EVALUATION OF METHODS IN MATHEMATICS EDUCATION

FRANK QUINN

Abstract. Some methodologies in education research and curriculum development seem almost designed to generate spurious findings and discourage deeper insights. As a result the current emphasis on “research–validated” methods is more likely to block good ideas than to promote them.

Contents

1. Introduction 1
2. Missing Criteria 2
2.1. Resources 2
2.2. Student effort 2
2.3. Procrastination 4
3. Process as a criterion 5
3.1. Multimedia 5
3.2. Technology as a goal 5
3.3. Trendy methods 6
3.4. NCTM Standards 6
4. Outcome measurement 6
4.1. Student variation and statistics 7
4.2. Goal selection 9
5. Teaching, learning and errors 12
5.1. Teacher–centered education 13
5.2. Error correction 13
5.3. Repetition 14
6. Summary 15

1. Introduction

The research community devoted to improving the learning of mathematics has grown significantly in size and professionalism but it has not grown in success. In this article we suggest reasons for this, and in particular argue that the way new ideas are developed and evaluated has channeled the enterprise in unproductive directions.

Many of these conclusions come from work in a large computer–learning facility (over 6,000 students, 500 computers) developing computer–based courses, computer–tested classroom courses, and computer–lab additions to traditional courses. The big issues came not from computers per se but from the environments they

Date: March 2006, revised October 2007.
provide. Traditional classrooms herd students down a single path and provide many tools to keep them on it. Computer environments necessarily give students more control. Factors that are locked together and invisible in classrooms come apart and must be understood separately. The author has spent most of the last decade trying to understand these new learning environments. Experience from 30 years of traditional teaching turned out to be a minefield of preconceptions rather than a useful guide.

This article describes flaws that seem to be common in math education research and curriculum development. §2 concerns important factors omitted from evaluations; §3 lists dangers of evaluation on the basis of method rather than outcomes while §4 describes dangers of incautious use of statistics when outcomes are analyzed. Finally, §5 concerns problems caused or obscured by a focus on teaching rather than learning.

2. Missing Criteria

Here we describe factors that play crucial roles in success or failure of educational programs but are not considered in evaluations.

2.1. Resources. The main reason “good” educational experiments flop in practice has to do with resources. Education has a limited budget and the task of educators is to do the best they can with the means provided. Since educators have no influence on resources there is generally no point in thinking about them. But this is a dangerous blind spot: new methods that require more resources are nonstarters, and neglecting costs in evaluations will miss this.

For instance imagine a new way to teach fractions is shown to significantly improve outcomes. Teachers are urged to adopt the method and parents and administrators are led to expect that great things can be accomplished. But as is often the case the improvements were accomplished in intensive sessions with fewer than fifteen students. There is not much hope the method would work in a typical large attention-span–challenged class, and no particular reason to think that lavishing such resources on other methods would not work as well. The high expectations are unrealistic and attempts to widely implement the method will lead to (yet another) failure.

If resource costs were tracked and factored into the evaluation then a small-scale study like the example above would be considered at best a pilot project needing large-scale field testing. Investigators should develop indicators for successful scaling, or do medium-scale trials before considering the project complete. This would be harder than current practice and there would be fewer “successful” small projects, but long-term impacts would be greater.

2.2. Student effort. Student effort—and a student’s willingness to expend effort—seems to be a crucial factor. This is an unexpected conclusion and we have no suggestion as to how these might be measured, but we describe the problem.

2.2.1. A behavior model. We offer a simple model to explain certain patterns in student behavior. This can at best be a rough guide with myriad exceptions, and in a later section we argue that assuming students have common features can be dangerous. Nonetheless the model encapsulates a lot of painful experience and offers insight into some difficulties in educational research.
The model is this: students see a grade as their objective in a course. They come to the course with a target grade and an “effort budget” they are willing to expend. They work until they either reach the target grade or expend their effort budget. If they hit the target grade first then they quit and get extra free time. If they hit the effort limit first then they quit and accept a lower grade.

Put another way, it is rare for a student to either put in extraordinary effort to get a good grade or, if a subject is easier than expected, keep working after reaching a target. Both do occur of course, but not often enough to invalidate the model as a predictor of bulk behavior.

This model has both explanatory and predictive value, and we explore it below.

