

In Defense of Robin Barrow's Concerns about Empirical Research in Education

JACK MARTIN

Department of Psychology, Simon Fraser University

*This paper offers a defense of Robin Barrow's main arguments in *Giving Teaching Back to Teachers*, including additional material concerning the inability of the aggregate data and statistical methods employed in research in education (and research on teaching) to speak to individual teachers and students or to particular classrooms. This defense and extension of Barrow's position is applied in a critique of a proposal made by Lorraine Foreman-Peck in her 2004 debate with Barrow, entitled *What Use is Educational Research?*, published in 2005 by the *Philosophy of Education Society of Great Britain*. A central confusion that attends and limits much empirical research in education and social science concerns conflation of two different senses of the concept general, as "common to all" or "on average." The havoc this confusion plays ought not be ignored or minimized by educational researchers and their advocates who tend to exaggerate the empirical regularity in social scientific data and therefore the generalizability of social science research in education and elsewhere.*

When I first read Robin Barrow's (1984) *Giving Teaching Back to Teachers*, I was in my first year of a new appointment at what then was named The University of Western Ontario, having spent the previous eight years of my academic life in Simon Fraser University's Faculty of Education. This move was related to my reading of Barrow's book because it was co-published by *The Althouse Press*, the in-house publisher attached to my new workplace at Althouse College, otherwise known as the *Faculty of Education at UWO*. The book was the focus of much spirited conversation in the faculty lounge of the College during my first year there because of its forceful arguments and because Barrow himself had spent a sabbatical at Althouse College a few years prior to my arrival. Having gotten to know Barrow, Faculty members of the College were eager to talk about their agreements and disagreements with his views and arguments.

My own reaction to Barrow's book on first reading was mixed. I generally agreed with what he said, but I wished he could have said it in a way that perhaps would be a bit more convincing to researchers themselves. In particular, I wanted additional arguments—or extensions of arguments—that conveyed a more intimate knowledge of what was problematic about the methods of inquiry typically employed by educational researchers. So, in this essay I want to both defend Barrow's main arguments in *Giving Teaching Back to Teachers* and add additional material concerning the inadequacies of the aggregate statistical methods employed in research in education (and research on teaching). My defense and extension of Barrow's position also includes a critique of a proposal made by Lorraine Foreman-Peck in her 2004 debate with Barrow entitled *What Use is Educational Research?* published in 2005 by the *Philosophy of Education Society of Great Britain*. Foreman-Peck's proposal is essentially the same as a call and plan for applied research in education issued by Lawrence Stenhouse in his 1979 Presidential Address to the

British Educational Research Association, entitled “The Study of Samples and the Study of Cases.” In this address, Stenhouse proposes that an appropriately insightful merging of quantitative studies of groups of individuals with case studies of particular individuals and/or events in education is the way forward for a progressive program of empirical research. In his words, “I believe that the description of cases and the analytic categorization of samples are complementary and necessary approaches in educational research, and it is high time that the superficial stylistic differences between their proponents were recognized as impediments to good sense in the research community” (p. 6).

Of course, almost thirty years later, Stenhouse’s wish has come true, as more and more researchers in education and other applied areas now champion mixed-methods inquiry: quantitative and qualitative, group and individual. In psychology, personality psychologist Gordon Allport is typically credited as fathering this idea as a way to make large-scale group difference studies of personality traits relevant to case studies of individual personalities. Allport considered group difference studies to be nomothetic, indicating they aimed to establish general laws that applied to all individuals. He considered case studies to be idiographic, indicating they aimed to establish facts and warranted interpretations about individuals. In the first chapter of his 1937 book *Pattern and Growth in Personality*, Allport states his purpose in developing and defending such a “bridging.” He states:

The separation of psychology into two opposing branches is too sharp. I agree with the French psychiatrist [Étienne Eugène] Azam, who many years ago wrote that the science of character cannot proceed by generalities, as does psychology, nor by individualities as does art. It occupies a middle position. ... There is no reason we should not learn from every generalization about human nature what we can. At the same time we need to be alert for concepts and methods that enable us to understand patterned individuality. As we have seen, the individual represents lawful order in nature. We should use *all* avenues of approach in the developing science of the psychology of personality (p. 12).

