HUGH BURKHARDT

13. METHODOLOGICAL ISSUES IN RESEARCH AND DEVELOPMENT

INTRODUCTION

This chapter builds on Alan Schoenfeld's seminal contributions on methodological issues (Schoenfeld, 1980, 1985, 1992, 1994, 2002, 2006, 2007, 2010) and on our discussions over many years of collaboration and complementary thinking: Alan with the priorities of a cognitive and social scientist with a concern for practice; I with those of an educational engineer who recognizes the importance of insight-focused research for guiding good design. Alan has primarily aimed to bring rigor to research in mathematics education – to move it toward being an "evidence-based" field with high methodological standards. The Shell Centre team has an approach to research that gives high priority to impact on practice in classrooms. The analysis here reflects the challenges that we have faced, individually and together, and their wider implications for research methods in education.

The chapter begins with issues of strategy, going on in the third section to look at qualities that enable studies to make contributions to the body of research-based knowledge that is reliable enough, for example, to guide design. But strategy is not enough; in research, as in design, the details matter, so the fourth section focuses down on the essential core of education: classrooms, and what can make teaching and learning more effective. It looks at the challenges of designing such research through three case studies, each based on custom-designed research tools that fit their complementary but very different purposes, and draws some general conclusions about the design of tools for research. The examples reflect one of Alan's major methodological themes: that research should be "inspectable" so that readers can follow the chain of inference from data to claims. The fourth section draws these elements together, setting out a vision of education research that would likely be more purposeful and effective – a vision that, I believe, Alan broadly shares.

STRATEGIC ISSUES FOR RESEARCH IN EDUCATION

Across the various departments of a university there are very different styles of research. This breadth is summarised in the definition used for the UK Research Assessment Exercise (RAE, 2001) in which all UK university departments are rated every five years or so:

> 'Research' for the purposes of the RAE is to be understood as original investigation undertaken in order to gain knowledge and understanding. It includes

Y. Li and J.N. Moschkovich (eds.), Proficiency and Beliefs in Learning and Teaching Mathematics, 203–235.

> work of direct relevance to the needs of commerce and industry, as well as to the public and voluntary sectors; scholarship; the invention and generation of ideas and, images, performances and artifacts including design, where these lead to new or substantially improved insights; and the use of existing knowledge in experimental development to produce new or substantially improved materials, devices, products and processes, including design and construction.

Unusually, in research in education all of these elements can be found. But, I will argue, the whole is less than the sum of the parts – and that it doesn't need to be so.

Styles of research in education

The breadth of the above definition may surprise people. It arises from taking seriously four different traditions, characteristic respectively of the humanities, the sciences, engineering and the fine arts. The focus of both the humanities and the science approaches is the search for improved *insights*; in education these cover learning, teaching, professional development, and the behaviour of education systems and their key constituencies. The engineering research approach has a rather different priority: *impact* on systems. In education this focus is on developing products and processes that will help teachers and other professionals move to more effective practices. Fine arts are similarly concerned with products as well as analysis. Let us look at each tradition in a bit more detail.

The "humanities" approach

This is the oldest research tradition, based on scholarly acquisition of knowledge and critical analysis of it, and of other people's work. From the RAE definition it is

> original investigation undertaken in order to gain knowledge and understanding; scholarship; the invention and generation of ideas ... where these lead to new or substantially improved insights.

In the humanities there is no tradition of empirical testing of the assertions made. The key product is critical commentary – as, for example, on works of art or literature.

There is a lot of this in education. The ideas and analysis, based on the authors' reflections on their experience, are often valuable. Without the requirement of further empirical testing, a great deal of ground can be covered. This is still the most influential approach, partly because it supports the general belief that anyone can play, "expert" or not. This allows politicians to choose their own "common sense" policies.

However, since so many plausible ideas in education have not in practice led to improved outcomes across the system, the lack of empirical support is a key weakness. How can you distinguish reliable comment from plausible speculation? This has led to a search for "evidence-based education" and the emerging dominance in the research community of the "science" tradition.

The "science" approach

This style of research is also focused on better *insight*, of improved understanding of "how the world works," through the analysis of phenomena, and the building of models that help to explain them. In the RAE definition, it is again

> original investigation undertaken in order to gain knowledge and understanding; scholarship; the invention and generation of ideas ... where these lead to new or substantially improved insights.

This is the same wording as for the humanities approach, but with an additional implied requirement for empirical testing of the assertions made, which are now called hypotheses or models. Such testing takes time and effort, and narrows the range of what can be covered in a single study.

The key products are, again, assertions but now supported by evidence-based arguments and evidence-based responses to key questions. The evidence is expected to be empirical. The products are research journal papers, books and conference talks.

This approach is now predominant in the research in science and mathematics education. Such research provides insights, identifies problems, and suggests possibilities. However, it does not itself generate practical solutions, even on a small scale; for that, it needs to be linked to the "engineering" approach.

The "engineering" approach

This is directly concerned with practical *impact* – not just understanding how the world works but helping it "to work better." It does this by developing solutions to recognised practical problems in the form of tools and processes that help professionals become more effective. It not only builds on science research insights, insofar as they are available, but goes beyond them. In the RAE definition it is

> the invention and generation of ideas ... and the use of existing knowledge in experimental development to produce new or substantially improved materials, devices, products and processes, including design and construction.

Again there is an essential requirement for empirical testing of the products and processes, both formatively in their development and in evaluation. The importance of science-based insights varies, depending how far the "theory" is an adequate basis for design.

The key products are not only new tools and/or processes that work well for their intended uses and users but also new insights that come from the development process. (Below we give examples of this.) With these elements, development *is* research. However, in the academic community it is often undervalued – in some places only "insight" research in the science tradition is regarded as true research currency. I come back to these issues in the fifth section.

Table 1. Four scales of R&D.

	Focal variables	*Typical Research and Development Foci*
Learning (L)	Student Task	R: Concepts, skills, strategies, metacognition, beliefs D: Learning situations, probes, data capture
Teaching (T)	Instruction Student Task	R: Teaching strategies and tactics, nature of student learning D: Classroom materials that are OK for some teachers
Representative Teachers (RT)	Teacher Instruction Student Task	R: Performance of representative teachers with realistic support. Basic studies of teacher knowledge and competency. D: Classroom materials that "work" for most teachers
System Change (SC)	System School Teacher Instruction Student Task	R: System change D: Tools for Change – i.e., materials for: classrooms, assessment, professional development, community relations

he "fine arts" approach

This is related to the "humanities" approach rather as "engineering" is to "science." In the RAE definition is it is "the invention and generation of ideas and, images, performances and artifacts including design, where these lead to new or substantially improved insights."

The key products are paintings, sculpture, musical compositions etc. I will say little about this approach because, though it enriches education and could do more, it is not central here.

I believe that all these research traditions have contributions to make in achieving reliable research insights in education, and in translating them into practical impact in classrooms and school systems, but that currently the balance among the four approaches is far from optimal. What balance, of effort and of "academic credit," would be most effective, and how does it differ from the current pattern? I will argue that that there should be more "engineering" research and that this needs reliable research insights to build on. The implications for "science" research in education are the focus of the third section.

Scales of research and development

My next strategic point looks at different foci of research, and the scale of research effort that is needed for each to contribute significantly to the overall challenge: *establishing a sound research-based path from insights to large scale implementation*.

I find it useful to distinguish four different foci: learning, teaching, teachers, and school systems. The distinctions are summarised in Table 1, with the different research and development foci in the third column. The very different scales needed for the four kinds of study may be summarized as: a laboratory; a classroom; many classrooms; and whole school systems.

There is a crucial difference between T, which is about teaching possibilities, usually explored by a member of the research team, and RT, which is about what can be achieved in practice by typical teachers with available levels of support. Design research is often confined to T, whereas impact on practice requires going further, at least to RT. In "engineering" research in education (Burkhardt, 2006), the process of design research at T is continued through further rounds of trialing in more typical classrooms, so the products work well for a well-defined target group of real users.

Currently, the great majority of research is confined to L and T. A better balance across these different kinds of work is needed, if research and practice are to benefit from each other as they could. This has big implications for research strategy, since it is evident that RT and SC research needs larger research enterprises and longer time-scales. We return to this, too, in the fifth section.

RESEARCH INSIGHTS FOR IMPROVING PRACTICE

In this section, I look at features of insight-focused research that make it useful for guiding practice and, in particular, the design of educational materials and processes. The analysis builds mainly on Alan's first Handbook paper on methods (Schoenfeld, 2002) which has further references. In section VI he remarks:

> A very large percentage of educational studies are of the type, 'here is a perspective, phenomenon, or interpretation worth attending to,' and that their ultimate value is both heuristic ('one should pay attention to this aspect of reality') and as catalysts for further investigation.

This shows a remarkably modest level of confidence in the products of the research enterprise – a level of confidence that I share. It is illuminating to review his reasoning.

Schoenfeld's dimensions

In the paper Alan suggests three dimensions that help us to think about research claims. Briefly, they may be summarized as:

- *Generalizability*: How wide a range of phenomena does a claim cover?
- *Trustworthiness*: How well substantiated is the claim?
- *Importance*: How much should we care?

Typically, any given research report contains assertions in different parts of this 3-dimensional space, illustrated in Figure 1. Let us focus on the first two variables, G and T. A typical research study looks carefully at a particular situation – for example, a specific intervention based on clearly stated principles tried out in a few classrooms, collecting and analysing the teacher and student responses to the intervention. If carefully done, the results are high on T but, because of the limited range of the variables explored, low on G, shown as the zone A on the graph.

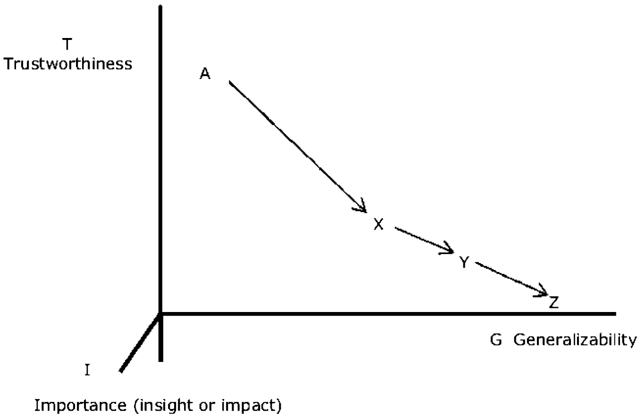


Figure 1. The trajectory of a typical research report.

However, the conclusions section of a typical paper goes on to discuss the "implications" of the study. These are usually much more wide ranging but *with little or no empirical evidence to support the generalisations involved.* These hopes, each a greater extrapolation with fewer warrants, are illustrated as X, Y and Z in the diagram. In this example, X might represents the suggestion that most students would respond similarly, Y that it would work for teachers at all stages of professional development, Z that the design principles would work across different topics in the subject. These are essentially speculations or, a little more kindly, plausible commentary in the humanities tradition.