2.2.2. Standards versus success. It is well–known and obvious that performance of successful students can be improved simply by raising standards for grades. The drawback is that fewer students will be successful. The model suggests that students who quit because they reached their target grade will work more if standards for that grade are increased. Students who reached the limit of their effort budget will accept lower grades. Educators try to balance standards and failure rates, although views on the proper balance has changed quite a lot over time.

Pressure to avoid dropouts and high failure rates has driven a lowering of standards. Most students hit their target grade before their effort maximum. This has consequences at the next level: not only have lowered standards left students less well-prepared, but ease of success has led them to revise their effort budgets downward.

Most students are capable of, and willing to do, better work. It follows that having several tracks with different standards could improve outcomes, particularly at the upper end where our system is weakest. However the argument above suggests track selection should be based on willingness to work rather than past performance or innate ability. It is certainly unclear how this might be measured but it is a worthwhile research topic since the potential payoff is enormous. Also tracking by ability or performance is politically unpopular, and tracking by willingness to work (if it can be measured) might be more acceptable.

2.2.3. Equivalent outcomes. A frequent experience with full-scale trials is that different methods have statistically equivalent educational outcomes. By this measure there is no basis for preferring one to any other. However the behavior model suggests that there is a reason outcomes are similar and that some methods may have benefits that are not being measured.

Suppose two method are being compared. One enables students to achieve a given level of mastery with less effort, but the trial is “fair” in the sense that fixed achievements are required for a given grade. According to the model, students who are already reaching their target grade won’t do better but will benefit through reduced effort. Since most students are in this group we expect very little overall outcome improvement. Thus equivalence of outcomes should be an expected consequence of student behavior, not a reflection on program quality. If we want a clear indicator of program value we apparently must figure out how to measure student effort.

The end objective is still better outcomes. In this view this would be accomplished in two steps. First compare methods but look for reduced effort (if it can be measured) rather than better outcomes. Then, after implementing the more
efficient method, increase standards to bring effort requirements back to previous levels.

2.2.4. Work ethic. It is obvious that students’ willingness to work is a major determinant of outcomes. The arguments above suggest that as a side effect it also tends to hide or distort other factors. Is this a fact of life, or does it just emphasize the old point that part of a teacher’s job is to engage students and get them to work?

Sadly, work ethic is one of the things eroding from the American character. Students are increasingly unwilling to work and harder to engage. On the whole teachers will be unable to slow this decline let alone reverse it. There have been examples of teachers who dramatically inspired their students, but they are rare enough to become subjects of major motion pictures. Blaming teachers is a prescription for failure.

2.3. Procrastination. Everyone knows procrastination is a problem and traditional courses are packed with preemptive measures against it: constant checked homework; frequent quizzes; lots of major tests; and generally so structured that students who aren’t working can be identified and hassled. On the other hand no one seems to measure procrastination directly and there has been little thought about how it might effect new approaches to education. The problem described here may seem obvious but recognizing it was a surprise outcome of years of data mining in a relatively unstructured course: the only measure that correlated strongly with failure turned out to be a proxy for procrastination.

Many new approaches to education lack traditional defenses against procrastination. For instance a student might be allowed to choose from a variety of tools to accomplish a task, rather than be forced to do it a particular way. But the consequence of not checking use of a particular tool is that it is hard to know if any tool is being used. This problem can be acute in computer– or web–based assignments where—for better or worse—students must play a more active role in the learning process. Procrastinators fare poorly in such environments.

For another example consider the common complaint that most homework is pointless busywork. An alternative when there is a well-defined task (e.g. a kind of math problem) is to provide plentiful examples, give instructions to work examples until they can be done reliably, and a deadline. The ability to quit when ready provides motivation and a payoff for fast and accurate learning, and many students respond well. Procrastinators tend to put it off until too late.

Generally we can expect that any increased reliance on student initiative or reduction of lockstep control will show mixed results: non-procrastinators may do better but procrastinators will do more poorly. This is a problem we are stuck with: chronic procrastination is either a character trait or so difficult to unlearn that it might as well be. It may be a limiting factor in how much choice or control can be given to students. Alternatively it might be possible to identify chronic procrastinators and provide them with a more supportive (i.e. constricted) environment.

Finally we contrast procrastination with the limited work budget problem discussed above. In a fixed environment the distinction is not useful: students who wait until they don’t have time to do the work may as well not be willing to do it at all. The difference becomes important when the environment changes. Giving
students more control should benefit non–procrastinators with low work budgets but be counterproductive for procrastinators.

