POPPERIAN RULES FOR RESEARCH DESIGN

... empirical science should be characterized by its methods: by our manner of dealing with scientific systems: by what we do with them and what we do to them. Thus I shall try to establish the rules, or if you will the norms, by which the scientist is guided when he is engaged in research or in discovery. (Popper 1959, 50)

INTRODUCTION

Although Karl Popper penned the words quoted above in the early 1930s, in the book that was to serve as the foundation of his reputation as a philosopher of science, over the next six decades he did not actually contribute to the technical literature in what most scientists would regard as the realm of methodology—by and large, he remained a philosopher and dealt with the logic or theory of methodology, not with the “nitty-gritty” of scientific research design and practice (certainly nor in the social sciences and related fields).

Nevertheless, despite this logical or philosophical emphasis in his writings, it appears that they have had a warmer reception among practicing scientists than with professional philosophers. For what Popper gave scientists was a way of thinking about their work qua scientists—an orientation or attitude that enables them to gain a “meta-level” perspective on the specific methodological, empirical, and theoretical problems that they are grappling with. This ultimately was more valuable for them—and probably more practically efficacious—than any detailed methodological insights he could have provided; as Magee puts it, Popper’s work has had “a notably practical effect on people who are influenced by it: it changes the way they do their own work” (Magee 1985, 4). Sir Peter Medawar and
Sir John Eccles, both Nobelists in the medical/biological sciences, are among the distinguished researchers who have been impressed by Popper in this way. (A longer list of admirers from science, art history, and politics is given by Magee 1985, chap. 1.) But physical, biological, and medical scientists considered as a group probably are less self-conscious about their work, and are less inclined to cite philosophers, than are social scientists and educational researchers, who—perhaps because of their concerns about their status as scientists and because of the well-known charge brought against those of them who subscribe to the naturalistic ideal that they suffer from “physics envy”—often are eager to discuss and try to defend their research studies in philosophical or quasi-philosophical language. And one of the authorities drawn upon in this context is Karl Popper. The following statement from a recent essay in the *Educational Researcher*, written by the prominent educational psychologist Nathaniel Gage, is not atypical; it occurs in a passage where Gage is assessing the work of the social psychologist Kenneth Gergen:

> What then demarcates scientific from nonscientific knowledge—from mere opinion, superstition, and the like? Popper’s criterion is falsifiability. Theory and observation give us the basis not for scientific truth but merely for conjectures, which scientists must try to falsify. . . . In the abstract, Gergen’s position is nonfalsifiable and hence not scientific. (Gage 1996, 8)

Although philosophical critics of Popper might blanch at Gage’s brief excursion into difficult territory, I will argue that this usage of the principle of falsification or falsifiability is—in practice—more beneficial than not. Of course, there is much else in Popper that ought to be drawn on by social scientists and educational researchers as they carry out (as well as assess and talk about) their work, but as space is limited I shall have to confine myself here to this one aspect of Popper’s work. In short, I propose to make an initial attempt to provide what Popper did not, namely an application of the idea that progress in our knowledge arises via falsification, to some of the practical methodological issues that arise in the course of social science (and especially educational) research; but I shall also try to make clear why this aspect of Popper’s work has been so appealing to those members of the educational and social science research communities who have come into contact with it. The following discussion, then, is not a detailed examination of the pros and cons of the philosophical writings of Karl Popper; rather, the focus is on some practical methodological issues in research, but the issues that are selected for examination shall be ones where, I shall try to show, Popper’s discussions of falsification have something to contribute. (I do not wish to claim that I am the first to have taken steps in this direction; Donald Campbell—a social scientist who also had notable philosophical skills—produced a variety of important work on research methodology that bears the stamp of Popper’s influence. See, for example, Cook and Campbell 1979.)

Before getting underway with my main discussion, there is an obvious objection to my plan that needs to be dealt with.

**POPPER AS TOUCHSTONE: AN OBJECTION**

Despite its high standing with many scientists, Popper’s work has been subjected to serious criticism by epistemologists and philosophers of science (see, for example, Newton-Smith 1981; Stove 1982; and Stokes 1998). Therefore, it might seem to be an exercise in futility to devote much time to discussing its practical influence or significance unless these criticisms can be answered. After all, would we waste any time considering the practical relevance, for scientists in the late twentieth century, of the phlogiston theory of combustion, which has long since been refuted?