Only large scale studies or metanalysis can move beyond this problem and establish "zones of validity" for research insights.[1] An example from the work of Alan Bell, Malcolm Swan and the Shell Centre team on "Diagnostic Teaching" illustrates this well (Bell, Swan, Onslow, Pratt, & Purdy, 1985; Bell, 1993). This approach, now often called "formative assessment" or "assessment for learning," is based on leading students whose conceptual understanding is not yet robust into making errors, then getting them to understand and debug these misconceptions through structured discussion. The early work showed learning gains through the teaching period (pre- to post-test) similar to those of the comparison group which had standard direct instruction teaching – but without the fall-away over the following 6-months that is so familiar to teachers ("They knew it when we did it"). This is illustrated in Figure 2.

The first study was for one *mathematics topic*, with the detailed treatment designed by *one designer*, taught by *one teacher* to *one class*. It was, in Alan's words, "worth attending to." Only several studies later, when the effect was shown to be stable across many topics, designers, teachers and classes could one begin to make reasonably trustworthy statements about "diagnostic teaching" as an approach. Even then, there remained further questions about its accessibility to typical teachers in realistic circumstances of support – an issue we return to later in this chapter.

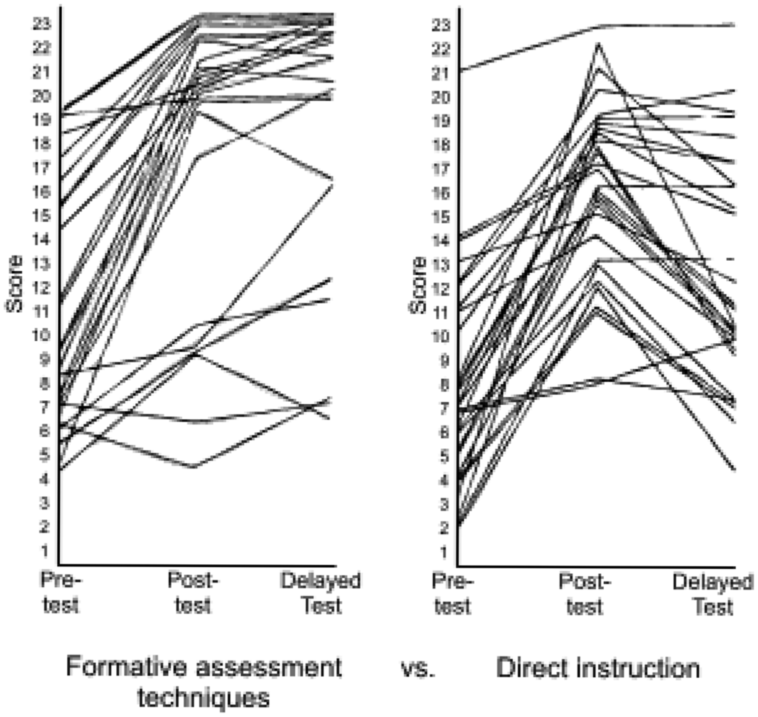


Figure 2. Typical results from work at the Shell Centre suggesting that teaching based on "cognitive conflict" techniques used in formative assessment improves long-term retention of learning.

The general point here is that much research is really about treatments, not about the principles the authors claim to study; to probe the latter one must check stability across a range of variables (student, teacher, designer and topic in this case). This typically *needs time and teams* beyond the scale of an individual Ph.D. or research grant. Other subjects arrange this; if it were more common in education, the research could have high G *and* T and, if the importance I were enough, be a reliable base to build further work on, in both the science and engineering traditions.

On importance, it is enough for the moment to say that a result can hardly be important unless it is generalizable beyond a specific study. My own perspective is that importance can come either from substantial impact on improving educational practice or from theoretical ideas of broad application with evidence for their generality.[2] Because of the scale of effort required to establish such evidence, the latter are rare.

Returning to Alan's comment with which this Section opened, very little insight-focused research has enough evidence of the generality and boundaries of its insights for them to provide a sound basis for design. Their conclusions may well be "worth attending to" but finding a range of validity is usually left to the engineers.

If you need statistics, forget it

What does this eye-catching heading mean? How can it be defended in a research field in which statistical analysis of data is so central? First, it is not a rejection of the importance of data; far from it. The key point is that:

> The large variability in implementation of educational innovations will wash out any small effects, however "significant" the gains may be from a purely statistical point of view. So only substantial clearly visible gains are likely to prove robust.

In medicine, by contrast, certain kinds of intervention, such as taking a prescription drug, can be implemented with little variation.[3] Even if the drug is only marginally effective, if used widely it can save (or, better, extend) thousands of lives; many drugs are of this kind. So randomized controlled clinical trials that show small improvements are valuable; it is these that need large samples and powerful statistical analyses. If the gains are substantial, as in the early research on antibiotics where people dying of septicemia were dramatically cured, you don't need statistics. Indeed, if it becomes clear during clinical trials that the control group is disadvantaged, the trial is immediately discontinued on ethical grounds and both groups given the treatment. My assertion is that the variability in implementation of educational initiatives is such that only where research shows clear and substantial gains are these likely to be robust and worth taking forward.

From a wider perspective, this is about the relationship between "systematic error" and "statistical error."[4] In most educational research, the systematic uncertainties are substantial. How far is the innovation actually happening, as designed? What range of strategies does the teacher use? How do teacher background, professional development, systemic support from principals/school district vary? How do all these affect outcomes? Large samples give data that is statistically more "reliable" – but uncertainties in the control of variables like those just listed, crucial to effective design and development, often make these error estimates delusory.

In education research, systematic errors dominate

This is not as despondent a message as it may seem. For example, in classroom research people say "every classroom is different"; true, but observations across mathematics classrooms, at least, show huge similarities in important ways. We have found that sample sizes of 3 to 7 are often optimal. This allows one to use always-limited research resources to collect and analyse richer data on each case, while distinguishing features that are probably generic from the idiosyncratic.

. . . but what about survey research?

There is one caveat to the theme of this chapter that I must mention. There *are* categories of research that yield results with well-established generality, discussed in detail in Schoenfeld (2002, 2007). For example, survey research, with all its sophistication and limitations, can be valuable in identifying widespread problems and suggesting provisional diagnoses. It is the epidemiology of education. In contrast, this chapter is focused on research that will lead to better "treatments":

intervention studies, and design and development of new or improved products and processes.

However, the many variables and the problems of their control that characterize education limit the diagnostic value of the data, which is of limited depth even in sophisticated surveys, making inference far more challenging than, for example, in the Doll studies that produced such a persuasive case for the harmful effects of smoking. Even there, establishing the causal connection was decisive to wide acceptance.

RESEARCH TOOLS FOR THE ZONE OF INSTRUCTION

Alan's analysis of mathematical problem solving (Schoenfeld, 1985) identified four levels of activity in the problem solving process: overall control, strategic plans, tactical decisions, and the technical skills in carrying them out. It offers a useful way to think about all problem solving, including our goal here: devising more effective methods for educational research. The argument so far has been about strategy; this needs to be complemented with something on tactical and technical aspects. Handling these well is crucial to the research enterprise. Details matter. This section seeks to exemplify that.

I have chosen to focus on research on the activities of teachers and their students in the classroom for several reasons. Elmore (2011) calls this, and those things that impinge directly on it such as teaching materials and professional development, "the zone of instruction." This is where educational improvement happens; the rest is, at best, merely supportive. Further, within classroom research, classroom observation is the most challenging single aspect. Of course, other kinds of information are important: student work, student and teacher responses to questionnaires probing their activities and attitudes, teacher logs and teaching materials, are all important sources of complementary information.

I shall also concentrate on observation, because of the richness of the information that is in principle available and the challenges of selecting and collecting what is significant in a form that can be analyzed to yield useful insights, both specific and general. I believe that there is no adequate substitute for structured observation, expensive though it is in time, and therefore resources. Equally, it is an area of research design with opportunities for improvement.

Classroom observation: three case studies in data selection

The flood of available information in a mathematics lesson is overwhelming. A television picture transmits millions of bytes per second, yet it can capture only a small part of what is visible in a classroom – missing, for example, most student discussion and written work. At a less information-theoretical level, the information flow is still unmanageable. Selection is inevitable. The research challenge is to understand what is going on, so as to select, to capture, and to analyse what is most cost-effective for the purposes of the research. There are inevitable tensions, and the necessary trade-offs, in optimizing the selection and collection of data in research. The theme here is “horses for courses” – that the optimal choices depend on the phenomena on which the research chooses to focus, and theoretical ideas it seeks to test. Cost-effectiveness is at the core of the design challenge.

Any research design also needs to look at how best to communicate the analysis to the “positive-thinking skeptics’ that form any good research community. To make the research process explicit, Alan has long argued (see e.g. Schoenfeld, 1980) that researchers should make their data available, along with rich enough descriptions of their research methods such that readers could themselves examine the data and evaluate the inferences. He has done so over his career, producing “inspectable” studies that make both substantive and methodological contributions.

From the myriad of published “observation schedules” (see e.g. Good & Brophy, 2002), I have chosen these three because they all seek to capture aspects of the richness that is present in mathematics classrooms, each combining breadth with attention to detail. These cases illustrate three very different approaches to capturing what happens, each with a different balance of priorities. The first emphasizes depth of understanding of teachers’ decision making, down to the level of their individual “moves’ in a lesson; the goal was to construct a theoretical model of a specific area of human problem solving: teaching. The second was designed to find how far the pattern of dialogue in classrooms changed when teachers used specific new materials; the complementary goals were to elicit some design principles, and to provide feedback for refining the materials, so the study needed to cover many lessons. Both achieved their very different goals. The last (still in development) has a balance of these priorities, covering many lessons with a focus on the mathematical nature of the discussion and teacher professional development over a year.

Teacher decisions focus

The first case comes from Alan’s long running “teacher modeling” program, published in a series of papers and brought together as the core of his book: *How we think: A theory of goal-oriented decision making and its educational applications* (Schoenfeld, 2010). This study is based on an extremely detailed analysis of video of three lessons, taught by very different teachers: two highly experienced and innovative, the third a recent graduate. The goal of the research was ambitious: to understand every move the teacher made in the lesson in terms of three dimensions: *their knowledge*, *goals* and *orientations* (earlier called beliefs). Knowledge is defined broadly, including mathematical knowledge and skills, pedagogical content knowledge, and knowledge of pedagogical strategies, tactics and skills. The meaning of goals and orientations will become clearer through the example below.

The data is presented in three parallel streams, the latter two subdivided, with time increasing down the page. The streams are increasingly analytic, namely:

- *a full transcript* of the dialogue
- *a parsing of the dialogue*, with levels of increasing detail, from the major activities of the lesson down to the smallest self-contained episodes.

– *a graphical representation of goals and orientations* in the form of continuous vertical bars, shaded to show the level of activity of each at that point in the lesson.

