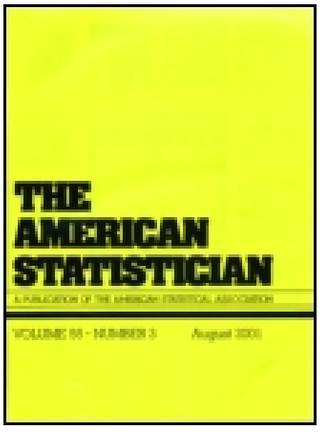

The American Statistician
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/utas20

The Design of Replicated Studies

R. Murray Lindsay [a] & A. S. C. Ehrenberg [b c]
[a] University of Saskatchewan , Saskatoon , Saskatchewan , Canada
[b] South Bank Business School , London , SE1 , England
[c] Stern School, New York University , New York , NY , 10003 , USA
Published online: 27 Feb 2012.

To cite this article: R. Murray Lindsay & A. S. C. Ehrenberg (1993) The Design of Replicated Studies, The American Statistician, 47:3, 217-228

To link to this article: http://dx.doi.org/10.1080/00031305.1993.10475983

COMMENTARIES

Commentaries are informative essays dealing with viewpoints of statistical practice, statistical education, and other topics considered to be of general interest to the broad readership of *The American Statistician*. Commentaries are similar in spirit to Letters to the Editor, but they involve longer discussions of background, issues, and perspectives. All commentaries will be refereed for their merit and compatibility with these criteria.

The Design of Replicated Studies

R. MURRAY LINDSAY and A. S. C. EHRENBERG*

Replication is little discussed in the statistical literature nor practiced widely by statistically minded researchers. It is needed not merely to validate one's findings, but more importantly, to establish the increasing range of radically different conditions under which the findings hold, and the predictable exceptions. This article describes how to design highly differentiated replications. The irrelevance and/or impossibility of identical replications are also discussed. Practical illustrations of the success and failure of replicated studies are given.

KEY WORDS: Close replications; Differential replications; One-off studies; Varying more than one condition.

"The glorious endeavour that we know today as science has grown out of the murk of sorcery, religious ritual, and cooking. But while witches, priests, and chefs were developing taller and taller hats, scientists worked out a method for determining the validity of their results: they learned to ask *Are they reproducible*?" Scherr (1983)

1. INTRODUCTION

Methodological authorities generally regard replication, or what is also referred to as "repeating a study," to be a crucial aspect of the scientific method. The right kind of repetition means that a previous result will have its scope extended. It leads to generalizable results, rather than merely to isolated and uncertain findings.

In the physical sciences important findings get repeated hundreds of times, first deliberately and then as a built-in part of subsequent work. But replication in the social sciences is rare, including in our own area of management studies (Abdolmohammadi, Menon, Olever, and Umapathy 1985; Burgstahler and Sundem 1989; Campbell 1969; Hubbard and Armstrong 1989). Being immature, the social sciences have little in the form of bodies of reliable knowledge or pedigrees of reliable methods from which to establish new findings. One might therefore think that replication would be treated as critical. Yet this is not the case (Chase 1970).

We believe three main reasons underlie the lack of replication:

1. *Successful* replication in the social sciences is widely thought to be unlikely or at least difficult ("Is social science a science?").
2. Replication is widely seen as very mundane.
3. Perhaps to more surprise, the precepts of modern statistics largely impede or even thwart replication: statistical methods mostly focus on how to analyze a *single* set of data ("Is the result significant?"), rather than how to handle and interpret *many* sets of data ("Does the result generalize?").

Whether generalizable results (that is, routinely successful replications) are possible in social science is in the end a matter of fact. We believe that there are enough well-established cases for the verdict to be positive. (One example is outlined in Section 4.2; see also Ehrenberg and Bound 1993.) But finding generalizable results depends on actively looking for them, which is the topic of this article. The other two reasons are examined in Section 2.

In Section 3 we present a theory of replication. This focuses on the suitable design of studies so that they will allow generalizations to emerge (if they in effect exist). A first replication crucially tells us whether or not the result is potentially generalizable at all. We also argue that there can be no such thing as *identical* replications ("mere repetitions"), although this is a popular supposition or ideal. Instead, there are bound to be differences in the conditions of the different studies and this is of the essence: a successful replication tells us that the result held again *despite these differences*. We examine in some detail the ways in which to vary the conditions of observation under which one repeats a study, the main aim being either to extend the scope of the result or to establish its limitations. Some replications are "close" (where we fully expect the result to recur). More telling are replications that are highly "differentiated," where one does not know beforehand (or can hardly even guess) whether the same result will recur or not.

*R. Murray Lindsay is Associate Professor of Accounting, University of Saskatchewan, Saskatoon, Saskatchewan, Canada. A. S. C. Ehrenberg is Research Professor of Marketing, South Bank Business School, London SE1, England, and Visiting Professor, Stern School, New York University, New York, NY 10003.

In Section 4 we briefly review two examples of empirical work from the management literature (our own applied specialization) to illustrate results that either do or do not repeat, and why. One illustration comes from accounting and organizational behavior, the other from marketing and economics. Section 5 concludes with some general comments.

2. WHY REPLICATION IS RARELY PRACTICED

In this section we note how both the lack of rewards to the researcher and the traditional orientation of statistical teaching run counter to the basic practice of replication.

2.1 Lack of Rewards for Replication

The academic environment often contributes to the lack of replications. Repetition is considered to be an "inferior" form of research (Umapathy 1987, p. 170) and of "low prestige" (Campbell 1986, p. 122). As Collins (1985, p. 19) noted, science reserves its highest honors for the originators. We accept that this has to be so. But without corroboration, the original result would not be meritorious (for example, the "discovery" of cold fusion). Nor can all of us win prizes all the time. Much scientific work is of the "normal science" variety (Kuhn 1970). It largely is (or should be) a professional activity concerned with developing "objective" (meaning "widely agreed upon") knowledge and understanding and also practical applications of this, rather than with the researcher's self-promotion and possibly tenured appointment (see Burgstahler 1987).

Because of the overemphasis on appearing "original," replications are thought to be more difficult to publish and seem seldom to be rewarded in terms of career recognition (Denzin 1970, Hubbard and Armstrong 1989; Kerr, Tolliver, and Petree 1977; Mahoney 1985; Rowney and Zenisek 1980; Smith 1970; Walster and Cleary 1970). This emphasis on "original research" has even become embodied in the doctoral dissertation (Mack 1951). In an audit of 265 accounting faculty with a Ph.D., only 2% described their work as having been a "replication" (Abdolmohammadi et al. 1985). Such reactions to replication are, however, wide off the mark.

First and foremost, repetitions need not and should not be *mere* repetitions. As we develop further in Section 3, they can be designed to extend the scope of the previous results, so as to lead to more powerful empirical generalizations (sometimes radically so). Such extensions of scope can be very demanding. Done well, they provide opportunities for doing something new, and for excitement.

Second, a successful repetition indicates competence —it can be a real achievement. Indeed, replication is a risky business, especially early on. Failure to obtain the same result means either that the previous investigator was wrong, or that the new one is incompetent, or that the topic is more complex than either investigator had thought. And answering questions of "Why?" in the case of failure—or even in the case of success—provides plenty of scope for originality. Publishing replications also reduces the chance of accepting false hypotheses ("Type I errors").

