MATHEMATICS EDUCATION RESEARCH: A Guide for the Research Mathematician

Curtis McKnight
Andy Magid
Teri J. Murphy
Michelynn McKnight

American Mathematical Society
Providence, Rhode Island

CONTENTS

PREFACE

Mathematics education research in undergraduate mathematics has increased significantly in the last decade and shows no signs of abating in the near future. Thus far, this research has often been associated with innovations in curriculum such as calculus reform, statistics education, and the use of computational and graphing technology in instruction. This association has had two consequences that were in some ways unfortunate. First, much of this research arose in efforts to support and defend proposed reforms and innovation. As a result, this entire domain of research has become, for many mathematicians, linked inextricably with reform agendas and demands for change. Second, since much of the research was concerned with changes in courses and teaching methods, much of the research has been in the form of "teaching experiments" in which one form of course organization, teaching methods, or technology use (or all of the above) has been taught in one or more sections of a course and compared, for the most part statistically, with other sections of the same course taught with more traditional organization and methods. While reform may be admirable and teaching experiments may have some value when properly done, it is unfortunate that this should be the main image of mathematics education research for so many research mathematicians.

Mathematics education research, carefully conducted, is something far more fundamental and widely useful than might be implied by its use by the advocates of innovation in undergraduate mathematics education. Most simply, mathematics education research is inquiry by carefully developed research methods aimed at providing evidence about the nature and relationships of many mathematics learning and teaching phenomena. It seeks to clarify the phenomena, illuminate them, explain how they are related to other phenomena, and how this may be related to undergraduate mathematics course organization and teaching. Among the phenomena investigated may be those most germane to proposed reforms, but the phenomena are not limited to those nor are they pursued with methodological bias towards one result or another (no matter what the preferences of those doing the research).

Even with this more general conception of mathematics education research, two major obstacles remain for research mathematicians who wish to evaluate the validity of such research or even to share in carrying it out. First, mathematics is essentially "proof-based." This should not be taken to be saying that all there is to mathematics is proof without room for insight, intuition, concept-building, and conjecture. Rather, the statement is meant to imply that the typical canon of legitimate methods is that a conjecture's proof is linked by correct deductive logic to accepted axioms. In contrast, mathematics educational research cannot proceed by strict logical deduction and seeks instead carefully to build up evidence relevant to answering questions about the phenomena with which it is concerned. That is, even the most careful mathematics education research is necessarily "evidence-based" rather than "proof-based."

Mathematics education research is research using the paradigms and methods of educational research and, more specifically, its applications to mathematics teaching and learning. Even applied to undergraduate mathematics education, this research uses methods and models not typically familiar to research mathematicians. Further, since it involves the application of the methods of an established discipline (mathematics education research), that research can come complete with that discipline's jargon and be as opaque as technical discussions often are to "non-specialists." Even limiting these research methods to a sub-domain (undergraduate mathematics teaching and learning) that is familiar to research mathematicians still may leave those mathematicians as "non-specialists" about the methods involved because the language, the assumptions, and the rationale of the methods are not made clear in each research study (as, for example, would also be true in a piece of mathematics research building on much previous work along a particular research line).

This book is intended to meet three goals in response to the situation outlined above. First, a non-jargon introduction is presented for educational research. Second, more commonly used research methods are surveyed, along with their rationales and assumptions. Third, careful discussions are provided to help research mathematicians read or listen to education research more critically as well as more informedly.

Chapter by Chapter Summary

Chapter 1. Evidenced-based Pedagogy

This introductory chapter introduces the approach of the book, expanding and discussing the rationale sketched above in more detail.

Chapter 2. Recognizing Research

What is often posed as mathematics education research runs the gamut from anecdote to isolated empirical findings to careful research studies. Well-conducted and well-reported research studies have strong common features (clear, significant questions; carefully chosen variables; careful description of the context; a clear statement of methods; sufficient details of analysis to allow its evaluation; careful, not over-generalized results; and the separation of findings from conjecture). This chapter uses this idea to discuss how to approach purported mathematics education research, either published or presented, to gain an overview of its design and form an initial evaluation of how seriously it is to be taken as legitimate research. The chapter also introduces the fundamental distinction between quantitative and qualitative methods that underlies the rest of the book and provides a "site map" for the remaining chapters.

