Building Confidence in Qualitative Research : Engaging the Demands of Policy

Harry Torrance

Qualitative Inquiry 2008 14: 507
DOI: 10.1177/1077800407309380

Qualitative Inquiry
Volume 14 Number 4
June 2008 507-527

10.1177/1077800407309380
http://qix.sagepub.com
hosted at
http://online.sagepub.com

Building Confidence in Qualitative Research

Engaging the Demands of Policy

Harry Torrance
Manchester Metropolitan University, United Kingdom

The quality of qualitative research has been subject to considerable criticism recently, partly driven by the development of an international movement for "evidence-based policy and practice." In the United States, randomized controlled trials (RCTs) are posited by some as the best way of producing reliable research knowledge. Also, responses to criticism of qualitative research is leading to the production of various "standards" and "guidelines" to control the production of qualitative research. This article argues that RCTs do not respond to policy makers' needs and timescales and, furthermore, that producing standards for qualitative research is more likely to restrict quality than enhance it. Rather, what is required of qualitative researchers is to engage with policy makers and research participants to acknowledge the limits of research knowledge while addressing issues of quality collaboratively.

Keywords: *qualitative research; applied research; knowledge production; research utility*

Recent U.S. legislation has privileged "scientifically based research" in decisions about funding educational programs and funding educational research; moreover, *scientific* is defined largely in terms of experimental design and methods, especially randomized controlled trials (RCTs), also sometimes known as "randomized field trials" (Eisenhart, 2006; Eisenhart & Towne, 2003). This seems to be related to prior criticism and review of the quality of educational research (National Research Council [NRC], 2002) has been interpreted as an attack on more qualitative approaches to educational research (Denzin & Giardina, 2006) and furthermore as an attack that

Author's Note: This article was originally presented to a symposium on "Standards of Evidence in Qualitative Inquiry," American Educational Research Association Annual Conference, Chicago, April 2007. Correspondence concerning this article should be addressed to Harry Torrance at h.torrance@mmu.ac.uk.

warrants urgent response from what might loosely be called the "qualitative research community" (St. Pierre & Roulston, 2006).

Attacks on the quality of educational research, and particularly the quality of qualitative educational research, have their parallels in the United Kingdom (Hargreaves, 1996, Hillage, Pearson, Anderson, & Tamkin, 1998, Tooley & Darby, 1998), have similarly affected the debate in Australia (Yates, 2004), and indeed are beginning to emerge in the European Union (Bridges, 2005; S. Brown, 2003). The argument has been that educational research (and in some respects social science research more generally) is too often conceived and conducted as a "cottage industry," producing too many small-scale, disconnected, noncumulative studies that do not provide convincing explanations of educational phenomena or how best to develop teaching and learning. There is not a cumulative or informative knowledge base in the field and it is characterized as being of both poor quality and limited utility.

Thus, it can be argued that those working in educational research in general and in qualitative traditions in particular are facing a global movement of neopositivist interest in so-called "evidence-based" policy and practice, where what counts as legitimate evidence is construed very narrowly indeed. This is manifest not only in country-specific initiatives and legislative action—for example, No Child Left Behind (2001) in the United States and the English National Curriculum and Testing system (Torrance, 2003)—but also in international assessment and evaluation activities, such as Trends in International Math and Science Study and Programme for International Student Assessment (Torrance, 2006) and the Campbell Collaboration, which seeks to review and disseminate social science knowledge for policy makers (Davies & Boruch, 2001; Wade, Turner, Rothstein, & Lavenberg, 2006). Clearly, these various manifestations differ in their origins, orientations, and specific intentions; they are not a coherent and homogenous movement. But equally, they do seem to represent a concerted attempt to impose (or perhaps re-impose) scientific certainty and system management on an increasingly complex and uncertain social world. It is apparent then that what is happening in the United States is not unique; it is almost certainly connected to movements elsewhere and should probably be understood in these terms (indeed, an exploration of links and policy flows would be a very interesting study). However, the legislative obsession with RCTs does seem to be peculiar to the United States, and I begin by exploring this before moving on to reflect on the British experience of responding to calls for better quality educational research and in particular better quality qualitative research.

The Case for and Against RCTs

The case for RCTs seems to derive from a combination of the methods of natural science with the supposed needs of policy. It is argued that randomized experiments mean that any systematic observed differences between the sample that has received the "treatment" and the "control" group, which has not, must be attributable to the treatment. The design reveals whether there is a "causal link" and the treatment can be said either to "work" or "not work," and some calculations can be made about the size of the effect: "The experiment is the design of choice for studies that seek to make causal conclusions, and particularly for evaluations of educational innovations" (Slavin, 2002, p. 18). Of course, experimental design can also be more subtle and complicated than this, with different elements of such designs potentially revealing different aspects of program impact (on different subsamples of students, for example, if the overall sample is large enough), but it is the appeal to certainty about "what works" that is claimed to attract policy makers: "If we implement Program X instead of Program Y, or instead of our current program, what will be the likely outcomes for children?" (Slavin, 2002, p. 18). Such attraction is easy to appreciate. It sounds seductively simple. When charged with dispensing millions of tax dollars on implementing programs and supporting research, one can understand that policy makers would value this sort of help.

This is not the place to discuss all the criticisms (and rejoinders) about the nature of causality and the place of RCTs in understanding social interaction and evaluating human services. They have been well rehearsed in recent issues of *Educational Researcher* (e.g., Burkhardt & Schoenfeld, 2003; Erickson & Gutierrez, 2002; Feuer, Towne, & Shavelson, 2002; Maxwell, 2004; Riehl, 2006; Slavin, 2002). The relevant point about RCTs and policy, particularly when comparing the United States with England, is that so much prior "qualitative" work has to be accomplished before any RCT might be designed; and much policy is in any case decided well before any RCT might be implemented, let alone the results become public. Thus, Slavin's (2002) deceptively simple question about Program X versus Program Y begs many more questions about where Program X and Y come from in the first place. Such decisions are already a long way down the road of policy development and implementation, often too far down the road for it to be worthwhile doubling back. Prior questions would include the following: How and why have Program X and Y been developed? What sort of perceived problem are they trying to fix? What is the research evidence that indicates the nature of the problem and the specific different approaches to fixing it? These are all

questions that require prior investigation of both outcomes (perhaps secondary analysis of test result data) and processes (e.g., ethnographic studies of why the test data are as they are), not to mention value judgments about whether particular practices and/or inconsistencies in the data do indeed indicate the existence of a "problem" that needs to be "solved."