A central aim of this article will be to explain why such a scientific synthesis across individuals and groups—or cases and samples, as Allport, Stenhouse and Foreman-Peck would have it—is incoherent. But first, I want to offer a brief summary of Barrow’s argument concerning the necessary limits of empirical research in education before turning to Lorraine Foreman-Peck’s invocation of the case-sample research synthesis of Allport and Stenhouse.

Barrow’s Argument Against the Hegemony of Empirical Research in Education

In his debate with Lorraine Foreman-Peck (Barrow & Foreman-Peck, 2005), Barrow provides a straightforward, mostly restrained, and succinct précis of what he had to say in his 1984 book. In making his case against empirical research in education, he argues that scientific research in education is not scientific in the manner of the established physical sciences. To support his argument, he notes the comparative conceptual clarity with which phenomena in physical science can be rendered, the relative ease with which they can be captured and isolated in experimental settings without changing them, the greater clarity in purpose and aim, and the greater ease of creating control conditions. The upshot is that ascertaining lawful regularities is possible in many branches of physical science, whereas the best that educational research has to offer might generously be posited as generalizations that are not very general.

Unlike laws, generalizations admit to exceptions, especially given the multitude of situational, personal, and interactional, normative aspects that attend educational contexts.

Nonetheless, Barrow is clear that generalizations from educational research ought not be dismissed out of hand, despite the fact that examples of tautology and strong conceptual entailment are not infrequent in the literature of educational research and the proclamations of those who value and conduct it. For example, influential educational researcher Jere Brophy (2006) developed a list of most commonly replicated results which included the “findings” that effective teachers make a difference, allocate most of the available time to activities designed to accomplish instructional goals, spend a great deal of time actively instructing their students, and provide a supportive learning environment. But rather than dwelling on such examples, Barrow offers a positive proposal that claims the only way to make sense of posited empirical generalizations in educational contexts is to consider them within and against the situational, personal, and interactive particulars of any given context of application (i.e., in consideration of specific teachers, students, classrooms, and socially sanctioned ways of interacting). The upshot is that it always is individual teachers familiar with such particulars who must do the heavy lifting of determining what is most likely to work for them and their students within their classrooms. There is no other option if generalizations from educational research are to be used sensibly. And, of course, such tasks are made ever more complicated by two overarching facts: (1) that educational success is a highly contested concept debated by teachers, administrators, parents, and communities and (2) that education’s usual form is one of interactions between teachers and pupils who all are active and reactive agents. In Barrow’s (2005) words:

methodologies designed for the study of material, defined and observable objects are not naturally suited to the study of human interactions, which are largely non-material and poorly defined ... Two people meeting just isn’t the same as two billiard balls meeting or two chemicals combining. The autonomy and the individuality of the human are ultimately what prevent there being a true science of human behavior. (pp. 27-28)

Ultimately, Barrow’s argument “is an argument to the effect that we need to emphasize other things in educational research than empirical inquiry on the model of the natural sciences” (p. 29). This is the exact sentence with which he concludes his contribution to the debate with Foreman-Peck.

Foreman-Peck’s Attempt to Salvage Educational Research

In response to Barrow’s position, Lorraine Foreman-Peck (2005) agrees that “to adopt research findings uncritically is as de-professionalizing as ignoring them” (p. 35). She states that “The criterion of strict reliability cannot be met ... since all possible variables cannot be controlled” (p. 37) and that “Recommendations for teaching drawn from such studies cannot have the authority of causal laws ... since ... the variables or conditions involved are too numerous” (p. 37). She also agrees that teachers matter greatly. In her more science-and-technology-friendly turn of phrase, she advises:

