurs analysent et illustrent deux méthodes faisant largement appel à l'informatique et qui conviennent dans de telles circonstances: les tests de randomisation, qui clarifient les effets de traitement, et la technique *bootstrap*, qui permet d'évaluer le caractère incertain des généralisations appliquées à des populations. Ces techniques sont utiles dans les études non aléatoires dans la mesure où elles fournissent aux chercheurs le meilleur scénario possible d'incertitude statistique.

Statistical analysis in educational research has a long history of technical success. There is a sense of timeliness in the concepts taught and learned in the compulsory classical statistics course that lies at the centre of most graduate programs in education. However, in the light of new directions in statistics, and stimulated by the work of Tukey (1958, 1977), we should like to reexamine the role statistics play in research. Often overshadowed by their more complex inferential counterparts in traditional applications, enhanced descriptive statistics, including graphical analysis, are now increasingly used for exploration and deep portrayal. Less well known are recent inferential developments which would liberalize the solution of problems in educational research where *randomness* is absent.

For researchers steeped in the tradition of Campbell and Stanley (1963) and, later, of Bracht and Glass (1968), there are two kinds of validity in experimental and quasi-experimental research. The first is *internal validity*:

“Did the independent variable alone elicit the particular outcome observed?” The second, *external validity*, of concern not only in experimental and quasi-experimental research but also in survey research, deals with the extent to which observed results may be generalized beyond the sample used. Random assignment is fundamental to internal validity; random selection is essential for aspects of external validity that apply to the population to which results are to be generalized.¹ But in educational research we almost never have randomly selected samples of any well defined population, and for many of our experiments we are unable to assign people randomly to treatment conditions.

Perhaps it is this inability to randomize that has contributed to the movement away from survey and experimental and quasi-experimental methods toward more qualitative strategies. A second possible contributing factor is the lack of knowledge about the sampling distribution of various sample estimators (e.g., correlations corrected for attenuation, generalizability coefficients, mean absolute deviations). One early attempt to solve these problems applied the delta theorem (see Rao, 1973, p. 387) to derive asymptotic expressions for the standard error of the sample estimator, subsequently combined with standard statistical testing procedures to construct confidence intervals (see Hakstian, Schroeder, & Rogers, in press, for examples). A second, early procedure employs cross-validation or replication analyses in which statistical results from two or more randomly equivalent sub-samples are compared to determine the degree of stability across replications (Shepard, Gamilli, & Williams, 1984; Stone, 1974). More recently, the availability of inexpensive high-speed computers allows the use of less widely known computer-intensive procedures to answer some of the statistical problems brought about in educational research by lack of randomness or of appropriate statistical inferential tests. Generally speaking these strategies, non-parametric in nature, use computing power in place of theoretical assumptions to develop indicators of uncertainty. We review two such procedures—randomization tests (Edgington, 1969), and the bootstrap (Efron, 1979)—to illustrate their role in educational research.

THE ROLE OF RANDOMIZATION TESTS IN DETECTING TREATMENT EFFECTS

Edgington (1969) noted that in “basic psychological experimentation, the principal interest in statistical hypothesis testing is to distinguish between treatment effects and effects arising from individual differences” (p. 97). As an initial check on our experimental results, we need some way of determining whether observed differences among groups are due to systematic influences or to the individual variability that characterizes much human behavior.

Generally, researchers in education must use nonrandom samples—in this case, samples constituted of subjects who happen to be available. Often, though, it is possible to assign these subjects randomly to two or more conditions of the experimental treatment. In these situations, where random selection of subjects from a population of interest is not possible but random assignment of available subjects to treatment conditions is, the use of parametric inferential statistical tables is theoretically not warranted (Edgington, 1980, p. 4). The condition of random sampling necessary for their use has not been met. Instead, Edgington (1969, 1980) and others (Bradley, 1968; Fisher, 1935; Kempthorne, 1955; Winer, 1971) recommend the use of randomized tests.