While I believe that the work of Miller and others (see Miller 1994) goes some way toward defending Popper against his critics and therefore leads to a softening of the previous objection, I shall not be able to defend this objection here (I should stress that I do not rate Miller’s spirited defense of Popperian “critical rationalism” as being a complete success). Instead, I shall take two different tasks: First, the aspects of Popper’s work that have drawn most critical attention are his theory of verisimilitude (approach to, or nearness to, truth) and the issue of whether or not the use by researchers of nonsuppressed hypotheses involves a “whiff” of induction (i.e., whether use of these hypotheses assumes—inductively—that they are likely to hold true in the future). These issues certainly relate to Popper’s ideas on falsification but do not seem to undermine the use of falsification as a significant aspect of normal research design. Second, I think it is important to note that the general objection I raised earlier assumes that a position that is subject to philosophical criticism is thereby rendered practically sterile—but this, of course, is far from being the case, for a position can have fruitful practical implications or applications while not getting matters quite straight at the theoretical or philosophical level.
An extreme example is provided by Newtonian theory; although we now see that this is far from getting the theoretical picture right, it is still a fruitful theory to use in most practical settings. (And successors to Newton's theory have to be able to account for its practical success.)

Building on this point a little, I wish to suggest (1) that Popper's philosophical writings on falsification can be translated into practical guidance in a way that other contemporary philosophies of science usually cannot, and (2) that this guidance is both benign and productive.

1. Philosophers of science often have as their focus the interests of philosophers, and they do not adopt the perspective of scientific researchers. (It is notable that most examples used by philosophers of science involve cases of past scientific work, where problems have at least been temporarily resolved), and they often set up their discussions by taking what Thomas Nagel has called “the view from nowhere” — which is helpful, perhaps, for philosophers but not for researchers.

Consider, for example, the following passage from a recent interesting book, in which the author is discussing progress in science:

Explanatory progress consists in improving our account of the structure of nature, an account embodied in the schemata of our practices. Improvement consists either in matching our schemata to the mind-independent ordering of phenomena (the robust realist version) or in producing schemata that are better able to meet some criterion of organization (for example, greater unification). (Kitcher, 1993, 106)

To be told that in order to progress we should match our theories with “mind-independent” nature is advice that, in the abstract, probably neither Popper nor I would disagree with; the problem, however, is that in practice this advice is sterile, for mere mortals cannot tell whether or not theories actually do match reality (William James, of course, made a similar point in his controversial writings on truth). So this account is not of much practical use to the working researcher unless it is supplemented by an account of how we are to settle this vital issue (hence Popper's belief that we only can make “progress” by locating and then eliminating our errors). The other criterion mentioned by Kitcher—greater unification—is practically more easily decided, but, of course, also is dubious as we cannot be sure that, simply on the basis of the fact that our theories hang together, they therefore must reflect the structure of nature (coherence or unification is quite a weak epistemic criterion). Only a person standing outside nature, standing “nowhere,” could determine whether or not our theories actually do “match” nature. Unfortunately, researchers do not stand “nowhere”; they are never in such a privileged position. (These remarks should not be taken as a decisive criticism of Kitcher, for his lengthy book has much of great worth in it, and I have taken an isolated quotation for illustrative purposes; nor do I wish to suggest that Popper is entirely free of similar limitations, but on the whole, much of his philosophy is fairly directly translatable into practice. On the particular issue raised in my example, Popper often stressed that ascertaining the truth of our hypotheses is a “regulative ideal” but that in practice we will never be able to determine whether or not we have attained this goal. See, for example, Popper, 1965, 226.)

2. Popper's philosophy is benign in two senses: first, it cannot lead researchers far astray — the errors to be made by following Popper (if any) are not serious; and second, in practice even those who disagree with his philosophy can often follow him with profit and without any logical inconsistency. Popperian guidance is productive in the sense that it will incline those who follow it to accept criticism, state their claims with clarity, cast their research designs in fruitful ways, open their theories and hypotheses to empirical test, and so forth (good things all). But more of this in a moment.

THE MANY FACES OF FALSIFICATION

It is well-known that the central idea of Popper's philosophy of science is falsification (and the related notion of falsifiability): Science is demarcated from nonscience by the fact that its hypotheses are, in principle, falsifiable; and science progresses, not by the proving or confirming of hypotheses, but as it were, negatively, by way of refutation or falsification. For Popper, a genuine test of a hypothesis is a serious attempt to falsify it; if the hypothesis withstands this attempt, it is corroborated (but not confirmed or verified)—which means that it survives, temporarily, perhaps to face refutation tomorrow. Popper's ideas here were based on the simple logical insight (technically known as modus tollens) that although no finite body of evidence can definitively prove or establish the truth of a universal hypothesis of the form “all X are Y” (except in the rare cases where we have been able to examine every X), acceptance of one piece of negative evidence (an X that is not a Y) can refute or falsify the generalization.

