Can there be “research in mathematical education”?

Herbert S. Wilf
Department of Mathematics
University of Pennsylvania
Philadelphia, PA 19104-6395

Abstract

We examine a number of papers and a book, all of which have been cited, by people who are knowledgeable in the field, as being good examples of “research in mathematics education.” We find specific serious flaws, indeed fatal flaws, in all of them, so that no conclusions of any interest follow as a result of any of the “research” that is reported in these works. We have found no evidence that the research paradigm, involving test and control groups, randomized trials, etc., which is invaluable in the life sciences, is of any use whatever in studying mathematics education and we urge that it be abandoned, in favor of human-to-human discourse about how we can improve curricula and teaching.

This article has evolved through a few stages. It began as a talk that I gave at George Andrews’s birthday conference at Penn State in the autumn of 1998. Its theme then was that although there appear to be some studies in the literature of mathematics education that support the ideas of cognitive-based teaching and learning of mathematics, in fact these studies are seriously flawed in their methodology, or else lacking in proved conclusions, so that in the final analysis there are not any studies that demonstrate the effectiveness of such instruction.

Since then, other studies have been pointed out to me as possibly having some content, but as we will show below, these also are designed with the most elementary sorts of flaws that

render them totally unable to help us to conclude anything at all. At this writing I don't know of any works in any of the numerous journals of this "subject," that prove anything at all that is of value in mathematics education at any level.

So now this essay is a critical study of various papers that have been written and which have been alleged, by various people, to be good examples of research in mathematics education. They aren't, and it's easy to see why, in each case. These are not "straw men" either, because I didn't choose them. In each case I looked at the paper because somebody with some credentials suggested that I ought to do so. You should look at them too, and see what you think.

Let's put it another way. If I had read 100 papers in this field and had then shown you five or six that were of very low quality, then you might indeed think that, well, every field has some good work and some not so good work, so what else is new? But I didn't. I read five or six papers, not selected by me but chosen by people with knowledge of the field, and every one of them has fatal deficiencies.

To prepare for the original talk I had a look at some recent educational research journals. I found an article entitled "Assessment of a problem-centered mathematics program: third grade," by Wood and Sellers, in the J. for Research in Mathematics Education (a journal of the NCTM) 27 (1996), 337-353. On its first page, this article states that

> "It is currently well acknowledged that the recent NCTM recommendations for reform in mathematics education emphasize a need to change the way mathematics is taught and learned .."

which I can scarcely disagree with. A few lines later we read that

> "Current evidence from existing research projects that were instigated prior to, or coinciding with, the release of the reform documents indicate

> that students in reform-based classes do have significantly better achievement in mathematics than those in traditional instruction."

(Please remember the phrase "do have significantly better achievement" because we'll come back to it later.) So there you have it. The claim is that students do better in classes with the reformed curricula. That sparked my interest, for now I was going to see some studies that would support the NCTM proposals.

There follow three citations to the literature. Hence by looking up those citations it is claimed that we will discover some research-based evidence for the effectiveness of these reform proposals. I want to emphasize that these three papers, which I am about to comment on in some detail, are being cited as the primary support for the contention that NCTM-backed curriculum reform proposals have value in enhancing the mathematical achievement of students.

So let's have a look at these three papers. Now, given the context here, you can probably sense that I am about to find fault with the work in these papers. You might imagine that since I am a mathematician, I am probably going to say that the F-test was used inappropriately when the G-test should have been used, or that some sample size was too small to support some conclusion, or that somebody's kurtosis was excessive, or something technical like that. In fact, the shortcomings of these papers are *colossal*, and no mathematical or statistical training is prerequisite to perceiving them.

The first one of these citations is to a very widely cited paper of Cobb, Wood, Yackell, Nichols, et al, called "Assessment of a problem-centered second-grade mathematics project," and it appeared in the same NCTM journal, the Journal for Research in Mathematics Education **22** (1991), 3—29. I emphasize that this paper is very widely cited.

So let's have a look at it. Its abstract states that

> "Ten second-grade classes participated in a year-long project in which instruction was generally compatible with a socioconstructivist theory of knowledge and recent recommendations of the NCTM. At the end of the school year, the 10 project classes were compared with 8 nonproject classes on a standardized achievement test and on instruments designed to assess students' computational proficiency and conceptual development in arithmetic, their personal goals in mathematics, and their beliefs about reasons for success in mathematics."

Let's see how the experiment was set up.

> "The students in the study attended three schools that contained both project and nonproject classes. The ratios of project to nonproject classes in these schools were 5:2, 3:2, and 2:4. Students within each school were heterogeneously assigned to second-grade classes by the principals on the basis of reading achievement scores. The schools each served an almost exclusively Caucasian student population with a wide range of socioeconomic backgrounds. Ten second-grade teachers volunteered to participate in the project and use the instructional activities. The nonproject teachers used the Addison-Wesley (1987) second-grade textbook as the basis for their mathematics instruction. Both project and nonproject teachers taught mathematics for approximately 45 minutes each day."