It might be argued that this view of how RCTs are developed renders such prior investigative work as somehow less "scientific" or even "prescientific" (Shavelson, Phillips, Towne, & Feuer, 2003, p. 28). But we know that RCTs are difficult and expensive to organize (Slavin, 2002) and therefore that the evidence on which the design is based has to be pretty secure in the first place. If various forms of qualitative work can be trusted in this respect, why can't they be trusted in their own right? Certainly in England, although there are criticisms of the quality of educational research and regular calls for more, and more rigorously trained, quantitative researchers in the social sciences generally (Hillage et al., 1998; Organisation for Economic Co-operation and Development [OECD], 2002), qualitative studies of the sort indicated earlier are still funded by government departments and the U.K. Economic and Social Research Council (ESRC). The value of qualitative studies in exploring the nature of a particular problem is well recognized by government and other research users/sponsors, although their link to policy making is often indirect; even when commissioned by government departments, research is not necessarily used in the policy making process in any straightforward way (see below).

Furthermore, the thrust of the evidence-based policy movement in England at present tends to favor reviews of research—synthesizing findings from multiple studies—rather than relying on the results of a single study, no matter how well conducted. Arguments in favor of conducting such reviews, particularly those known as "systematic reviews," derive from the critiques of social and educational research outlined earlier: that the findings of empirical studies are often too small scale, noncumulative, and/or contradictory to be useful (Gough & Elbourne, 2002; Oakley, 2000, 2003). Advocates are closely associated with the Cochrane Collaboration in medical and health care research and the Campbell Collaboration in social science, both of which favor the accumulation and dissemination of research findings based on scientific methods, particularly randomized control trials. As such, systematic reviewing is often associated with criticisms of qualitative research and is very much located within the "evidence-based policy and practice" movement (Davies, 2004). For these reasons and others, systematic reviewing has its critics in the United Kingdom (e.g., Hammersley, 2001, MacLure, 2005).[1] Nevertheless, for the purposes of the argument here, the development and

use of systematic reviewing points to a policy caution about relying on single studies, experimental or otherwise, and further demonstrates the divergent manifestation of apparently similar policy concerns in the United Kingdom and the United States. Moreover, even one of the leading proponents of systematic reviewing and the use of RCTs in England concedes that

> RCTs generally find smaller effects than other designs, and the effects of most interventions, whether medical or social, are modest. (Oakley, 2006, p. 76)

This, of course, links with another set of issues: that of the timeliness, cost, and utility of research. It is often incumbent on policy makers to be seen to be doing something about a perceived problem or be acting in response to what seems to be a good idea (be it research based or not) without waiting for the definitive results of science (even supposing these can be produced). Even the results of systematic reviews can take months to appear, let alone the results of newly commissioned studies, and policy makers in England are as likely to ask for very rapid reviews of research, or "rapid evidence assessment" (Boaz, Solesbury, & Sullivan, 2004, p. 12), to be conducted over a few days or weeks and possibly assembled via an expert seminar, as to commission longer term systematic reviews. An investigation of research reviewing by the U.K. ESRC-funded Centre for Evidence-Based Policy and Practice noted that the shortest period devoted to producing a commissioned research review was 15 days, whereas the longest period was 30 months (Boaz, Solesbury, & Sullivan, 2007, p. 8).

Similar issues pertain to commissioned studies and evaluations. Colleagues at Manchester Metropolitan University (MMU) are currently involved in several projects evaluating the impact of ICT on schools and teaching in England, including the impact of placing tools such as interactive whiteboards in classrooms (e.g., Somekh et al., 2007). Introducing the latest computer technology into schools is a politically "good thing": There is no way that the U.K. Department for Education and Skills is going to rip out all the whiteboards that have been put into schools if the evaluation is not particularly positive. At the same time, schools and local authorities (school districts) that have not yet had what they would consider to be their "fair share" of this investment are unlikely to be satisfied by a blunt report that simply says it "doesn't work." They want the hardware and the chance to try it out for themselves.[2] Thus, the MMU evaluations are expected to report, through survey, observation and interview, and action research activities with participating teachers, on what seems to work and what problems have been encountered, how and why it works or not, and what lessons can be learned for future

"roll out" of the program to other local authorities, initial training, and continuing professional development (CPD) activities.

Thus, policy builds, one initiative on another, incrementally over time, with many more issues other than the "scientific evidence" coming into play. It might be argued that policy should not develop like this, but it does, and in a democracy (rather than a "scientocracy"), it is not clear how it could be otherwise. Scientific evidence is but one element in a democratic policy making process. Public values and interests influence matters at the macrolevel of the decision-making process, and the professional judgment of innumerable local actors mediate policy at the microlevel.[3]

It is interesting that much of this sort of incremental, practice-oriented research activity seems to be reflected in current debates about "design experiments." And just as criticism of qualitative research is not unique to the United States, although the focus of attention on RCTs seems much sharper than elsewhere, so also the debate about design experiments seems to indicate that RCT advocacy in the United States is not uniquely an attack on the specific field of qualitative research. All approaches to research that do not employ RCTs seem to be subject to critical scrutiny. Thus, a recent special issue of *Educational Researcher* devoted to exploring "Design Experiments" (Vol. 32, No. 1, 2003) included a response from Shavelson et al. (2003) that dismissed the approach by asserting that "an entirely different conceptualisation of 'evidence-based' education has captured the imagination of federal policymakers" (p. 25). They further asked, "Should we believe the results of design experiments?" (p. 25). Thus, they invoke political power as the determining factor in methodological debate, while simultaneously undermining the claims of one particular methodological approach.