Finally, successful replication provides the basis for further and deeper explanatory studies and theory. It is hardly worth asking why something occurs, or how to apply it in practice, if we are not sure whether it can be observed at all, let alone routinely.

2.2 Statistical Precepts

Statistical methods and papers are dominated by the "cult of the isolated study" (Nelder 1986). They are not geared to deal with, let alone emphasize, the problems of combining information from many sets of data (MSoD) (see also Ehrenberg 1975, 1982, 1990; Ehrenberg and Bound 1993). They therefore pay little or no attention to the generalization process that lies at the heart of the scientific method. As a result of focusing on single sets of data (SSoD), statistically minded researchers have unfortunately come to be obsessed with their *methods* rather than with their *results*. Three examples of this focus on analyzing single sets of data are tests of significance, regression analysis, and correlation coefficients, as we now outline.

Statistical Significance. Running a test of significance is the single most common method for inference in the social sciences (Johnstone 1986; see also Acree 1978; Gigerenzer 1987; Gigerenzer et al. 1989; Guttman 1985; Oakes 1986; Walster and Cleary 1970). Obtaining a statistically significant result is often considered to eliminate the need for repeating the study (Carver 1978; Finifter 1972, pp. 156–166; Guttman 1985; Lykken 1968). Along with a bias against publishing *negative* findings (see Lindsay 1990), this has led to statistical significance being regarded as the ultimate objective of a study (Yates 1951), rather than more properly as just a first step.

A test of significance tells us no more than that the observed result is in fact probably real—as if the whole population in question had been measured—and that it is unlikely to have been the result of a sampling error. But it remains an isolated result for that particular population. Significance cannot and does not tell us whether the same result would hold again in a different population or under different conditions. To establish that would require much explicit replication.

Tests of significance are not taken very seriously in areas where replicable results have already been obtained. An example is the relationship between smoking and lung cancer. As Guttman (1985) notes, this has been studied repeatedly. Each study dutifully reported its significance level to show that the apparent correlation in that sample was actually there. But researchers continued to replicate and extend the work, to establish how far the correlation was a *general* phenomenon, to eliminate or pin down the influence of other factors (for example, pollution, heredity, and the amounts or recency of smoking), and also to delve into the direction and nature of any causation. Given 10, 20, or 30 studies all showing correlations of smoking and lung cancer,

the p value of any single study quickly paled into insignificance. Similarly, the current debate about the possible effects of *passive* smoking does not turn on the statistical significance in any single set of data.

Statistical significance can matter in a first study (at least if its design is so humdrum as to contain no internal replications). But after that, questions of significance arise only as and when subsequent studies show *discrepancies* from the earlier relationships or subpatterns ("Are they real"?).

Regression Analysis. Regression analysis is a second example where traditional (modern) statistical teaching is not aimed at establishing generalizable results. In ordinary least squares regression, a new equation is estimated for *every* new set of data. Little or no attention is paid to whether that equation might also fit any other data and hence possibly be generalizable, an outcome that is in fact technically highly unlikely if not impossible for a "best fit" equation to bivariate data (see Ehrenberg 1975, p. 243). Nor is attention traditionally paid to what equations have already been fitted to *previous* data.

It is therefore no wonder that quantitative social and management science based on current statistical methods has in fact come up with few lasting, that is, generalizable, findings (Ehrenberg and Bound 1991). We aim to discuss the technicalities more fully elsewhere (Ehrenberg and Lindsay 1993).

The Correlation Coefficient. A third example is the correlation coefficient, when used on its own to reflect a relationship.

Reporting a positive value of $r = .6$, say, shows that x and y are in some way positively related (with a "significant" r merely meaning that the population value r is probably not zero). But reporting $r = .6$ does not tell *how* the variables are related (for example, by how much y varies with x). If another study again reports $r = .6$, the actual relationship and residual scatter there could either be the same as before, or different (that is, have the same slope and intercept as before, or not).

If the second study yields $r = .8$ rather than .6, the actual relationship could still either be the same as before, or different (just the *relative* degree of scatter might differ). There is no way of telling from the reported correlations in two or more studies, .6 or .8 or whatever, whether the quantitative relationships agree and hence generalize, or not (Ehrenberg 1975, p. 237; 1982, p. 164). It is therefore once more not surprising, on reflection, that merely reporting correlations on their own has produced no quantitative findings of any generalizable and lasting value, as in much psychometrics and sociology, or in the widespread use of factor analysis, which is correlations based on correlations.

These are three major examples that illustrate the apparent gulf between current statistical methods and the idea of replication, that is, focusing on the search for generalizable or routinely predictable results. The traditional methods mostly make no allowance for combining and analyzing different data sets. Unfortunately, these methods are very popular since they are very easy to use (for those who know the basic how-to). They are even considered to provide the figleaf of scientific respectability (see Ravetz 1971). But one is not likely to find a generalizable result (if it should exist) if one does not *actively* look for it.

Apparent Exceptions: Bayesianism and Meta-Analysis. Many scientists, and no doubt some statisticians, do in fact engage in replicated studies or the like. But this is *despite* traditional statistical teaching, rather than because of it.

Nonetheless, two modern statistical movements explicitly relate to the analysis of more than an isolated set of data. One is the Bayesians' attempts to employ prior knowledge probabilistically. This, however, uses the new data only to adjust the "probability" of the result (that is, one's "degree of belief" in it). It does not tell us in what way its scope (that is, its generalizability) has increased with a successful replication, or how it has been circumscribed or negated with an unsuccessful one, and what, in either case, one might therefore want to do next. Not surprisingly for us therefore, the Bayesian endeavor to mate new and old knowledge has in practice been pretty sterile.

The other movement tackles certain aspects of the secondary analysis of earlier data under the grandiose new title "meta-analysis," for what should be the basic scientific activity of comparing different studies (for example, Hedges and Olkin 1985; Wachter and Straf 1990). Meta-analysis, however, still largely concentrates on merely reducing the sampling errors and increasing the statistical significance of the result, rather than on establishing its generalizability under the different conditions of observation.

Neither movement seems to have led to notable empirically grounded generalizations, nor do their protagonists themselves claim to have done so.

3. A THEORY OF REPLICATION: THE DESIGN OF STUDIES

We now outline three basic features in the design of replicated studies. These are the role of a first replication, the impossibility of identical replications, and then, in some detail, the nature and the degree of the differentiation between different replications. We illustrate the arguments with a notional study of the demotivating effects of dominating bosses, based on 46 executives in General Refrigeration Inc (GRI) in 1981.

3.1 The One-Off Study and a First Replication

If the characteristic feature of scientific knowledge is that a result has to be repeatable (Lindsay 1992; Ravetz 1971), it must involve more than one set of data. Hence replication becomes a key consideration in the design of scientific studies. An isolated study remains virtually meaningless and useless in itself (for example, "Student," cited in Pearson 1938; Ehrenberg 1975, p. 370; Guttman 1985; Kempthorne 1978, p. 10; Nelder 1986; Popper 1959, p. 45; Ravetz 1971, pp. 174, 374).