This graphical representation can be seen both as a fine-grained description of the lesson as it unfolded and as the basis for a model of the teacher's decision making: a model equipped with the knowledge, goals, and orientations found in the graphical representation would produce decisions consistent with those of the teacher.

These elements are illustrated for two short sections of the lesson (from Schoenfeld, 2010, chapter 5) in Figures 3 and 4. The teacher is a distinguished science teacher and the lesson is about criteria for choosing "the best number" from a set of measurements. The teacher motivates the discussion in terms of tests of blood alcohol concentration; the students then make multiple measurements on something more accessible – the length of a table.

The way the analysis is structured and communicated exemplifies Alan's belief, noted above, that readers must be able to follow the data and its analysis in enough detail to allow them to critically review the author's thinking – the opposite of "trust me" styles of commentary in some research. Here I can give only a flavor of the way this is done and the tools he developed to do it.

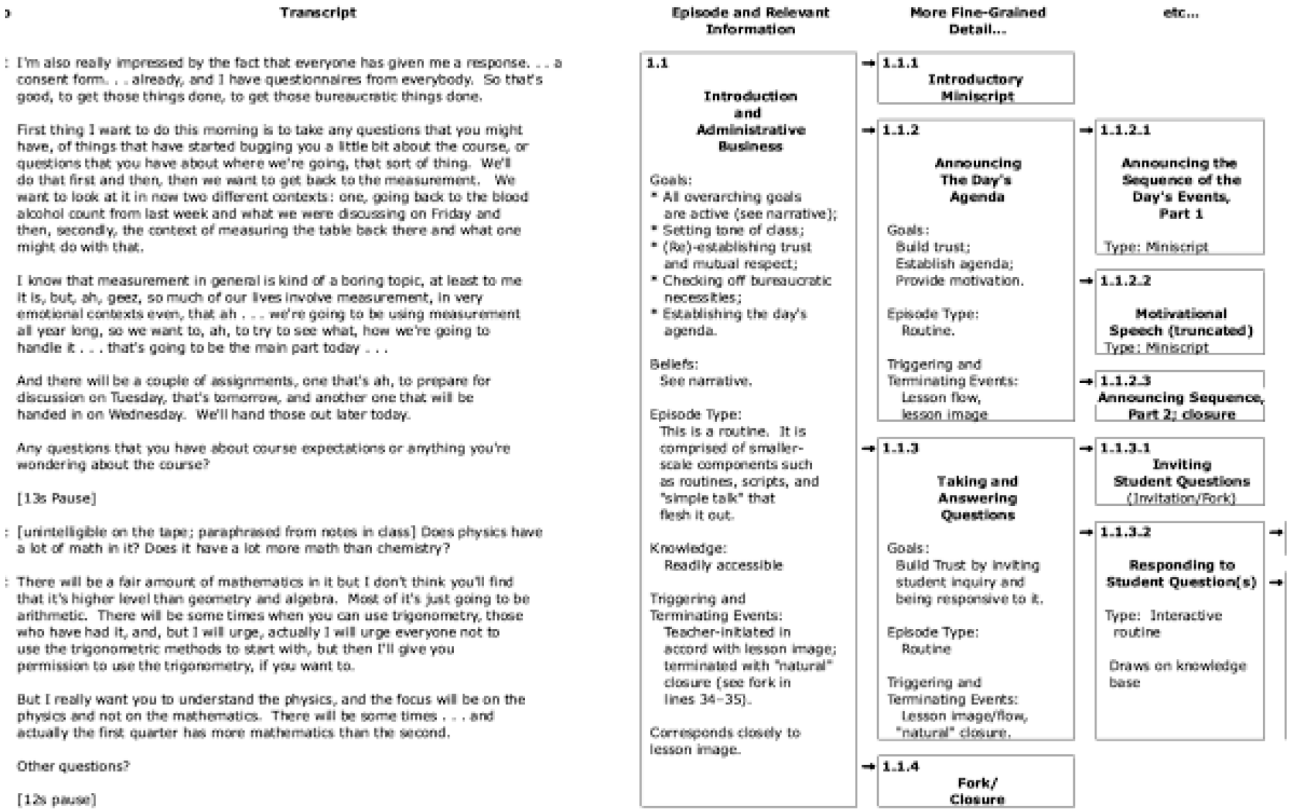
Figure 3 shows how the transcript of the opening of the lesson, which is largely organizational, is parsed at increasing levels of detail. The focus here and throughout is the detailed attention given to understanding the raw data, epitomised by the transcript but enriched by other aspects of the video.

Figure 4 shows how the parsing of the more complex dialog in the core of the lesson, on choosing an appropriate summative measure, is analysed into the bar notation which shows the active goals and orientations at each point in the dialogue. The main goals and orientations, noted at the bottom of the figure, show that, while goals g and l are concerned with the key content to be learned, the other goals are focused on the classroom dynamics that will support the learning processes, reflecting the teachers orientations rather as tactics support strategy.

From analyses of this kind, Alan and his students and collaborators, have built up a theoretical model of teachers' decision making. I hope this brief sketch will encourage the reader to enjoy the rigor of the analysis by reading *How we think*, which goes on to apply a similar methodology to other areas of human real-time decision making, including medical diagnosis.

Classroom discussion focus

In this case the goals and the context were quite different. The study complemented and supported the ITMA program of design and development of educational software. ITMA (Investigations on Teaching with a Microcomputer as an Aid) focused on a single computer with a large (TV) display in each classroom – an approach that realistically reflected the level of hardware provision at the time.[5] The project leader, Rosemary Fraser, had found in her own classroom that simple non-routine problem solving software of this kind promoted student engagement and mathemat-



Transcript

: I'm also really impressed by the fact that everyone has given me a response. . . a consent form. . . already, and I have questionnaires from everybody. So that's good, to get those things done, to get those bureaucratic things done.

First thing I want to do this morning is to take any questions that you might have, of things that have started bugging you a little bit about the course, or questions that you have about where we're going, that sort of thing. We'll do that first and then, then we want to get back to the measurement. We want to look at it in now two different contexts: one, going back to the blood alcohol count from last week and what we were discussing on Friday and then, secondly, the context of measuring the table back there and what one might do with that.

I know that measurement in general is kind of a boring topic, at least to me it is, but, ah, geez, so much of our lives involve measurement, in very emotional contexts even, that ah . . . we're going to be using measurement all year long, so we want to, ah, to try to see what, how we're going to handle it . . . that's going to be the main part today . . .

And there will be a couple of assignments, one that's ah, to prepare for discussion on Tuesday, that's tomorrow, and another one that will be handed in on Wednesday. We'll hand those out later today.

Any questions that you have about course expectations or anything you're wondering about the course?

[13s Pause]

: [unintelligible on the tape; paraphrased from notes in class] Does physics have a lot of math in it? Does it have a lot more math than chemistry?

: There will be a fair amount of mathematics in it but I don't think you'll find that it's higher level than geometry and algebra. Most of it's just going to be arithmetic. There will be some times when you can use trigonometry, those who have had it, and, but I will urge, actually I will urge everyone not to use the trigonometric methods to start with, but then I'll give you permission to use the trigonometry, if you want to.

But I really want you to understand the physics, and the focus will be on the physics and not on the mathematics. There will be some times . . . and actually the first quarter has more mathematics than the second.

Other questions?

[12s pause]

Episode and Relevant Information

1.1
Introduction and Administrative Business

Goals:
* All overarching goals are active (see narrative);
* Setting tone of class;
* (Re)-establishing trust and mutual respect;
* Checking off bureaucratic necessities;
* Establishing the day's agenda.

Beliefs:
See narrative.

Episode Type:
This is a routine. It is comprised of smaller-scale components such as routines, scripts, and "simple talk" that flesh it out.

Knowledge:
Readily accessible

Triggering and Terminating Events:
Teacher-initiated in accord with lesson image; terminated with "natural" closure (see fork in lines 34–35).

Corresponds closely to lesson image.

More Fine-Grained Detail...

→ **1.1.1**
Introductory Miniscript

→ **1.1.2**
Announcing The Day's Agenda

Goals:
Build trust;
Establish agenda;
Provide motivation.

Episode Type:
Routine.

Triggering and Terminating Events:
Lesson flow, lesson image

→ **1.1.3**
Taking and Answering Questions

Goals:
Build Trust by inviting student inquiry and being responsive to it.

Episode Type:
Routine

Triggering and Terminating Events:
Lesson image/flow, "natural" closure.

→ **1.1.4**
Fork/ Closure

etc...

→ **1.1.2.1**
Announcing the Sequence of the Day's Events, Part 1

Type: Miniscript

→ **1.1.2.2**
Motivational Speech (truncated)
Type: Miniscript

→ **1.1.2.3**
Announcing Sequence, Part 2; closure

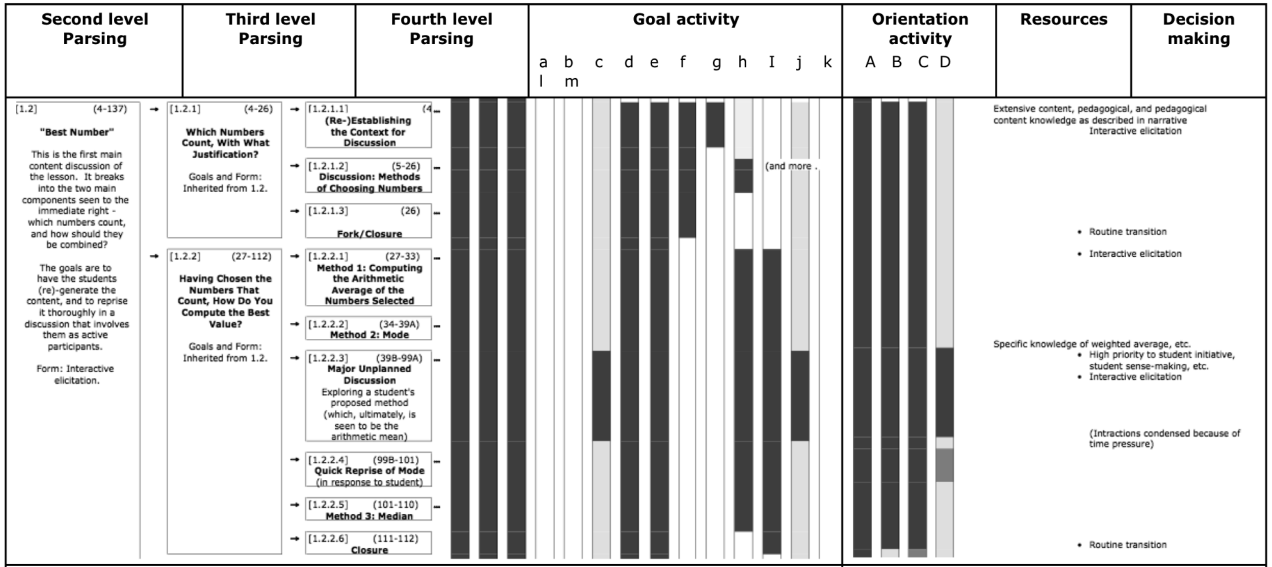
→ **1.1.3.1**
Inviting Student Questions
(Invitation/Fork)

→ **1.1.3.2**
Responding to Student Question(s)

Type: Interactive routine

Draws on knowledge base

Figure 3. A multi-level parsing of the introductory episode of the lesson.