This process was brilliantly captured in the title of his book Conjectures and Refutations; readers of this work sometimes fail to notice the two
quotations at the front that serve to drive home the moral—one from Oscar Wilde, “Experience is the name every one gives to their mistakes,” and the other from J. A. Wheeler, “Our whole problem is to make the mistakes as fast as possible. . . .” In his first great book, The Logic of Scientific Discovery, falsification served as the glue that bound the whole work together—his discussion of the empirical base of science, of the demarcation of science, of simplicity, of degrees of testability, and of corroboration (his replacement for the notion of confirmation). It is no wonder that this is the aspect of Popper’s work that springs most readily to mind for some practicing researchers (as my earlier citation of Nathaniel Gage demonstrated).

It is important to note that Popper was not a naive falsificationist (see Lakatos 1970); he realized that, just as our decision that a hypothesis has not been refuted is revisable in the light of later experience, so, too, is our decision that it has been refuted; all of our knowledge-claims (whether positive or negative) are merely tentative hypotheses. Furthermore, Popper knew full well that a hypothesis can be saved from refutation by a number of stratagems—for example, by claiming that one of the auxiliary or supplemental hypotheses that we accepted in order to carry out our tests was mistaken (see Popper 1959, 42, 50). The scientific attitude, however, consists in accepting the methodological principle that one should avoid saving hypotheses in this way (unless some further warrant for so doing is available).

It is time to turn to some points about the practicalities of falsificationist methodology in educational research.

A. We do not have to look far to see why falsification is so important for educational researchers; the chief clue was given by Popper when he remarked in the early pages of Conjectures and Refutations that it is easy to find apparent “confirmations” of any theory if one looks for them (Popper 1965, 37)—but what counts is whether or not our conjecture survives a strenuous attempt to refute it, by our failure to find evidence that is incompatible with it. We can, for example, apparently “verify” the theory that the world is flat by citing some “confirmations” (“it looks flat,” “balls placed on the flat ground do not roll away,” and so forth), but nevertheless the theory is wrong. Now, the relevant point here is that the field of education is beset by conflicting theories and viewpoints, all of which were inspired by some observations or data and which are held by their adherents thereby to be established; therefore, carrying out studies that merely add to the stock of reasons that can be offered as to why a theory is right
journeyed only a few miles from campus to another local school district, with virtually an identical ethnic makeup and which was similarly impoverished and had the same kinds and degrees of social problems as the district in which the theory had been formulated! At the oral exam there was no answer to my remark that if the theory had worked in the first school setting, it was highly likely to work in the second, and thus it had not been given a genuine, searching test. As the work in the dissertation was otherwise quite competent, however, we settled on a compromise—the candidate would drop reference to “testing” the theory and would speak instead of “replicating” it. (Replication, of course, serves a purpose, but it is not as challenging to the theory as a specifically designed test; but there is always the chance that the replication could fail, although in this particular case the selection of the site and the subjects had maximized the theory’s chances of success.)

Experience has taught me that a diagram is helpful here, although it somewhat oversimplifies matters. (See figure 8.1.) If one depicts as a circle the “universe of potential evidence or experience” that is pertinent to the theory under test, then the evidence actually used to generate the theory comes from a very small segment. If one then imagines that nature might have hidden potentially falsifying evidence somewhere else in the “universe,” the task of the researcher is to design a study to locate this; it would be good to find it as soon as possible (we need, as Wheeler noted, to make our mistakes as quickly as possible, or rather, to discover that we have made a mistake as quickly as possible—for there is little point in prolonging our allegiance to a flawed theory). To use the figurative language of the diagram, searching in a small segment contiguous to the one where the theory worked is not likely to be efficacious. Popper has pointed out that we must use our background knowledge to find “the most probable kinds of places for the most probable kinds of counter examples—most probable in the sense that we should expect to find them in the light of our background knowledge” (Popper 1965, 240).

C. The tendency to search for “confirming” rather than disconfirming or refuting evidence is particularly strong in research that uses qualitative methods. (There are many varieties of qualitative research, and there is a remarkable range of views about how such a researcher should go about collecting evidence or data; as I do not wish to be taken as rashly generalizing my remarks here to apply to all schools of thought, I will use qualifiers throughout. Chapter 10 makes some further points about qualitative inquiry.) The situations under which much qualitative work is done fos-ters this attitude; either as participant or nonparticipant observers, these researchers are quite often studying complex social settings where there is an overwhelming supply of material that potentially is worth noting. A simplifying hypothesis about what is taking place in a particular setting will be eagerly embraced, for this will prevent the observers from drowning in data by providing guidance about what to note and what to ignore (Popper’s often-cited point that all observation is theory-laden springs to mind here).