For example, why should anyone care about the finding that executives who have a dominating boss are less motivated in their work, if this has been found to hold only once, for the 46 GRI executives in just the one company and the one year. In and of itself, this is a one-off result. Under what conditions, if any, will it hold again? Did it also hold in GRI in the following year? Has it held in any *other* company? Or not? Without knowing any of this, why should anyone pay attention, or act on that finding?

But if the 1981 GRI result were in fact found to hold again for these same 46 managers a year later (in as far as they were still there in 1982), we would have learned that the result was not a total flash in the pan. It *could* potentially become more widely generalizable. It is perhaps just beginning to be worth noting.

More unambiguously still, if the result does *not* repeat in 1982, the initial 1981 finding has to be virtually disregarded since no generalization is possible, at least not any straightforward one. The initial result probably did not even mean what it seemed to say, that is, that bossy bosses are demotivating, since in the next year that was no longer so. The first replication is therefore the most dramatic. Compared with the initial finding in isolation, the first replication shows whether or not a wider or lawlike generalization is possible.

The contrast between a single study and replicated studies—that is, analyzing a SSoD or MSoD—is akin to Deming's distinction between an *enumerative* study, which describes a sample from one particular population (where a 100% sample would provide the complete answer), and *analytical* studies, which compare many populations (for example, Deming 1950, 1975). Here the supposed ideal of 100% coverage of some undefined "super-population" of all the past, present, and future locations, researchers, and methods of investigation is impossible. Deming therefore concluded that it is neither feasible nor efficient in any analytic study to attempt to sample such a universal kind of "super-population" statistically.

However, Deming has posed little by way of actionable alternatives (as we attempt to do here). At the same time he has been more ambitious—hoping to understand *causes* and to forecast *change* (mainly as improvements in quality). In contrast, our focus here is merely on being able to predict successfully that the given result (whether the effect of bossy bosses, or the square law $d \propto t^2$ for the distance d traveled by falling bodies in time t, say) will hold again in further studies, that is, that the same model holds again within the previous limits of scatter (see Ehrenberg and Bound 1992, sec. 5.1). Understanding the causal mechanism and/or inducing change would come later.

3.2 The Impossibility of Identical Replications

One reason for the low regard for replication is the mistaken view that it merely involves repeating the original study *exactly*. Researchers often do not recognize, or even deliberately ignore, the fact that replication must always involve some variation in the conditions of the study. The notion that replications can in effect be identical repetitions is, however, widespread even in the physical sciences. For example, Harré (1990) quoted Galileo on his famous experiment on the square law for falling bodies as saying "repeated a full hundred times, we always found that the spaces traversed were to each other as the squares of the times." Taken literally as near-identical replications, this would merely have established the small errors of measurement involved (which Galileo had in fact already done previousy). The crux of Galileo's experiment was that he repeated it for *different distances*. But this was treated by Harré in the quoted passage as a mere aside.

The supposed ideal of identical replications—"repeating it a hundred times"—is wrong on several counts. First, no repetition of an earlier study can be totally alike, if only because of the difference in time (and all that can vary with that, for example, the temperature and our knowledge of the previous results). Hence an identical replication is virtually *impossible.*

Second, an identical replication would in any case be *pointless*: if all conditions were literally or effectively identical, the results would have to be so too. There would be no need to do the same study again.

Third, and most important, we need to cash in on such differences in the conditions of observation as do occur (or are *made* to occur), rather than to try to sweep them under the carpet: "The same result was obtained in the two studies *despite* the differences in the conditions" (for example, that d/t^2 = constant for Galileo, for markedly different distances d).

In general, the more explicit, differentiated, and/or deliberate such variations in the conditions of observation are while still obtaining the same result (for example, that bossy bosses appear to be demotivating), the more telling and exciting the outcome. Thus compared with successfully repeating the initial GRI 1981 study in General Refrigeration Inc a year later, the replication is more powerful if it is successfully repeated on managers in *another* company, in another country, and *five* years later. The result is not merely shown to be "repeatable," but to be repeatable under markedly different conditions. And we now would surely like to know how else the second company differed—was it larger or smaller, did it have centralized or decentralized management, and so on.

A Hold-Out Sample. Although deliberate replications in social science studies are quite rare, some investigators do incorporate a "hold-out" sample in their design (or at least in their analyses), especially when fitting a model.

For the 46 executives in the 1981 General Refrigeration Inc study, for example, one might first analyze only 36 of them, and then check the results for the other 10, the hold-out sample. Or in analyzing a time-series of annual data from 1974 to 1989, one might estimate the model for the first 14 years and then check its fit separately for the last two.

This approach generally amounts to a form of *very close* replication—the hold-out sample being usually

very similar to the main sample. Thus 1988 and 1989 were probably not that dissimilar to the preceding 14 years 1974 to 1987 (whereas both 1973 and 1990/91 with their OPEC and Gulf War oil crises, might have been regarded as too different to use for a "hold-out" check, but would in fact for that reason have provided a real test, if included).

Some analysts even use a supposed hold-out sample which was *randomly selected* from the full data in hand (as the 10 GRI executives might have been). But a random sample *must* agree with the data from which it is sampled, other than for random sampling errors. It is not a test of either the substantive results or the fitted model, but only of the process of drawing a random sample.

But even if one uses a *systematic* subsample (for example, all GRI executives with more than 20 years' service), the degree of generalizability that can be achieved with just *one* hold-out sample is not great. At best, using a hold-out sample is only tinkering with the problem of replication.

3.3 Close Versus Differentiated Replications

In practice, repeating a study should mean carrying out a different study and seeing whether the results are the same. But how different should the replication be? It can, we think, be helpful to distinguish two broad forms of replication, "close" and "differentiated." This distinction can apply both to studies as a whole and to specific factors in the conditions of each study, although the distinction will vary over time.

Close replication attempts to keep almost all the known conditions of the study much the same or at least very similar (for example, the population or populations in question, the sampling procedure, the measuring techniques, the background conditions, and the methods of analysis).

A close replication is particularly suitable early in a program of research to establish quickly and relatively easily and cheaply whether a new result can be repeated at all. All that may be explicitly varied is the timing of the study, plus that the investigator is more experienced (and in particular already knows the outcome of the first study!).

There will, however, also be many other conditions of observation that may, or indeed will, vary even between two close replications (Collins 1984). Some of these factors will be known at least implicitly (for example, that one study was done in the summer, the other in the winter), others will not. If the result of the two studies is the same, we will have learned (or confirmed) that the result is robust to all these other variations in the situation. The more explicit the description of these varied factors (even if only with hindsight), the more powerful the result.

If, however, any of these variable factors *do* matter, the study will not replicate successfully. In other words, the results themselves will tell us if the replication was in effect close. The strength of such a disconfirming result—any *failure* to replicate—increases the closer the explicit design of the replicated study is to that of the original one (Collins 1984). We cannot easily blame the failure on some potentially major factor that varied, because with a close replication we had not deliberately included any major differences. The implication is instead that there was something seriously wrong with the underlying concept of the study, or with its planning, or its execution.