Main goals in this episode

a. Have the class interact as a community of inquiry with freedom to explore, conjecture, reason through

b. Have the students experience physics as a way of making sense of the world.

c. Provide a warm positive atmosphere in which students feel valued, encouraged to speak, etc.

f. Answer any questions student might have

g. Have students conceive of "best number" as a whole – which data, combined in what way?

h. Have students re-generate content via interactive elicitation

i. Work through choosing and justifying ewhich numbers count

k. Elaborate on specifics of content that arise in dialogue via interactive elicitation

l. Have students address three measures of central tendency (mean, median, mode)

Orientations

A. Doing physics is and should be a sense-making activity.

B. Where possible, ideas should come from students.

C. Class discourse should minimize teacher "telling"

D. Student sense-making activity should be given highest priority.

Figure 4. Deciding on the best number to summarise a set of measurements.

ical discussion. The ITMA team of teacher-programmers designed and developed many examples of such software, along with lesson notes for the teacher.

The research goal was to understand better what happens when this material is used by a variety of teachers in their classrooms, and to exploit that understanding in the design and development of such software and accompanying curriculum support. Seventeen teachers agreed to choose and use 10 lessons from the draft collection, and to be observed in their normal teaching and in using these new lessons.

We found that structured classroom observation was essential to capture the changes in the pattern of interpersonal dynamics that the team had found in their own classrooms. How far would the materials lead other, more typical teachers to work in similar ways? We needed a lesson observation protocol that would enable observers to capture key information within the time and effort we could afford. With about 200 lessons to study, we decided on one hour for "live" observation and a brief post-lesson discussion with the teacher, with about one further hour for the analysis of that lesson.

We decided to design our own observation system, based on an intense open-minded study of 10 examples of lessons on video. Three of us viewed these lessons many times, discussing what we could see that seemed to us significant for our purposes. Terry Beeby, the graduate student, got to know the lessons so well that, whenever in our discussions a type of event was suggested as significant, he could quickly find similar examples for discussion.

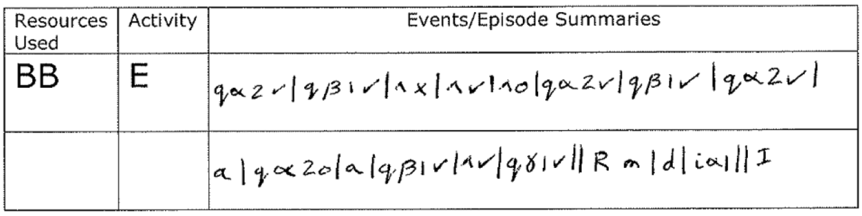
We were particularly interested in those things that differed from teacher to teacher, and from conventional mathematics lessons to those using ITMA software. Of the things we saw in the video lessons, the variation in the patterns of discourse were particularly striking, with profound changes from the teacher-directed nature of most British mathematics lessons. The outcome of this tool design process was *SCAN – a systematic classroom analysis notation for mathematics lessons* (Beeby, Burkhardt, & Fraser, 1980). Key features include:

- Use of a shorthand notation (rather than box-ticking or diagrams)
- Three timescales: *activities* within the lesson, self-contained *episodes*, linguistic *events*
- *Events* include: **q**uestion, **e**xplanation, **i**nstruction, **h**ypothesis, **m**anagement, social **g**ambit with qualifiers for:

  Initiator: assumed to be the teacher; pupils **p** or numbered
  Depth: α recall of a single fact, β familiar exercise, γ extension
  Guidance: **1** detailed, **2** specific, **3** open
  Correctness: $\surd$ correct, **x** wrong, **?** unclear

This method of developing observation tools, through the intensive study of a sample of videos to identify and classify events that are significant from the point of view of the study, is of general value. In this respect, it is rather like the previous example, though covering many lessons made cost-effectiveness more important.

As with any shorthand, it takes time to become fluent in the notation. But, for example, 20 teachers after three hours training on video produced very consistent



Resources Used	Activity	Events/Episode Summaries
BB	E	qα2✓|qβ1✓|^x|^✓|^o|qα2✓|qβ1✓|qα2✓|
		α|qα2o|α|qβ1✓|^✓|qγ1✓||R m|d|iα||I

Figure 5. A SCAN record of a simple lesson opening.

Table 2. Comparative statistics on 3 teachers working with JANE.

Lesson		*A*	*B*	*C*
questions asked				
(resolved)	α	17(15)	7(5)	3(0)
	β	15(15)	17(14)	10(10)
	γ	7(6)	9(7)	15(9)
explanations		8	12	6
assertions/instructions		1	2	6
student questions		1	0	0
student explanations		0	4	11

"live" SCANs of a simple conventional lesson on polygons. Figure 5 shows a SCAN record of the first five-minute exposition (E) activity at the blackboard (BB).

The teacher launches the lesson with an exposition activity **E** by checking that the students know some basic definitions (a revision episode, **R**). Note the linguistic style, dominated by short questions **q**, mostly of single facts α with fairly close guidance **1** that elicited correct responses $\surd$. (^ signals a repeat of the question.) The teacher then initiates a second activity (I) of individual student work (W1 on the following line, not shown); in this he gives the formal definition of polygons **d**, then gives detailed instructions i for a simple activity: working some similar cases. This is a teacher who uses the Q&A mode of exposition, which is common in the UK. He gives a lot of support to his students, while keeping them on a short leash intellectually. The SCAN provides detailed semi-quantitative evidence of this.

The three lesson extracts in Figure 6 are more interesting, in themselves and for the purpose of the study. They show different teachers working with the same piece of software: a simple "function machine" program. In using it you give JANE a number; when you press the answer key, she gives you one back. The question is "What does JANE do to numbers?" (There are six girls, who multiply, and six boys, who add different numbers. You can go on to a "function of a function" investigation, involving a boy then a girl or vice versa.) The mathematical purpose is to develop students' hypothesis generation and an awareness of the implications and limitations of evidence – that a counter example kills a conjecture but many examples are not a proof. You also practice mental arithmetic.

Table 3. Classroom roles distributed among teacher, students and microcomputer.

Directive roles	*Facilitative roles*
Manager	Counsellor
Explainer	Fellow Student
Task-Setter	Resource

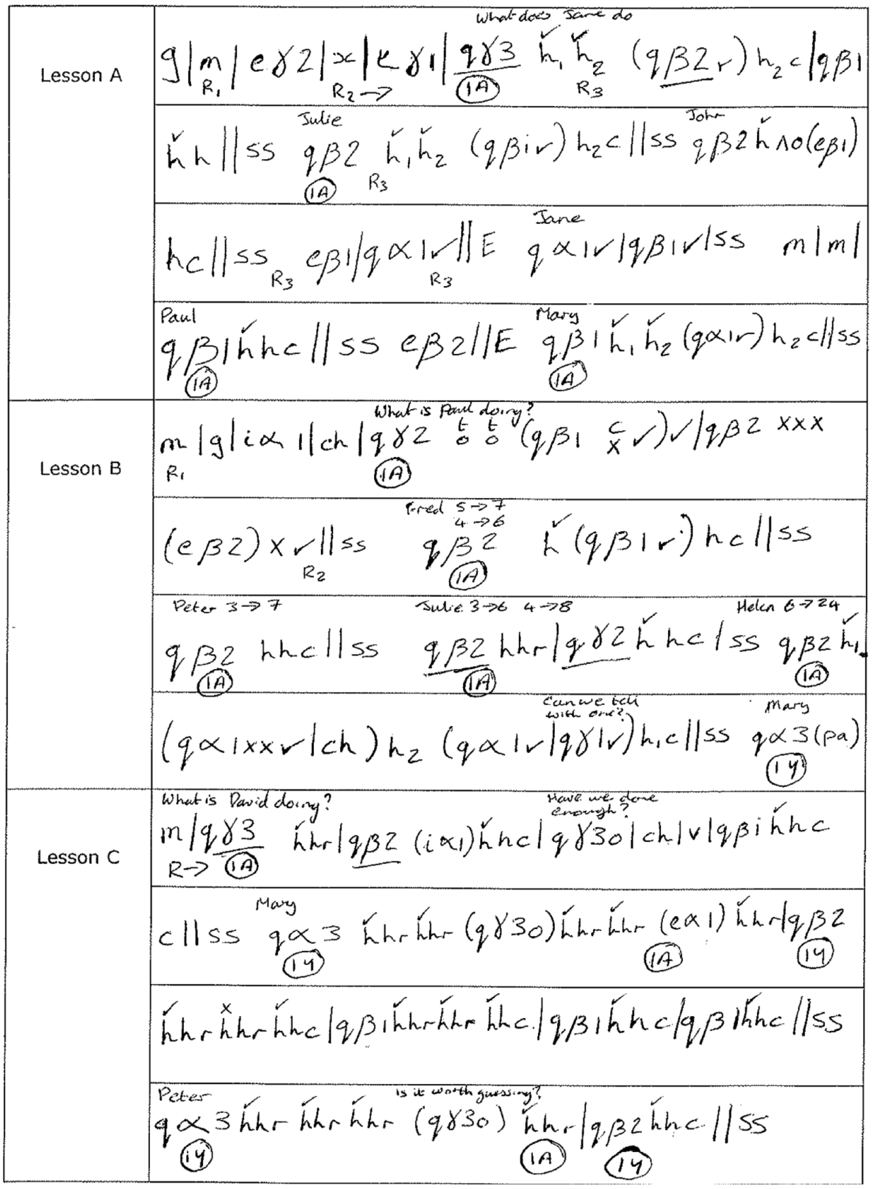
What did we learn from the full SCANs, along with the lesson materials, some student work and the less-structured notes of the observer? The lesson worked well for all three teachers, with nearly all students focused throughout. The three teachers worked in very different ways. These show in the simple statistics in Table 2. Note, for example, the differences in the distributions of α, β and γ questions and the number of pupil explanations across the three lessons

Looking at the rhythm of each lesson, even these short extracts show that Teacher A established a rhythm of very short "search successful" (**SS**) episodes; these continued through the lesson, exhaustively repeating the same pattern before going onto the two children challenges, which then became exercises in combining operations the class had already worked out. In contrast, Teacher C had much longer episodes, collecting multiple alternative hypotheses and delaying closure. Later, after collecting much confirmatory evidence on one hypothesis, he asks "Can we be sure?"; after a long pause with no response, he squares his mathematical and pedagogical consciences with "Well, we can be pretty sure" – which seems fair, in the universe of 11 year-old students who are too young to have the concept of rigorous proof (Bell, 1976).