Any generalized finding from research needs to be held by the teacher as a tentative possibility or as a hypothesis since it may be that in most cases X is the case, but it could be that her teaching situation is exceptional. How is she to tell if her case is exceptional? This requires a willingness to try new approaches

and the skills to evaluate them” (p. 39). She continues with “Knowing how your teaching is received is a form of self evaluative research. It does require an open mind, commitment and wholeheartedness ... Humility in teaching is a hard won virtue.” (p. 39)

Despite all this, Foreman-Peck chooses to package and channel her agreements with Barrow in a way I find particularly interesting because it reminds me so much of moves and strategies I have encountered time and again in the history of psychology, including the previously mentioned attempt to bridge nomothetic and idiographic methods and approaches in personality psychology by Gordon Allport. Her transitional move is to recruit teachers not as Barrow has done by positioning them as sensible, well-educated people, but by writing that “Teachers have to see themselves as users and producers of research. ... If teachers are to be enabled to use research creatively, they need to be trained to understand the nature of research, how to evaluate it, and engage with it” (pp. 42-43). Then, after dismissing Barrow’s ‘case against empirical research’ as “only a case against accepting poor empirical research” and therefore as “confused and unconstructive”, she plays her trump card, which turns out to be singularly Allportian. Repeating that “Scientific causal laws are not appropriate ways of describing human actions,” she asks: “How then can we have a social science of education and what connection does this have to the classrooms and teaching?” (p. 47).

Her debt to Stenhouse and strong similarity to Allport are revealed when the answer she proposes to her question takes the form of a plea for the use of a nomothetic group approach combined with an idiographic case approach. In her words:

The study of samples can reveal regularities in groups or categories of people that they may not be aware of themselves ... However it does not specify what kind of action is warranted and it may be that more detailed case study work, i.e. work with individuals, is necessary to unearth what the constraints are, and the ways ... pupils think. What is suggested here is that a study of samples is complemented by the study of cases.” (pp. 47-48)

Unfortunately, and for reasons I shall now elaborate, group or sample research in most areas of social science, including educational research, can tell us nothing whatsoever about any particular individual or situation. Such research, on occasion, can provide useful group demographic information for purposes such as planning and locating the physical facilities of schools or predicting how many students with particular exceptionalities might be expected in a particular school district or even what teaching approaches might tend to be useful or harmful *on average*. However this research is completely silent about whether or not any particular student will or will not benefit from any educational or teaching approach or strategy.

What Can and Cannot be Claimed on the Basis of Traditional Group Research and Statistical Analysis in Psychology and Education

It is especially important to recognize that what is judged to be beneficial or harmful on average need not be beneficial or harmful to all or even the majority of students, let alone to any particular student. To illustrate the irrelevance of educational research using group designs and statistical analysis to how any individual student might respond to teaching, Lamiell (2003) discusses several variations of a basic

research scenario that involves the random assignment of 20 children to one of two spelling classes employing different teaching strategies: either “formal drill” or “silent reading.” At the end of the school year, all students completed a spelling test. The purpose of these studies was to explore the question: What can be learned from such studies that might explain the learning and test performance of individual pupils? The overall mean of all 20 students on the end-of-year spelling test was 10.90, the mean for the 10 “formal drill” students was 1.30 points lower at 9.60, and the mean for the 10 “silent reading” students was 1.30 points higher at 12.20. This difference in test performance between the two groups of students was highly statistically significant, or reliable. Nonetheless, when the individual test results of students in the formal drill group (the group performing more poorly) were examined, the test scores of 4 of the 10 students in that group were above the overall mean, with two of these pupils scoring higher than any other students in either group. When individual test results of students in the silent reading group (the group performing better) were examined, a full half of those children scored below the overall mean. A simple regression equation using these data revealed that this equation predicted none of the specific scores of any of the students. It should be noted here that regression equations of this kind are frequently treated by some researchers as law-like and said to reveal important empirical regularities. Lamiell goes on to show that although a greater percentage of overall variation in the student scores could be accounted for statistically when student gender and achievement motivation were added to subsequent regression equations, individual students in each of the smaller groups created by the addition of one or both of these additional “independent variables” still differed noticeably from each other and aggregate mean scores, thus defying the purported “law-like” predictability of the more complex regression equations. There is nothing here to support Foreman-Peck’s assertion that the study of samples can reveal important and useful *regularities* in the ways in which students respond to specific educational interventions.