3. Process as a criterion

Educational approaches are often evaluated on the basis of methods used rather than outcomes. This is easier and faster, and is a reasonable proxy for outcomes if similar methods have been carefully evaluated in similar areas and have had good outcomes. It is an appropriate way to design new approaches as a starting point for development and testing. Unfortunately process evaluations are misused more often than not. It is particularly dangerous to have an ideological attachment to a Good Thing and “know” it will improve any program.

3.1. Multimedia. Video and animations are considered a Good Thing. They have wide and satisfactory use in some subjects, and any proposal that includes them gets extra credit. However the successful uses are in subjects with low expectations for testable outcomes. The generalization that these are Good Things in any area has turned out to be false: they are much less unsuccessful in areas that require concentration and have high testable outcomes.

Our students are very accomplished spectators. The entertainment industry has shortened their attention span and trained them to suspend critical thinking while in spectator mode. The advertising industry has forced them to avoid learning while in spectator mode. Consequently spectator mode is an enemy of serious learning. Anything that triggers it, including almost anything that moves on a screen, is likely to seriously degrade math learning.

3.2. Technology as a goal. A common process goal is that a course should “use technology.” The process has been taken as the outcome goal as well, and there are no educational outcomes to be assessed. This guarantees positive evaluation of the course, but probably also guarantees that nothing of real value will come of it and any bad consequences will be overlooked. Some examples:

- Internet use is a common process goal. The justification offered is “it is important for students learn to use the tools of the information age.” This might make sense if the goal was “learn to use the tools well.” The internet has a low signal to noise ratio and the ability to filter and critically assess information would indeed be valuable. Students are not learning this, and will not as long as use alone is the goal. It is true that even professionals have difficulty with critical assessment, and an attempt to teach school children these skills would almost certainly fail an honest evaluation. But the difficulty of finding realistic outcome goals does not justify dodging the issue.

- Computer (as opposed to calculator) use in math is sometimes taken as a goal with justification “computers have transformed real-life use of mathematics and we should prepare our students for this.” However computers have made math more powerful, not easier. In fact effective computer use requires quite a bit more discipline and sophistication than standard by-hand work. Computers can solve many standard problems in a few keystrokes. But learning keystrokes instead of tedious hand methods puts students further away rather than closer to the sophistication needed for effective computer work.
Visualization is a particularly attractive use of computers and a prominent feature of programs with “computer use” as a goal. It seems to have few benefits beyond pretty pictures. At the high end, an NSF-funded institute was established to determine if direct visualization could be a useful tool in mathematical research. The answer was “no” and when the institute was unable to find a more productive focus it was disbanded. At lower levels the experience is that people have to know what they are looking at before they can see it. In mathematics, at least, visualization seems to have limited use as a primary learning tool, but this may not become apparent until evaluation criteria graduate from any use to effective use.

A problem common to these examples is that the error-correction part of learning has been completely omitted. See the final section for further discussion.

3.3. Trendy methods. “Discovery”, “Reform” and “Standards-based” methods seem to be popular now. The danger here is that exciting ideas, sometimes demonstrated in pilot projects but not tested on large scales may be prematurely adopted. In some cases there have been large-scale adaptations but such extensive changes were needed that the name is the main similarity to the pilot program. Or the materials were used in a course with different methods. All too often these modifications are overlooked and success is taken as validating the original vision. Even though they have actually failed to scale, the methods are used as the basis for process evaluations.

3.4. NCTM Standards. The most remarkable example of the use of process standards is the evaluation of math curricula according to how well they conform to the 1989 National Council of Teachers of Mathematics (NCTM) standards. This has been so widespread that curricula and standards documents nationwide have been profoundly influenced.

Use of process standards as a proxy for outcomes is reasonable when the processes have been shown to have good outcomes. Thus if the NCTM standards had been a distillation of best current practice then a push for general conformity would have made sense. However large parts of the NCTM document were a bold attempt to chart the way to the future rather than a distillation of the past. It was a research agenda rather than a finished product, and pushing for implementation before large-scale evaluation was a procedural error.

These standards have been at least partially implemented on the widest possible scale over the last decade and a half. For better or worse, a large-scale evaluation is now possible. Many outcomes, for instance degree of preparation for university work, have declined significantly. By presenting it as a finished product the promoters raised the stakes and narrowed the outcome to pass/fail. By the rules they themselves have established it seems to be a failure and continued promotion has triggered a backlash. Unfortunately the process of correcting large-scale failure may also sweep away any potential value as a research agenda.