One of these teachers also taught the polygon lesson of Figure 6; the reader is invited to guess which one from the evidence in the SCANs on their styles.

The outcomes of this work included both improved lesson units and their associated software, and some insights with wider implications for design. I will mention one: the *roles analysis* (Burkhardt, Fraser, Coupland, Phillips, Pimm, & Ridgway, 1988). In analysing the SCAN data, the researchers were struck by the various roles played in the classroom dynamics by the teacher, the students and the computer. Far from the computer being an inanimate tool, it was clear from the reactions that each piece of software gave it a personality,[6] as with "What does Jane do to numbers?" Detailed study identified about 30 roles, which we boiled down to 6 main groups, shown in Table 3. Most of the names are self-explanatory. Counsellors advise, they do not direct or explain. A *Resource* supplies information, but only when asked.

In regular mathematics lessons, most teachers take the directive roles, the students *(Fellow) student*, and the resources are inanimate – typically textbook and worksheets. In lessons with the ITMA software, the software on screen took over much of the *manager*, *explainer* and *task-setter* roles, and the teachers moved to play *counsellor* and *fellow student*. Somewhat to our surprise, without any prompting teachers moved from the front of the class, talking with students about "What's



<table>
<tr><td rowspan="4">Lesson A</td><td>g|m R_1|eγ2|x R_2→|eγ1|What does Jane do qγ3 (1A) ✓h_1 ✓h_2 R_3 (qβ2r) h_2c|qβ1</td></tr>
<tr><td>✓h h||ss Julie qβ2 (1A) ✓h_1 ✓h_2 R_3 (qβ1r) h_2c||ss John qβ2 ✓h∧o(eβ1)</td></tr>
<tr><td>hc||ss R_3 eβ1|qα1✓ R_3||E Jane qα1✓|qβ1✓|ss m|m|</td></tr>
<tr><td>Paul qβ1 (1A)|✓hhc||ss eβ2||E Mary qβ1 (1A)|✓h_1 ✓h_2 (qα1r) h_2c||ss</td></tr>
<tr><td rowspan="4">Lesson B</td><td>m R_1|g|iα1|ch|What is Paul doing? qγ2 (1A) $\overset{t}{o}$ $\overset{t}{o}$ (qβ1 $\overset{c}{x}$ ✓)✓|qβ2 xxx</td></tr>
<tr><td>(eβ2)x✓ R_2||ss Fred 5→7 4→6 qβ2 (1A) ✓h (qβ1r) hc||ss</td></tr>
<tr><td>Peter 3→7 qβ2 (1A) hhc||ss Julie 3→6 4→8 qβ2 (1A) hhr|qγ2 ✓h hc|ss Helen 6→24 qβ2 (1A) ✓h_1</td></tr>
<tr><td>(qα1xx✓|ch) h_2 (qα1✓|Can we tell with one? qγ1✓) h_1c||ss Mary qα3(pa) (1Y)</td></tr>
<tr><td rowspan="4">Lesson C</td><td>What is David doing? m|qγ3 R→ (1A) ✓hhr|qβ2 (iα1)✓hhc|Have we done enough? qγ3o|ch|v|qβ1 ✓hhc</td></tr>
<tr><td>c||ss Mary qα3 (1Y) ✓hhr ✓hhr (qγ3o) ✓hhr ✓hhr (eα1) (1A) ✓hhr|qβ2 (1Y)</td></tr>
<tr><td>✓hhr ×hhr ✓hhc|qβ1 ✓hhr ✓hhr ✓hhc|qβ1 ✓hhc|qβ ✓hhc||ss</td></tr>
<tr><td>Peter qα3 (1Y) ✓hhr ✓hhr ✓hhr Is it worth guessing? (qγ3o) ✓hhr (1A)|qβ2 (1Y) ✓hhc||ss</td></tr>
</table>

Figure 6. SCAN records of three teachers working with JANE.

it doing?" Sixteen of the seventeen teachers made these role shifts naturally. (The exception stood proudly next to the screen throughout the lessons, sharing its role!) Alan has never liked SCAN, primarily because it does not record the specific mathematical content of the discussion in each episode. (The teaching materials and student work samples do, of course, partly fill this need but they are not linked to specific events in the lesson by the SCAN record.) The next case describes our current efforts to meet this concern in a protocol that combines something of the economy of SCAN with a deeper look at the mathematical structure of the discourse.

Mathematical discussion focus

The Mathematics Assessment Project (MAP) is developing lesson materials that support formative assessment for learning in US classrooms. The power of formative assessment for learning, when it is done well, was summarized in the metanalysis of Black and Wiliam (1998). Their and others' subsequent work has approached the challenge of making formative assessment happen through professional development; they find long-term and intensive work with teachers is needed, making the challenge of "going to scale" something between very expensive and unrealistic. The MAP lessons are a product of the first engineering research on supporting formative assessment for learning primarily through teaching materials (MAP, 2012).

The Shell Centre design team is led by Malcolm Swan, with Alan as PI of a Shell Centre-Berkeley collaboration. The previous emphasis on professional development reflects the fact that these lessons take most teachers of mathematics well outside their pedagogical and mathematical comfort zone. The lessons provide support for teachers in this broadening of their professional capacity. They are being used in school systems across the US to support the implementation of the Common Core State Standards for Mathematics. The initial reception has been enthusiastic.

Structured classroom observation of trial lessons has guided two iterations of revision of each lesson. Now we need to learn in more detail and more depth about what happens as teachers gain experience in using these materials. Alan is leading the team in a program of research in which the design of an appropriate protocol for observation and analysis will play a central role.

We plan to observe 20 teachers, each using 10 of the formative assessment lessons in the course of a school year, along with some of their normal teaching. Each lesson will be videoed. Nonetheless, as in 3.3■■AU: is this reference to a section, to a figure or to a table?■■, for an analysis of around 200 lessons, cost-effectiveness is a prime consideration. The development of the protocol is ongoing but the current version has the following features.

In the large, the goal of the research is to produce an analytic scheme that captures the things that research indicates are the essential aspects of a lesson – the goal being to document the relationship between the presence and frequency of those classroom behaviors and the depth of student learning. The former will be captured by the analytic scheme, the latter by robust tests of student understanding such as the Balanced Assessment/MARS tests.

The characteristics of such a research-in-practice analytic scheme must be radically different than those of the scheme in the section on "Teacher decision focus." The analyses in Schoenfeld's book took years to produce; in contrast, a SCAN coding can be done in real time. The goal of the current analyses is to produce a coding of a lesson in no more than twice real time (a real-time observation plus the same amount of time to convert one's observational notes into a formal coding record), while at the same time being directly sensitive both to important classroom activities and the quality of the mathematics being discussed.

After much experimentation, the MAP team converged on a scheme that has five "process- or practice-related" dimensions and one focused content-related dimension. Ultimately, these are coded in five different types of classroom activity.

First, the dimensions for analysis. The research team believes that each of the following dimensions are central in examining classrooms:

1. Mathematical focus, coherence, and accuracy. Is the mathematics discussed rote and mechanical, or are procedures connected to underlying concepts? Do the students have the opportunity to do sense-making? If the students do not have the opportunity to engage with meaningful mathematics, they are not going to learn it.

2. Cognitive Demand. Classroom observation shows that, when students encounter difficulty, many teachers provide "help" that actually removes the main challenges from the task, lowering the level and depriving the students of the opportunity for productive struggle. Are classroom interactions structured so that students have the opportunity to grapple meaningfully with the mathematics?

3. Access. Which students get to participate. Are most of the students involved, or only a select few?

4. Agency: Accountability and authority. Do students have the opportunity to speak and write mathematics, to become expert and share that expertise?

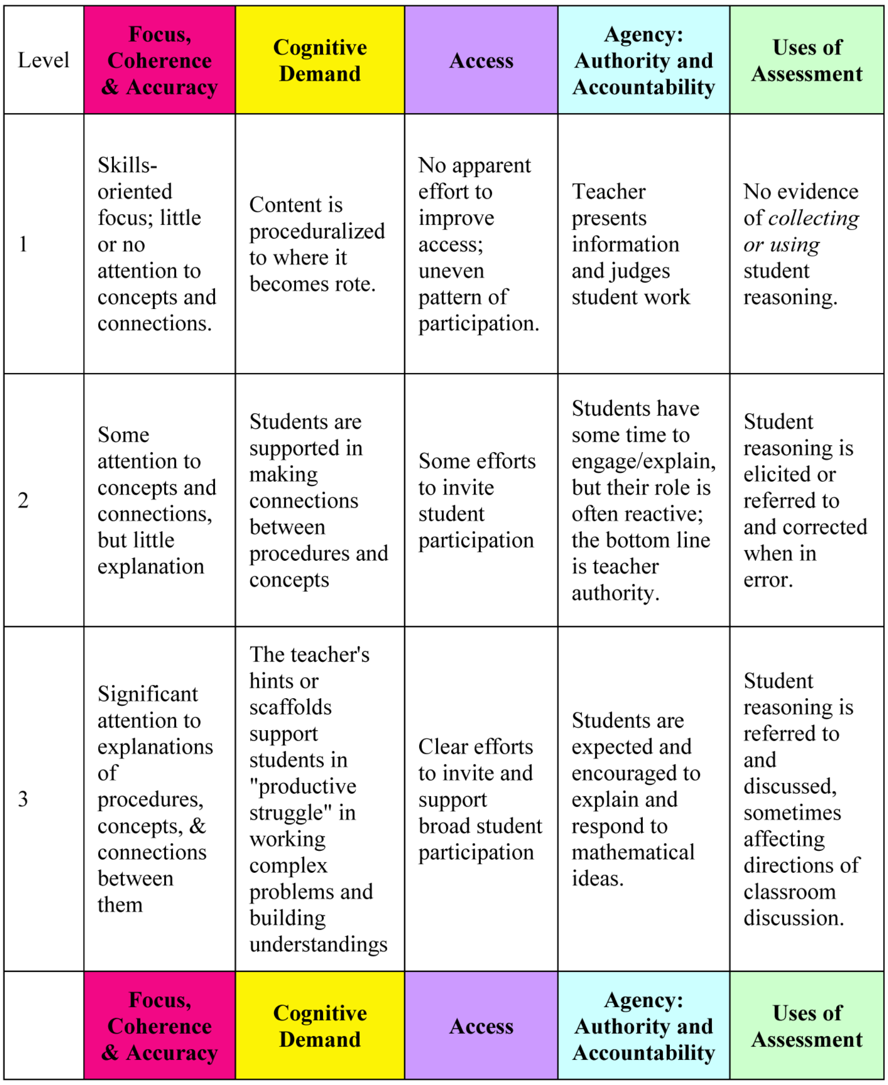
5. Uses of assessment. Does the teacher obtain information about student understandings, formally or informally, and use that information in ways that allow the lesson to build on student understandings and address misunderstandings?