Part I. Quantitative Research

Chapter 3. Critiquing Quantitative Research

All quantitative research may be critiqued on the basis of features common to all such research. The approach of the chapter involves identifying variables, classifying them (dependent, independent, intervening, contextual variables, etc.), and using this understanding of the variables presented to judge how appropriate the research presented is to the questions it seeks to address. The chain of research questions to variables to methods to formal results to stated conclusions is used to suggest an organized approach to evaluating the significance and legitimacy of the evidence developed and conclusions drawn.

Chapter 4. Reliability and Validity in Quantitative Research

Two central issues in all evidence-based research are reliability and validity. Would the research methods used be likely to lead to the same results in similar experiments and studies? Do the methods used change the phenomena studied from what naturally occurs in mathematics teaching and learning or do they investigate them in forms close to what would spontaneously occur ordinarily in mathematics education. An attempt is made to discuss these issues clearly and use that discussion to introduce a small number of terms (face validity, inter-rater reliability, etc.) which a critical reader of educational research is likely to encounter. Finally, an additional basis for evaluating a research study's adequacy is introduced by discussing how to think about threats to its reliability and validity.

Chapter 5. A Survey of Statistical Methods

Statistical methods are frequently used in mathematics education research in an effort to produce hard, credible evidence pertaining to the questions addressed. The chapter presents a brief "taxonomy" of common statistical methods, their strengths and weaknesses, and what to examine in their use. No method can be discussed in detail and this is not a handbook of how to use these methods. Rather, it is an attempt to provide guidance on understanding the approach of each method and allowing critical readers to judge how well and appropriately the methods are used.

Part II. Qualitative Research

Chapter 6. Critiquing Qualitative Research

In recent years, excessive reliance on quantitative research methods and the search for deeper understanding of what is being investigated have lead increasing numbers of mathematics education researchers to use qualitative research methods (case studies, observations, etc.) either alone in some research studies or in combination with quantitative, statistical methods. Qualitative methods are introduced in this chapter and, in the sense of Chapter 3, guidelines are presented for critically reading or listening to research that makes at least some use of qualitative methods.

Chapter 7. Reliability and Validity in Qualitative Research

Many are prone to dismiss what are here called qualitative research methods as producing nothing more than anecdotes or examples. However, when carefully used, established qualitative methods can produce results that are both reliable and valid. The basis for this reliability and validity is discussed in this chapter and used to provide a basis for critically "consuming" research that makes use of qualitative methods.

Chapter 8. A Survey of Qualitative Methods

Many research mathematicians have either a true familiarity or a general sense of familiarity with quantitative methods. Qualitative methods are more alien to them. This chapter introduces a brief "taxonomy" of qualitative methods including observations, video- and audio-taping, analyzing written artifacts (syllabi, student work, etc.), case studies, and so on. The strengths and weaknesses of each are discussed as well as what should be watched for in their use. The potential validity and reliability of each method is discussed.

Part III. Using Mathematics Education Research Appropriately

Chapter 9. Teaching Experiments, Quasi-experimental Research, and Threats to Validity

The most common form of mathematics education research encountered by research mathematicians is likely to be teaching experiments in which experimental methods and innovations are tried in one class along with a presumably carefully controlled other class. Since it is almost impossible for teaching experiments to achieve the randomization assumed in research designs, they are at best "quasi-experimental" research approximating but not completely exemplifying traditional experimental designs. This does not in itself render their findings worthless. However, it warrants a carefully evaluation of the threats to their validity. Since this is such a widely encountered form of educational research, these issues and how to evaluate threats to validity and the legitimacy of results are discussed in this chapter.

Chapter 10. Evaluation, Assessment, and Research

It is now a commonplace of funded projects, as well as of careful local innovations, to build formal evaluation approaches into the design of projects. This evaluation often includes assessment of students and whether their performance in certain learning tasks improves. Since this is a major context in which many research mathematicians may encounter the use of and need for mathematics education research methods, this chapter is devoted to the special issue of the relationships between project evaluation, assessment, and educational research. This is done both to help in evaluating proposed or completed evaluations and as an introduction to how project evaluations may be used to produce research results when they use careful, empirical methods to produce evidence related to research questions of interest.