As an area becomes better understood (that is, some successful close repetitions have been obtained), the value of further close replications diminishes and we need more differentiated ones. The notion of close replication will, however, still apply to *specific factors* in the design of the study. For example, one will usually keep the measured variables themselves very similar for quite a number of studies, that is, continue to investigate the relationship between executives and their bosses, say, rather than that between pupils and teachers.

Differentiated replication involves deliberate, or at least *known*, variations in fairly major aspects of the conditions of the study. The aim is to extend the range of conditions under which the result (for example, the demotivating effect of bossy bosses) still holds—for example, for large and small firms, in different industries (of specified types), for well-paid and for underpaid executives, and so on. Exploring a result with deliberate variations in the conditions of observation is the essence of generalization.

There are three basic reasons for running differentiated replications. First, there is the notion of "convergent validity" or triangulation, namely that by performing studies that use different methods in the broad sense (for example, different measurement instruments, methods of analysis, experimental setups, and/or investigators), one obtains confidence that the result is due to the conceptual variables under study and not just an artifact either of the persons conducting the study (see Rosenthal 1976) or of the particular manner in which the original study had been conducted. The supposed meaning of the result (that bossy bosses demotivate initially normal executives) must be queried if it lacks resilience, that is, if small variations in the method change the result markedly and unexpectedly (Mulkay and Gilbert 1986). This is a crucial observation: that failure to replicate casts doubts not merely on the existence of the result, but also on whether it actually meant what it was thought to mean.

Second, heavily differentiated replication leads to extensions of the scope of the result and hence its subsequent practical applicability, that is, to other firms, other industries, different types of executives, other years, or whatever.

Third, differentiated replication is a search for *exceptions*, a route towards establishing the conditions under which the generalization in fact does *not* hold (see also Ehrenberg 1966, 1975, 1982; Ehrenberg and Bound 1993).

With close replications one generally expects the same result to recur and therefore is "just checking" and would usually be surprised if it did not hold again. With

differentiated replications one is not so confident. One may even expect the result to break down but wants to establish this empirically, especially if varying that factor is a new departure, or the factor is being varied by a large amount.

The distinction between differentiated and close replications should be regarded as no more than a helpful framework for thinking about the conditions of a study. It is subjective, as are most decisions in designing a study, including whether to do the study at all ("objectivity" comes from whether the *results* are the same). The distinction is in any case not an absolute one—it is relative to what is known at each point in time. If a study shows that what had previously been thought to be a potentially important factor does not in fact affect the result (for example, whether or not the boss is vertically a bit disadvantaged), it then becomes a *minor* factor in this context, that is, part of *close* replications in future.

3.4 Varying More Than One Condition

Many of the conditions that vary between different replications can be controlled to a greater or lesser extent by deliberate selection, that is, what firms, industries, or countries one studies, what ages and types of executives, whether the different studies are carried out at the same time or in different years, and so on. But the traditional experimental precept of varying only one condition at a time is seldom feasible. Nor is it usually desirable. The aim is to see whether the same result occurs *despite* differences in the conditions The more differences that are covered in a successful replication, the greater the generalization that accrues to the result. However, conducting such a study entails a gamble: if different results were to occur, the investigator would not know which of the differing variables or combinations of variables are responsible for the results. One would then have to go back and retrace one's steps (Barlow and Hersen 1984), but this is needed only if the results differ.

In his highly innovative work on factorial designs, Fisher (1935) stressed the need to check for interactions between the main factors in a study, for example, whether the effect of dominating bosses is larger (that is, "different") for younger executives than for older ones. This would mean that there is no single generalizable result. But despite this emphasis on interactions, Fisher and the statistical literature on experimental design since have mainly stressed the role of factorial designs in reducing experimental errors rather than in increasing the degree of generalizability of any findings.

There are broadly two ways of letting such different conditions or factors vary (for example, the firms that are investigated and the age of the executives): either independently of each other, or together ("confounded").

Varying the conditions or factors of an experiment independently of each other (known broadly as "factorial design") means, for example, first choosing differing firms (two or more), and then separately choosing younger and older executives within each. This allows one to assess (1) differences between the firms, (2) differences between the age groups, and (3) variations of the age-group differences for the different firms ("interactions").

The alternative approach—varying two or more conditions *together*—is often inevitable, for example, repeating the 1981 GRI study in 1982 and in a different firm, and with one firm perhaps having older executives than the other. Such "confounding" of different factors or conditions matters little in replication studies aimed at establishing generalizability (unlike studies that seek to assign *causes*), since the aim is negative, to establish that the observed effect is the same despite differences in firms, years, ages, and so forth. The possibility that some of the confounded conditions accidentally had compensating effects is generally covered by subsequent replications.

As the number of conditions and/or the levels of each condition is increased (for example, the number of firms and age groups), the number of cells of the study increases rapidly, along with its potential complexity, especially if many positive interaction effects were to occur. At the same time the sample size in each cell will decrease, for a study of a given size. One way to simplify this situation is to ensure that some of the replications are rather close, that is, ones that should not affect the results unless one's intuitions were all wrong (for example, for readings taken before or after lunch), let alone should they interact with *other* factors. If successful, this means that one will have established that these supposedly close replications were *in fact* very close, and one will have done so at a very low cost and without affecting sample sizes for the rest of the analysis.

Another possible design simplification is not to cover all possible combinations of the conditions ("partial" or "fractional" replication). This can be done at the cost of not being able to check some of the possible interactions between the different factors covered. Checks on interactions may not always be needed with a high degree of priority (especially perhaps in cases where the main effects are not large or even nil, as would be expected for "close" replications).

Examples of partial replications occurred in the pricing study that we discuss in Section 4.2. Prices were *increased* for two product categories (Soup and Tea) and *decreased* for two others (Cereals and Confectionery), but not the other way around (for example, no price increases for Cereals). This was acceptable since the size of the sales effects was the same (as hypothesized beforehand) for the increasing and decreasing prices in those product categories that were covered. Hence there was no need to check directly a price increase for Cereals, for example.

3.5 Condition to Vary

In principle one might want to vary an infinite number of conditions in any program of replication studies. In practice, it is somewhat simpler, in that only two kinds of conditions seem to matter:

1. Conditions or factors that are thought might affect the hypothesis under study. In investigating the relationship between dominating bosses and executives' motivation, the length of time these people have worked together might affect it. So might the "culture" of the firm. These factors should therefore be checked at some stage in a research program. But there seems little point in checking the effect of interest rates, or of earthquakes in China (except possibly if the latter affects any GRI-type trade with Pacific Rim countries).

2. So-called "confounding variables," that is, conditions or factors that are known or thought to be related to just one or other of the separate variables under study (for example, executive motivation). Examples would be the criteria used in selecting the executives, or the firms' reward system. Here one needs to see if these factors have any effect on the relationship with the *other* main variable under examination (in our example, dominating bosses).

The distinction between (1) and (2) can subjectively guide an investigator in deciding what factors to investigate. But they cannot be used to justify the validity of the study. For this, it is the *results* that matter. The touchstone in considering what factors to vary remains "When in doubt, find out." If anyone suggests that a certain factor has been wrongly omitted in a series of studies (that is, that it might affect the results), then in principle it should be checked.