Some qualitative approaches make use of predetermined observation categories or checklists; others make a virtue of the fact that hypotheses are not determined in advance of the fieldwork and stress that the categories should emerge as the data are being collected (for example, the “grounded theory” approach of Glaser and Strauss 1967). In either case, however, theories or hypotheses are made use of—either before the observations commence or during the course of the study; the fact that these then have a directive influence on the study (the hypotheses or theories guide the researchers as to what is relevant and what is not, and what is not is usually not paid attention to and often fails to be recorded) is, at best, often only paid lip service. The development of a so-called “mental set” that directs the observations is known to be an important “threat to validity” of qualitative research (Sadler 1982).

Qualitative inquirers, however, have not been entirely insensitive to these issues; the literature contains a number of suggestions about how the “credibility” or “believability” of the products of qualitative inquiry can be assessed, but unfortunately, many of these can be seen to be defective when examined from a Popperian perspective (see chapter 10 for further discussion). It should be noted at the outset that many qualitative
inquirers are reluctant to use the terms *true or truth*, even as a Popperian “regulative ideal”; this failing has reached almost epidemic proportions among those recent workers who use the so-called “narrative method” (see chapter 4). But the tendency to replace *true or truth* by *credibility* and so forth simply will not suffice, and for a simple reason—a study can be credible or believable but not true, just as a swindler’s story is untrue but usually is quite believable. Many accounts of “exotic cultures” by early anthropologists (who often were missionaries) were judged to be highly credible by educated circles in the Europe of the day, but now are regarded as quite wrongheaded (i.e., they have been falsified). And, to complicity matters, the truth is often quite incredible: a president of the United States was involved in at least the cover-up of a burglary; and a fur-covered mammal that is egg laying, and that has webbed feet and the bill of a duck, actually does exist in the antipodes (the platypus)! The plain fact is that there is no reliable connection between believability (a subjective or historically and culturally located criterion) and truth.

The major criteria used by qualitative researchers boil down to four (a similar list is discussed in chapter 10, but the issues here are so important that a little redundancy may not be amiss):  

1. Coherence or “structural corroboration.” The idea here is that a qualitative study is to be believed, or is credible, if its segments “hang together.” Popper put the objection to this in a nutshell:  

   Thus while coherence, or consistency, is no criterion of truth, *simply because even demonstrably consistent systems may be false in fact, incoherence or inconsistency do establish falsity; so, if we are lucky, we may discover inconsistencies and use them to establish the falsity of some of our theories.* (Popper 1965, 226, emphasis added)

   Of course, inconsistency within a theoretical system does not indicate that *every* item is false (although all might be), but it does indicate that there is an error somewhere.  

2. Inter-researcher agreement or consensual validation. This criterion amounts to saying that a qualitative account is to be believed if several different researchers say the same thing or acknowledge that a particular account squares with their own experience. But once again, the fact that an account is agreed to by several (or even many) individuals does not mean that it is true. All of us can provide examples of theories or accounts that once were widely agreed to but that were later shown to be untrue (i.e., were refuted); wide consensus in the Middle Ages that the world is flat did not make it true that it is actually flat.  

3. Cross validation, or triangulation. The criterion here is a little more complex—researcher A collects evidence that “validates” one aspect of a theory or account, while researcher B, who is taking a different approach, gets evidence “validating” some other aspect of the theory. Thus, the theory is established as credible and probably even true! Certitude supposedly increases even more if researcher C turns up another confirming line of evidence. Ignoring the Popperian point that no evidence can ever completely “validate” or “confirm” a theory, what we have here is a situation that relates to what is called in logic and philosophy of science “the inference to the best explanation” and that also directly parallels the use of circumstantial evidence in criminal trials—the “truth” of the prosecution’s case is established by several different (and independent) groups of facts that all converge on the guilt of the accused. The logical situation really is this: evidence A is *compatible with* the theory under examination (X is guilty); but so is evidence B and C; therefore, the theory is true (it is true that X is guilty). Certainly, this form of argument can often establish that it is credible to believe the theory (the theory that X is guilty may be the best explanation we can think of for all these facts—at the moment), but, of course, the truth of the theory has *not* been established, and the line of argument is not logically valid (the leap from “compatible with” to “true” is an enormous one). And there is little need to point to the fact that many a person found guilty on the basis of circumstantial evidence has later been exonerated when new evidence came to light that refuted this judgment.  