Over many years of conducting and supervising empirical research in education and psychology, I can testify that the statistical methodological assumption of empirical regularities in aggregate data—an assumption that, upon closer examination of results for individuals, turns out to be non-existent—is more the norm than exception in work of this kind. Individual results of exposure to different teaching methods, strategies, and styles often differ greatly from results for groups to which these individuals have been assigned or belong. A few years ago, one of my students, Wanda Power (2014), was able to obtain, after much difficulty and many approaches to authors of published research in psychology, raw data from two highly cited empirical studies. When she examined individual data from these studies, like Lamiell (2003), she found almost as many deviations of individual results from the directional results reported for aggregated group data to which the individual results contributed as she did regularities in direction of outcome. For her work, Wanda was awarded the Mary Wright Award for Best Student Paper at the annual conference of the Canadian Psychological Association.

A moment’s reflection will suffice to inform us why such discrepancies between empirical research results for individuals and the groups to which they belong are legion. As Barrow has argued, teachers, pupils, and others participating in empirical research in education and psychology are persons with agentive capability. Their interpretations of research instructions and situations inevitably reflect their life experiences. In a study I conducted many years ago while in SFU’s Faculty of Education (Martin, 1979) another of my students who acted as teacher and research assistant, Bill Marble, was puzzled by the very different responses high school students had to an experimental intervention we were studying. This intervention assumed that asking students higher-order questions would help them to think and learn

more deeply. When Bill eventually got around to talking to individual students, he discovered that some students did not respond well to higher-order questions intended to increase their interest and enthusiasm because they assumed the teacher was attempting to show them up by asking such questions of them, or catch them out, or otherwise reprimand them for some reason. Others were confused and did not understand how they were supposed to respond. Still others didn't notice that anything different was going on.

Of course, an understanding of what is happening, which tends to happen in a great deal of educational research, follows from exactly the points Barrow has made in *Giving Teaching Back to Teachers* and in his later debate with Lorraine Foreman-Peck. In Barrow's words:

“We ... cannot [know and] even isolate and control for the various different dimensions to the activity of the individual teacher [and pupil] and we also must recognize that each different student-teacher interaction and “combination creates a new possibility” (Barrow & Foreman-Peck, 2005, p. 25).

Tellingly, Barrow had the insight to add, “I am not suggesting that all researchers are oblivious to these or other problems. But I do think that, generally speaking, the magnitude of the problem has been vastly underestimated.” Exactly so! However, I think it might not be inappropriate to add “and/or purposefully ignored.” My own interaction with empirical researchers in education and psychology tells me that these problems are indeed well known but quietly tolerated in the interests of getting on with careers and opportunities, and in doing so, convincing ourselves that these problems will somehow not stand in the way of ‘progress.’ If so, assuming that the addition of individual case research will somehow allow group research that need not concern itself with results of the individuals studied is a bit like offering group researchers a “get out of jail” card.

On the Incompatibility of Nomothetic and Ideographic (or Sample and Case) Inquiry

To support this line of criticism, a bit more needs to be said about the incompatibility of what Allport called nomothetic and ideographic research and Stenhouse referred to as sample and case research. As both a historian and a statistician, Jim Lamiell, in his 2003 book *Beyond Individual and Group Differences*, traces a central confusion about the interpretation of results from group difference research to historical changes in the meaning “general” as employed as a concept in social science, especially psychology. Lamiell, who learned to read and speak German fluently so that he might conduct historical work on German psychologists of the late 1800s, informs us that one of the fathers of modern psychology, Wilhelm Wundt understood the German equivalent to the term *general* to mean “common to all.” Thus, a general law governing the behavior of falling bodies would “hold for, or apply to, each particular one of many discrete instances of falling bodies” (Lamiell, 2003, p. 182). For Wundt, experimental psychology was a discipline seeking general laws in the sense of laws common to all individuals. In his famous laboratory at Leipzig, Wundt and his assistants tested one subject at a time, meticulously recording all responses of each individual and eschewing any combining of results into higher-order aggregates. Only when results were common to each and every subject tested did they give them serious consideration. Not surprisingly, given this requirement, Wundt restricted his experimental research to very basic tasks