Chapter 11. Finding Research: the Literature Search

While the tone of much of the above has been that of research mathematicians "encountering" mathematics education research, there are other cases in which they might want to actively seek such research studies, whether to satisfy their own curiosity, to think carefully about issues raised by others, or to deal informedly with proposed changes at their university. This chapter offers a brief introduction to the computerized databases that may be searched for educational research studies relevant to questions about teaching and learning undergraduate mathematics. It discusses the main features of several databases, features of the "search engines" that are readily available for searching those databases, and provides some helpful hints to effective on-line searching.

Chapter 12. From Consumer to Producer

Much of this book has taken the approach of helping research mathematicians to be critical, "enlightened" consumers of mathematics education research. More and more often, some mathematicians find themselves involved or wishing to be involved in carrying out such research themselves. This brief chapter discusses how to organize the information from the preceding chapters into an approach for planning and carrying out a mathematics education research study.

The authors are grateful to the American Mathematical Society for the opportunity for this volume to appear as an AMS publication. In the University of Oklahoma Department of Mathematics, where we (CM, AM, TM) are employed, and where research mathematicians and mathematics education researchers work side by side, both have learned how to learn from each other (especially the former from the latter). Our hope with this volume is to share some of that opportunity with the mathematical community at large.

Curtis McKnight, Andy Magid, Teri Murphy, Michelynn McKnight
Norman, Oklahoma, February 2000

CHAPTER I

EVIDENCE–BASED PEDAGOGY

Mathematicians and Proof

Simply by offering to guide research mathematicians into the world of mathematics education research, we are making many assumptions, not the least of which are that we know who research mathematicians are and why they want, or should want, to know what mathematics education research is. So well start by clarifying these two basic assumptions.

"Research mathematician" is a synonym for "mathematician". We've chosen the redundant adjective to emphasize that the readers we intend to guide are highly skilled in the methods of thinking and speaking about the creation, appreciation, and assessment of new mathematics. They've created their own, officially usually beginning with their Ph.D. thesis, and have continued to do so either on a track towards academic tenure or for a government or industry employer; they regularly look at journal articles and reprints in their field, as well as hearing lectures on new research developments at seminars, colloquia, meetings and conferences; and they scrutinize the work of others while acting as referees, reviewers, and members of hiring, tenure, and promotion committees.

Creation of new mathematics, like creation in arts, humanities, or sciences, uses all sorts of modes of thought: intuition, careful study of key examples, leaps of insight, and chains of informal reasoning about plausible truths. But *demonstration* of new mathematics requires a rigorous proof. This is one of the hardest things for nonmathematicians to grasp about the discipline: illuminating examples, visually useful illustrations, common–sense plausibility, even engaging rhetoric, is simply not a proof. (Incidentally, the hardest thing for nonmathematicians to grasp is often that there can be creation of new mathematics at all!). But of course the reliability of mathematics depends on its theorems being correctly established by a logically valid proof. Consequently, a central part of the training of a mathematician is the development of the ability, or rather abilities, to formulate a logically valid proof and to detect logical flaws in the proposed proofs of others.

Not all mathematicians, or potential mathematicians, are fans of this latter activity, which can lead to the tension filled seminar room where the stressed speaker is being pressured to defend what was a moment before declared "obvious" to sharpshooting colleagues bent on detecting gaps in the speakers proof. But all agree that something is either a proof or it isn't and that what makes it a proof is that every assertion in it is correct. This is a key point, and we repeat it for emphasis: a mathematical proof is all wrong unless it is all right in every step and detail.

Mathematicians know much more than what constitutes a mathematical proof, of course. They also know mathematics, and how to use it to solve problems, not to mention the life they lead outside of the world of mathematics. In that outside world, "proof" means something quite different, and, generally speaking, mathematicians have no trouble deciding when the "all wrong unless it's all right" standard doesn't apply. For example, in criminal trials juries are asked to find the accused innocent unless the case made by the prosecutor is proven "beyond a reasonable doubt". No mathematician, acting in mathematician mode, and thinking about real probabilities, could ever vote to convict on such a standard. Juries in civil trials, by the way, are asked to find for the plaintiff if the "preponderance of the evidence" so suggests. A true mathematician should conclude that standard had been met if the probability that defendant caused the damage is any real number exceeding 0.50.