4. Outcome measurement

Standard protocols for experimental design and assessment have developed as educational research has grown as a discipline. The most sophisticated have taken clinical trials in medicine as a model. This seems appropriate since both medicine and education have to deal with the worst possible experimental subjects, people.
The ideal educational trial begins with a method to be tested and specific outcomes (e.g., question types) that the method is expected to influence. To avoid questions about “changing the target after the arrow is shot” there are usually advance decisions about how the data is to be analyzed and how various outcomes will be interpreted. The trial itself involves two groups of students, one using the new method and one with a standard or control method. Data is gathered. Frequently there is an analysis of student characteristics to check for bias in the division into groups. Finally there is a statistical analysis of the outcomes. The care and sophistication of the statistical analysis is often taken as a key indicator of the value of the study, so this analysis is often quite elaborate.

It is unfortunate that this tidy experimental design will almost never give useful conclusions. Even if it worked exactly as expected the outcomes would be problematic because crucial parameters have been neglected, as explained in the first section of this article. But the experimental design itself is unreliable. The medical community discovered and adjusted for this long ago. The educational community needs to do so as well.

One conclusion is that educational research is still in its infancy. Recent attempts to be more “scientific” have not actually made the field more mature and effective.

4.1. **Student variation and statistics.** Statistics only gives an accurate picture when all students will be effected in roughly the same way: “one size fits all.” This is rarely the case, and we describe instances where modest overall improvement resulted from big improvement in one subpopulation of students canceling a big decline in another. The medical analog is the strangely belated realization that males are different from females and that there are environmental and genetic differences between ethnic groups.

The point made above is that a different approach to data analysis is needed when there are subpopulations with different educational needs or responses. It is also unclear how the data would be used. Rejecting methods that favor one subpopulation over another would probably leave us with no educational program at all. Offering different methods to serve different subpopulations might be effective but would require great care in how people are assigned to, or allowed to choose, different methods.

4.1.1. **Modes of thinking.** People learn through three main channels: visual, auditory and kinetic/tactile. They tend to have dominant learning channels just as they have dominant hands (left or right). At one time this was well known and a principle topic in educational research. The awareness has all but vanished, possibly because it conflicts with the homogeneity hypotheses needed for statistical analysis. The fact has not vanished however, and educational methods that emphasize one channel will still favor one of these groups over the others.

For example not long ago reading was taught largely visually through pattern recognition of letters and words. Proponents of this approach were either unaware of the needs of auditory and kinetic/tactile students or expected them to adapt much like left-handed students were once required to write with their right hands. The program may have been effective in some average way but outcomes for primarily–auditory students were unsatisfactory. There were significant numbers of high–school students with very weak reading skills. This drove a resurgence of phonics
approaches that favor these students. If phonics turn out to be unsatisfactory for primarily–visual students we may see a swing back to visual methods.

At this point it might seem obvious that the solution is to offer both approaches and let students use the one that works best for them. This also has problems, with resources and placement for instance, but the show–stopper is the conflict with dominant dogma.

How do strongly kinetic/tactile students fit into the various reading programs? Generally poorly. Some adapt and go on to become surgeons, or in the author’s case, a topologist. In the past many who could not function “right-handed” gravitated to trade schools or apprenticeship programs, or dropped out to work. These alternatives are now considered “left behind” and the students are retained in standard school programs. Consequently we expect to still see significant numbers of high-school students with weak reading skills. Any group likely to have kinetic/tactile orientations, athletes for instance, should be heavily represented in this group. Maybe athletes are not dumb after all, but are just not served by current programs.

Learning-channel differences may pose serious problems for web or computer-based educational materials. Current materials are intensely visual. Does this render them inaccessible to significant numbers of students with auditory or tactile primary learning channels? Current analytical techniques can’t even see the question.

4.1.2. Placebo effects. Small-scale trials almost always have positive outcomes even when large-scale trails with the same methods are unsuccessful. Students do better in small groups and they respond to extra attention and the instructor’s expectation that something good should happen. The medical analog is the placebo effect and it regularly swamps information about the drug or procedure being tested. The medical solution is the double–blind trial, but this is rarely feasible in education.

If small trials have predictable and meaningless evaluations do they have any point? Perhaps they should be formative rather than summative. Instead of conclusions like “the method works” perhaps “with the following modifications we feel the method is ready for large–scale trials.” Perhaps the problem is not the use of statistics but insufficient caution, wisdom, and humility about what the numbers tell us.