6. Domain specifics. If a lesson focuses on a particular topic, what is the most important mathematics in that topic? Does the lesson grapple with that content? The sixth dimension is handled separately. For each of the first five we have a general rubric on a 3-point scale, outlined in Table 4.

This is a broad summary. In fact, we employ context-specific versions of this rubric for each of the following classroom activities:[7]

- teacher giving directions (setting up or modifying tasks for student work)
- teacher exposition of mathematical ideas (this may be in the form or lecture or classroom summary)
- classroom discussion of mathematical ideas, in which there are student contributions;
- students seek to clarify mathematical ideas and/or reveal confusion

Table 4. Dimensions and levels for the MAP observation protocol.



Level	**Focus, Coherence & Accuracy**	**Cognitive Demand**	**Access**	**Agency: Authority and Accountability**	**Uses of Assessment**
1	Skills-oriented focus; little or no attention to concepts and connections.	Content is proceduralized to where it becomes rote.	No apparent effort to improve access; uneven pattern of participation.	Teacher presents information and judges student work	No evidence of *collecting or using* student reasoning.
2	Some attention to concepts and connections, but little explanation	Students are supported in making connections between procedures and concepts	Some efforts to invite student participation	Students have some time to engage/explain, but their role is often reactive; the bottom line is teacher authority.	Student reasoning is elicited or referred to and corrected when in error.
3	Significant attention to explanations of procedures, concepts, & connections between them	The teacher's hints or scaffolds support students in "productive struggle" in working complex problems and building understandings	Clear efforts to invite and support broad student participation	Students are expected and encouraged to explain and respond to mathematical ideas.	Student reasoning is referred to and discussed, sometimes affecting directions of classroom discussion.
	Focus, Coherence & Accuracy	**Cognitive Demand**	**Access**	**Agency: Authority and Accountability**	**Uses of Assessment**

– connecting to prior knowledge (can be during set-up, or when discussing work on problems)

This scheme is still under development, but preliminary testing indicates that it has some face validity with teachers, and meets the constraints discussed above – lessons can be coded in no more than twice real time, and with some degree of consistency. Time will tell with regard to the scheme's utility. "Watch this space."

In summary

The purpose of this section has been to make and illustrate three points in the challenging context of designing tools for classroom observation:

- Designing reasonably efficient methods of data selection, capture and analysis is at the heart of good research design.
- The design will always involve trade-offs, with the balance determined by the project's research priorities – this implies the design should normally be custom-tailored and, of course, "mixed methods."
- The earlier in the process that redundant data can be discarded, the lower the cost – provided, of course, that you don't throw away essential data.

The three cases outlined here reflect different priorities. Each was a choice that suited the purpose in hand. All three could be improved and extended with additional resources.

I have featured this detailed technical aspect of research for several reasons:

- the central importance of capturing rich data from the classroom;
- the interesting challenges of doing observation well;
- the potential that technology offers in this area.

There are already devices that link written notes to an audio recording, so that the touch of the special pen on a note replays the audio from the moment it was made, allowing easy reconsideration and expansion of interesting events. Apps for both tablets and smartphones will allow us to show on screen a rich analytical framework for observation, so that observers' input can more easily be made in real time, and captured automatically for analysis. As ever, we will have to be vigilant that the technology does not impose standard solutions that undermine the research quality.

TOWARDS MORE PRODUCTIVE RESEARCH: A "SYSTEMS" PERSPECTIVE

This section brings together the strategic and tactical issues discussed so far into a set of suggestions on changes in the grand strategy for research in education that would enable it to make a greater contribution.

The argument builds on previous sections and the synthesis in the paper "Improving educational research: towards a more useful, more influential and better-funded enterprise" (Burkhardt & Schoenfeld, 2003). Looking at education in comparison with other fields, this paper identifies six elements that are needed for a research program to have impact on practice. These are shown in Table 5.

The paper goes on to look in more detail at the various barriers to such change, and ways in which they might be overcome. Here we discuss the implications for various key communities – researchers of various kinds, teachers, schools, school systems and policy makers.

Table 5. Elements needed for research to improve practice.

1. Robust mechanisms for taking ideas from laboratory scale to widely used practice. Such mechanisms typically involve multiple inputs from established research, the imaginative design of prototypes, refinement on the basis of feedback from systematic development, and marketing mechanisms that rely in part on respected third-party in-depth evaluations. These lab-to-engineering-to-marketing linkages typically involve a strong research-active industry (for example, the drug companies, Bell Labs, Xerox PARC, and IBM).
2. Norms for research methods and reporting that are rigorous and consistent, resulting in a set of insights and/or prototype tools on which designers can rely. The goal, achieved in other fields, is cumulativity – a growing core of results, developed through studies that build on previous work, which are accepted by both the research community and the public as reliable and non-controversial within a well-defined range of circumstances. (Work on the cutting edge is something else, of course, with some uncertainties and controversy in every field of research.)
3. A reasonably stable theoretical base, with a minimum of faddishness and a clear view of the reliable range of each aspect of the theory. Such a theory base allows for a clear focus on important issues and provides sound (though still limited) guidance for the design of improved solutions to important problems.
4. Teams of adequate size to grapple with large tasks, over the relatively long time scales required for sound work of major importance in both research and development.
5. Sustained funding to support the Research-to-Practice process on realistic time scales.
6. Individual and group accountability for ideas and products; do they work as claimed, in the range of circumstances claimed?

Table 6. Current academic priorities tend to favor.

new results *over* replication and extension
trustworthiness *over* generalizability
small studies *over* major programs
personal research *over* team research
first author *over* team member
new ideas *over* results that can be relied on
disputation *over* consensus building
journal papers *over* products and processes

"Importance" – For whom?

In most societies, the long-term goal for education is to improve the outcomes for children in terms of performance and attitude – the range of things they can do well, know about, use effectively, and enjoy. How to achieve this is a high-profile issue of policy and politics.

The educational research community surely shares these goals. How well is it structured to focus on them? Like any community, it has its own agendas and inward-looking concerns. The great majority of researchers are in academic institutions, so the community needs systems for evaluating work and selecting people for appointment, tenure and promotion. Research in education has a value system that guides these judgments, outlined in Table 6.

It will be clear from the argument so far that these are not the priorities that are likely to lead to building a body of reliable detailed research that can underpin design, and thus build a direct link from research to improved practice. Indeed the second and third sections and Table 5 suggest that they are likely to have the opposite effect. How has it come to be this way?

First, how do these priorities serve the internal needs of the research community? If you look for a fundamental measure of quality in research in any field, it is self-referential.

Impressing key people in your field

is the prime criterion. Each field turns this into a set of quasi-objective criteria. How did education come to the pattern in Table 6?

There is a pattern of pressures on researchers that helps explain. Researchers, being human, tend to like research similar in style to their own. Academics are usually only part-time researchers, with substantial loads of teaching, and administration of courses. Yet, to be seen as successful, they are expected to produce several journal publications a year. Acceptance by journal reviewers depends on the studies being seen as "trustworthy." Ph.D. students need to be trained in research and to produce publishable work within three or four years. Assigning credit is more difficult with multiple authors, let alone large teams. As explained in the third section, all these factors encourage neat small-scale "science" studies. Partly because of the limited empirical warrants that such studies provide, there is a continuing acceptance of commentary in the humanities tradition – interesting and plausible new ideas get published, noticed and cited, despite the paucity of evidence on their validity and generalizability. Replication, a key element in scientific research, is simply not sexy.

All this does much to explain why education lacks a body of generally accepted research results; in other research-based fields there is often intense disputation, but only at the cutting-edge of new research. There *is* a modest body of research in education that is beyond dispute within fairly well-defined boundaries. To take one example, there is a "common sense" policy in some US states of making students who fail repeat a grade; yet many studies have shown that this produces little or no improvement in performance for most students and a large drop-out rate. Many design principles, like those mentioned in 3.3,■■AU, do you refer to a section, figure or table?■■ are supported by a solid body of evidence from design research (though much of it is unpublished). There are other examples. But building a growing body of reliable evidence requires careful work, with replication across a variety of circumstances to establish boundaries of validity of the insights involved. Because such work does not fit the current academic value system, little of it is done.[8]

How is this avoided in other fields?

What can education learn from science, engineering and medicine that would mitigate these pressures and improve the value system for research? There are various elements. In every field of research, significant new ideas and discoveries always have the highest prestige – but they have to earn it. Because there is an established body of research results, and theoretical models that reflect it, any new suggestion will have implications – so new results must be tested. Other researchers in the same area will seek to replicate the ground-breaking study, to probe its research design and analysis for weaknesses and alternative explanations of the result. There is prestige in being active in these sub-communities.

In many fields, the key experiments can *only* be done well by substantial teams over periods of years.[9] (The core of my argument is that education is such a field.) Mechanisms have been developed for giving appropriate credit to individuals, according to their contributions to the work of the team. Ph.D. students are given specific jobs of experimental design, construction or analysis to carry through, and to write up in the wider context of the whole experiment as their dissertation.[10]

Underlying all this is, of course, money. In science, engineering and medicine it is accepted that serious research needs explicit funding, for the salaries of research team including the time of leading academic researchers, and for the equipment and running costs of the enterprise. This has led to billion dollar budgets in science, engineering and medicine with government-funded initiatives that, if successful, are taken over and developed further by research-based industries. Antibiotics, nuclear energy, electronics, the internet and the world wide web are only some of a broad spectrum of examples where this has happened.

What is the situation in education? Tens of thousands of people in universities around the world do research as part of their academic work. While there is little or no marginal funding for most people, the total cost of their research time is substantial.[11] Could the impact be increased by a more coherent system?

There are agencies that fund research in education, but they have budgets that are an order of magnitude smaller than for science, engineering and medicine. History may help us to understand why. Research budgets in science and medicine were small a century ago; they boomed only during and after the second world war, when these fields produced results with a practical payoff that society recognized and wanted, including the notable examples just mentioned. Though the need in education is well-recognized, educational research has yet to make that breakthrough. *To do so, it will need to have a direct beneficial impact that society recognizes*.

Which brings us back to "importance," the third dimension in Alan's classification of research studies. The discussion so far implies that criteria for assessing importance should take impact on practice very seriously. For this the engineering research approach provides the cutting edge of the research enterprise, turning reliable insights from other research into design principles, tools and processes of direct use in practice. Equally, this needs reliable insights from science research to build on. It is encouraging that funding agencies in education tend to put most of their money into studies that they believe will have direct impact on practice. They are still far from achieving the kind of coherent support that is summarized in Table 5.

Why doesn't it happen? A key reason is the absence of serious evaluation. There are few substantial studies of widely available materials. Those there are tend to be profoundly inadequate, often looking only at student learning outcomes – usually scores on tests that assess only a subset of the learning goals. The ambitious *What Works Clearinghouse* review of mathematics curricula illustrates many of the problems, both in methodology and in lack of adequate research input. Schoenfeld (2006) vividly tells the unhappy story.