Mathematicians, of course, serve on juries all the time without being confused by the nonmathematical use of the word "proof". It is a key assumption of the jury system that a citizen with no special training be able to understand and assess the presentation of the case and make a correct decision. Unfortunately we can make no such assumption about presentations in human sciences research. It is our goal in this work to provide the background that enables the mathematician to understand and assess presentations in mathematics education research. Central to that understanding is the distinction between evidence and proof. It's clear from the discussion so far that "evidence" is not proof in the mathematical sense. Exactly what it is, and how to assess it, will occupy much of our attention later in this chapter and indeed throughout the book. But first, we still haven't said why we think mathematicians should want to be able to understand and assess this literature. We do so now.

Mathematics Education Research and Mathematics Education Reform

The majority of research mathematicians, and, we believe, the majority of our readers, are employed in institutions of higher education and spend a considerable fraction of their time and effort teaching. For many, their training as teachers was through an apprenticeship system, in which they served as graduate teaching assistants with varying levels of classroom responsibility as their experience as teachers increased. Some had additional formal training as teachers as graduate students; a somewhat larger number may have participated in specific professional development experiences ("workshops") designed to enhance particular teaching techniques in the course of their subsequent employment. It is unlikely in any of this training or professional development experience, however, that they had any exposure to the research literature in education.

We believe, incidentally, that despite the seemingly casual nature of this training, the system is functional and the instruction delivered by mathematicians engaged in university teaching is quite good overall. We further believe that there is value added to undergraduate mathematics instruction when it is delivered by research mathematicians with Ph.D.s. And we believe that there is room for improvement.

This brings us to a source of potential confusion. In the past decade and a half, efforts have been made to improve undergraduate mathematics instruction through innovation in curriculum, text materials, and teaching methods, including the use of technology. These innovations, which are often referred to as "reform", have been the subject of passionate debate in the mathematics community by both their advocates and their detractors. The member publications of the mathematical organizations have carried a number of articles on this subject, and the meetings of these organizations have included presentation sessions on the subject as well. Many such reports and presentations have consisted of what we will call here "teaching experiments": a form of course organization, teaching method, or technology use (or all of the above) had been used in one or more "reform" sections of a course, and the results in terms of student performance had been noted. Sometimes there has been comparison to a control group of students taught in a traditional course, including the case of historical controls. Almost without exception, the purpose of the "experiment", and the purpose of the report, has been to advocate the change agenda.

This discussion has been healthy for the mathematical community, and has contributed positively in terms of a renewed focus on the importance to mathematicians of improving the quality of undergraduate mathematics instruction. Nonetheless, it may also have had the unfortunate side effect of identifying mathematics education research, in the perception of many research mathematicians, with advocacy for a reform agenda. Mathematics education research, however, is something far more fundamental and widely useful than those whose exposure to the field consists of reports of "teaching experiments" might suspect. We have put the words "teaching experiments" in quotes so far to indicate that what we are describing rarely meets the standards of experiments: comparability of subjects, the isolation of the effect of the intervention, and even simple methodological considerations of statistical power of the experimental design are in general lacking in these investigations and their reports. They do constitute research, and they may have value when properly done, but they rarely rise to the standard of experiments.

Before we depart this topic, we also wish to note that we have benefitted from trying out innovations in curriculum and instruction in our own classrooms, and have ended up adopting a number of aspects of the reform agenda that work for us. We've also tried to help our colleagues locally and nationally by sharing the results of our successes and failures through reports and presentations, and through the preparation of curriculum materials. We also think that having informed access to the literature of mathematics education research is useful to the mathematician, and we are ready to explain why.

Mathematicians and Mathematics Education Research

In our view, mathematics education research is inquiry by carefully developed research methods aimed at providing evidence about the nature and relationships of many mathematics learning and teaching phenomena. It seeks to clarify the phenomena, enlighten them, explain how they are related to other phenomena, and how this may be related to undergraduate mathematics course organization and teaching. Among the phenomena investigated may be those most germane to proposed innovations ("reforms"), but the phenomena are not limited to those nor are they pursued with methodological bias towards one result or another (no matter what the preferences of those doing the research).