4.1.3. Cultural and behavioral bias. “Avoid cultural bias” is usually taken to mean “replace ‘cow’ and ‘chicken’ with ‘dog’ and ‘pidgeon’ because the population is now largely urban.” Or “be sure ‘guns’ and ‘dolls’ are evenly represented in word problems.” Avoiding behavioral bias takes forms like “behavioral difference are normal and teachers should deal with it.” However some biases are not education–neutral and avoiding or ignoring them can compromise the validity of a study.

There are subcultures that highly value education and whose students are willing and well-prepared, and subcultures that find education irrelevant and demeaning and whose students are unwilling if not disruptive. There are students who need medication before they can control their behavior. Does “avoid bias” mean “do not report the effects of disruptive students”? Or worse, does it mean “do not report that disruptive students were excluded from the experimental group”? In any event there is essentially no discussion of disruption in the math–ed literature even though math is probably one of the most vulnerable subjects.
Disruptive students degrade any educational environment and will completely defeat some approaches. Other cultural or individual differences may impose other limits. It is not clear what, if anything, could be done about this. The point here is that we cannot know the real potential or limits of an educational method if these factors are ignored, and current methods ignore them.

4.1.4. Sample size. Many trials reported in the education literature are small, with as few as a dozen students. The hope is that sense can be extracted through careful statistical analysis. This would be doubtful in a physics experiment with uniform particles and is silly when dealing with people. How big should a statistically analyzed trial be?

The author analyzed a multi-section college calculus course with enrollment between 900 and 1400 per semester, divided roughly evenly between two teaching methods and tested with a common final exam. This was done for six semesters with a total of around six thousand students. Amazing variation was seen. Three subpopulations with different characteristics were identified and there were differences due to class size and teacher effects. However most of the variation remains mysterious and considerably exceeds what would be expected from random distribution. For instance one class of over a hundred students had an unusual outcome pattern, significantly different from other classes with the same teacher or the same size. There are group effects such as an attitude—good or bad—“infecting” a class, but we are wary of this as an explanation for such large differences in such large classes. We do not have an explanation, much less a way to anticipate or correct for this.

The conclusion is that there is far more variation in educational trials than would be expected if the underlying assumptions of statistical analysis were valid. A thousand students may not be enough to ensure reproducible results.

4.1.5. Limited imagination. The final criticism of the standard measurement protocol concerns deciding beforehand what to measure and what it should mean. It is always hard to find something one is not looking for but this practice makes it impossible. Education is too complex and human imagination too limited for this to be acceptable. Most of what the author has learned about education—nearly everything in this article—was originally well beyond his imagination. It was not clever prediction borne out by trials, but ignorance slowly and often reluctantly dispelled by confrontation with data.

The alternative to prior decisions is data mining after the fact. The problem with this is artifacts. Randomly comparing lots of things always produces coincidences, so lots of unexpected connections will emerge but most of them are bogus. Further trials or analysis are required to eliminate bogus conclusions and sharpen real ones. This is much more trouble than running data through a statistical program but it may be necessary for real progress.

4.2. Goal selection. Returning to the theme of problems with measuring outcomes, we consider how goals are selected. Courses and lessons are part of an intricate whole, not free-standing entities. Goals that make sense in a limited context may either advance or undermine work at later levels, and bad choices can give great immediate outcomes but greater damage later on. Mathematics is particularly highly interconnected and vulnerable to goal-selection errors, and we describe some common ones.
4.2.1. *Calculators.* Memorizing the multiplication table is a pain and long multiplication and division are dreary. Calculators offer relief. If “accurate arithmetic” is taken as a goal then calculators are a winner. If student joy is factored in then calculators look like the best things since sliced bread. Unfortunately the view from the college level is that calculator-trained students often have significantly weaker number sense and other deficits, described below. Calculator use urgently needs to be reconsidered in spite of the glorious short-term outcomes.

Does this mean we should go back to multiplication tables? Not necessarily. For instance calculators now use keystrokes and connect directly to the motor/tactile learning channel. Perhaps tactile thinking is bad with numbers. Many students learn the multiplication table by verbal repetition so perhaps we need a connection to the auditory channel. If so then calculators with verbal data entry might solve the problem. Or if an extended expression could be entered and visually checked and edited before execution perhaps it would connect with the visual channel.