While if they were well done, such comparative reviews might help client school systems make better informed choices, they give no guidance on how to improve the products. For that one needs to know, in detail like that discussed above:

- what actually happens in classrooms
- with teachers at various levels of professional development
- using specific materials of various kinds
- with students of various abilities and backgrounds, as well as
- outcomes across the whole range of goals.

The skills needed for such work are in the mainstream of insight-focused research in education but the scale means that it needs large teams, and is therefore expensive. I estimate that to get enough high-quality information to guide the next round of improvement to the NSF mathematics curricula would cost around $100 million, comparable to what was spent in their development. Such knowledge in depth would move the field forward. It looks expensive but we will show that such costs are trivial in the context of the education system.

As it is, published curricula are evaluated the same way that movies, plays and restaurants are reviewed. The differences between well-presented draft materials and a well-engineered product that works well are not obvious on inspection. So it is not surprising that publishers see no need to pay the higher costs of research-based development. As a result, education has no research-based industry of the kind that, in other research-led fields, takes much of the engineering load of turning prototypes into robust products.

The new balance – A vision for an effective research community

Let us look in a bit more detail as to the sort of pattern of research that would make educational research the "go to" community for policy makers seeking to improve education, as medical research is when health issues arise. Table 5 makes it clear that major changes are needs, leading to coherent ongoing programs of research and development. There are many ways this might be achieved. Here I outline a model that draws together the diagnoses of system problems so far into an explicit "solution" that might provide, at least, a basis for useful discussion.

The changes I envisage include three strands, listed here with their aims in terms of the knowledge, goals and beliefs behind government decision making:

- *Evaluation*, so that both current problems and the impact of initiatives can be recognized and understood.
  This will enhance government knowledge of the current situation and, more importantly, provide evidence to encourage their currently-intermittent belief that these problems need well-engineered solutions.
  This will include both survey research and the collecting of much more detailed information on the implementation and outcomes from specific initiatives, independently carried out but on a basis, and using research tools, agreed

with the developers and their funders. These studies will have a formative focus, as well as providing summative information to guide policy.

- *Development*, so that, in recognised problem areas, well-engineered products and processes are developed to help professionals realize their and the system's goals more effectively.

  This will reinforce government knowledge of effective change processes, and gradually undermine their belief that "the profession" will be able to find good solutions to any problem (a belief that all professions encourage). This will be based around ongoing programs in specific areas by established research teams, with two or three working in parallel on major challenges (again as in frontier science and engineering).

- *Cumulative research*, so that the community builds a body of research, with established reliability and bounds of validity, that goes beyond "worth paying attention to," providing a solid foundation: for design, better than authors' experience; for policy, better than politicians "common sense."

This will encourage governments to ask for advice from the research community, and to take it, recognizing that there is a zone of reliable knowledge they do not own. (Advice to government in other fields is always based on the accepted body of research results, plus warnings of uncertainties.)

This approaach will require building research collaborations in each important area, with groups doing parallel studies on important issues in varied but related circumstances using common treatments and instruments. The challenges of tool development (as in the fourth section) and the collecting and analysis of adequate data sets will be shared in a co-ordinated way. The goal of each group is a set of results that can only be challenged at the boundaries.

Note that there is no mention here of changing the educational goals of government. There are, of course, disagreements – for example, about the appropriate balance between general education and specialised study and training. However, much the largest mismatch is between current shared intentions and actual outcomes in practice. Finally, one must never forget the prime goal of democratic governments: to get re-elected, which militates against controversial change and spending money. However, many governments have made a commitment to evidence-based policy, at least at the rhetorical level. There are some intermittent signs that they will move forward with this on some fronts.

Funding will be needed for most of these things to happen. The next section estimates the costs of doing research-for-practice reasonably well. However, it is worth looking at what might be achieved within the enormous existing resources represented by the research time of the academic community. There are opportunities.

Evaluation of the kind sketched above lies within the skill set, if not the current practice, of educational researchers. Given its crucial role in convincing politicians that research pays off in their terms, it is here that the best route to bootstrapping a substantial investment in research may lie.

Building an accepted research base offers a major opportunity to the research communities to undertake longer-term research with replication to explore the generality, and the boundaries, of interesting results.

Engineering and design research teams enjoy relatively good financial support from government, reflecting their perceived value in developing robust solutions to difficult challenges. However, their funding is rarely even medium-term, each project being a one-off; closer links with the research enterprise in their institutions could "bridge" the funding gaps more effectively than at present, if more of their colleagues saw design and engineering research as of value.

These things all imply that collaboration must be recognized as positive, requiring changes in the current academic value system. This remains a major challenge. Money can help: even modest amounts of funding would allow these things to happen, and give researchers some feeling of recognition that is different from acceptance of their papers by journals. At least as rewarding is for researchers to see their work having beneficial impact on children and teachers; specific mechanisms for this should be part of research designs. Most academic researchers will continue doing what they do but there are enough of them for even a modest shift in the balance of research styles to have real impact.

What would all this cost?

A research-based approach costs much more than simple authorship – the standard approach in which experienced professionals write down and publish what has worked for them, without thorough developmental testing. Research-based design and development normally needs several rounds of trials, with rich and detailed feedback from a variety of classrooms guiding the revision and refinement of the products. It becomes part of a continuing program of formative feedback, which contributes both new insights and new products to the overall program.

One can get a rough estimate of costs from some examples; in current terms, adjusted for inflation:

- NSF mathematics curricula in the US were funded in 1990 at rather more than $1 million per school year of 180 lessons; the second round of implementation funding plus inflation raises the cost to around $15,000 per lesson.[12]
- Shell Centre development of 3 week "replacement units" in the 1980s cost £100,000 for 15 lessons, around $30,000 per lesson now.
- The formative assessment lessons in our current development are costing around $30,000 per lesson.
- If we accept $30,000 as a typical estimate, what would the cost implications of this approach be for the whole curriculum in the US? Let us err on the high side:
    - 25 hours a week for 40 weeks a year for 13 years ~ 13,000 hours;[13] double this for children with different needs
    - $30,000 per hour lesson ~ $800 million

A round of total re-development would take at least 10 years ~ \$80 million per year The annual running cost of the US K-12 education system ~ \$400 billion.

Cost of high-quality materials development ~ 0.02% of turnover!

If a country won't spend that, it isn't serious – or it doesn't believe a research-based approach has significant advantages. In practice, not everything will need to be redeveloped every time . Clearly, cost should not be an issue; selling the concept of research-based development, and its more effective organization, are the challenges.

Technology may possibly offer a way forward here, because the costs of programming justify the costs of systematic design and development. So far technology has had minimal impact on modes of learning in mainstream mathematics and science education, which have become seriously out of line with the way mathematics and science are done outside school. Small scale work over the last 30 years has shown enormous potential in many diverse modes of use of technology, but no curricula in which technology is fully embedded have yet been developed. This is largely because of a mismatch of timescales: a seriously innovative curriculum takes 10–25 years to develop while the technology changes every few years, so there has been no stable "platform" for which to design. That situation may be changing. There are exciting current initiatives that are developing curricula without printed materials, where every student has a tablet computer. However, the challenge of doing this well tends to be underestimated. Realizing the potential of the technology will need fine designers who have explored and absorbed its affordances, so they can again focus on students and teachers.

The status and roles of "theory"

Finally, as a coda to this chapter, some comments on theory. Theory is seen as the key mark of quality in educational research. I am in favour of theory. (Indeed, in my other life, I am a theoretical physicist.) However, in assessing its roles in any field, it is crucial to be clear as to how strong the theory is. From a system point of view (Burkhardt, 1988), the key question is:

How far is this theory an adequate basis for design?

Again, it useful to look across fields. In aeronautical engineering, for example, the theory is strong; those who know the theory can design an airplane at a computer, build it, and it will fly, and fly efficiently. (They still flight test it exhaustively.) In Medicine, theory is relatively weak, but getting stronger. Despite all that is known about physiology and pharmacology, much development is not theory-driven. The development of new drugs, for example, is still often done by testing the effects of very large numbers of naturally occurring substances; they are chosen intelligently, based on analogy with known drugs, but the effects are not predictable and the search is wide. However, as fundamental work on DNA has advanced, and with it the theoretical understanding of biological processes, designer drugs with much more theoretical input have begun to be developed. This process will continue –

indeed there is now work, for example, on cancer drugs tailored to an individual's specific tumour.

In the range and reliability of its theories education is a long way behind medicine (perhaps 100 years), let alone engineering (at least 350 years). The much-quoted theories in education are ambitious. By overestimating their strength, damage has been done to children – for example, by designing curricula based on behaviorist theory. The current dominance of constructivism is similarly inadequate, though less dangerous. Its incompleteness is more obvious, since it is impossible to design a curriculum built only from constructivist principles. It is not that behaviourism or constructivism are wrong; indeed, they are both right in their core ideas, but they are incomplete and an inadequate basis for design. Physicists would call them "effects." The harm comes from overestimating their power, ignoring other effects.

Let me illustrate this distinction with an example from meteorology. Air flows from regions of high pressure to regions of low pressure sounds and is good physics. It implies that air will come out of a popped balloon or a pump. It also implies that winds should blow perpendicular to the isobars, the contour lines of equal pressure on a weather map, just as water flows downhill, perpendicular to the contour lines of a slope. However, a look at a weather map shows that the winds are closer to parallel to the isobars. That is because there is another effect, the Coriolis Effect. It is due to the rotation of the earth which twists the winds in a subtle way, clockwise around low pressure regions. (They go round the other way in the Southern Hemisphere.) In education there are many such effects operating. We have mentioned some of them but, as in economics, it is impossible to predict just how they will balance out in a given situation.

Some more modest theories have a better track record. "Teaching to the test" in systems with high-stakes testing is a good example; it summarises a general reality. The first two cases in the fourth section also exemplify this. Alan's studies of teaching, outlined earlier, provide solid evidence that knowledge, goals and beliefs are key variables to focus on – a valuable theoretical guide in the design of professional development, which has often chosen a much narrower agenda, often just knowledge of mathematics. The concept of "role shifting" and the way it deepens mathematical discourse in the classroom emerged from the study in the section "Classroom discussion focus"; it has since proven a robust design principle. Table 7 shows an example (Swan, 2008) of theory in the design of teaching materials in mathematics focused on conceptual understanding.

These more modest theories, sometimes called heuristics, are *phenomenological* in that they may be seen as summarizing a body of data on a group of phenomena. Every research field relies on such theories. An example from physics and engineering is Young's theory of elasticity. It says that how much a body stretches is proportional to how hard you pull it, with a constant of proportionality "Young's modulus" that is a property of the material. This phenomenological theory also covers what happens if you pull it too hard, notably when it breaks. The fundamental theory underlying this is quantum mechanics. (Young's modulus for metals is one of the few cases where you can actually calculate the coefficient from

Table 7. An example of phenomenological theory (Swan, 2008).