Essentially, randomization tests, or permutation tests as they are sometimes called, are strategies for testing whether observed differences in an experiment are simply the consequence of the random assignment of subjects to groups. The rationale advanced is that even with random assignment there is no direct control over all sources of between-subject variability for independent groups experiments or within-subject variability in repeated measures studies. Randomization tests provide a means of distinguishing between effects and noise, thereby shedding light on the validity of the treatment effect.

The basic steps of randomization testing are:

1. Specify the test statistic that describes the effect of interest. This could be a difference between means or medians, a difference between between variances or regression slopes, or even a standardized value like a t or F statistic.
2. List every equally probable assignment of units to treatment. If this list of combinations is very long, develop a computer algorithm that will create a random sample of these assignments. In either case, the configuration that was actually used in the study should be included as one of the assignments. (We shall return in a moment to the situation in which there are too many combinations to list conveniently.)
3. For each assignment, substitute the measurement associated with each unit and obtained in the actual experiment, and compute the test statistic.
4. Determine the proportion of test statistics with absolute values as large as the absolute value of the statistic obtained in the study. This proportion can be used as the probability corresponding to a two-tail test of the null hypothesis.

As an illustration, consider the following example. Suppose that we have six students and we randomly assign three to each of an experimental group and a control group. Following treatment, we collect measurements on the dependent variable. As we examine these results, reported in the second panel of Table 1, we consider the magnitude of the difference in the means

TABLE 1
*Random Assignments:
 Six Subjects Randomly Designed to Two Equal-Size Samples*

1	<i>C</i>	<i>E</i>	2	<i>C</i>	<i>E</i>	3	<i>C</i>	<i>E</i>	4	<i>C</i>	<i>E</i>	5	<i>C</i>	<i>E</i>
	1	4		1	3		1	3		1	2		1	3
	2	5		2	5		2	4		3	5		2	4
	3	6		4	6		5	6		4	6		6	5
	$d = 3.0$			$d = 2.3$			$d = 1.7$			$d = 1.7$			$d = 1.0$	
	$t = 3.67$			$t = 1.87$			$t = 1.12$			$t = 1.12$			$t = .61$	
6	<i>C</i>	<i>E</i>	7	<i>C</i>	<i>E</i>	8	<i>C</i>	<i>E</i>	9	<i>C</i>	<i>E</i>	10	<i>C</i>	<i>E</i>
	1	2		2	1		1	2		1	2		2	1
	3	4		3	5		3	4		4	3		3	4
	5	6		4	6		6	5		5	6		5	6
	$d = 1.0$			$d = 1.0$			$d = .3$			$d = .3$			$d = .3$	
	$t = .61$			$t = .61$			$t = .20$			$t = .20$			$t = .20$	
11	<i>C</i>	<i>E</i>	12	<i>C</i>	<i>E</i>	13	<i>C</i>	<i>E</i>	14	<i>C</i>	<i>E</i>	15	<i>C</i>	<i>E</i>
	1	2		2	1		2	1		1	2		2	1
	4	3		3	4		4	3		5	3		4	3
	6	5		6	5		5	6		6	4		6	5
	$d = -.3$			$d = -.3$			$d = -.3$			$d = -1.0$			$d = -1.0$	
	$t = -.20$			$t = -.20$			$t = -.20$			$t = -.61$			$t = -.61$	
16	<i>C</i>	<i>E</i>	17	<i>C</i>	<i>E</i>	18	<i>C</i>	<i>E</i>	19	<i>C</i>	<i>E</i>	20	<i>C</i>	<i>E</i>
	3	1		3	1		2	1		3	1		4	1
	4	2		4	2		5	3		5	2		5	2
	5	6		6	5		6	4		6	4		6	3
	$d = -1.0$			$d = -1.7$			$d = -1.7$			$d = -2.3$			$d = -3.0$	
	$t = -.61$			$t = -1.12$			$t = -1.12$			$t = -1.87$			$t = -3.67$	

of the the groups. Is this difference reliable, or is it attributable to chance? How likely is it that we would get these results under the hypothesis of no treatment effect?