Ironically there may be problems due to an insufficient connection to the motor/tactile channel. Graphing calculators give students quick and accurate access to graphs of functions. They see these graphs many more times than students once did, and become adept at picking out a particular graph from alternatives on a test. On the other hand they have never drawn these curves with a pencil. When they get to multivariable calculus and have to sketch solids, or work out regions of integration, they cannot draw graphs. They also cannot articulate qualitative features (e.g. an exponential function swoops up really fast). Does this mean we should give up graphing calculators? Not necessarily. It may be sufficient to require students to draw a picture as part of coursework and testing.

The points are that solutions at one level may lead to problems downstream; and that seeing this and finding long-term solutions might require thinking (or following data) far outside our preconceptions. If we cannot rise to this challenge then we probably should return to multiplication tables, at least for students we want to be capable of pursuing technical careers.

4.2.2. *Over-simplification.* Over-simplification of problems used to train and test students at one level can cause problems later. We describe three examples.

- Multiplying polynomials or other compound expressions is a standard task. When this is first encountered the focus is on the simplest case: two binomials, \((a + b)(x + y)\). It is common for teachers to introduce the mnemonic “FOIL” for the algorithm in this case. This increases speed and accuracy with binomials but formulates it in a way that does not generalize to larger expressions, and these students often have trouble multiplying trinomials. Describing the process in terms of associativity and distributivity may require more practice and be slower but it would make the step up to bigger problems completely routine.

- Finding roots of a quadratic corresponds to factoring it as \((x - r)(x - s)\). For simplicity this is usually illustrated and tested with quadratics with integer roots. But most quadratics do not have integer roots, and students taught this way often have trouble dealing with these when they come up in later courses. This is a much bigger problem than might be apparent from the K-12 perspective. Many methods and applications of mathematics require solving for a variable. Quadratics are one of the very few families of equations that students can easily solve and so are heavily used in examples
and problems in college courses. Consequently any student who has trouble with quadratics is at a serious disadvantage.

- At a higher level, most high-school calculus courses are oriented toward preparation for the AP calculus test. Problems on this test are simplified and routine, so the course goal is to deal quickly and accurately with routine problems. Lots of mnemonics and tricks are used and the whole thing is rather mechanical. High school teachers, and more to the point AP calculus test designers, probably do not know that quite a lot of a science/engineering college calculus course is devoted to getting students to unlearn some of this.

4.2.3. Symbols and numbers. In K-12 work in recent years there has been a substantial increase in decimal numerical tasks and a corresponding decrease in symbolic, integer, or rational problems. Calculators drive some of this. Teachers may believe that numbers are the real goal; symbols are just placeholders for numbers, and now that we have calculators we can do the real thing. Students can work with circles of radius 5.687 rather than “r”. Or a fraction like $\frac{2}{3}$ is a frustrated division, and we can now carry it out to get a “real” answer., $\frac{2}{3} = 0.6666$. Perhaps using numbers is supposed to convince students that the problems have “real-life” significance. Maybe teachers have such an investment in developing calculator skills that they want to use these at any opportunity. In addition to any of this, teachers are also responding to pressure from high-stakes tests as we describe below. In any case calculator use has blossomed.

In college courses we now see students who have trouble dealing with problems when the answer is not a number. They can handle circles with radius 5.687 but not with radius “r”. They have trouble with expressions with two or more symbols, and generally have weak symbolic manipulation skills.

The connection between calculators and weak symbolic skills is this: fractions and “numbers” like $\pi$ and $\sqrt{2}$ are more than half-way to being symbols in the way they are handled. People who have learned to deal with $3 + \sqrt{2}$ as a root of a quadratic can routinely deal with $3 + r$. People who work with 4.414 as the root see $3 + r$ as a completely different thing. The painful algorithms used to do long division and multiplication are the same algorithms used in multiplication and division of polynomials, so the step to symbols is a minor one for students who do arithmetic by hand. It is a whole new—and complicated—world for calculator users.

Goal selection on high-stakes K-12 tests has also contributed to the decline in abstract thinking. Most tests are multiple choice, and answers are mostly numerical. Numerical answers are partly a matter of convenience for test developers. Symbols in answers provide clues: $\pi r^2$ is identifiable as the area of a circle of radius $r$, while 8.5 is not obviously the area of a circle of radius 2.7. Numerical answers are therefore an easy way to keep students from identifying the correct answer without working the problem. However this has consequences. It has become the primary goal of many courses to prepare students for these tests, and they therefore emphasize numbers over symbols.