Design experiments involve testing out hypotheses about learning, embedded in specific materials and pedagogic approaches, in small-scale "real life" situations (classrooms, after school clubs, etc.); learning the lessons of how the materials and pedagogies work; and trying to "scale up" for more general testing and application. The approach now seems to be associated with psychologists who wanted to "get out of the lab" and conduct field-based experiments (A. Brown, 1992; The Design-Based Research Collective, 2003). However, similar approaches have been associated with curriculum research and development, and action research, for many years (Elliott, 1989; James, 2006; Stenhouse, 1975) and have parallels in related endeavors such as "mixed method," "deliberative," and "realist" approaches to evaluation (Chatterji, 2005; House & Howe, 1999; Pawson & Tilley, 1997). A more practical and policy-friendly set of approaches to applied educational research and development is hard to imagine. Yet such work is dismissed because it

relies "on narrative accounts to communicate and justify [its] findings" (Shavelson et al., 2003, p. 25). Shavelson et al. (2003) invoke their membership of the NRC Committee (which, interestingly, they describe at one point as "our committee" p. 28) to "scrutinize the knowledge claims from design studies through the lens of the [NRC Report] guiding principles" (p. 26). They conclude that "experiments should [be used] for choices among important design alternatives" (p. 28), and by "experiment," in this statement, they mean "randomized experiments" (p. 28). To be fair, design experiments are treated seriously by Shavelson et al.; they are not arbitrarily dismissed, but ultimately, they are treated as just another "prescientific" preparatory stage (p. 28) before the "real science" of RCTs begins.

So what is going on here? Many different research perspectives and approaches to applied research, curriculum development, and program evaluation cast doubt on RCTs as the "one best way" to conduct educational research and offer convincing evidence that educational progress can be made by other more pragmatic and incremental means (Burkhardt & Schoenfeld, 2003; Chatterji, 2005; The Design-Based Research Collective, 2003; Erickson & Gutierrez, 2002; Maxwell, 2004; Riehl, 2006). Yet such arguments appear to be making little headway. Indeed, even those who get directly involved and collaborate in good faith with the "What Works" agenda may be censored when their findings do not support an apparently previously decided "party line" (Schoenfeld, 2006). There is not much science in censorship. Equally, exclusive reliance on RCTs is not only not necessary for policy making, but in many key respects, it is not desirable, given the diverse constituencies and interests that policy making must reconcile, the contingent nature of the process, and the contingent nature of local development and implementation of innovative programs. The NRC (2002) report, subsequent reiterations of its main arguments (Feuer et al., 2002; NRC, 2005; Shavelson et al., 2003), and concomitant legislation (Eisenhart & Towne, 2003) seem more like a general attempt to discipline educational research and researchers, to produce a general shift in the problematics and topography of educational research, than to produce better evidence for policy making.

The Response of Qualitative Research

To recap, the specific focus on RCTs seems peculiar to the United States, and advocacy of RCTs seems directed at many different approaches to educational research, not just qualitative research. Nevertheless, criticism of the quality of educational research in general and of qualitative

research in particular is widespread internationally and can certainly be understood as part of a more general move to reassert the preeminence of a natural science model of causality, what counts as evidence in social science, and the primacy of outcome measures in debates about efficiency and effectiveness in human services (Thomas & Pring, 2005; Yates, 2005). In policy terms, the basic issue is that of justifying the overall level and specific content of government expenditure on public services. How can policy makers come to know which programs to invest in and whether or not they are effective?

As argued earlier, research evidence can (and should) only ever be one element of such a policy making process. Equally, there are many good reasons apart from serving policy for qualitative researchers to continue to reflect on the strengths and weaknesses of their field. Nevertheless, the relationship of research to policy is what seems to be driving current concerns and being manifested in reports such as the NRC (2002, 2005), a recent "Workshop on Scientific Foundations of Qualitative Research" (Ragin, Nagel, & White, 2004), discussion of doctoral programs (Eisenhart & DeHaan, 2005), the new American Educational Research Association (AERA, 2007) "guidelines" for reporting research, and so forth. The response to criticism has been to start trying to "set standards" in qualitative research, and particularly qualitative educational research, to reassure policy makers about the quality of qualitative research and to reassert the contribution that qualitative research can (and should) make to government-funded programs.

The problem, however, is that the field of "qualitative research" or "qualitative inquiry" is very large and diverse, and there is unlikely to be easy agreement about core standards. Recent meetings of the International Congress of Qualitative Inquiry (University of Illinois, 2005, 2006, 2007) have attracted up to 1,000 participants on each occasion from 55 different countries, working in and across many different disciplines (anthropology, psychology, sociology, etc.), different applied research and policy settings (education, social work, health studies, etc.), and different national environments with their different policy processes and socioeconomic context of action. It will be difficult to reach agreement, and indeed, it is not self-evident that such agreement is desirable even if it were possible. Nor is this simply a matter of scope and scale, of what might be termed *practical complexity*, whereby agreement might eventually be reached, at least in principle. Different disciplines and contexts of action produce different readings and interpretations of apparently common literatures and similar issues. It is the juxtaposition of these readings, the comparing and contrasting within and across boundaries, that allows us to learn about them and reflect on our own

situated understandings of our own contexts. Multiplicity of approach and interpretation, and multivocalism of reading and response, is the basis of quality in the qualitative research community and, it might be argued, in the advancement of science more generally. The key issue is to discuss and explore quality across boundaries, thereby continually to develop it, not fix it, as at best a good recipe, at worst a government-issue straightjacket.