To answer these questions, we need to consider all the possible results that could occur by partitioning scores into all possible sets of groups of three. These are shown in panels 1–20, Table 1. We can express the treatment effect as a t -ratio or, alternatively, and more simply, as d , the difference between the means of the scores of the two groups. Adopting the latter approach, we see that the value of d for the observed sample, 2.3, is not an uncommon value based on the model of random allocation. The probability of obtaining a

$|d| \leq 2.3$ is .20, a value insufficiently compelling to reject the hypothesis of "no treatment effect." If, on the other hand, the probability of the value of the test statistic for the obtained sample allocation were small (for example, .05 or less) in the empirical distribution of randomized test statistics, then we would reject the null hypothesis in favour of the hypothesis of systematic influence.

Randomization tests can be applied in principle to a broad range of situations including unbalanced designs, single subject designs, correlational studies and multivariate designs (see Edgington, 1980). In practice, though, a problem arises in enumerating all possible sample allocations. In most experimental situations encountered in educational research, the development of the empirical sampling distribution of the test statistic would be a formidable task. For example, with 50 subjects distributed to four groups of 10, 12, 13, 15, the total number of combinations is 2.15×10^{27} . Fortunately, a computer-generated empirical sampling distribution of the test statistic can be produced using a random sample of the possible partitions. Edgington (1980, p. 43) recommends a minimum of 1000 replications, including the obtained samples.

Edgington has pointed out that the probabilities calculated using classical hypothesis testing procedures are entirely comparable to those obtained using randomization tests. He argues that these classical procedures yield interpretable results even when the data are not from random samples *selected* from some defined population. These interpretations, according to Edgington (1985), depend upon the fact that the samples were randomly *assigned* to groups at the beginning of the study.

To illustrate the observation that, in situations where conventional tests exist, they give good approximations to randomization tests, consider the data displayed in Table 2. Two classes of 28 students each were observed for 20 hours, and the number of responses made by students to teacher questions was recorded for each student (Macdonald, 1972). Although they were not randomly assigned to classes, assume for the moment that the 56 students had been. (Later we shall argue that the procedure possesses validity without the random assignment assumption.) The t -value associated with the difference between the two means is 9.24. Referring to Student's t -distribution, the probability that $|t| < 9.24$ is less than .001. In the empirical distribution formed from 1000 randomization t s, the probability that $|t| < 9.24$ is less than .001.²

Table 3 shows the further comparability of the two solutions. The distribution of tabled t -values closely approximates the distribution of randomized t -values at the percentile values typically used in inferential testing. It should be noted that the variability and skewness of the two sets of class scores seem quite different; with more similar raw score distributions, the fit between the distributions of tabled and randomized t s will be even closer than that shown in Table 3.

TABLE 2
*Stem and Leaf Display of Student
Responses to Teacher Questions in
Two Classes*

<i>Class 1</i>	<i>Class 2</i>
87542 : 0 :	
97777654421 : 1 :	
99777665111 : 2 : 5	
0 : 3 : 8	
: 4 : 279	
: 5 : 346	
: 6 : 04478	
: 7 : 1488	
: 8 : 489	
: 9 : 0256	
: 10 : 2	
: 11 :	
: 12 : 8	
: 13 :	
: 14 :	
: 15 : 8	
: 16 :	
: 17 : 4	

TABLE 3
Randomized and Tabled t-Values at Selected Percentile Points

<i>Percentile</i>	<i>Randomization t-value</i>	<i>Tabled t-value</i>
.005	-2.76	-2.68
.010	-2.47	-2.40
.025	-1.99	-2.01
.050	-1.58	-1.68
.100	-1.27	-1.30
.900	1.25	1.30
.950	1.63	1.68
.975	1.97	2.01
.990	2.49	2.40
.995	2.91	2.68

Given the similarity of solutions, the question naturally arises "Why not just use conventional procedures with their already available tables?" Traditional procedures tend to focus on those properties of a sample that can easily be manipulated analytically, such as the mean, variance, and correlation coefficient. But, as Tukey illustrated (1977), many other properties of a sample are of interest but beyond the reach of exact mathematical analysis. The randomization procedure provides a solution in these cases.