We have not argued that either calculators or high-stakes tests are inherently evil, only that they are very powerful and may unintentionally have bad consequences.

4.2.4. The College/K-12 divide. We have observed that goal changes may lead to short-term success and long-term failure. If both are located in K-12, or both at
the college level, then there is an educational research community that should, in principle, notice and correct the problem. However the examples above are of changes in K-12 that cause problems at the college level, and there is almost no communication across that divide. We briefly describe the situation.

The “new math” debacle of the 1960s drove a wedge between the K–12 and university communities. Since then research in K–12 has become quite professional (whether or not it is on target) and the leadership has been focused and effective. College preparation is a big part of the job but they are completely confident they can accomplish it without much input from college teachers. Also, the input offered seems to be largely unsupported personal opinion. These opinions (e.g., calculators are causing problems) often conflict with articles of faith (calculators are Good) so they are seen as Wrong as well as unsupported.

Communication requires a receiver and a transmitter, and at the K-12 end the signal sounds like noise and the receiver is turned off. The problem at the college end is that there is no transmitter. More precisely, there is no mechanism for collating and sharpening individual concerns to arrive at a “conclusion of the community” let alone any good way to present such a thing to others.

The college community certainly has keen awareness of shortcomings in school math preparation but for them the important question is how to deal with it. There are strong opinions on “how did it get this way?” but these opinions are rarely carefully thought out; there is no well-developed educational research community that might extract a useful signal from the noise; and there is no leadership that might organize some other way of getting this done. The consequence is that “input” from the college community really is rarely more than individual opinion. Mathematicians certainly know that mathematical ideas must be tested with care and most will be wrong. If they approached educational ideas the same way their individual opinions might be pretty good. Unfortunately they seem to be wrong as often as anyone else.

The lack of communication is a serious problem. Here is an analog: many college math courses prepare students for work in other subjects. The needs of these subjects provide an anchor for content. Material cannot be weakened or omitted (too much, anyway) just because it does not fit well into a new educational approach because this evokes negative feedback. In principle preparation for college work should provide the same sort of content anchor for K–12 math programs. In practice the lack of communication keeps this from happening.

The conclusion is that we have no collective mechanism for dealing with problems that straddle the K-12 – college divide, and prospects for one developing any time soon are gloomy.

5. Teaching, Learning and Errors

Some problems seem to be due to, or at least hidden by, an increasing focus on teaching rather than learning. The focus itself is understandable: growing pressure for results puts attention on things that can be directly influenced, mainly teaching. Most education researchers are located in teacher preparation programs and see development of teaching techniques as their mission. The very phrase “teacher preparation” invites a focus on teaching. However in the end learning is the objective and teaching is only effective if it supports learning.
5.1. **Teacher–centered education.** When teachers are considered the main actors, students tend to be regarded as essentially all the same, if not “blank slates”. The central problem is taken to be getting students to conform to the teacher’s direction and expectations. This vision has difficulty accommodating a variety of learning styles (c.f. the discussion of tactile, visual and auditory styles above) and has many other drawbacks.

Emphasis on teaching also locates responsibility in a problematic way. Teachers, as the main actors, are responsible if students don’t learn and students are absolved of accountability. In this view the way to better performance is more pressure on teachers. This degrades the attractiveness of the profession and chases away teachers whether it improves student performance or not.

5.2. **Error correction.** The most serious problem in teacher-oriented education concerns the way errors are handled. The problem was revealed by study of new educational environments in which teachers play smaller roles, or even without a teacher in the traditional sense. Teachers are clearly not the main actors. A small amount of human mentoring is vital, but this is not teaching and experienced teachers often have difficulty doing it effectively. Mentoring is focused on learning, not teaching.

We describe the role of error correction in learning, and why it causes trouble. The context is that people look for, and find, patterns in their experiences. The usual explanation is that this developed as a survival skill in dangerous situations. In any case “natural learning” is a strong and largely innate part of our intelligence.

The problem is with mistakes. Any single person’s experience will have coincidences and bogus patterns and the natural learning drawn from these are wrong. But critical thinking is apparently not a survival skill: our error-correction abilities are much more primitive than the natural–learning ability. Superstitions are born easily and are notoriously hard to root out. Effective error-correction must be learned.