consisting of perceptions of stimulus change, reaction times, and so forth, tasks he hoped might reveal the most basic laws of psychology. Anything else, he assumed, would require non-experimental methods and involve the study and interpretation of human actions and interactions in everyday contexts using methods of observation, interview, and textual, sociocultural, and historical analysis.

When psychology migrated to America, most American psychologists, even those who had studied with Wundt, began to adopt a different conception of “general”: one influenced by the statistical methods of Galton, Pearson, Fisher, and others. They mostly eschewed Wundt’s experimental methods that involved testing one individual after another and meticulously reporting results for each and every individual tested. Americans like Edward Thorndike began to act as if aggregate-level statistical knowledge was a window into individual level phenomena. The statistically aided study of large groups of individuals and the reporting of data as aggregated across individuals replaced adherence to Wundtian insistence on replication across individuals and restriction of experimentation to simple, basic tasks. With inferential statistics, psychologists gained a methodological expertise that seemed both scientific and easy to apply within and outside the laboratory. In a short period of time “general” as *common to all* was replaced with “general” as what is found to be the case *on average*.

In passing, it is worth noting that not all psychologists went quietly along for the ride. Some, like Kurt Lewin, Egon Brunswik, and later David Bakan, emphatically rejected the idea that statistical regularities drawn from aggregated data could be understood as constituting nomothetic laws of human behavior or mental life that could be applied to all individuals. Nonetheless, despite such cautions, the scientific and practical riches to be reaped by this conceptual shift and the avalanche of statistical innovations and techniques that followed were too powerful for most psychologists and other social scientists to resist.

The detailed story of all of this, such as supplied by Lamiell (2003), is fascinating. However, for my purposes this historical sidebar, added to previous examples and illustrations, hopefully suffices to make the following points. Statistical tests of aggregated data drawn from groups of individuals do not supply information that allows for the interpretation or prediction of the behavior or experience of any particular individual. If an understanding of particular individuals is what researchers are after, their empirical inquiries must be carried out at the level of the individual. Any claim that is true on average is not true in the particular.

Empirical research in education is not without value. Aggregated group data often can be used validly to monitor and predict overall trends in student achievement, general efficacy of teaching methods, and so forth, so long as “overall” and “general” are not interpreted as applying to any particular individuals. When asked if this particular teaching method will work or be good for this particular pupil, or even asked to offer a probabilistic estimate based on knowledge of aggregated group data, it is just wrong for researchers to claim any expertise with respect to answering such questions. As we all ought to know, the fact that 60% of students have benefitted from a teaching approach does not mean that any particular student has a 60% chance of doing so. There simply is no way of knowing, even if the result of a particular empirical study reporting that 60% of students benefitted from treatment X was entirely generalizable to other situations and classes (which it almost never the case in empirical research in education), whether or not any particular student might belong to the 60% group that benefits or to the 40% group that does not. Contra Foreman-Peck, Allport, and Stenhouse, samples do not complement cases. Cases do work that samples cannot do. Cases and samples are not, as Stenhouse (1980) would have it, separated by “superficial stylistic differences,” (p. 4) but by the unbridgeable fact that they are suited to entirely

different purposes, no matter how much the latter are allowed to masquerade as complementary to the former.

The further fact is that social science research, in education or psychology or anywhere else, has yet to produce a single law that can be stated with precision and is clearly supported by statements of conditions of application that make it exceptionless. General truths of the kind enshrined in well-known scientific laws, such as Bernoulli's laws of fluid dynamics or Fourier's law of heat conduction, hold true in that they apply to each and every relevant instance governed by these laws. For example, in all instances of heat conduction, the time rate of heat transfer through a material is proportional to the negative gradient in the temperature and to the area, at right angles to that gradient, through which the heat flows. Statistical truth, on the other hand, does not hold true across all relevant instances. To speak statistically is to speak about what is true on average, and something that is true on average is not true of all the instances that contribute to the average.