For example, Maguire (1986) used the randomization test to assess the difference in variability between the two data sets shown in Table 2. Because of the difference in the shape of the two distributions, the interquartile range was selected as the index of variability. The question of interest was, given that the interquartile ranges were 13 and 36 for the two classes, how plausible is the difference of 23 between these two values under the model of random allocation to groups? Alternatively, is there evidence that the observed difference is systematic?

To answer these questions, a randomization test must be used; classical testing procedures for difference between interquartile ranges are not available. Shown in Table 4 are some percentile values of the frequency distribution for the absolute value of differences between ranges for 500 random partitions of the data.³ If we consider the absolute values of differences, then 5% of them are greater than or equal to 28. This finding suggests that the observed difference of 23 is likely attributable to non-systematic influences.

When it is not possible to assign subjects randomly to groups as in the classroom interaction study, what is to be done? We contend that classical procedures and randomization tests can aid in making an interpretation in these circumstances. Would these results be expected if only a random process were operating? If so, then the hypothesis of no systematic influences is tenable. Otherwise, this null hypothesis is rejected in favour of one in which one or more systematic influences (the treatment, or one or more unwanted nuisance variables, or some combination) reliably account for the observed differences.

We justify this contention by construing the situation in terms of competing models. On the one hand, the observed partition could have been influenced by treatments (and other systematic influences) while, on the other, it may be the consequence of random (non-systematic) processes. To be valid, the statistic used to represent the treatment effect must have two characteristics: it must be sensitive to treatment influences, and monotonically related to the strength of these effects. These two characteristics ensure that strong effects align themselves with extreme values of the statistic. We can then separate values that are consistent with the random partition model from those that are relatively improbable. Because of these characteristics,

TABLE 4
*Percentile Values of the Sampling Distribution of
Absolute Differences Between Interquartile Ranges*

<i>Percentile</i>	<i>Magnitude of Difference</i>
.750	17.5
.900	23.5
.950	28.0
.990	32.5
.995	35.0

values of the test statistic that are improbable under the random model lie in the direction of consistency with the model of systematic influence.

In our second example, the probability of getting a difference between interquartile ranges of 23 or larger is about .11. Thus, if a random process is operating, then we can expect to obtain differences of this magnitude about 11% of the time. It would be nice to know what the equivalent probability is under the model of systematic influence, but this is not possible. All we know is that the more extreme the value, the greater the likelihood of a systematic effect. If the random model is rendered implausible, then our degree of belief in the systematic model is enhanced.

We cannot, however, make claims about the treatment *per se* based only on statistical analysis. Our analysis relates to the plausibility of two competing models of influence. The statistical procedures are silent on the identification of a specific cause. They are devices for detecting effect uncertainty, that is, for separating apparently systematic influences from those that behave like random processes. Identification of the *source* of reliable effects is a function of correct specification of dependent and concomitant variables together with careful design.

THE ROLE OF THE BOOTSTRAP IN ESTABLISHING GENERALIZABILITY

As previously noted, in educational research we can seldom select random samples from well defined populations. Instead, we must work with what may be called samples of convenience. But we know that such samples, like random samples, represent something beyond themselves. Paraphrasing Cornfield and Tukey (1956, p. 913), we cannot point to a specific population from which these people were selected, yet the final conclusion is certainly not just applicable to them. It is reasonable to introduce or otherwise create a population "like" the sample observed, and then draw inferences about this population from our observations.

The bootstrap (Efron, 1978, 1982) and its predecessor the jackknife (Tukey, 1958) are two procedures for creating a population from a sample of observations and subsequently drawing an inference about that population from these observations. Both procedures answer the question "What can we do to go beyond our data when we do not have a random sample?"