Experience in the United Kingdom

Some attempt at just such "fixing" has been made in the United Kingdom, and the results are instructive. Recently, for example, independent academics based at the National Centre for Social Research (a not-for-profit organization) were commissioned by the Strategy Unit of the U.K. government Cabinet Office (2003b) to produce a report on *Quality in Qualitative Evaluation: A Framework for Assessing Research Evidence*. The rationale seems to have been that U.K. government departments are increasingly commissioning policy evaluations in the context of the move toward evidence-informed policy and practice, and it was considered that guidelines for judging the quality of qualitative approaches and methods were necessary.

The report is in two parts: a 17-page summary, including the "Quality Framework" itself (Cabinet Office, 2003b), and a 167-page full report (Cabinet Office, 2003a), including discussion of many of the issues raised by the framework. The summary report states that the framework has been

> designed primarily to assess the *outputs* of qualitative enquiry . . . and. . . . It is also hoped that the framework will have a wider educational function in the preparation of research protocols, the conduct and management of research and evaluation and the training of social researchers. (Cabinet Office, 2003b, p. 6)

So the framework is a guide for the commissioners of research when drawing up tender documents and reading reports, but it also has ambitions to influence the conduct and management of research and the training of social researchers.

The problem, however, is that in trying to cover everything, the document ends up covering nothing, or at least nothing of importance. The basic "Quality Framework" begs questions at every turn, and the full 167-page report reads like an introductory text on qualitative research methods. Paradigms are described and issues rehearsed, but all are resolved in a bloodless, technical, and strangely old-fashioned counsel of perfection. The reality of doing

qualitative research and indeed of conducting evaluation, with all the contingencies, political pressures, and decisions that have to be made, is completely absent. Thus, in addition to the obvious need for "findings/conclusions [to be] supported by data/evidence" (Cabinet Office, 2003a, p. 22), qualitative reports should also include the following:

> detailed description of the contexts in which the study was conducted (p. 23)
> discussions of how fieldwork methods or settings may have influenced data collected (p. 25)
> descriptions of background or historical developments and social/organizational characteristics of study sites (p. 25)
> description and illumination of diversity/multiple perspectives/alternative positions (p. 26)
> discussion/evidence of the ideological perspectives/values/philosophies of the research team (p. 27)

And so on and so forth across 6 pages and 17 quality "appraisal questions."

No one would deny that these are important issues for social researchers to take into account in the design, conduct, and reporting of research studies. However, simply listed as such, they compose a banal and inoperable set of standards that beg all the important questions of conducting and writing up qualitative fieldwork: Everything cannot be done, choices have to be made, how are they to be made, and how are they to be justified?

To be more positive for a moment, it might be argued that if qualitative social and educational research is going to be commissioned, then a set of standards that can act as a bulwark against commissioning inadequate and/or underfunded studies in the first place ought to be welcomed. It might also be argued that this document at least demonstrates that qualitative research is being taken seriously enough within government to warrant a guidebook being produced for civil servants. This might then be said to confer legitimacy on civil servants who want to commission qualitative work, on qualitative social researchers bidding for such work, and indeed on social researchers more generally, who may have to deal with local Research Ethics Committees (Institutional Review Boards in the United States) that are predisposed toward a more quantitative natural science model of investigation. But should we really welcome such "legitimacy"? The dangers on the other side of the argument, as to whether social scientists need or should accede to criteria of quality endorsed by the state, are legion. In this respect, it is not at all clear that, in principle, state endorsement of qualitative research is any more desirable than state endorsement of RCTs. Defining what counts as science is not the state's business.

Another arena in England where research meets policy is that of "systematic reviewing," mentioned earlier (Oakley, 2003, 2006; Wade et al., 2006). Initially, findings based on RCTs were considered the "gold standard" of systematic reviewing, but this position has been significantly modified, as it has encountered considerable skepticism in the United Kingdom and work is now under way to integrate different kinds of research findings, including those of qualitative research, into such reviews. This may be construed as progress of a sort, but it also involves attempts to appraise the quality and thus the "warrant" of individual qualitative research studies and their findings: Are they good enough to be included in a systematic review or not? This in turn can lead to absurdly reductionist checklists as the complexity of qualitative work is rendered into an amenable form for instant appraisal. Thus, for example, Attree and Milton (2006) report on a "Quality Appraisal Checklist . . . [and its associated] . . . quality scoring system . . . [for] the quality appraisal of qualitative research" (p. 125). Studies are scored on a 4-point scale:

A No or few flaws
B Some flaws
C Considerable flaws, study still of some value
D Significant flaws that threaten the validity of the whole study (p. 125)

Only studies rated A or B were included in the systematic reviews that the authors' conducted, and in the article, they attempt to exemplify how these categories are operationalized in their work. But as with the Cabinet Office example above, their descriptions beg many more questions than they answer. Thus, lengthy appraisal (the Cabinet Office reports) leads to a counsel of perfection—researchers are extolled to do everything—while rapid appraisal (Attree & Milton, 2006, in the context of systematic reviewing) leads to a checklist of mediocrity. Even the most stunning and insightful piece of qualitative work can only be categorized as having "no or few flaws." Again, to try to be fair to the authors, they indicate that

> the checklist was used initially to provide an overview of the robustness of qualitative studies . . . to balance the rigor of the research with its importance for developing knowledge and informing policy and practice. (Attree & Milton, 2006, p. 119)

But this is precisely the point at issue: Standards and checklists cannot substitute for informed judgment when it comes to balancing the rigor of the research against its potential contribution to policy. This is a matter of judgment both for researchers and for policy makers.