As a finding becomes established, the rate of replication studies in their own right will diminish. Studies aiming to link the result to other findings and to delve into causal mechanisms will take over. In such programmatic research, close replications will, however, be performed continuously: if a finding is worth studying more deeply, one will be obtaining new observations on it as a matter of course. Occasionally there will also still be deliberate extensions to some new set of conditions altogether, as a highly differentiated and perhaps rather original replication.

3.6 Systematic Replication Studies

Many replication studies occur on a fairly ad hoc basis, as the various researchers' interests and opportunities might dictate, and without any overarching master plan.

At times a more systematic approach is, however, needed, with a range of differing replications being planned together. Only in this way can one detect any relatively minor but consistent deviations or subpatterns in the results, that is, where the previous model was consistently biased, even if perhaps only marginally so.

One current example is a simultaneous and hence systematic study of as many as 3,600 distinct sets of data worldwide on children's average heights $\overline{h}$ and weights $\overline{w}$ at different ages (Bound and Ehrenberg 1993). This showed that the relationship $\overline{h} = 55 \ln \overline{w} - 53 \pm 2$ which had been established in some previous more or less isolated studies of children aged 2 to 18 in half a dozen countries (other than for girls postpuberty) still largely held. But it had to be marginally adjusted to $\overline{h} = 53 \ln \overline{w} - 46 \pm 2$ for Indo-European boys and to $\overline{h} = 51 \ln \overline{w} \pm 2$ *for other boys* and also for all girls up to age 13. Such a need to "fine-tune" the previous relationship could not have been determined without a systematic set of new replications.

Conversely, another systematic replication study of a previously well-established generalization concerning a "Double Jeopardy" type of brand-loyalty pattern across 34 different U.S. product categories (Uncles, Hammond, Ehrenberg, and Davis 1993) has shown that overall there was no systematic bias in the previous model. But it also set clearer limits on the extent of certain irregular subpatterns, deviations, and outliers (that is, the influence of "excluded variables").

4. EXAMPLES FROM THE MANAGEMENT LITERATURE

In this section the theory of replication previously developed is illustrated with two examples taken from the management literature. The purpose of replication is to see whether or not a degree of generalization can be achieved. The first example illustrates an apparent failure for the result in question to generalize, and why this occurred. The other example illustrates an apparent success.

4.1 A Failure to Obtain Generalization: Accounting Performance Measures

The first example is taken from the management accounting literature and consists of a series of related studies that involved high degrees of differentiated replication. The example illustrates the need for much more basic, *closer*, replications in the initial stages of an enquiry because the more differentiated ones mostly did not, in fact, replicate successfully. (There are also other such instances of this kind in the managerial literature, for example, see Dall'Olmo Riley 1992.)

The studies in question examined the effects of superior managers' reliance on accounting performance measures (APM) to evaluate subordinate performance. They did so by, put very broadly, contrasting high and low reliance on APM styles. This is one of the relatively few areas in management accounting research where there has been any sequence of "repeated" studies (see Brownell 1987 for an overview).

The enquiry began with Hopwood's (1972) pioneering work. His results indicated that a certain type of high reliance on APM was observed to lead to *negative outcomes*, such as subordinates feeling more job-related tension, having poorer relations with both their peers and their superior, and tending to misreport or to take undesirable actions that merely improved the short-term accounting measures.

Hopwood's study was then replicated by Otley (1978). Otley chose a company to which he had access and whose environmental characteristics would, he also thought, be more appropriate to the use of a high reliance style. He also extended his study to examine the

impact of the effects of evaluation style on managerial performance. Otley's results differed markedly from Hopwood's: He found little effect of evaluation style on job-related tension or on the manipulation of accounting information. His evidence in fact indicated that a high reliance style was associated with better subordinate performance.

Brownell (1982) then attempted to reconcile the two previous studies by putting more weight on the moderating variable of *budgetary participation*. He posited that high reliance on APM would be associated with higher job performance when budgeting participation is high, and less with low participation. His results supported this hypothesis.

In parallel, Hirst (1981) also attempted to reconcile the Hopwood and Otley studies. He conceptualized *task uncertainty* to be a moderating variable underlying the different results of the two studies. He argued that a high reliance style is inappropriate when task uncertainty is high (as apparently in Hopwood's sample), but appropriate when task uncertainty is low (as probably in Otley's sample). Hirst (1983) empirically tested his hypothesis and found it held for job-related tension. But relations with superiors were positively related with reliance on APM only in low task-uncertainty situations.

Following these studies, Govindarajan (1984) proposed that *environmental* uncertainty (a different conceptual variable to *task* uncertainty) would moderate the relationship between reliance on APM and business unit effectiveness. His measures of evaluation style and of business unit effectiveness, however, differed from those in the preceding studies (which themselves differed in certain potentially important operational ways). He reported a significant, positive correlation between environmental uncertainty and the low reliance style in the highest performing units, and no relationship in the lowest performing units.

Brownell (1985) undertook a similar study to Govindarajan but used a different formulation yet again. He hypothesized that management *function* (specifically, marketing versus research and development) would moderate the relationship between evaluation style and job performance because of the differing environment facing a unit manager. However, his results did not support this hypothesis. He therefore dropped managerial function as a criterion and introduced a new variable, *complexity* (the number of critical elements facing a unit manager). Brownell observed that managers facing more complex environments performed better with reduced reliance on APM.

Finally, in an Australian study designed to reconcile and integrate the previous studies, Brownell and Hirst (1986) failed to replicate the findings of Brownell (1982) and Govindarajan (1984) with respect to job performance, and those of Hirst (1983) with respect to job-related tension. This occurred even though the measurement instruments were the same as in Brownell (1982), though different from Hopwood (1972), and the job-related tension instrument was the same as in Hirst (1983). Otherwise different measures were used.

The authors could not fashion convincing arguments to explain the failure to replicate the previous findings, other than perhaps yet even higher order interactions or possibly cultural differences between Australian and American managers. Obtaining a generalizable result was therefore becoming at best more complex. It seems that a simplifying breakthrough would still need to be firmly established.

Taken as a whole, this body of research, although typically "interesting" in seeking to explain discrepancies, does not add up to a coherent body of knowledge or understanding. Obtaining any simple generalizable result seems still to have eluded the researchers, as is the case in budgeting research more generally (Umapathy 1987). Nor yet does the work provide applicable management advice (although this may not really have been expected). In his overview, Brownell (1987, p. 188) admittedly went as far as to conclude that "there is now sufficient evidence to permit the tentative conclusion that a heavy use of accounting information in performance evaluation should be confined to situations characterized by low levels of environmental uncertainty." But the emphasis, it seems to us, has to be on the word "tentative," whatever that may really mean. The results add up to little by way of robust, generalizable findings. Typically, one may therefore doubt whether the concepts and measures used even mean what they say.

The major design flaw in these studies seems to us that they consisted primarily of extensions into new measures and conditions and even constructs before it had been shown that the earlier results were in fact directly and routinely replicable. Close, rather than mostly highly differentiated, replications were needed early on. For example, before a measurement technique is to be varied (and this is ultimately desirable in order to establish the robustness of the results), it is essential to repeat the use of the original technique in at least one or two studies. The analyses of the data would then in turn have needed to focus on any "significant sameness" of results between the close replications, rather than just on statistically significant differences within any one of the various data sets.