4. Checking with the individuals who were studied. This criterion is widely touted in the qualitative research community; the following exposition of it is from a book by two well-known educational researchers who also write quite broadly on social science methodology (the “flip-flopping” over “truth” and the use of what they seem to regard as synonyms stand out):  

   The determination of credibility can be accomplished only by taking data and interpretations to the source from which they were drawn and asking directly whether they believe—find plausible—the results. This process of going to the sources—often called “member checks”—is the backbone of satisfying the truth-value criterion. (Guba and Lincoln 1982, 110)

   The criterion, then, is this: After you have done some descriptive work in a social setting—and even better, if you have come up with an explanation for the events that you have observed—you then obtain validation
by finding out if the people you observed find your descriptions and/or theory plausible. If so, your work satisfies the “truth-value criterion.”

Taken as a general criterion by which to judge qualitative inquiry, this seems to me to be so patently implausible that I am amazed it has such broad currency. The crucial flaw is that no distinction is made here between types of problems that might be being pursued, and in particular it is cavalier with respect to the emic/etic distinction. (1) If the problem is the emic one of cataloguing the beliefs or actions of the individuals you have observed, then checking your description with them is certainly a wise step to take, but even here there can be serious complications—your native “informants” may not be telling you (and may not want to tell you) the whole truth about what they believe, and so their assent to your account is not a fool-proof indication that you have accurately described their beliefs; and, of course, their actions can be described in a variety of ways, and the way they would want to describe what they have done is not necessarily the way you or others would want to describe what had happened. I may think that I was responding to provocation by my spouse, but an observer might describe it as a case of attempted male suppression of a woman—once again, Popper’s point that all observation (and related description) is theory-laden springs to mind. (2) But when it comes to the explanation you have crafted for the events you have seen (usually an etic endeavor if you are a social scientist and those whom you have observed are not), the criterion of “member checks” completely falls apart: Do we really think that our explanations of psychological pathologies, or of cultural practices such as female genital mutilation (or even much less horrendous practices), or of a teacher’s reactions to students of different social backgrounds to her own, should be made hostage to whether or not those whose actions or practices they are actually agree with—or find credible—how we have explained these practices?

Some methodologists of qualitative research, such as Miles and Huberman (1984), do mention in passing that “looking for negative evidence” is an important tactic; the thrust of my discussion thus far has been to suggest that—in practice—it is about the only one of any substance. Finding a credible or believable theory or account is sometimes (but not always) the first step, but by itself it is an extremely halting and weak step for the reasons discussed earlier. The qualitative researcher should be disciplined enough to keep working after this step has been accomplished (if it is accomplished at all)—he or she should actively seek data that, if found, will refute the hypotheses or accounts or descriptions that have been developed. And while the qualitative researcher is engaged in this search, he or she should not be misled by spurious criteria or defective methodological rules.

D. The strong Popperian emphasis on refutation raises a set of issues familiar to educational researchers and statisticians under the guise of “Type I” and “Type II” errors, which in turn relate to the use of the so-called “null hypothesis” method in the conduct of true (i.e., randomized) experiments and the use of inferential statistics. The discussion should start with the null hypothesis.

At first blush the use of the null hypothesis method is quite Popperian, although when I once wrote to Sir Karl about this he indicated that he was not aware of any specific connection between it and his own work. Nevertheless, the logical resemblance is startling. Highly simplified, the essence is this: A randomized experiment using a control group and a treatment group is instituted to throw light on the question of whether the treatment produces an effect. But here is the trick—rather than attempting to establish that the treatment does have an effect, the null hypothesis method inverts the process and assumes that there is no difference and therefore that the treatment has had no effect; and the data are analyzed in an attempt to refute this null hypothesis. In short (and crudely), instead of trying to confirm the hypothesis that is of interest (the experimental hypothesis), the researchers attempt to refute—or show as highly unlikely—the null hypothesis; for if they are able to do this, they apparently have established the likely truth of the experimental hypothesis! This ingenious procedure was devised early this century by the noted statistician Sir Ronald Fisher. (It should be noted that descriptions of the logic of this method given in many elementary statistics books are quite sloppy; accounts differ in subtle but important ways from book to book. The chief variation is in the way the null hypothesis is defined; some works say it is the hypothesis that the treatment does not have an effect; others say that it is the hypothesis that the means of the scores of the control and experimental groups will be the same; and still others conflate the two of these. My own discussion, alas, is also somewhat sloppy but, it is hoped, not fatally so; precise discussion of the issues here would take a monograph, not merely a portion of a chapter in a book.)