Proponents of systematic reviewing still try to insist on expelling judgment from the process however and rendering qualitative work in quantitative terms. As their focus of attention has expanded from a concentration on RCT studies, they have come to bemoan the fact that different reporting traditions and practices in different fields restrict their capacity to evaluate studies and "extract data" easily. Reporting "guidelines" have come to be produced in a manner with which all empirical studies should accord so that they can more easily be assessed for "quality:" for example, "draft guidelines for the REPOrting of primary empirical Studies in Education (the REPOSE Guidelines) . . ." (Newman & Elbourne, 2004, p. 201). Such guidelines are argued to be "relevant to the reporting of any kind of primary empirical research using any type of research design" (Newman & Elbourne, 2004, p. 208). This is an extraordinarily ambitious claim with obvious homogenizing intent. The actual guidelines comprise a two-page checklist of note-type subheadings, including supposedly generic and all-encompassing categories such as "sampling strategy," "data collection," "data analysis," and so forth (p. 211). Individually, they are unobjectionable; taken together, they constitute yet another counsel of perfection that would require a book-length report to fulfill and, if applied in practice, will always lead to the conclusion that anything short of a book is of poor quality. It is a strange product for a movement ostensibly concerned with utility, as policy makers routinely deal in memos, not books.

Developments in the United States

Similar standards and guidelines and checklists are starting to appear in the United States, with, I would argue, similar results. Thus, for example, Ragin et al. (2004) report on a "Workshop on Scientific Foundations of Qualitative Research" conducted under the auspices of the National Science Foundation and with the intention of placing "qualitative and quantitative research on a more equal footing . . . in funding agencies and graduate training programs" (p. 9). The report argues for the importance of qualitative research and thus advocates funding qualitative research per se, but equally, by articulating the "scientific foundations," it is arguing for the commissioning of not just qualitative research but of "proper" qualitative research. Thus, for example, they argue that

> considerations of the scientific foundations of qualitative research often are predicated on acceptance of the idea of "cases." . . . No matter how cases are

> defined and constructed, in qualitative research they are studied in an in-depth manner. Because they are studied in detail their number cannot be great. (pp. 9-10)

This is interesting and provocative with respect to the idea of standards perhaps acting as a professional bulwark against commissioning inadequate and/or underfunded studies. A quick and cheap survey by telephone interview would not qualify as high-quality, "scientific," qualitative research. But when it comes to the basic logic of qualitative work, Ragin et al. (2004) do not get much further than arguing for a supplementary role for qualitative methods:

> Causal mechanisms are rarely visible in conventional quantitative research . . . they must be inferred. Qualitative methods can be helpful in assessing the credibility of these inferred mechanisms. (p. 15)

And in the end, their "Recommendations for Designing and Evaluating Qualitative Research" also conclude with another counsel of perfection:

> These guidelines amount to a specification of the *ideal* qualitative research proposal. A strong proposal should include as many of these elements as feasible. (p. 17)

But again, that's the point, what is feasible is what is important, not what is ideal. How are such crucial choices to be made? Once again, "guidelines" and "recommendations" end up as no guide at all, but rather, they are a hostage to fortune whereby virtually any qualitative proposal or report can be found wanting.

Perhaps the exemplar *par excellence* of this tendency is the AERA (2007) "Standards for Reporting on Empirical Social Science Research in AERA Publications." All 15 closely typed pages are devoted to "educational research grounded in the empirical traditions of the social sciences . . . other forms of scholarship . . . e.g. history, philosophy, literary analysis, arts-based inquiry . . . are beyond the scope of this document" (p. 1). So already, we are alerted to what is really important. Even this truncated version of what counts as educational research spawns "eight general areas" (p. 2) of advice; each of which are subdivided into a total of 40 subsections, some of which are subdivided still further. Yet only one makes any mention of the fact that research findings should be interesting or novel or significant and that is the briefest of references under "Problem Formulation," which we are told should answer the question of "why the results of the investigation would be of interest to the research community" (p. 2; though intriguingly, in this

context, not the policy community). So in this case, we are confronted by both a counsel of perfection and a checklist of mediocrity. The standards may be of help in the context of producing a book-length thesis or dissertation, but no 5,000-word journal article could meet them all. Equally, however, even supposing that they could all be met, the article might still not be worth reading. It would be "warranted" and "transparent," which are the two essential Standards highlighted in the preamble (p. 2), but it could still be boring and unimportant.

It is also interesting to note that words such as *warrant* and *transparency* raise issues of trust. They imply a concern for the very existence of a substantial data set as well as how it might be used to underpin conclusions drawn. Yet the issue of trust is only mentioned explicitly once, in the section of the Standards dealing with "qualitative methods": "It is the researcher's responsibility to show the reader that the report can be trusted" (AERA, 2007, p. 11). No such injunction appears in the parallel section on "quantitative methods" (pp. 10-11), and in fact, the only four uses of the actual word *warrant* in the whole document all occur in the section on "qualitative methods" (pp. 11-12). The implication seems to be that quantitative methods really are trusted—the issue does not have to be raised—whereas qualitative methods are not. Standards of probity are only of concern when qualitative approaches are involved.

As is typical of the genre, the Standards include an opening disclaimer that

> the acceptability of a research report does not rest on evidence of literal satisfaction of every standard. . . . In a given case there may be a sound professional reason why a particular standard is inapplicable. (p. 1)

But once again, this merely restates the problem rather than resolves it: We are confronted by 15 pages of Standards that do not offer any real guidance on how actually to conduct and report empirical research. The issue, each and every time, is how to choose between alternative courses of action and how to justify that choice.

Toward a Different Approach

It is not that qualitative research has no standards or even poorly articulated standards. Far from it, the library shelves are stacked with epistemological discussion and methodological advice about the full range of qualitative approaches available, along with what is at stake when fieldwork choices

are made and what are the implications of following one course of action rather than another. Reading such sources iteratively and critically, in the context of designing and conducting a study, and discussing the implications and consequences with doctoral supervisors, or colleagues or project advisory groups, is what maintains and develops standards in qualitative research.