Research must ultimately go well beyond replication as such. In this regard, the broad policy in these studies of looking for explanations of discrepancies was not inappropriate and provides plenty of scope for innovation. But by itself it led to increasing complexity and, we feel, ultimate doubt. The reason, we believe, is the lack of solid, generalizable results somewhere at the center. This program of research would, we think, have been more effective if each separate (and "different") study had very deliberately started with a close replication of at least a part of what had gone before. As a start, the question is whether, keeping conditions and measures very similar, Hopwood and Otley's original results could have been obtained again, or not. And not only just once, but routinely.

It may, of course, be that there are no generalizable results to be had in this area, or at least not any rea-

sonably simple ones. But that would also be worth knowing, as background to further academic research, and also in any search for practical applications. But at present even such a negative outcome has not been clearly enough established.

4.2 A Generalizable Result in Pricing

Our second example is chosen to illustrate that generalizable results based on extensive replication can in fact occur. We also note briefly that what was involved in this particular instance was a high use of both close and differentiated replications.

The broad topic is price elasticity. Specifically, it is about the size of the sales response of a competitive brand to a change in its price, in some otherwise fixed market context. The traditional question, "What is the price elasticity of brand X?" implies that different brands and products categories would have different elasticities. But is this so? And if so, how and why? There is remarkably little discussion in the literature, let alone factual evidence.

The study briefly reported here (Ehrenberg and England 1990) set out to investigate this question under just one particular set of experimental conditions, using in the first instance a wide range of otherwise more or less "close" replications. Typically, the price of Kellogg's Corn Flakes (KCF) was dropped experimentally over a period of several months from just over +15% above its normal price to just less than −15%. Three competitive and closely substitutable brands were also available for purchase, with prices that were kept fixed at their normal levels. An increase in KCF's sales share from 34% to 84% was observed, which was both highly significant and in the expected direction. This is, however, merely an isolated fact, obtained under certain experimental conditions, and of no general interest or importance in itself.

The question was how the size of this sales response would compare with the sales responses to a +15% to −15% price change for other brands and product categories. How far would the responses differ, and if so in what ways? Or rather, how far would the relative sales response *not* vary from brand to brand, or category to category?

The +15% to −15% price changes were therefore replicated for *each* of the three other brands of breakfast cereals (some with large and some with small sales share) in three separate split-samples, with the other three brands having prices fixed at their normal levels. This design was also repeated for each of the four brands of a different, and lower priced, product, namely confectionery. There were thus eight close replications, with only the brand or product—the factors under investigation—varying in a more differentiated way. The data were collected by a semilaboratory form of experimentation, with consumers buying the products for real money at a succession of 12 fortnightly sales calls over five months in their homes, and with interspersed consumption in the home.

The results showed virtually the same relative sales responses, in that a model with the same price elasticity of −2.6 fitted the 16 sales shares for the eight brands at +15% or −15% prices, ranging from 0% to 84%, to within an average of about three percentage points, a correlation of the order of .98. (The Kellogg's Corn Flakes result typically implied a directly fitted elasticity of −2.8, close to the average of −2.6.) A robust result had therefore been established by eight close replications. But it was still only of limited generality (that is, price always going down from 15% to −15%; for just eight brands; in only two product categories; and under certain circumscribed experimental conditions).

To extend this result, more differentiated replications on certain factors were also introduced, all as part of one large study, otherwise using *close* replication. In brief, the sales responses to price changes were measured in the same basic way for the following:

1. Price going *up* from −15% to +15%, instead of *down* from +15% to −15%. But this was for two *other* product categories (again one higher priced, tea, and one lower priced, soup).
2. Minimal price changes also of +1p or −1p (for all four products).
3. Price changing either fast (at each successive purchasing opportunity), or more slowly (after two sales calls at a steady price).
4. Some of the +15% to −15% price changes occurring after one (or more) similar price changes (that is, within the same +15% or −15% range) had already been imposed previously, thus giving consumers differing prior experience of such changing prices.

The model with a common price elasticity of −2.6 continued to fit, with "locally" fitted elasticities varying between only −2.5 and −2.9. (For two or three brands with somewhat discrepant results, close replications on new samples of consumers were run a year later, which then gave findings more like −2.6.)

The study also included one set of more sharply differentiated replications for a fifth product category (biscuits). Here each variable-price brand was made to compete with three somewhat less closely substitutable fixed-price brands than was the case in the other four product categories. This still led to similar elasticities for the four different variable-price biscuit brands. But as expected (for less substitutable items), this common elasticity was lower, at about −1.5 rather than −2.6. This part of the study shows that while price elasticities can be the same for different brands and products in otherwise close replications, elasticities *do*, it seems, vary systematically with the choice of the competitive context. This extension is, however, as yet less widely based than the preceding result, being for four brands in one product category only.

The study also illustrates the use of "confounded designs" and "partial replications" for the sake of economy, as noted in Section 3. Thus not all factors were varied independently or in a fully balanced way, as in

Item 1 above for example (see Ehrenberg and England 1990 for further details).

Discussion. This example demonstrates how it is possible to establish a regular and predictable regularity of a possibly surprising kind—here that the relative sales response to a given price change (the "price elasticity") can be the same for different brands (for example, with large and small sales shares, and different identities) and different product categories, and for other varying conditions such as price going up or coming down, doing so fast or more slowly, and consumers having differing previous experiences of such price changes. The key is that the design used close replications except for the specific factors under study (such as a different brand or product category, or price going up rather than down).

If any of these factors, like a brand's previous pricing history, or a brand having a large rather than a small sales share, had markedly affected its sales response to the +15% to −15% price changes, one would not have found a single consistent elasticity of −2.6, but a great variety of elasticities. This would also have been a valid and worthwhile outcome to have established: that sales of different brands and products all respond differently to the same price change, that is, that there is no single generalizable result to be had. But the results turned out to be much simpler (as had been expected).

One limitation of the results is that virtually all the results came from a single if large study (effectively quite a number of simultaneous smaller studies). This strongly suggests the need for at least one rather close replication at another point in time, but perhaps also in another country and involving other products and different observers, to see whether at least the same *qualitative* result does in fact recur, such as price elasticities for different brands and products being the same (and perhaps even the value −2.6). Even just one quite small (or "partial") replication of this kind—close (very similar in design) but independent—would at this stage be very telling. Suppose it did *not* give the same result! At this stage there is almost no such evidence either way.

Despite its apparently quite wide degree of generality (brands, products), the finding summarized here is also quite limited in *other* respects. For example, it dealt always with a situation where only one brand had its price changed, with (just three) competitive brands remaining at their fixed prices. Would the same kind of result occur with more complex pricing changes? Would it also occur in other, perhaps less "experimental," situations? The desirability of knowing whether the result would then repeat is underscored by supposing that it did in fact not do so, even on repeated attempts. The result reported here would then be no more than an isolated and perhaps rather obscure, and ultimately even dubious, finding (*why* did it happen just once?).

A further limitation of the finding reported here is that it was established only with packaged goods and not with durables, services, or industrial goods. With the latter, other factors affecting consumers' response to price—such as incomplete information, risk, search behavior, transaction costs, and probably income effects (for high-price items)—would come into the picture. The result here therefore needs more highly "differentiated" replications or extensions to other such areas. Would the price elasticities for different brands still be somehow "the same"? There is room for highly imaginative designs and measurement techniques in conducting such replication studies.