Even within this general conception of mathematics education research, two major obstacles remain for research mathematicians who wish to evaluate the validity of such research or perhaps to participate in carrying it out. First, as already noted, mathematics is essentially "proof-based." While this should not be taken to be saying that all there is to mathematics is proof without room for insight, intuition, concept-building and conjecture, the statement is meant to imply that the typical canon of legitimate methods is that a conjecture's proof is linked by correct deductive logic to accepted axioms. In contrast, mathematics educational research cannot proceed by strict logical deduction and seeks instead to build up careful *evidence* relevant to answering questions about the phenomena with which it is concerned. That is, even the most careful mathematics education research must be "evidence-based" rather than "proof-based." We will have much more to say in this chapter and throughout the book about what "evidence" means and how to judge it. For now, we again want to stress that evidence is not proof, as mathematicians understand the latter. Second, mathematics education research is research using the paradigms and methods of educational research and, more specifically, its applications to mathematics teaching and learning. Even applied to undergraduate mathematics education, this research uses methods and models not typically familiar to research mathematicians. Further, since it involves the application of the methods of an established discipline (education research), that research can come complete with the discipline's jargon and be as opaque as any technical discussion is to non-specialists. Even limiting these research methods to a sub-domain (undergraduate mathematics teaching and learning) that is familiar to research mathematicians still may leave those mathematicians as non-specialists as far as the methods involved are concerned, because the language, the assumptions, and the rationale of the methods are not made clear in each research study (as, for example, would also be true in a piece of mathematics research building on much previous work along a particular research line).

So why should the mathematician bother with this somewhat inaccessible literature? There are of course any number of reasons, from intellectual curiosity to the need to evaluate a departmental colleague specializing in education research, but the one we would like to focus on, indeed advocate for, is so that mathematicians can practice evidence–based pedagogy.

Our term is borrowed from the health sciences: a physician who selects treatments based on their effectiveness as established through published trials is said to be practicing evidence–based medicine. This sounds as if it should be obvious standard procedure. But in fact, treatments may be selected because the physician saw it done that way as an intern, because a drug or medical equipment salesperson came by urging adoption of a product, because a colleague has had success with a treatment plan, or because this is the way the physician and others in the area always treat this kind of case. Just because it was selected by one these methods doesn't mean that the treatment is not successful, of course: physicians abandon things that don't work. However, advocates for evidence–based medicine would argue, and we think most observers would agree, that using the results of the scientific literature to formulate treatment plans must overall lead to more treatment successes.

Analogously, a mathematician faced with a teaching assignment can proceed by the way they saw it done as a graduate assistant, could try the latest textbook or graphing calculator innovation that the publishers or manufacturer's representative gave them, borrow and follow a colleague's syllabus, or carry out the assignment the way they always do in the department. Most people do one of these, and most of the time it works out. But just as in the case of the physician, consulting the literature for the evidence of effectiveness should, overall, lead to more success. This is our ideal of evidence–based pedagogy: the instructor consults the mathematics education research literature for evidence of effectiveness in selecting and evaluating curricular materials and pedagogical methods.

Just how to do this consulting the literature is part of what we are explaining in this book. But just as important is understanding what "evidence" means in this context.

Evidence

Evidence, simply put, is fact presented in support of a hypothesis. Of course its not simply that. In the usual model of the scientific method, one has a theory, from which one deduces predictions, and then one collects data from observation or experiment which, if they agree with the predictions, are taken as evidence for the hypothesis that the theory is true. That there must be something more to it is immediately apparent from elementary epistemological conundra like the following: the hypothesis is that all crows are black (if $X \in \{\text{Crows}\}$ then $X \in \{\text{Black Things}\}$). Observing a black crow is evidence for the hypothesis. But the hypothesis is logically equivalent to $X \notin \{\text{Black Things}\}$ implies $X \notin \{\text{Crows}\}$. And observing a yellow canary is evidence for this latter. For that matter, so is observing a green bath towel. Moreover all of these observations are evidence of the same sort for hypotheses like "all pianos have fur".