Did you notice the word that invalidates this entire "study"? Let me show you once more just the sentence that contains the crippling word: "Ten second-grade teachers volunteered to participate in the project and use the instructional activities." What's the key word? "*volunteered*."

The classes that followed the NCTM reform model, the "test group," in other words, were taught by teachers who *volunteered* to do so. That means that those teachers were the kind of people who are receptive to change and to trying something new. That makes the

teachers who taught the control group, the non-reform curriculum, a group of *non-volunteering* sorts of teachers; just the kind that you might expect to be not so good at inspiring the young with the beauties of mathematics. In other words, the instructors were *self-selected* to match the division into groups; a classical violation of the most elementary principles of the design of such experiments.

So I submit to you that what this elaborate and expensive experiment proved was that teachers who are receptive to new ideas and suggestions are better able to inspire our children. I'm glad we've proved that. But really, I don't think I needed all of that expensive convincing.

The second citation advanced by Wood and Sellers to back their assertion that the NCTM reforms have significantly improved things was to a paper of Carpenter, Fennema, Peterson, Chiang and Loef, in the American Educational Research Journal **26** (1989), 499-531. It is called "Using knowledge of children's mathematics thinking in classroom teaching: an experimental study."

Here, twenty first-grade teachers were assigned randomly to an experimental "treatment." They attended a "month-long workshop in which they studied a research based analysis of children's development of problem solving skills in addition and subtraction." Another twenty teachers were assigned randomly (they got that right!) to a control group. The students were then tested for their achievement using a "TBS Level 7" achievement test, and here are the results: The scores in the test group were 20.95 with a standard deviation of 2.08, and in the control group were 20.05 with a standard deviation of 1.81. *No statistically significant difference in achievement scores was found, between the two groups.*

The authors of the paper agree with this assessment. They don't come right out and say that the main object of the study looked the same in the test and control groups. They do say it in one half of one sentence that begins with the word "although": "Although students in [the test group] classes and students in control teachers' classes did not differ significantly in their performance on the [achievement test], …" The sentence goes on to say that the

students remembered number facts better in the test group. Please now recall the phrase that I asked you to remember for later use: these studies were cited as showing how reform curricula enhance *student achievement,* specifically. Yet the authors say that this one did not show that. Nevertheless their paper was cited for its results on the possible enhancements, via curriculum reform, of students' mathematical achievement .

But there weren't any.

So here we have a study called "Using knowledge of children's mathematics thinking in classroom teaching: an experimental study," which finds that using knowledge of children's mathematics thinking in classroom teaching makes no measurable difference at all in the achievement of the students.

How can it be that if we use "knowledge of children's mathematics thinking" in our teaching then we get no measurable improvement in their learning? Easy. We don't have any knowledge of children's mathematics thinking. My own guess is that there are almost as many different modes of mathematical thinking among children as there are children. I'm happy with that. I wouldn't particularly like to live in a world where children all approach mathematics by thinking about it in the same or similar ways.

Since the authors and I agree that nothing of any significance, statistical or otherwise, to the matter at hand happened here, let's go on to the third and last citation advanced by Wood and Sellers to back their assertion that the NCTM reforms have significantly improved student achievement. This was to a paper of Hiebert and Wierne, in the American Educational Research Journal **30** (1993), 393-425. It is entitled "Instructional tasks, classroom discourse, and students' learning in second-grade arithmetic."

Here the analysis of the claimed results is complicated by the fact that *the authors claim nothing at all*. In their abstract, after a description of exactly what the controlled experiment was that they did, there is only one sentence that announces anything like a conclusion. It reads as follows:

> "The results suggest that relationships between teaching and learning are a function of the instructional environment; different relationships emerged in the alternative classrooms than those that have been reported for more traditional classrooms."

That is a very modest claim indeed. I could not be so boorish as in any way to disagree with it. So I agree. Yes indeed. Relationships between teaching and learning *are* a function of the instructional environment.

To make sure that I hadn't missed anything, I read through the description of the method, the analyses that were done, the results, the discussion, the whole paper, with the proverbial fine toothed comb, expecting to reach a section of "conclusions," or something like that. I did find, near the end, a paragraph that begins with the words "To summarize, …" So let's have a look at that one.

"*To summarize*," it says, "*the data suggest that teaching and learning can be related through the kinds of instructional tasks provided and the nature of the classroom discourse*." Period. End of "to summarize." If I didn't miss something, what that says is that teaching and learning can be related to what goes on in the classroom.

Well, I certainly hope so. But, as Horatio said in Shakespeare's Hamlet (I, v, 125-126),

> "There needs no Ghost, my lord, come from the
> Grave to tell us this."