Although historically the jackknife came first, the bootstrap more clearly exposes the central idea underlying these procedures. We discuss the bootstrap; more detailed statistical descriptions of both the bootstrap and the jackknife can be found in Efron (1982). The rationale for the bootstrap procedure begins with conceptualization of a population consisting of a very large number of replications of the observed sample values. This hypothetical population can be thought of as the population which the observed sample perfectly represents. Bootstrap samples are then drawn from the population and the statistic of interest is calculated from the bootstrap sample. Confidence intervals can then be constructed from the distribution of statistics calculated for the bootstrap samples. Instead of actually producing the hypothetical population, bootstrap samples can be created by sampling with replacement in the following procedure:

1. Specify the statistic of interest. This could be a mean, interquartile range, correlation, generalizability coefficient, t or F test statistic, or any other index that describes the effect.
2. Using the sample data (n people), calculate the statistic.
3. Select a bootstrap sample by randomly drawing n elements *with replacement* from the set of sample values.
4. Calculate an estimate of the statistic from the bootstrap sample using the same formula or procedure as in step 2.
5. Repeat steps 3 and 4 many times (e.g., 1000). Order the resulting values.
6. Construct confidence intervals (e.g., 95%) by counting in 2.5% from each end.

To illustrate the procedure, consider the data for Class 2 shown in Table 2. The interquartile range for this class is 36 ($Q1=55$, $Q3=91$). To build a 95% confidence interval for the interquartile range, we begin by producing a bootstrap sample. Since there are 28 people in Class 2, we drew a random sample of 28 values from Class 2, making sure to replace each value before the next one was selected. (Thus it would be possible, although very unlikely, to get 28 values of 174.) We then calculated the interquartile range on the first bootstrap sample. When this was done, it was found to be 31.5. We repeated this process 500 times, and rank ordered the 500 values. Selected percentile values for this sample of 500 values are shown in Table 5. To obtain a 95% confidence interval, we counted in 2.5% from either end. That is, we looked for the 12th and 13th smallest and largest values. By averaging the 12th and 13th values at the lower end, we obtained a value of 23, and at the upper end,

TABLE 5
*Percentile Values of Bootstrap Distribution
of Interquartile Ranges for Class 2*

<i>Percentile</i>	<i>Value of interquartile range</i>
.005	18.0
.01	20.0
.025	23.0
.05	24.5
.10	27.0
.50	35.5
.90	47.0
.95	52.0
.975	60.0
.99	70.0
.995	74.0

60. Since 23 falls at the 2.5 percentile, and 60 falls at the 97.5 percentile, then 95% of the samples have interquartile ranges between 23 and 60. Recalling that the observed interquartile range is 36, the broad confidence interval suggests that the observed value is very unstable indeed. The stem-and-leaf display reveals the reason for this. In Class 2 there are several values that are relatively unusual or extreme (128, 158, 174). The presence of these values leads to several bootstrap samples with high interquartile ranges. A population which is exactly manifested in Class 2 has a great deal of sampling variability, hence uncertainty, associated with it.

How valuable is this interval? The answer is in two parts. Conceptually the jackknife solves one problem of external validity in situations where we do not have random samples. The solution is to create a population (a) about which we know all of the important details, and (b) of which our obtained sample is perfectly representative. The 95% bootstrap confidence interval allows us to say that we are at least this uncertain about replications to other samples randomly selected from this population.

Statistically, how do intervals constructed using classical procedures compare to bootstrap intervals? Would we get nearly the same confidence intervals? The answer appears to be yes, particularly in the case of large well-behaved samples. Returning to Class 2 (Table 2) and considering the mean, the 95% confidence interval, constructed using classical procedures ($\bar{X} \pm t(27; p \leq .025)$) is [65.1, 90.9]. This interval agrees closely with the interval constructed using the bootstrap method, [65.1, 89.8].

When, then, is the bootstrap to be preferred over conventional procedures? There appear to be two sets of circumstances: (1) when the obtained sample is not large and/or the distribution of scores in the sample contains extreme outliers (see Bradley, 1968, for additional comments) and (2) when the sampling distribution of the statistic is not well known—for example, correlation coefficients corrected for attenuation, generalizability coefficients, robust regression, estimators, factor loadings, interquartile ranges, and scaling coefficients.