Setting standards in qualitative research, however, is a different enterprise. It implies the identification of universally appropriate and applicable procedures, which in turn involve documentary and institutional realization and compliance. And as we have seen, the results of such efforts are not helpful to a deliberative process such as research. Moreover, it is not that the committees and research teams that produce such documents are incompetent or malicious but rather that the discursive nature of the problem is not resolvable in terms of written standards. Language cannot settle matters of judgment. It can only ever open up more questions (of ambiguity and specificity: "But what do you mean by . . . ?"). In turn, the impulse of the committee discussion or the policy workshop is to attempt to answer such questions, but all criteria, pursued in this way, simply "multiply like vermin."[4]

Thus, we cannot legislate judgment out of the process of quality control; rather, our judgments must be educated by discussion, debate, and the testing of ideas and findings in public forums through the various processes of academic life both formal and informal. Hitherto such processes have been largely internal to the scientific community, producing self-regulated quality in the long term, though with the possibility that any individual study may fall short of appropriate standards at any particular point in time. This is a situation that governments (and some researchers themselves) no longer seem to want to tolerate. Every study must now be "quality assured" by being "standardized."

At the same time, however, it has been recognized from many different perspectives, including that of the empowerment of research participants on one hand and policy relevance and social utility on the other, that an assumption of scientific disinterest and independence is no longer sustainable. Other voices must be heard in the debate over scientific quality and merit, particularly with respect to applied, policy-oriented research. Thus, for example, Gibbons et al. (1994) distinguish between what they term *Mode1* and *Mode 2* knowledge, with Mode 1 knowledge deriving from what might be termed the traditional academic disciplines and Mode 2 knowledge deriving from and operating within "a context of application:"

> In Mode 1 problems are set and solved in a context governed by the, largely academic, interests of a specific community. By contrast, Mode 2 knowledge is carried out in a context of application. (p. 3)

Mode 2 knowledge will thus generate solutions to problems as they emerge in much the same way as the "Design Experiments" or action research approaches reviewed earlier. Such knowledge is "transdisciplinary . . . [and] involves the close interaction of many actors throughout the process of knowledge production" (p. vii). In turn, quality must be "determined by a wider set of criteria which reflects the broadening social composition of the review system" (Gibbons et al., 1994, p. 8).

These arguments have been used to underpin a discussion document commissioned by the U.K. ESRC on "Assessing Quality in Applied and Practice-Based Educational Research" (Furlong & Oancea, 2005). Although the document falls into the category of yet another "Framework" or set of "Standards" and largely retains the distinction between scientific merit defined in terms of theory and methodology and social robustness defined in terms of policy relevance and utility, it nevertheless does not simply retreat into science or, perhaps more accurately, a narrow scientism (as does the U.S. advocacy of RCTs, for example). Its production is an acknowledgement that other sources of legitimacy and criteria of quality are important. Thus, the report articulates four dimensions of quality—epistemic, technological, use value for people, and use value for the economy—and argues strongly that a restricted, traditional view of scientific quality is no longer tenable.

In practical terms, this means designing studies with collaborating sponsors and participants, including policy makers, and talking through issues of validity, warrant, appropriate focus and trustworthiness of the results. A significant amount of such work is under way in the United Kingdom at present (James, 2006; Pollard, 2005, 2006; Somekh & Saunders, 2007; Torrance & Coultas, 2004; Torrance et al., 2005). The process is not without its problems or critics, but in essence, the argument is that if research is to engage with policy, then research and policy making must progress both theoretically and chronologically in tandem. Neither can claim precedence in the relationship. Research should not simply "serve" policy; equally, policy cannot simply "wait" for the results of research. Research will encompass far more than simply producing policy-relevant findings; policy making will include far more than research results. Where research and policy do cohere, the relationship must be pursued as an iterative one, with gains on both sides.

Governments, and some within the scholarly community itself, seem to be seeking to turn educational research into a technology that can be applied to solving short-term educational problems, rather than a system of enquiry that might help practitioners and policy makers think more productively about the nature of the problem and how it might be addressed. The latter

process will be as beneficial to policy as it is to research. Producing research results takes time, and they are unlikely to be completely unequivocal in any case. Drawing policy makers into a discussion of these issues is likely to improve the nature of research questions and research design, while also signaling to them that the best evidence available is unlikely ever to be definitive.

The U.S. policy focus on RCTs is all the more puzzling in light of these developments and arguments in the United Kingdom. Similarly, the more general scholarly retreat into trying to define the "scientific" merit of qualitative research simply in terms of theoretical and methodological "Standards" rather than in wider terms of social robustness and responsiveness to practice seems to betray a defensiveness and loss of nerve on the part of the scholarly community. We need to acknowledge and discuss the imperfections of what we do rather than attempt to legislate them out of existence. We need to embody and enact the deliberative process of academic quality assurance, not subcontract it to a committee. Assuring the quality of research and particularly the quality of qualitative research in the context of policy making must be conceptualized as a vital and dynamic process that is always subject to further scrutiny and debate. The process cannot be ensconced in a single research method or a once-and-for-all set of Standards. Furthermore, it should be oriented toward risk taking and the production of new knowledge, including the generation of new questions (some of which may derive from active engagement with research respondents and policy makers), rather than supplication, risk aversion, and the production of limited data on effectiveness for system maintenance ("what works"). Thus, researchers and, particularly in this context, qualitative researchers must better manage their relationships with policy makers rather than their research activities per se. This will involve putting more emphasis on interacting with policy and policy makers, less emphasis on producing "guidelines" and "standards" that will only ever be used as a stick with which to beat us.

In the conclusion to a new book, *The Work of Educational Research in Interesting Times*, my colleague Bridget Somekh argues that "educational research communities . . . have been socially constructed as powerless . . . and have colluded in this process . . . through an impetus to conformity rather than transgressive speculation" (Somekh & Schwandt, 2007). She further argues that engagement with policy and policy making should include the discussion of "speculative knowledge" (i.e., future possibilities emerging out of research) "to improvise the co-construction of new visions" (p. 340). This seems to me to be a much more productive ground for engagement with policy making. It is not without its threats and challenges, especially with

respect to cooption and collusion, but if it is speculative of new policy (and research) and properly cautious about the provisional nature of research knowledge, rather than promising a false certainty and legitimacy for policy, then the dialogue could be productive on both sides.