Learning divides roughly into corresponding stages: first getting information and seeing patterns; and then error correction in the patterns seen. Teacher-oriented education focuses on information delivery. This is the easiest part of the task and is often mechanical enough to be done by computers. Diagnosis and correction of errors is more subtle and for the foreseeable future will depend on teachers or mentors.

In brief: the really essential role of teachers is not information delivery, but diagnosis and correction of errors. This is also the hardest part of independent learning, so the best way a teacher can help a student “learn how to learn” is to be clear and deliberate about error correction.

We illustrate how this plays out in practice with example responses to a student making mistakes in a math problem:

“This is wrong. I’ll watch while you go through your work; let me know if you spot the error and I will watch for it too.”

The student is the main actor. It is clear that a mistake has been made and must be corrected, and that the student has significant responsibility for the correction. The student learns to find errors as well as getting the specific error fixed.

“This is wrong. I see your mistake and will show you how to avoid it in the future.”
The student is more passive. However the information delivered is targeted and the student sees the diagnosis process in action.

“This is wrong. Let me show you how to do it.”

The student is passive. The mentor sees an error was made, the student’s work is discarded rather than diagnosed, and the response is repetition of information. Clues about the specific error are buried in the general picture and frequently no more accessible than the first time the information was delivered.

“This isn’t quite right. I see you have the right idea but I’ll show you again.”

This phrasing is common even when they don’t have the right idea, as a gentle and encouraging way to tell them they are wrong. However it undercuts the learning process by suggesting there is no need to locate and correct an erroneous “idea”. The tiresome technical error may not seem important enough to need correction. The student is disengaged as well as passive.

These responses range from effective to unproductive, with the latter being more common today. The conclusion is that weaker critical thinking and independent–learning skills may be due in part to teaching methods in K–12: emphasis on presentation and information delivery; neglect of error diagnosis; kinder, gentler ways to deal with errors; even suggesting that untestable “understanding” may be as valuable as testable skills.

An error-diagnosis approach to learning requires cooperative students. Not only must they be willing to participate individually, but in a class setting the rest of the class must be able to work independently while one student is getting individual attention. The approach is vulnerable to disruption by unwilling or disinterested students. Consequently for researchers to even consider this approach would require them to both shift emphasis from teaching to learning and to either exclude or account for the effects of disruptive students.

5.3. Repetition. International comparisons reveal that in the US material is presented in shorter segments and there is much more repetition. Outcomes are also weaker. It is often suggested that to improve outcomes we should lengthen segments and reduce repetition. However short repeated segments may be a symptom rather than the root cause, and changing this without addressing root causes may worsen the situation.

Short lesson segments may work better for classrooms with little error–correction. Once a significant number of students have become dysfunctional it makes sense to take a break, give them time to forget, and try again with a new round of information delivery. Students prone to errors may have to go through the process several times before they stumble on the correct approach. It is, of course, inefficient for students who got it right the first time.

Lack of error correction may itself be a symptom rather than a root cause. Error diagnosis and correction is a one-on-one activity so in a classroom setting it requires the cooperation of all students in the class. Focus on information delivery may work better for classrooms with discipline problems.

This chain of connections, tracing bad outcomes back through repetition and lack of error correction to discipline problems, is highly speculative. The point here is not whether or not it is correct but that current mindsets and evaluation procedures prevent the educational research community from investigating such things.
6. Summary

The 1983 “Nation at risk” report described the K–12 situation at the time in dire terms:
“If an unfriendly foreign power had attempted to impose on America the mediocre educational performance that exists today, we might well have viewed it as an act of war. As it stands, we have allowed this to happen to ourselves. [...] We have, in effect, been committing an act of unthinking, unilateral educational disarmament.”

This proclamation sparked enormous activity in education research and upheavals in curriculum design, all intended to address the problem. But in many ways the problem has gotten worse. In the 1990s ambitious graduate programs in mathematics were largely populated by immigrants: we were importing high-quality K–12 and undergraduate education, critical thinking, and work ethic. Now we can no longer meet demand through imports and the high-tech jobs that require these skills are beginning to be exported.

Not only has the enormous activity been largely unproductive, but attitudes, assumptions and methodologies employed seem more likely to accelerate the decline than arrest it. Finally the problematic attitudes and methodologies seem to be locked in place by political pressures and incorporation into policies of the NSF and other funding sources. It seems likely that our educational disarmament will continue for the foreseeable future.