Besides pointing out that there are philosophical problems in the theory of knowledge, examples like this make clear that there are levels of evidence. Most people are probably intuitive Bayesians on this issue; that is, we start with an *a priori* estimate of the probability that a hypothesis is true, and using the evidence then formulate an *a posteriori* estimate. Of course most people are also intuitive Aristoleans: they understand that the hypothesis is either true or false, and that the probability estimate is some attempt to quantify, on a scale of 0 to 1, of how convinced they are of its truth.

What this means in the context of evidence–based pedagogy is then that we have some question, we want to know what's known about the answer, so we consult the literature. Assume that we find some relevant studies. Now we have some evidence. Is it convincing, or rather, how convincing is it? Much of what we have to say throughout this book is intended to help the reader answer this question. For now, we are going only to present a small, and incomplete, taxonomy of types of education research studies and how convincing their conclusions might be. We will do this with a hypothetical example.

For the purposes of the example, we will speak as if we were considering conducting various types of investigations of hypotheses, while in the evidence–based pedagogy paradigm, what we should be doing is consulting the results of studies.

We will take for granted the point that all undergraduate mathematics majors, regardless of future career plans, should understand the Riemann integral, and that a part of this is understanding that a bounded real function on a closed interval is Riemann integrable if and only if it is continuous off a set of measure zero. We will further assume that the typical student meets this theorem in a first undergraduate course in real analysis for which the prerequisites are primarily Calculus. In their precalculus and Calculus studies, students have primarily dealt with elementary functions (remember these are typical students) and probably have understood them in terms of their analytic expressions and continuous, or piecewise continuous, graphs. They are going to need to understand even these familiar functions in a new way in order to understand the class of integrable functions. How can they get to this point?

There are lots of possible questions and hypotheses embedded in this example. We'll start with the one that purports to describe the understanding of the function concept by students who have completed Calculus. Let's say that we have devised some way to measure and assess this understanding, say by a simple examination or interview (and already this is a big assumption). Now we're ready to measure and assess. This will be a *descriptive* study, and *observational* as opposed to experimental. What population are we interested in here? Is it all students who have completed Calculus? Probably we are focused on just the mathematics majors. Still, we have to decide whether we will measure all mathematics majors who have completed Calculus, or simply a sample. And we have to decide if the sample is to be random, and if so how it is to be randomized. So far, tracking over time is not an issue. We simply want to know how mathematics major who have completed calculus understand real functions, so we observe (measure) some. This is a *cross sectional* observational study, because we are not following a specific cohort over time.

Another hypothesis implicit in our example is that the typical students do expand their concept of function in the process of understanding the theorem on Riemann integrable functions. Again, we suppose we have a way to assess and measure the understanding of the function concept. Suppose our selected students are measured after they complete Calculus, and then followed and measured again after they complete Real Analysis. This is still observational; since the study involves following a group forward through time, it is a *prospective* study. If we instead only measure selected students after they have completed Real Analysis and then used some other device (say looking up their Calculus grades) to see their earlier status, we would have a *retrospective* study.

So far, we are only talking about gathering data. Now let's suppose that what we are actually interested in is how students make the transition from understanding generic functions as being (piecewise) continuous to being integrable. We suspect that this transition is more difficult for students who complete Calculus with strong graphic inclination than those with a weak one. So we decide to compare a group of students who have taken a Calculus course in which they were exposed to a strong emphasis on graphics with a control group who took one with a weak emphasis. One way to do that would be to follow two groups of students from their entry into Calculus, one of which was exposed to the Calculus emphasizing graphics and one of which wasn't, and see how hard it was for the students in each group to master the expanded function concept in Real Analysis. This is still a prospective observational study, but now it is being used to test a hypothesis. We are comparing the difficulty in Real Analysis of the exposed group to the control group. If instead we selected a group of students who had difficulty in Real Analysis (cases) and a group who didn't (controls) and looked at the fraction of each who had taken graphics intensive Calculus, we would also be collecting information about the same hypothesis, but this time from a retrospective observational study, also called a case/control study.