A word of caution: the meaningfulness of bootstrap solutions depends upon the characteristics of the sample obtained. This sample must represent interesting and meaningful variation among the sample units—students, classes, schools. The basis of choice and a full bio-demographic description of the obtained sample must be made public so that readers can assess whether to apply the results to their own particular situations.

CONCLUDING REMARKS

Computer-intensive statistical developments have had a liberalizing effect on the application of statistics. The comparability of solutions produced using classical procedures, where they exist, with solutions produced by randomization tests and by the bootstrap and the jackknife, reveals further the robustness of classical procedures. Where classical procedures do not exist or cannot appropriately be applied, we now have available procedures for testing hypotheses and constructing confidence intervals. But we must be careful to note that despite these developments, it is still essential to design our studies carefully and describe fully how we obtain our samples in order to identify the actual sources of systematic variation detected through statistical inference.

NOTES

- ¹ Bracht and Glass (1968) distinguished two types of external validity. Population validity addresses the question "When a study has been completed and a relation found, to what populations can it be generalized?" Ecological validity concerns issues like "If the social setting in which the study is conducted is changed, will the relation still hold?" and "Are the variables representative of the behaviour of interest?"
- ² Programs for this and the other examples presented in this paper were written in APL and are available from the first author on request.
- ³ Part of this example is from Maguire (1986).

REFERENCES

- Bracht, G. H., & Glass, G. V. (1968). The external validity of experiments. *American Educational Research Journal*, 5, 437-474.
- Bradley, J. V. (1968). *Distribution-free statistical tests*. Englewood Cliffs, NJ: Prentice-Hall.
- Campbell, D. T., & Stanley, J. C. (1963). Experimental and quasi-experimental designs for research on teaching. In N. L. Gage (Ed.), *Handbook of research on teaching* (pp. 171-246). Chicago: Rand McNally.
- Cornfield, J., & Tukey, J. (1956). Average values of mean squares in factorials. *Annals of Mathematical Statistics*, 27, 907-947.
- Edgington, E. S. (1969). *Statistical inference: The distribution-free approach*. New York: McGraw-Hill.
- Edgington, E. S. (1980). *Randomization tests*. New York: Marcel Dekker.
- Edgington, E. S. (1985). Random assignment and experimental research. *Educational Administration Quarterly*, 21, 235-246.
- Effron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, 7, 1-26.
- Effron, B. (1982). *The jackknife, the bootstrap, and other resampling plans*. Philadelphia, PA: Society for Industrial and Applied Mathematics.
- Fisher, R. A. (1935). *The design of experiments*. Edinburgh: Oliver and Boyd.
- Hakstian, A. R., Schroeder, M. L., & Rogers, W. T. (in press). Inferential procedures for correlation coefficients corrected for attenuation. *Psychometrika*.
- Kemphorne, O. (1955). The randomization theory of experimental inference. *Journal of American Statistical Association*, 50, 946-967.
- Macdonald, E. (1972). *A comparison group and dyadic interaction as measured by the Flanders' interaction analysis categories*. Unpublished doctoral dissertation, University of Alberta, Edmonton.
- Maguire, T. O. (1986). Applications of new directions in statistics to educational research. *Alberta Journal of Educational Research*, 32, 154-171.
- Rao, C. R. (1973). *Linear statistical inference and its applications*. New York: Wiley.
- Shepard, L., Gamilli, G., & Williams, D. M. (1984). Accounting for statistical artifacts in item bias research. *Journal of Educational Statistics*, 9, 93-128.
- Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society, Series B*, 36, 111-147.
- Tukey, J. W. (1958). Bias and confidence in not quite large samples. *Annals of Mathematical Statistics*, 29, 614.
- Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.
- Winer, B. J. (1971). *Statistical principles in experimental design*. New York: McGraw-Hill.

Thomas O. Maguire and W. Todd Rogers are in the Department of Educational Psychology, Faculty of Education, University of Alberta, Edmonton, Alberta, T6G 2G5.