Although the logic here is Popperian, there actually is a strong Popperian objection to this procedure (for a lively exposition of this point by a psychological researcher, see Meehl 1991): More is involved in the testing of a hypothesis than merely adopting the form of an attempted
The hypothesis should be a bold conjecture, with lots of content; and we should not know beforehand either that it is untrue or that it is almost certain to hold in the context in which we are testing it—the test should be a genuine one and one from which we can learn something, whether the hypothesis survives the test or not. Early on, Popper wrote (in a passage in which I have omitted some sentences not pertinent to the issue at hand):

According to the view that will be put forward here, the method of critically testing theories, and selecting them according to the results of tests, always proceeds on the following lines. From a new idea, put up tentatively, and not yet justified in any way—an anticipation, a hypothesis, a theoretical system, or what you will—conclusions are drawn by means of logical deduction.... And finally, there is the testing of the theory by way of empirical applications of the conclusions which can be derived from it. (Popper 1959, 32-33, emphasis added)

None of the conditions Popper mentions here are met in the “testing” of the null hypothesis. First and foremost, it is not an informative and bold hypothesis; it merely states (in one of its formulations discussed earlier) that at the conclusion of the experiment the means of the scores of the two groups (or some similar measure) will be the same. This is quite a bland conjecture. Second, not only is the null hypothesis “not justified in any way,” but, according to many authorities, it is not justified at all in virtually any educational or social setting (see Morrison and Henkel 1970)—in fact, it is known with virtual certainty beforehand that the scores of the two groups will differ. This line of reasoning is as follows (see, for example, Meehl 1991, 24): No two human groups, even when (or perhaps especially when) selected randomly, will be exactly the same in their mean score on some dimension of interest—no two such groups are going to be exactly alike with respect to average weight, or height, or IQ, or ability to learn new material in math or science or whatever, or in their responses to a treatment or to some set of test items. In short, the null hypothesis is virtually certain to be false!

There are added complexities here, of course, but these do not affect the point I am making. Thus, although there will always be a difference between the control and experimental groups, there is the important issue of whether the difference has been produced by chance or has been produced by the treatment. The answer to this cannot be established with certainty; instead, statistical tests of significance are used to guide researchers in deciding between these two explanations. It is common for researchers to say (in my view, somewhat sloppily), if it seems likely that there is only a random difference, that the null hypothesis has been supported—for what really is of interest in this whole procedure is not merely whether there is a difference between the two groups, but whether there is a difference that is likely to be due to the treatment that is being tried out. But, as we have seen, this hypothesis has not been tested directly. A final complexity is that a researcher can always obtain highly statistically significant results indicating that the treatment did make a real difference (i.e., that the null hypothesis has been falsified), by using a large sample size. Paul Meehl sums all this up rather nicely:

from the fact that the null hypothesis is always false in soft psychology, it follows that the probability of refuting it depends wholly on the sensitivity of the experiment—its logical design, the net (attenuated) construct validity of the measures, and, most important, the sample size, which determines where we are on the statistical power function. Putting it crudely, if you have enough cases and your measures are not totally unreliable, the null hypothesis will always be falsified, regardless of the truth of the substantive theory. (Meehl 1991, 25)

The basic point in all this, however, is that refuting an uninformative null hypothesis that is wrong anyway amounts to an incredibly weak “test” of the actual hypothesis that we are interested in and runs quite contrary to the spirit of Popper’s work—for, according to Sir Karl, what counts is the testing of “risky predictions” (Popper 1965, 36). (Paul Meehl makes much of the fact that in sciences such as physics, statistical testing of this sort is virtually never used. The hypotheses of interest are tested directly, not by the inverted logic of the null hypothesis test; this is made possible by the fact that the theories or conjectures here are very precise—Popper would say they have lots of content, and so they can yield quite specific, testable predictions.)

E. The previous discussion leads neatly into the topic of types of errors. In carrying out a study or running an experiment, researchers run into two dangers when analyzing their data. First, they can decide that the experimental treatment produces a real effect when in fact there is no effect at all but only chance differences between the groups in the study—they can accept a theory or hypothesis when actually it is false. This is called a Type I error. Second, they can reject a hypothesis or conjecture
that actually is true. This is called a Type II error. In sum, the errors are as follows:

Type I: Acceptance of a false theory or conjecture
Type II: Rejection of a true theory or conjecture

Most books on research design identify Type I errors as being of greater consequence—it is worse to accept error than it is to reject truth. (This is a practical and not a logical or epistemological point—for epistemically, both are actually errors; but the argument often made is that in real-life situations the consequences of Type I errors are usually, although not always, worse than those of Type II errors. I am dubious about this, for in both cases one’s actions are being guided by faulty hypotheses.)