OK, now it's my turn to summarize.

The paper of Wood and Sellers cites three papers in support of its contention that proposed NCTM reforms can significantly improve students' mathematical achievement. The first one of the studies cited had teachers who *volunteered* to teach the experimental sections, while

non-volunteering teachers taught the control sections, thereby violating one of the most elementary and fundamental principles of randomized trials: *thou shalt not self-select*.

The second one of the papers cited did a very careful experiment and found, by their own admission, no significant differences in achievement between the test and control groups.

The third one of the papers cited found only that *what is learned is related to what goes on in the classroom*, which, it seems to me, doesn't support, or for that matter, invalidate, the NCTM position very much, or indeed affect anything very much.

On such pillars does the edifice of research into curriculum reform rest.

In January of 1999 it was suggested to me that the book of Alan Schoenfeld, "Mathematical Problem Solving" (Academic Press, 1985) might contain some valid conclusions based on some experimental evidence. So let's look at it.

In the "Introduction and Overview" it is stated that Part I "is an attempt to outline and flesh out a theoretical framework for investigating mathematical thinking," while Part II "offers some of the detailed evidence upon which the hypotheses advanced in Part I are based."

So the detailed evidence lies in Part II, and we should look there to find it.

Within that part, the introduction continues, "Chapter 6 describes a small-scale laboratory study," and "Chapter 7 offers the first full-fledged documentation of the results of one of my problem-solving courses."

So we should look at Chapter 7 for the full-fledged documentation. OK, let's.

I find there a description of a study of two groups of students at Hamilton College. The experimental group "consisted of the 11 students who had enrolled in Mathematics 161, `Techniques of Problem Solving,'" while the control group "consisted of eight students with

comparable backgrounds (five with one semester of calculus, three with three semesters) who were recruited from a concurrent course in structured programming."

The book then goes on to measure a number of things that relate to the students' success at solving mathematical problems, and the statistics show a number of differences between the two groups. On page 240 three measures of differences in performance are described, and the differences are said to be "quite substantial." The implication is left, though it is never actually stated, that the differences in performance are related to the difference in the teaching methods that were used on the two groups.

But there is another hypothesis that explains the data equally well: *students who register themselves in a course in "Techniques of Problem Solving" and are then selected for a study about problem solving, will do better than students who register themselves in a course in "structured programming" and are then selected for a study in problem solving.* (!) Surely this is the most elementary type of error that one could make, loading the dice by not matching the test and control groups. Not only not matching them, but actually choosing for the test group students who have self-selected for their interest in the very subject matter that is allegedly being tested by the study.

It would have been interesting to see what would have happened if those in the structured programming course had been the experimental group and those in the problem solving course had been the control group, but the choice was made the other way around.

In October of 2000, another paper was suggested to me as an “excellent example” of Research in Mathematical Education. This is “The influence of semantic content on algorithmic behavior,” by Robert B. Davis and Curtis McKnight, which appeared in the Journal of Mathematical Behavior, **3** (1980), 39—89. This article does not begin with any kind of abstract, so the following summary of its contents is my own.

The authors begin with a nice observation. After interviewing a number (unspecified) of 3rd and 4th grade students, they found none who correctly answered a particular problem in

subtraction of whole numbers, namely “7002-25=?”. They also learned that students who get answers that are egregiously wrong seem not to be able to recognize that something has gone wrong. These students do not use such potentially helpful equipment as analogies with money-changing, Dienes’s MAB blocks, or approximate calculation to do estimates.

The authors then wished to explore in depth the reasons for this failure to use available artillery in recognizing errors, so the next 24 pages of the paper are all about an extensive interview with *one* particular 3rd or 4th grade student named Marcia. In that interview Marcia was asked about every possible step on her road down the primrose path to the wrong answer. She was asked for her reasoning in every step that she took -- she was even asked if she wanted to write in a different color of ink – to no avail. She was resolutely unable to pay up her borrowed digits and get the right answer, despite extensive cajoling, directions and hints. This interview begins on page 51 and ends on page 74.

On page 75 we read the conclusion of this exercise, which is that although students “possess other knowledge that could alert them to the presence of an error … yet it is almost certain that these children will not use this knowledge in this way.” This appears to be the major result of the paper, namely that 3rd and 4th graders do not use their other knowledge, or ideas of approximation, to alert themselves to errors in arithmetic problems.

The same conclusion in virtually the same words was stated before the extensive interviewing of Marcia, so it is unclear what the authors hoped to gain from grilling her. It is also unclear to me how the interviewing of just one 3rd or 4th grader can possible be considered “research” into anything at all. Surely this is a new world’s record for inadequate sample size! However since nothing emerged after the interview with Marcia that was not stated before the interview also, one cannot accuse the authors of extrapolating conclusions from one interview. One can only accuse them of boring me utterly with endless details about a pointless interrogation of one little girl.