Notes

1. For example, it is variously argued that the view of knowledge production and accumulation in the social sciences on which systematic reviewing is based is epistemologically flawed and that such reviews are in any case not fit for purpose, taking too long to complete, costing too much, and producing too little by way of useful material for policy. MacLure (2005) further argues that the technologically driven database searches that are employed "degrades the status of reading and writing as scholarly activities" (p. 393) and that the overall approach is animated by a fear of the unknowable (and hence unaccountable) interpretations of researchers inherent in the use of language itself.

2. There are also issues of commercial contracts, government investment in and support of the Information and Communication Technology sector, and so on that need not concern us here but obviously tie in to long-term investment in these and other programs in ways that make a simple "what works" answer untenable and unusable.

3. For a more extensive discussion of these and similar issues, see Hammersley (2005).

4. This exquisitely apposite phrase comes from an article written by Margaret Brown (1988, p. 19) about initial attempts to produce the English National Curriculum and Testing system, which resulted in every member of the subject group writing teams wanting to include everything that they considered important to the subject (Torrance, 2003).

References

American Educational Research Association. (2007). *Standards for reporting on empirical social science research in AERA publications*. Retrieved September 27, 2007, from http://www.aera.net/opportunities/?id=1850

Attree, P., & Milton, B. (2006). Critically appraising qualitative research for systematic reviews: Defusing the methodological cluster bombs. *Evidence & Policy*, *2*(1), 109-128.

Boaz, A., Solesbury, W., & Sullivan, F. (2004). *The practice of research reviewing 1: An assessment of 28 review reports*. London: U.K. Centre for Evidence-Based Policy and Practice, Queen Mary College. Available at http://evidencenetwork.org

Boaz, A., Solesbury, W., & Sullivan, F. (2007). *The practice of research reviewing 2: Ten case studies of reviews*. London: U.K. Centre for Evidence-Based Policy and Practice, Queen Mary College. Available at http://evidencenetwork.org

Bridges, D. (2005, December). *The international and the excellent in educational research*. Paper prepared for the "Challenges of the Knowledge Society for Higher Education" Conference, Kaunus, Lithuania.

Brown, A. (1992). Design experiments: Theoretical and methodological challenges in creating complex interventions in classroom settings. *Journal of the Learning Sciences*, *2*, 141-178.

Brown, M. (1988). Issues in formulating and organising attainment targets in relation to their assessment. In H. Torrance (Ed.), *National assessment and testing: A research response* (pp. 15-25). Cheshire, UK: British Educational Research Association.

Brown, S. (2003, September). *Assessment of research quality: What hope of success?* Keynote address to European Educational Research Association Annual Conference, Hamburg, Germany.

Burkhardt, H., & Schoenfeld, A. (2003). Improving educational research: Toward a more useful, more influential and better-funded enterprise. *Educational Researcher*, *32*(9), 3-14.

Cabinet Office. (2003a). *Quality in qualitative evaluation: A framework for assessing research evidence. A quality framework* (full report). London: Author.

Cabinet Office. (2003b). *Quality in qualitative evaluation: A framework for assessing research evidence. A quality framework* (summary report). London: Author.

Chatterji, M. (2005). Evidence on "What Works": An argument for extended-term mixed-method (ETMM) evaluation designs. *Educational Researcher*, *34*(5), 14-24.

Davies, P. (2004). Systematic reviews and the Campbell Collaboration. In G. Thomas & R. Pring (Eds.), *Evidence-based practice in education* (pp. 21-33). Maidenhead, UK: Open University Press.

Davies, P., & Boruch, R. (2001). The Campbell Collaboration. *British Medical Journal*, *323*, 294-295.

Denzin, N., & Giardina, M. (Eds.). (2006). *Qualitative inquiry and the conservative challenge*. Walnut Creek, CA: Left Coast Press.

The Design-Based Research Collective. (2003). Design-based research: An emerging paradigm for educational enquiry. *Educational Researcher*, *32*(1), 5-8.

Eisenhart, M. (2006). Qualitative science in experimental time. *International Journal of Qualitative Studies in Education*, *19*(6), 697-708.

Eisenhart, M., & DeHaan, R. (2005). Doctoral preparation of scientifically based education researchers. *Educational Researcher*, *34*(4), 3-13.

Eisenhart, M., & Towne, L. (2003). Contestation and change in national policy on "scientifically-based" education research. *Educational Researcher*, *32*(7), 31-38.

Elliott, J. (1989). *Action Research for Educational Change*. Buckingham, UK: Open University Press.

Erickson, F., & Gutierrez, K. (2002). Culture, rigor and science in educational research. *Educational Researcher*, *31*(8), 21-24.

Feuer, M., Towne, L., & Shavelson, R. (2002). Scientific culture and educational research. *Educational Researcher*, *31*(8), 4-14.

Furlong, J., & Oancea, A. (2005). *Assessing quality in applied and practice-based educational research*. Swindon, UK: Economic and Social Research Council.

Gibbons, M., Limoges, C., Nowotny, H., Schwartzman, S., Scott, P., & Trow, M. (1994). *The new production of knowledge*. Thousand Oaks, CA: Sage.

Gough, D., & Elbourne, D. (2002). Systematic research synthesis to inform policy, practice and democratic debate. *Social Policy and Society*, *1*(3), 225-236.

Hammersley, M. (2001). On systematic reviews of research literature: A narrative response. *British Educational Research Journal*, *27*(4), 543-554.

Hammersley, M. (2005). The myth of research-based practice: The critical case of educational inquiry. *International Journal of Social Research Methodology*, *8*(4), 317-330.