It is interesting to look at these errors through a Popperian lens. First, Popper would argue, I think, that the brief definitions of Type I and Type II errors given herein need to be reformulated, for he would be displeased with the terms accept and reject. All our knowledge is conjectural, and we need to revise continuously the previous decisions we have made about what has been falsified and what has not. Researchers working on a problem formulate a conjecture that they hope is true; they carry out a study that is designed as a test—the aim is to detect whether this conjecture is erroneous. If the evidence indicates that indeed it is, then the conjecture is discarded (but tentatively, for all decisions are revisable in the light of later experience). If, however, the conjecture is not falsified by the study, the researchers will maintain it; but, as Popper insists, they should not accept it in the sense of trusting it fully—for probably in some future study it will be falsified. Popper wrote:

The fact that, as a rule, we are at any given moment taking a vast amount of traditional knowledge for granted ... creates no difficulty for the falsificationist or fallibilist. For he does not accept this background knowledge; neither as established nor as fairly certain, nor yet as probable. He knows that even its tentative acceptance is risky, and stresses that every bit of it is open to criticism. (Popper 1965, 238)

Our “background knowledge,” of course, includes those items we believe at present to be unfurtable, but it also includes those things that we think have been refuted. (Popper only occasionally discussed the issue that our refutations as well as our confirmations are only tentative; but early in The Logic of Scientific Discovery 1959, 50, he did state that “In point

of fact, no conclusive disproof of a theory can ever be produced; for it is always possible to say that the experimental results are not reliable...”) Thus, for a start, a Popperian reformulation of the types of errors would be as follows:

Type I: Tentative maintenance of a false hypothesis
Type II: Tentative rejection of a true hypothesis

The second thing that Popper might say about all this is that we can never completely insulate ourselves against either of these types of errors; all of our knowledge is fallible, and much of what we think to be true is, in fact, likely to be erroneous. So long as we keep an open mind about what we are taking as “background,” so long as we are prepared to revise our tentative beliefs in the light of future experience, and so long as we strenuously test (try to refute) those beliefs that we are using to guide our actions, eventually we will be able to detect our current errors—we will see that we have falsely judged some conjecture to be true or some treatment to be effective, or that we have falsely rejected a conjecture or we have falsely judged some treatment to be ineffective. Of course, in the process we will make other mistakes, but that is the human condition! Popper made this point by way of a verse by Xenophanes (which he translated himself):

But as for certain truth, no man has known it,
Nor will he know it; neither of the gods,
Nor yet of all the things of which I speak.
And even if by chance he were to utter
The final truth, he would himself not know it:
For all is but a woven web of guesses. (Popper 1965, 26)

F. Researchers in education and the social sciences who design experimental studies often think in terms of “threats to validity,” but this is a notion that is more broadly applicable to research and certainly is relevant even to nonexperimental, qualitative work. A Popperian slant can be given to this topic.

Broadly speaking, a threat to validity is a flaw either in the design or the execution of a study that weakens/threatens the validity of the conclusions that might otherwise have been drawn. Consider three common examples: (1) If extraneous influences are allowed to impinge upon the treatment group but these do not affect the control group in a randomized
experiment, then the inference that the experimental treatment caused whatever differences were noted between the two groups is not valid. A key principle of the experimental method is that there should be only a single difference between the two groups, namely, that one group received the treatment and the other group did not; but in this example, there was more than one difference—the experimental group not only received the treatment, it also was affected by the extraneous factors (and it is these that might have produced the results that were noted). (2) After the two groups have been formed by random assignment of individuals, and the experiment has started to run, nonrandom attrition might occur that unevenly affects the control and experimental groups; this poses a serious challenge to the inference that the experimental treatment was responsible for the difference between the scores of the groups on the post-test. Thus, the experiment might involve preschool children, but during the experiment the older ones might be withdrawn by their parents (in order to start school), and this nonrandom withdrawal might be greater in one of the groups in the study than in the other—thus skewing the results by magnifying the difference between the scores of the two groups. (3) In qualitative research, the investigator might quite early on form a hypothesis about what is happening in the group that is being studied; this "mental set" poses a threat to the validity of the conclusions because subsequently the investigator is highly likely to pay attention only to factors that are compatible with this guiding conjecture—and so may not notice conflicting events.

It is regarded as good practice in educational research to anticipate the major threats to the validity of a study and, so far as is practicable, to try to neutralize these. Thus: An experiment running for a long period in a field setting can be monitored to detect any extraneous events as early as possible, and perhaps these can be deflected; attrition may not be preventable, but at least it can be monitored so that its impact on the study can be assessed; qualitative researchers can be trained to record their guiding hypotheses and to document the fact that they have searched for evidence that would be disconfirming. Now, at first sight it might seem that attempting to insulate your research from threats to validity is contrary to the spirit of Popperian philosophy and is an attempt to bolster yourself against refutation. Deeper reflection will show, however, that it is quite the reverse.