Or we might try to study the above hypothesis by an *experiment*. We would select a group of students, some of whom would be assigned to graphics based Calculus (the intervention group) and some of who would be assigned to low graphics Calculus (the control group) and then all would be measured in their difficulty in understanding integrable functions and the results of the two groups would be compared. To make sure that we were measuring only the effect of the intervention, we would either have to establish that our two groups were comparable in every way or, more typically, assign the students to the groups at random.

Experiments are hard to arrange, among other things because of the involuntary nature of the assignment to the intervention and control groups. We could, instead, just try some things out. For instance, we could decide to switch to a new Real Analysis syllabus that begins with a new unit designed to expose students to functions that were not piecewise continuous. Suppose we do this and discover that over the next few semesters students have an easier time with integrable functions. This is evidence for an intervention (the new unit) but not compared to anything. What we have is a series of instances, or a *case series*. If we just try this one semester, and it works, we have a *case study*.

We want to point out, of course, that our example is completely hypothetical. We think that intensive use of graphics in Calculus, including with the aid of technology, helps students understand the subject better and deeper. And we have not observed any deterioration in student's ability to understand integrable functions attributable to the use of graphics technology in Calculus.

In the hierarchy of convincing evidence for assessing interventions, the results of true experiments have pride of place. Next on the scale would come prospective observational studies, and after them retrospective observational studies. Case series come next, and case studies come last. All, of course, are evidence. In the case study, or the case series, the assertion is that something was done, and some outcome was observed, but without any comparison to what would have happened if it wasn't done, leaving open the possibility that the same outcome would be observed. In the retrospective study, we observe some successes and some failures, and ask what was done previously to each group. We are trying to conclude that the observed outcome is caused by the previous intervention, or lack of it, with the temporal order reversed. It's possible that trying the intervention will not cause the outcome in the future because something else changed. In the prospective study, we are tracking a group of those who experienced the intervention and a group who didn't and observing the outcome. This is cause and effect in the right temporal order, but there still may be differences between the two groups other than treatment that are causing the outcomes. To eliminate that possibility, an experiment is necessary.

Bias

Mathematics education researchers need not be disinterested observers. And just because the researchers have an agenda doesn't mean that the evidence produced by their studies is unusable by others not sharing their orientation. Typically such authors are not shy about disclosing how they hope their results help advance their cause, of course. Similarly, studies that were supported by publishers or technology manufacturers, and that use their products, need not be discarded by readers who prefer other brands of materials or equipment. Moreover, potential conflicts of interest of this later type are almost always prominently disclosed. Tendentiousness of both these sorts is a standard part of the human sciences enterprise, and usually doesn't have any effect on the data collected or its analysis.

There are other ways, however, that mathematics education research studies can be skewed into being misleading, which generally go under the name of *bias*. In this context, "bias" means flaws of data collection or analysis, or of study design, that undercut the basic assumptions of the study. Generally speaking, these flaws have to do with confusions about the representativeness of subjects or measurements. There are three general types of bias we want to identify.

First is *selection bias*. Selection bias occurs any time the procedure for selecting the subjects for the investigation produces a group that have some additional characteristics not typical of the population being studied, characteristics that may have an effect on the hypothesis being investigated. For example, a study of formation of function concepts among Calculus students that selected all of its students to assess from an 8AM section might come up with a group of students who choose early classes because they have afternoon and evening jobs, and consequently have less evening study time than others. An investigation that used only such students but wanted to make conclusions about students who had ample individual study time could come to inaccurate conclusions due to this inadvertent selection bias. While it is up to the investigator to be on guard for selection bias, it is not always detected, and a user of a study should always look to see if it's present. Incidentally, selection bias doesn't necessarily make the study useless. It may only mean that the population on which the study is being conducted needs to be restricted. If the restricted population fits the user's situation, then the conclusions (if any) will apply.