Hargreaves, D. (1996). Teaching as a research based profession (Teacher Training Agency annual lecture). London: Teacher Training Agency.

Hillage, J., Pearson, R., Anderson, A., & Tamkin, P. (1998). *Excellence in research on school* (Department for Education and Employment Research Report 74). London: Department for Education and Employment.

House, E., & Howe, K. (1999). *Values in evaluation and social research*. Thousand Oaks, CA: Sage.

James, M. (2006). Balancing rigour and responsiveness in a shifting context: Meeting the challenges of educational research. *Research Papers in Education*, *21*(4), 365-380.

MacLure, M. (2005). "Clarity bordering on stupidity": Where's the quality in systematic review? *Journal of Education Policy*, *20*(4), 393-416.

Maxwell, J. (2004). Causal explanation, qualitative research and scientific enquiry in education. *Educational Researcher*, *33*(2), 3-11.

Newman, M., & Elbourne, D. (2004). Improving the usability of educational research: Guidelines for the REPOrting of primary empirical research studies in education (the REPOSE guidelines). *Evaluation and Research in Education*, *18*(4), 201-212.

National Research Council. (2002). *Scientific research in education*. Washington, DC: Author.

National Research Council. (2005). *Advancing scientific research in education*. Washington, DC: Author.

Oakley, A. (2000). *Experiments in knowing*. Cambridge, UK: Polity.

Oakley, A. (2003). Research evidence, knowledge management and educational practice: Early lessons from a systematic approach. *London Review of Education*, *1*(1), 21-33.

Oakley, A. (2006). Resistances to new technologies of evaluation: Education research in the UK as a case study. *Evidence and Policy*, *2*(1), 63-88.

Organisation for Economic Co-operation and Development. (2002). *Educational research and development in England* (OECD Review, CERI/CD(2002)10). Paris: Author.

Pawson, R., & Tilley, N. (1997). *Realistic evaluation*. London: Sage.

Pollard, A. (2005). Challenges facing educational research. *Educational Review*, *58*(3), 251-267.

Pollard, A. (2006). So, how then to approach research capacity building? *Research Intelligence*, *97*, 18-20.

Ragin, C., Nagel, J., & White, P. (2004). *Workshop on scientific foundations of qualitative research*. Retrieved September 27, from http://www.nsf.gov/pubs/2004/nsf04219/start.htm

Riehl, C. (2006). Feeling better: A comparison of medical research and educational research. *Educational Researcher*, *35*(5), 24-29.

Schoenfeld, A. (2006). What doesn't work: The challenge and failure of the What Works Clearinghouse to conduct meaningful reviews of studies of mathematics curricula. *Educational Researcher*, *35*(2), 13-21.

Shavelson, R., Phillips, D., Towne, L., & Feuer, M. (2003). On the science of education design studies. *Educational Researcher*, *32*(1), 25-28.

Slavin, R. (2002). Evidence-based education policies: Transforming educational practice and research. *Educational Researcher*, *31*(7), 15-21.

Somekh, B., Underwood, J., Convery, A., Dillon, G., Jarvis, J., Lewin, C., et al. (2007). *Evaluation of the ICT TestBed project: Final report June 2007*. Coventry, UK: Becta.

Somekh, B., & Saunders, L. (2007). Developing knowledge through intervention: Meaning and definition of "quality" in research into change. *Research Papers in Education*, *22*(2), 183-197.

Somekh, B., & Schwandt, T. (Eds.). (2007). *Knowledge production: The work of educational research in interesting times*. London: Routledge.

Stenhouse, L. (1975). *An introduction to curriculum research and development*. London: Heinemann Books.

St. Pierre, E., & Roulston, K. (2006). The state of qualitative inquiry: A contested science. *International Journal of Qualitative Studies in Education*, *19*(6), 673-684.

Thomas, G., & Pring, R. (Eds.). (2005). *Evidence-based practice in education*. Buckingham, UK: Open University Press.

Tooley, J., & Darby, D. (1998). *Educational research: A critique*. London: Office for Standards in Education.

Torrance, H. (2003). Assessment of the national curriculum in England. In T. Kellaghan & D. Stufflebeam (Eds.), *International handbook of educational evaluation* (pp. 905-928). Boston: Kluwer.

Torrance, H. (2006). Globalising empiricism: What if anything can be learned from international comparisons of educational achievement? In H. Lauder, P. Brown, J. Dillabough, & H. A. Halsey (Eds.), *Education, globalisation and social change* (pp. 824-834). New York: Oxford University Press.

Torrance, H., Colley, H., Ecclestone, K., Garratt, D., James, D., & Piper, H. (2005). *The impact of different modes of assessment on achievement and progress in the learning and skills sector*. London: Learning and Skills Research Centre.

Torrance, H., & Coultas, J. (2004). *Do summative assessment and testing have a positive or negative effect on post-16 learners' motivation for learning in the learning and skills sector*. London: Learning and Skills Research Centre.

Wade, C., Turner, H., Rothstein, H., & Lavenberg, J. (2006). Information retrieval and the role of the information specialist in producing high-quality systematic reviews in the social, behavioural and education sciences. *Evidence and Policy*, *2*(1), 89-108.

Yates, L. (2004). *What is quality in educational research?* Buckingham, UK: Open University Press.

Yates, L. (2005, July). *Is impact a measure of quality? Producing quality research and producing quality indicators of research in Australia*. Keynote address for the Australian Association for Research in Education Conference on "Quality in Educational Research: Directions for Policy and Practice," Cairns, Australia.

Harry Torrance is a professor of education and director of the Education and Social Research Institute, Manchester Metropolitan University, in the United Kingdom. His substantive research interests are in the interrelation of assessment with learning, program evaluation, and the role of assessment in education reform. He has undertaken many applied, qualitative, and mixed-method investigations of these topics funded by a wide range of sponsors. He is an elected member of the U.K. Academy of Learned Societies for the Social Sciences.