Teaching design for conceptual understanding is more effective when we:

- Use rich, collaborative tasks. The tasks we use should be accessible, extendable, encourage decision-making, promote discussion, encourage creativity, encourage "what if?" and "what if not?" questions. Students should not need to start or finish at the same point, enabling everyone to engage with the activity.
- Develop mathematical language through communicative activities. Mathematics is a language that enables us to describe and model situations, think logically, frame and sustain arguments and communicate ideas with precision. Students do not know mathematics until they can "speak" it. Interpretations for concepts remain mere "shadows" unless they are articulated through language. We find that many students have never had much opportunity to articulate their understanding publicly.
- Build on the knowledge learners already have. This means developing formative assessment techniques so that we may adapt our teaching to accommodate learning needs. Lessons do not follow the traditional pattern for explanation followed by exercise. Instead, the teacher asks expose and assesses existing ways of thinking and reasoning before explaining. The teacher listens to the discussions before joining in, then attempts to prompt students to articulate their thinking and reasoning. Teacher explanation follows this discussion, it does not pre-empt it.
- Confront difficulties rather than seeks to avoid or pre-empt them. Effective teaching challenges learners and has high expectations of them. It does not seek to "smooth the path" but creates realistic obstacles to be overcome. Confidence, persistence and learning are not attained through repeating successes, but by struggling with difficulties. Conceptual obstacles are part of design, deliberately included to provoke discussion.
- Expose and discuss common misconceptions and other surprising phenomena. Learning activities should expose current thinking, create "tensions" by confronting learners with inconsistencies and surprises, and allow opportunities for resolution through discussion. The activities encourage misconceptions and alternative interpretations to surface so that they may be discussed. Conflicts originate both internally, within the individual, and externally, from an individual's interpretation of another person's alternative viewpoint.
- Use higher-order questions. Questioning is more effective when it promotes explanation, application and synthesis rather than mere recall. Teachers are encouraged to prompt students to reflect and explain through the use of open prompts that begin "Explain why ... "; "Show me an example of ... "; "How do you know that ... ?"
- Make appropriate use of whole class interactive teaching, individual work and cooperative small group work. Collaborative group work is more effective after learners have been given an opportunity for individual reflection. Activities are more effective when they encourage critical, constructive discussion, rather than argumentation or uncritical acceptance. Shared goals and group accountability are important. Teachers are advised to gradually establish "ground rules" for discussion among students and then behave in ways that encourage dialogic and exploratory talk.
- Encourage reasoning rather than "answer getting." Often, learners are more concerned with what they have "done" than with what they have learned. Aim for depth rather than for superficial "coverage," telling students that comprehension is more important than completion. The teacher's role is to prompt deeper reasoning by asking students to explain, extend and generalize.
- Create connections between topics both within and beyond mathematics. Learners often find it difficult to generalise and transfer their learning to other topics and contexts. Related concepts remain unconnected. Effective teachers build bridges between ideas, so design in multiple connections between different representations.
- Recognise both what has been learned and also how it has been learned. What is to be learned cannot always be stated prior to the learning experience. After a learning event, however, it is important to reflect on the learning that has taken place, making this as explicit and memorable as possible. Allow students to share their findings through the public display of their work. Encourage students to extend and generalise their ideas by making small changes to the examples, and then to explicitly formulate rules for equivalence. This helps the teacher recognise and value the contributions of students, extending and institutionalising them.

the underlying theory.) *However, such phenomenological theory is key in airplane design*.

What do phenomenological theories in education look like. Like the examples in the fourth section, they are specific and well-defined. The set of design principles in Table 7 builds on Malcolm Swan's own research (Swan, 2005) and earlier work by the Shell Centre team and by other design researchers. They are an example of phenomenological theory that has developed and proven robust over many years of application to the design of materials; nonetheless they and the field could benefit from further replicative studies.

I believe that the research enterprise should devote more effort to developing solid reliable phenomenological theories for specific areas, reflecting the balance of research in other fields. The growth of design research, which has this agenda, is encouraging. Such phenomenological theories build evidential warrants through further testing of their robustness and limitations, by their creators and by other designers. This process will, over time, build a knowledge base that others can rely on.[14]

However, it would be to repeat the common mistake to overestimate the completeness of theory. *In design, details matter* – they have important effects on outcomes that are not determined by theory. For the foreseeable future, design skill and empirical development will remain essential for turning research into tools to support practice, with theoretical input providing useful heuristic guidance.

ACKNOWLEDGEMENTS

This chapter has developed over the last decade from the concluding talk that I gave at the ICMI Algebra Study conference in Melbourne in 2001. I am grateful to Kaye Stacey for the stimulus of that invitation and for many discussions on issues of methodology, before and since. My thinking has developed through conversations over many years with Malcolm Swan and our colleagues at the Shell Centre, and with others including Paul Black, Phil Daro, Glenda Lappan, Günter Törner and, of course, Alan Schoenfeld.

NOTES

[1] All theoretical models in science have limits of validity. “Universal theorems are for mathematics, certainly not for mathematics education” (Henry Pollak).

[2] In other fields, these carry comparable prestige. The physicist John Bardeen won two Nobel Prizes, one of each kind, for the invention of the heart of modern electronics, the transistor, and for the theory of superconductivity.

[3] Even here, there is variation; some patients do not take their drugs as prescribed.

[4] “Uncertainty” is a better term; “error” often implies that “somebody made a mistake.”

[5] Thirty years later, electronic whiteboards are now widely available – and perfect for this mode of use.

[6] That is why we described this as a “teaching assistant” mode of computer use, hence ITMA.

[7] Although there are myriad variants of classroom activity structures, we have found that the following five types span most of the activities of interest, and that almost every classroom episode is one of these types.

[8] The Campbell Collaboration http://www.campbellcollaboration.org/ECG/Education/index.php, modelled on the Cochrane Collaboration in medicine, seeks to establish a body of accepted results through metanalysis but, in my view, the absence of a stream of replication studies means that it lacks the "feedstock" for such an approach. The Bush administration's "What Works Clearinghouse" suffered both from that and from a deeply flawed methodoology.

[9] Particle physics is an extreme example of "big science." The experiments at the CERN Large Hadron Collider have involved thousands of Ph.D. physicists, engineers and computer scientists over two decades, costing billions of dollars, with more to come. Papers will have hundreds of authors. This may be unattractive to some but, when it has to be done, it can be.

[10] It is worth recalling that the Ph.D. was created as a research training degree, in contrast to other doctorates (D.Sc., D.Litt, etc.) that reflect substantial professional achievement.

[11] 10,000 people on salaries of $50,000 spending 40% of their time on research, probably an underestimate, totals $200 million a year.

[12] These are order-of-magnitude estimates, avoiding "spurious precision."

[13] *Fifteen thousand hours* is the title of a famous UK study of schools (Rutter et al., 1982).

[14] This is one of the strategic goals of ISDDE, the International Society for Design and Development in Education. http://www.isdde.org/isdde/index.htm.

REFERENCES

Beeby, T., Burkhardt, H., & Fraser, R. (1980). *SCAN: A systematic classroom analysis notation for mathematics lessons*. Nottingham: Shell Centre.

Bell, A. W. (1976). *On the understanding of proof*. Ph.D. Thesis, University of Nottingham

Bell, A. (1993). Principles for the design of teaching. *Educational Studies in Mathematics*, *24*(1), 5–34.

Bell, A., Swan, M., Onslow, B., Pratt, K., & Purdy, D. (1985). *Diagnostic teaching for long term learning*. Report of ESRC Project HR8491/1: Shell Centre for Mathematical Education, University of Nottingham.

Bell, A., Swan, M., Crust, R., & Shannon, A. (1993). *Awareness of learning, reflection and transfer in school mathematics*. Report of ESRC Project R000-23-2329. Shell Centre for Mathematical Education, University of Nottingham.

Black, P. J., & Wiliam, D. (1998) Assessment and classroom learning. *Assessment in Education*, *5*, 7–74; see also *Inside the black box: Raising standards through classroom assessment*. London: King's College London School of Education (2001).

Burkhardt, H. (1988). The roles of theory in a "systems" approach to mathematical education, article in honor of prof. Hans-Georg Steiner's 60th birthday. *International Reviews on Mathematical Education, ZDM*, *5*, 174–177.

Burkhardt, H. (2006). From design research to large-scale impact: Engineering research in education. In J. Van den Akker, K. Gravemeijer, S. McKenney, & N. Nieveen (Eds.), *Educational design research*. London: Routledge.

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

Burkhardt, H., Fraser, R., Coupland, J., Phillips, R., Pimm, D., & Ridgway, J. (1988). Learning activities & classroom roles with and without the microcomputer. *Journal of Mathematical Behavior*, *6*, 305–338.

Elmore, R. (2011). The (only) three ways to improve performance in schools. http://www.uknow.gse.harvard.edu/leadership/leadership001a.html; see also *Instructional rounds in education: A network approach to improving teaching and learning*. Cambridge: Harvard Education Press.

Good, T., & Brophy, J. (2002). *Looking in classrooms*, 9th edition. Allyn and Bacon.

MAP. (2012). Material from the Mathematics Assessment Project may be found at http://map.mathshell.org.uk/materials/.

RAE. (2001). http://www.rae.ac.uk/2001/Pubs/1_98/1_98cd.html.

Rutter, M., Maugham, B., Mortimore, P., & Elston, J. (1982). *Fifteen thousand hours: Secondary schools and their effects on children*. Cambridge: Harvard University Press

Schoenfeld, A. H. (1980). On useful research reports. *Journal for Research in Mathematics Education*, *11*(5).

Schoenfeld, A. H. (1985). *Mathematical problem solving*. Orlando, FL: Academic Press.

Schoenfeld, A. H. (1992). On paradigms and methods: What do you do when the ones you know don't do what you want them to? Issues in the analysis of data in the form of videotapes. *Journal of the Learning Sciences*, *2*, 179–214.

Schoenfeld, A. H. (1994). A discourse on methods. *Journal for Research in Mathematics Education*, *25*, 697–710.

Schoenfeld, A. H. (2002). Research methods in (mathematics) education. In L. English (Ed.), *Handbook of international research in mathematics education* (pp. 435–488). Mahwah, NJ: Erlbaum.

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

Schoenfeld, A. H. (2007). Method. In F. Lester (Ed.), *Handbook of research on mathematics teaching and learning* (second edition, pp. 69–107). Charlotte, NC: Information Age Publishing.

Schoenfeld, A. H. (2010). *How we think: A theory of goal-oriented decision making and its educational applications*. Routledge.

Swan, M. (2005). *Improving learning in mathematics: Challenges and strategies*. Sheffield: Teaching and Learning Division, Department for Education and Skills Standards Unit.

Swan, M. (2008). Designing a multiple representation learning experience in secondary algebra. *Educational Designer*, *1*(1), http://www.educationaldesigner.org/ed/volume1/issue1/article3/index.htm.

AFFILIATION

Hugh Burkhardt
Shell Centre, University of Nottingham, UK
University of California, Berkeley, USA