Next is *information bias*. Information bias occurs when study subjects are misclassified. Some of this may be due to simple errors, for instance using an algebra placement test to divide subjects into groups who are to identified as well–prepared or poorly–repared for Calculus, when the test is actually only assessing those who reviewed for it prior to administration. Or it may be due to something more subtle: for instance, in classifying students by whether they have had a graphics intensive Calculus course or not, we may be lumping together a number of levels of graphic technology usage into an undifferentiated group. If for the effect being measured only the highest level of usage has any impact, this sort of classification will disguise that fact. Again, the investigator should have watched out for information bias, but the user of the study should also be aware of the possibility. In particular, if the user had reasons to expect an effect to be present and found a study that implied there was no such effect, the user might ask themselves whether the effect might be masked by information bias.

The third type of bias we want to caution about is *confounding*. Confounding occurs when two groups are being compared that are not really comparable, because of additional factors besides the one(s) being studied that have an effect on the outcome. For instance, consider a study that tried to assess the value of an inquiry based approach to teaching modern algebra by comparing a section taught during the fall semester by traditional methods and a section taught in a summer semester taught with inquiry based methods. It might end up being biased because the fall course drew its enrollment entirely from mathematics majors while the summer course included science and engineering majors as well. In most comparison studies in the human sciences, some confounding is inevitable: it is almost never the case that two groups of individuals are comparable in every aspect. To truly be a confounder, in the sense we are using it here, an aspect should effect the outcome being studied, even when the exposure (cause) being looked at is not present; it has to be associated with the exposure among all those potentially in the study, not just those exhibiting the effect; and it should not just be an intermediate step on the way from cause to effect, which would make it inevitably associated with the effect. Investigators generally try to control for confounding, either in their study design or in their data analysis. For example, if a students major is expected to have some influence on the outcome being measured, both for those who are exposed to the study factor and for those who arent, the investigator might try to assemble groups from various majors both for exposure and control. Or they might simply keep records of the various majors of their subjects and use an appropriate analysis technique which allows them to assess their results after "controlling" for student major. From the user's point of view, however, there may not be much they can do to extract information from a study biased by confounding, except possibly to be on the alert for the presence of a confounding factor.

Theory

Suppose an evidence–based pedagogue has identified a question, gone to the mathematics education research literature, and found one or more research studies that speak to the situation, are bias–free, and are evidence for a causal relationship between doing something and having a desired outcome. Should the methods be adopted? It depends. There are some additional criteria to consider.

First, we should look at the strength of the causal relationship. In a study using statistical methods, for example, just how likely is the result to be due to chance? Obviously "just barely not", by whatever standard the investigators used (typically, just less than one out of twenty), is not as strong evidence as "highly unlikely" (say one in a million).

Next, we should look and see if more cause produces more effect, or more precisely what the functional relation between cause and effect is. Generally speaking, we would expect cause to measurable effect to be a locally monotone function if the causal relationship were real and practically useful.

We would like the cause to consistently be associated with the effect in all, or at least the majority, of the studies we examine. We would also like the cause to be specific to the effect.

And finally, there should be some plausible theoretical explanation for the causal relationship. This is one of the places where the instincts of the research mathematician are often far from the practice of mathematics education research. Obviously a putative causal relationship, no matter how strong, monotone, consistent and specific studies have shown it to be, will only be an artifact if it is inconsistent with known fact. And obviously education researchers undertake studies because they have reasons to believe that there are associations and would like to see if there is evidence for these or not. But there are many (equally convincing) competing theories about how people learn mathematics, and about which materials and methods are most effective. When studies provide evidence for particular causal relationships, it is usually possible to find a theoretical pedagogical basis to explain the cause and effect. The careful user of studies, however, will generally focus on the evidence. It is quite possible that usable studies uncovered in a search for answers to a particular question may offer good evidence, but inconsistent theoretical explanations, for the same causal relationship. On the basis of the evidence, the user could conclude that the relationship is there, although the explanation for it is still open.

Theoretical pedagogy, or what some might term philosophy of mathematics education, is of course an important area of scholarship, and it has received some attention and contributions from research mathematicians. We don't think it has achieved definitive form, and perhaps it never will. We have not tried to summarize any of the current theories in this Guide. Indeed, we hope that regardless of whatever philosophy of mathematics education, however informal it may be, that our readers may be aligned with, they will be able to go the mathematics education research literature, locate possible sources of information, assess the evidence those sources provide, and use the results in their mathematics classrooms. In other words, they can practice evidence–based pedagogy.