Concluding Remarks

What I find remarkable but not surprising, given my knowledge of Barrow and his work, is how adept he was at identifying the basic problem with empirical research in education so many years ago. Of course, I'm sure he would acknowledge that there were all sorts of others before him who had picked up the whiff of something amiss in the pretensions and exaggerated claims of educational researchers to be able to right the good ship Education. Yet, Barrow has a gift for writing clearly, concisely, and convincingly about such matters that is exceedingly rare. He has the good sense and decency to steer clear of the overly technical language and pedantry that one frequently senses in those who embrace what they claim as a more expert manner of conducting themselves. Barrow sticks to ordinary language and his analysis of concepts and their use is not fettered by methodolatry. As a matter of fact, my main concern in drafting this essay is that Barrow will find my somewhat more technical elaboration of some of his arguments to be both unnecessary and unhelpful. If so, so be it. Such a response from him would do little to dampen my respect and affection for him as a deservedly renowned scholar, long-time colleague, and close friend. What Barrow has to say always deserves a hearing. Even if you disagree, he will make you think more fully and deeply about the matters under discussion and about what we might consider doing in relation to them. He does what a philosopher does and he does it exceedingly well.

Acknowledgments

This paper was presented at *Philosophical Issues in Education: A Symposium in Honour of Dr. Robin Barrow*, held at Simon Fraser University, Vancouver, B.C. on November 2, 2018.

References

- Alexander, P. A., and Winne, P. H. (2006). *Handbook of educational psychology* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.
- Allport, G. W. (1937). *Pattern in growth in personality*. New York: Holt, Rinehart and Winston.

- Barrow, R. (1984). *Giving teaching back to teachers: A critical introduction to curriculum theory*. Sussex, UK: Wheatsheaf Books and London, ON, Canada: The Althouse Press.
- Barrow, R., and Foreman-Peck, L. (2005). *What use is educational research? A debate*. London, UK: Philosophy of Education Society of Great Britain.
- Brophy, J. (2006). Observational research on generic aspects of classroom teaching. In P.A. Alexander & P. H. Winne (Eds.), *Handbook of Educational Psychology* (2nd edition, pp. 755-780). Mahwah, NJ: Lawrence Erlbaum.
- Lamiell, J. T. (2003). *Beyond individual and group differences: Human individuality, scientific psychology, and William Stern's Critical Personalism*. Thousand Oaks, CA: Sage.
- Martin, J. (1979). Effects of teacher higher-order questions on student process and product variables in a single classroom study. *Journal of Educational Research*, 72, 183-187.
- Power, W. (2014, June). Self-determination theory: A critical analysis. In K. L. Slaney & T. P. Racine (Chairs), *History and philosophy of psychology across the spectrum*. Symposium presented at the annual meeting of the Canadian Psychological Association, Vancouver, BC, Canada.
- Stenhouse, L. (1980). The study of samples and the study of cases: Presidential address to the Annual Conference of the British Education Research Association, September 1979. *British Educational Research Journal*, 6(1), 1-6.

About the author

Jack Martin is Professor Emeritus of Psychology at Simon Fraser University. His research interests are in the philosophy and history of psychology, social-developmental psychology, educational psychology, narrative, biographical psychology, with particular emphasis on the psychology of selfhood, human agency, and personhood. His books include *The Wiley Handbook of Theoretical and Philosophical Psychology* (2015, Wiley), *The Education of Selves: How Psychology Transformed Students* (2013, Oxford), *Persons: Understanding Psychological Selfhood and Agency* (2010, Springer), *Psychology and the Question of Agency* (2003, SUNY), and *The Psychology of Human Possibility and Constraint* (1999, SUNY).