For direct practical applications, very wide ranges of further replications would also be needed, to map out the different price elasticities that will, it seems, almost certainly occur in different competitive conditions, if generalizable patterns can in fact be found. The results summarized have shown—in just two competitively differing pricing contexts—that results that are generalizable within each context are in fact possible. But how does this vary across yet further pricing contexts, both qualitatively and especially also quantitatively? How do "fixed" price elasticities such as −2.6 and −1.5 vary for these different contexts? Are there any patterns in this?

Conceptually the finding here, however, already seems to have some import: any theorizing about price must now, it seems to us, take into consideration that in the case of directly substitutable brands of grocery products, the price elasticities need not vary between different brands and products.

5. CONCLUSIONS

The precept to replicate—"If a study is worth doing at all, it's worth doing twice"—is easy to apply, perhaps surprisingly so. It is always possible to introduce some deliberate variation in one's study design, for example, readings taken in two different firms rather than just in one, or with both "younger" and "older" executives. The simple but crucial question is whether the replications give the same result. And the more differentiated and explicitly described a successful replication is (for example, Japan versus the West, or a *large* firm versus a *small* one), the more powerful and exciting the conclusion will be, namely that the same result still holds.

Some degree of replication can be introduced without having to repeat a study at all, just by having designed the single study suitably in the first place, that is, by imposing some "internal" differentiation. Instead of treating the study as an undifferentiated whole (as in sampling just one population), it is always possible to differentiate one's observations, for example, analyzing separately those taken in the morning versus those obtained in the afternoon.

If varying such a factor seems trivial, it is an indictment of the investigator, namely of being not knowledgeable or competent enough to have introduced a more telling factor. But even such an apparently trivially close replication as morning versus afternoon is better than none (and can often be done at virtually no extra cost). Faced with just a single set of results in the morning, are we sure that they can be repeated at all, say in the afternoon? And how certain were we be-

forehand that either the respondents or the investigator having lunch, or perhaps being more tired in the afternoon, really does not matter? Ultimately, however, it is the more differentiated replications that will pose the real challenge. They also provide the big payoffs in terms of wider generalizations.

Successful replication is the bedrock of scientific knowledge. It tells us whether we have a result at all. It also tells us the range of conditions under which it is so far known to hold, under which it is therefore becoming routinely predictable, and under which it can be applied to practical problems. In addition, varying the conditions between different replications not only extends the scope of the generalization and determines its limits, but also tells us about some of the factors that do, or do not, affect the result *causally*.

In contrast, an isolated finding from a single undifferentiated study does not tell us whether or not the result is potentially generalizable to any other set of conditions. This remains so however statistically significant the result was, or however sophisticated the analysis. Even a first and close replication can begin to tell us much more, especially a *failure* to replicate.

The need for replication is in fact brought home by supposing that the result for our smaller firm, or for our younger executives, does *not* replicate for our second and larger firm, or for the older executives. What sort of result do we, in fact, then have? Would we not have to ask whether it *always* replicates for smaller firms and/or younger executives, and *never* for larger or older ones? Replication still seems to be the name of the game.

On looking at our journals one sees enormous numbers of articles and results. But to us they seem seldom to add up to anything much. We believe this is because they are mostly more or less isolated studies—testing some original idea of the investigator's—rather than building directly on something that he or she or others have already previously shown to be so.

In publishing empirical results, our journals should, we think, always insist on some degree of replication—at least some *internal* replication. If highly differentiated replications have been carried out, they are worthwhile describing in some detail, at least in terms of describing the conditions and context that have varied greatly. Close replications, on the other hand, where nothing much was changed, can be reported very briefly, if the outcome was successful.

Emphasizing the general need for replication does not presume that replications will always be successful. Many observable phenomena may simply not be generalizable. In that case the results of different replications will differ, and will remain discrepant however much investigators persevere with more and more studies. Such a negative outcome is well worth knowing and publishing: at the very least there is plenty of scope for showing how resourceful the investigators have been in failing to find a more positive outcome!

Increased emphasis on replication will affect how investigators set about their studies. In terms of design, there will have to be more by way of *programs* of research (at least implicit ones), seeking always to build on previous results, and at the same time extending and deepening them—a feature underlying Fisher's philosophy of experimentation (Fisher 1935). The analysis of the resulting data will also change. It will be more concerned with results than with techniques, that is, with comparisons between different studies to establish significant sameness rather than significant differences.

[*Received February 1991. Revised August 1992.*]

REFERENCES

Abdolmohammadi, M. J., Menon, K., Olever, T. W., and Umapathy, S. (1985), "The Role of the Doctoral Dissertation in Accounting Research Careers," *Issues in Accounting Research*, 59–76.

Acree, M. C. (1978), "Theories of Statistical Inference in Psychological Research: A Historic-Critical Study," unpublished Ph.D. dissertation, Clark University.

Barlow, D. H., and Hersen, M. (1984), *Single Case Experimental Designs*, New York: Pergamon Press.

Bound, J. A., and Ehrenberg, A. S. C. (1993), "Model Extension and Model Tuning: A Case Study," Working Paper.

Brownell, P. (1982), "The Role of Accounting Data on Performance Evaluation, Budgetary Participation, and Organizational Effectiveness," *Journal of Accounting Research*, 20, 12–27.

——— (1985), "Budgeting Systems and Control of Functionally Differentiated Organizational Activities," *Journal of Accounting Research*, 23, 502–512.

——— (1987), "The Use of Accounting Information in Management Control," in *Management Planning and Control* (rev. ed.), eds. K. R. Ferris and J. L. Livingstone, Beavercreek, OH: Century VII, pp. 177–196.

Brownell, P., and Hirst, M. (1986), "Reliance on Accounting Information, Budgeting Participation, and Task Uncertainty: Tests of a Three-Way Interaction," *Journal of Accounting Research*, 24, 241–249.

Burgstahler, D. (1987), "Influence From Empirical Research," *The Accounting Review*, LXII, 203–207.

Burgstahler, D., and Sundem, G. L. (1989), "The Evolution of Behavioral Accounting Research in the United States, 1968–1987," *Behavioral Research in Accounting*, 1, 75–108.

Campbell, D. T. (1969), "Reforms in Experiments," *American Psychologist*, 24, 409–429.

——— (1986), "Science's Social System of Validity-Enhancing Collective Belief Change and the Problems of the Social Sciences," in *Metatheory in Social Science: Pluralisms and Subjectivities*, eds. D. W. Fiske and R. A. Shweder, Chicaco: University of Chicago Press, pp. 108–135.

Carver, R. P. (1978), "The Case Against Statistical Significance Testing," *Harvard Educational Review*, 48, 378–399.

Chase, J. M. (1970), "Normative Criteria for Scientific Publications," *American Sociologist*, 5, 262–265.

Collins, H. M. (1984), "Discussion: When Do Scientists Prefer to Vary Their Experiments," *Studies in the History and Philosophy of Science*, 15, 169–174.

——— (1985), *Changing Order: Replication and Induction in Scientific Practice*, London: Sage Publications.