The authors state “An Unsolved Question,” namely “whether this phenomenon is essentially inevitable for most children of this age, or whether it is a consequence of an excessively

`algorithmic' (and meaningless) program of instruction." I can think of perhaps a thousand other reasons why these children do not apply other things that they know in order to spot errors in subtraction. Having taught at the college level for many years, I have always been struck by how few students can apply material that they learn in one course to the problems of another course. It seems that people compartmentalize what they learn, and it seems very hard for them (us) to break down those compartment walls and use things that we learn in, say, calculus, to do things in, say, physics. So I think the "unsolved question" is childish in the extreme, the observed phenomenon being neither of the two possibilities that are offered, but instead a complex mixture of about one million human factors that couldn't possibly be analyzed in a meaningful way with questionnaires and sample problems and logic and other irrelevancies like that.

The next section is "Implications for School Curricula," and I shudder at the consequences that might flow from these shaky foundations. Fortunately the authors seem to want only to use MAB blocks and to stress the parallelism between "MAB block representations and our paper and pencil calculations." They reach that decision despite details such as having no evidence that such use of MAB blocks and stressing of parallelism would in any way make a difference in the outcomes of these tests. The only justification they give for wanting to do this to children is that it is "our own practice in the design of curricula," although they do not state why it is their practice.

The final section is "A warning." It seems that the phenomena reported in this paper [and well known to everyone who has ever taught a class ..HW] should raise some disquieting notions about the future of mathematics education in the United States. [Marcia is really quite influential, you see … HW] "If, as we believe, meaning is hard to teach successfully, if meaning is something students are disinclined to use, and if at the same time, *meaning is the only effective foundation for truly powerful algorithmic performance* [Italics theirs ..HW] these "back to basics movements can easily prove harmful.

I must disagree. I think that the phenomena reported in this paper as so "old hat" that they are quite unable to raise any disquieting notions about anything at all. Did I need an almost

50 page paper to convince me that people have difficulty applying lessons that they learned in one place to problems that arise in another? I didn't. I'm convinced. I'm also convinced that not a shred of any activity that one might call research took place anywhere in this paper.

The authors would have made their points much more effectively by writing a thoughtful five page essay describing their thesis. The points that they have to make are good ones. Those points benefit not at all from being forced into a pseudo-research straitjacket. There's no need to vex Marcia for 24 pages. Leave her alone and tell us what you think and what you feel and what you think we might do about it. Give us some sample exercises which you think address the perceived problem directly, and let's try them out. Don't publish in a journal like this one. Groucho Marx once said "I wouldn't want to be a member of any club that would have a guy like me as a member." Well, this paper shouldn't go into any journal that would publish a paper like this. It should be rewritten as a human-to-human essay, devoid of the research trappings, and published in some outlet for essays. It would be much stronger for that. It wouldn't incite people like me to point out why it isn't research. It wouldn't pretend to be. That way we could relax, read it, and think about it – a big improvement.

Now you will have noticed that all I have done here is to recite what happened when I followed a few threads in the literature of research into mathematics education. I found one paper that referred to three papers and in all of them I found no enlightenment. Then I looked into one book that had been suggested as a good model, and found a fatally flawed experimental design. Maybe there are other research efforts out there that do better. Well, maybe there are. I don't claim to have done any research into research. I just pulled on these threads to see where they went, and they unraveled. I am not sanguine about the totality of enlightenment that lies out there.

Am I saying that we should not be receptive to new ideas in curricula? Of course not. I am saying that I don't know of any *research work* in this area that can help us to decide the merits of the proposals, and I seriously doubt that such research is even possible. There are just too many variables to control when we're talking about the study of a complex

interpersonal process like teaching and learning. We therefore have no alternative but to think about these issues for ourselves, and to ask ourselves what conditions do in fact encourage students to learn, to enjoy learning, and to work hard, and how can we, in fact, improve the quality of our teaching.

Finally, I want to emphasize that I am not an enemy of curriculum reform. I have been teaching for more than thirty years, and have reformed my own curriculum many times. I have written quite often about some of my own reforms, in case there is anything in them that might be of value to others. I value very highly the anecdotes and stories that fine human beings and gifted and sensitive teachers tell, about what worked for them in their classrooms. True reforms can be proposed and debated without the necessity of doing double-blind studies.

I am very dubious about the ability of reformers to *prove* that their proposed tinkering is of any universal value. It is possible that somewhere, someplace, somebody will do an intelligently designed experiment that will show that one of these proposals has some value. I haven't seen one yet, but it's possible. I'm waiting.

This article is dedicated, with respect and affection, to George Andrews, without whose example I probably never would have gotten interested in this subject – and I probably would have been better off as a result.