In essence, preventing the operation of a threat to the validity of a study makes interpretation of the conclusions that are reached less ambiguous. In a study where the treatment has been confounded by other factors, any gain that is found might, but just as well might not, have been caused by the treatment; or if the study shows no significant difference between the experimental group and the control, this could be because the treatment was ineffective, but it equally might result from the countervailing influence of the extraneous factors that threatened the study’s validity. In neither of these possible scenarios can we decide whether or not the hypothesis that the treatment is effective has actually been refuted. Therefore, it has not faced the most challenging test that was possible. If, however, the threat to validity has been prevented from operating, any gain on the outcome measure (or lack of it) must be due solely to the effectiveness (or lack thereof) of the treatment. Thus, rather than hampering refutation, guarding against threats to validity actually greatly fosters it. This leads directly to the next (and last) point.

G. It should be clear from the preceding discussion that, in Popper's view, ambiguity, vagueness, and obscurity or general lack of clarity are the enemies of progress in human knowledge—for these hamper the giving of pertinent criticism and the bringing to bear of severe empirical tests (criticism, of course, being a type of test or attempted refutation). For Popper, content, clarity or precision, and testability were all related; he claimed, counter-intuitively and controversially, that the bolder and clearer our hypotheses, the more improbable they were for they ruled out or “forbade” more:

There are, moreover... degrees of testability; some theories expose themselves to possible refutations more boldly than others... A theory which is more precise and more easily refutable than another will also be the more interesting one. Since it is the more daring one, it will be the one which is less probable. But it is better testable, for we can make our tests more precise and more severe. (Popper 1965, 256)

Given this strong emphasis on opening one’s conjectures to criticism and potential refutation, it is no surprise that Popper was a bitter opponent of the intellectual fashion that makes a virtue of obscure, but impressive-sounding, prose. In an attack on Habermas and the members of the so-called Frankfurt School of “critical theorists,” he wrote:

Many years ago I used to warn my students against the widespread idea that one goes to university in order to learn how to talk, and to write, impressively and incomprehensively... There is little hope that they will ever understand that they are mistaken... that the standard of impressive incomprehensibility actually clashed with the standards of truth and rational criticism. For these standards depend on clarity. One
cannot tell truth from falsity, one cannot tell an adequate answer to a problem from an irrelevant one, one cannot tell good ideas from trite ones, one cannot evaluate ideas critically, unless they are presented with sufficient clarity. (Popper 1976, 294)

Educational researchers have as much to learn from this as have philosophers! And we all must take responsibility for fostering this lesson—the valuing of clarity, and the willingness to give and receive rational criticism, can only thrive in a professional community that takes active steps to achieve these things.


POSITIVISM

Nowadays, the term _positivist_ is widely used as a generalized term of abuse. As a literal designator it has ceased to have any useful function—those philosophers to whom the term accurately applies have long since shuffled off this mortal coil, while any living social scientists who either bandy the term around or are the recipients of it as an abusive label are so confused about what it means that, while the word is full of sound and fury, it signifies nothing.

The antipositivist vigilanties, who realize nothing of this, still claim to see positivists everywhere. (When one is confused or suffering from delirium, it is possible to see _anything_.) Displaying what often amounts to an embarrassing degree of philosophical illiteracy, the vigilanties rarely bother to distinguish between classical (or Comtean) positivists, on the one hand, and the even more nefarious logical positivists, on the other. Furthermore, they use a number of faulty criteria, either singly or in combination, to identify their illusory foe. The general fantasy is that anyone who is impressed by the sciences as a pinnacle of achievement of human knowledge, anyone who uses statistics or numerical data, anyone who believes that hypotheses need to be substantially warranted, anyone who is a realist (another unanalyzed but clearly derogatory word) is thereby a positivist. The following discussion has as its aim the clarification of these delusions.

As a preliminary, however, it needs to be affirmed that the corpus of the sciences _does_ constitute a magnificent achievement. The issue is, for those who want somehow to learn from—and to emulate—the sciences: What precisely has been the source of their success? Before those who wish to copy the sciences in their own investigations are helped with outright condemnation, it seems a counsel of wisdom to examine the analysis of science that they give—for while some accounts may be narrow and deficient, and deserving of abuse, other accounts might be such that we are
THE EXPANDED SOCIAL SCIENTIST'S BESTIARY

A Guide to Fabled Threats to, and Defenses of, Naturalistic Social Science

D. C. PHILLIPS

ROWMAN & LITTLEFIELD PUBLISHERS, INC.
Lanham • Boulder • New York • Oxford