Dall'Olmo Riley, F. (1992), "Consumer Promotions: A Literature Review," London Business School, CMaC Working Paper.

Deming, W. E. (1950), *Some Theory of Sampling*, New York: John Wiley.

——— (1975), "On Probability as a Basis for Action," *The American Statistician*, 29, 146–152.

Denzin, N. K. (1970), *The Research Act*, Chicago: Aldine.

Ehrenberg, A. S. C. (1966), "Laws in Marketing: A Tail-Piece," *Applied Statistics*, 15, 257–267.

Ehrenberg, A. S. C. (1975), *Data Reduction*, New York: John Wiley.
——— (1982), *A Primer in Data Reduction*, New York: John Wiley.
——— (1990), "A Hope for the Future of Statistics: MSoD," *The American Statistician*, 44, 195–196.
Ehrenberg, A. S. C., and England, L. R. (1990), "Generalising a Pricing Effect," *Journal of Industrial Economics*, 39a, 47–68.
Ehrenberg, A. S. C., and Bound, J. A. (1993, in press), "Predictability and Prediction," *Journal of the Royal Statistical Society*, Ser. A.
Ehrenberg, A. S. C., and Lindsay, R. M. (1993), "The Analysis of Replicated Studies," Working Paper.
Finifter, B. M. (1972), "The Generation of Confidence: Evaluating Research Findings by Random Subsample Replication," in *Sociological Methodology*, ed. H. L. Costner, San Francisco: Jossey-Bass, pp. 112–175.
Fisher, R. A. (1935), *The Design of Experiments*, Edinburgh: Oliver and Boyd.
Gigerenzer, G. (1987), "Probabilistic Thinking and the Fight Against Subjectivity," in *The Probabilistic Revolution* (Vol. 2), eds. L. Kruger, G. Gigerenzer, and M. S. Morgan, Cambridge, MA: MIT Press, pp. 11–34.
Gigerenzer, G., Swijtink, Z., Porter, T., Daston, L., Beatty, J., and Krugger, L. (1989), *The Empire of Chance: How Probability Changed Science and Everyday Life*, Cambridge: Cambridge University Press.
Govindarajan, V. (1984), "Appropriateness of Accounting Data in Performance Evaluation: An Empirical examination of Environmental Uncertainty as an Intervening Variable," *Accounting, Organizations and Society*, 9, 125–135.
Guttman, L. (1985), "The Illogic of Statistical Inference for Cumulative Science," *Applied Stochastic Models and Data Analysis*, 1, 3–10.
Harré, R. (1990), *Great Scientific Experiments*, New York: Oxford University Press.
Hedges, L. W., and Olkin, I. (1985). *Statistical Analysis in Meta-analysis*, Orlando: Academic Press.
Hirst, M. K. (1981), "Accounting Information and the Evaluation of Subordinate Performance," *The Accounting Review*, LVI, 771–784.
——— (1983), "Reliance on Accounting Performance Measures, Task Uncertainty, and Dysfunctional Behaviour: Some Extensions," *Journal of Accounting Research*, 596–605.
Hopwood, A. (1972), "An Empirical Study of the Role of Accounting Data in Performance Evaluation," *Journal of Accounting Research* (suppl.), 156–182.
Hubbard, R., and Armstrong, J. S. (1989), "Replication and the Development of Marketing Science," *1989 Summer Educators Conference*, Chicago: American Marketing Association.
Johnstone, D. J. (1986), "Tests of Significance in Theory and Practice," *The Statistician*, 35, 491–498.
Kempthorne, O. (1978), "Logical, Epistemological and Statistical Aspects of Nature-Nurture Data Interpretation," *Biometrics*, 34, 1–23.
Kerr, S., Tolliver, J., and Petree, D. (1977), "Manuscript Characteristics Which Influence Acceptance for Management and Social Science Journals," *Academy of Management Review*, 20, 132–141.
Kuhn, T. S. (1970), *The Structure of Scientific Revolutions* (2nd ed.), Chicago: The University of Chicago Press.
Lindsay, R. M. (1990), "An Examination of the 'Negative Result' Bias: Some Empirical Evidence and a Statistical Redress," *Canadian Academic Accounting Conference*, Victoria.
——— (in press), "Achieving Scientific Knowledge: The Rationality of Scientific Method," in *Justifying Accounting Standards: Some Philosophical Dimensions*, eds. C. A. Lyas, M. Mumford, K. V. Peasnell, and P. C. E. Stamp, London: Routledge.
Lykken, D. T. (1968), "Statistical Significance in Psychological Research," *Psychological Bulletin*, 70, 151–159.
Mack, R. W. (1951), "The Need for Replication Research in Sociology," *American Sociological Review*, 16, 93–94.
Mahoney, M. J. (1985), "Open Exchange and Epistemic Progress," *American Psychologist*, 40, 29–39.
Mulkay, M., and Gilbert, G. N. (1986), "Replication and Mere Replication," *Philosophy of the Social Sciences*, 16, 21–37.
Nelder, J. A. (1986), "Statistics, Science and Technology," (Presidential Address), *Journal for the Royal Statistical Society*, Ser. A, 149, 109–121.
Oakes, M. (1986), *Statistical Inference: A Commentary for the Social and Behavioral Sciences*, New York: John Wiley.
Otley, D. T. (1978), "Budget Use and Managerial Performance," *Journal of Accounting Research*, 16, 324–335.
Pearson, E. S. (1938), "Student as Statistician," *Biometrika*, 30, 210–250.
Popper, K. R. (1959), *The Logic of Scientific Discovery*, London: Hutchinson.
Ravetz, J. R. (1971), *Scientific Knowledge and Its Social Problems*, New York: Oxford University Press.
Rosenthal, R. (1976), *Experimenter Effects in Behavioral Research* (enlarged ed.), New York: Irvington.
Rowney, J. A., and Zenisek, T. J. (1980), "Manuscript Characteristics Influencing Reviewers' Decisions," *Canadian Psychology*, 21, 17–21.
Scherr, G. H. (1983), "Irreproducible Science: Editor's Introduction," *The Best of the Journal of Irreproducible Results*, New York: Workman Publishing.
Smith, N. C. (1970), "Replication Studies: A Neglected Aspect of Psychological Research," *American Psychologist*, 25, 970–975.
Umapathy, S. (1987), "Unfavorable Variances in Budgeting: Analysis and Recommendations," in *Management Planning and Control* (rev. ed), eds. K. R. Ferris and J. L. Livingstone, Beavercreek, OH: Century VII, pp. 163–176.
Uncles, M. D., Hammond, K. A., Ehrenberg, A. S. C., and Davis, R. E. (in press), "A Replication Study of Two Brand Loyalty Measures," *European Journal of Operational Research*.
Wachter, K. W., and Straf, M. L. (1990), *The Future of Meta-Analysis*, New York: Russell Sage Foundation.
Walster, G. W., and Cleary, T. A. (1970), "A Proposal for a New Editorial Policy in the Social Sciences," *The American Statistician*, 24, 16–19.
Yates, F. (1951), "The Influence of *Statistical Methods for Research Workers* on the Development of the Science of Statistics," *Journal of the American Statistical Association*, 46, 19–34.