



Not just “implementation”: the synergy of research and practice in an engineering research approach to educational design and development

Hugh Burkhardt¹ · Alan Schoenfeld²

Accepted: 17 November 2020 / Published online: 4 December 2020
© The Author(s) 2020

Abstract

This paper builds on a range of traditions in educational research and design to argue, with empirical evidence, that constructing powerful instructional materials and approaches that work at scale requires a grounding in theory and a commitment to engineering practice, including rapid prototyping and multiple development cycles. Specifically, we claim that improving practice within a reasonable timescale requires replicable materials that integrate: (1) grounding in robust aspects of theory from prior research, (2) design tactics that combine these core ideas with a design team’s creativity, along with (3) flexibility in the draft materials that affords adaptation across contexts, (4) rapid prototyping, followed by iterative refinement cycles in increasingly realistic circumstances, with (5) feedback from each round of trials that is rich and detailed enough to inform revision, and (6) continued refinement on the basis of post-implementation feedback ‘from the field’. Examples of successful implementation are analysed and related to the various roles that research-based theory and programmatic research-based methods of development can and should play in the complex process of turning insights from research into improvements in practice. In contrast, we shall argue that materials which are written and published without the development processes (4) to (6)—still the great majority—lack research validity for use at scale.

1 Introduction

We begin with our title: *Not just “implementation”*. This Special Issue is about *implementability*, which implies a focus on implementation. Indeed it is widely believed, in the insight-focused research community and beyond, that turning an exciting new research result into products and processes that work well in practice is a straightforward process that might reasonably be called implementation (see e.g. Royal Society/British Academy 2018). This is hopelessly naïve. Even in a relatively simple case—the development of a revolutionary new drug in medicine—it took well over 10 years and concerted efforts due to the pressures of war to turn Alexander Fleming’s observation that a speck of mold on his Petri dish appeared to kill the bacteria growing there into the first usable dose of penicillin. Many other people made crucial Nobel-prize winning contributions along the

way, notably Howard Florey and his team, Kane and others in designing and developing a method for the essential transition from growing the mold on surfaces to producing it in significant quantity by using fermentation tanks. Many other research results—some established, some new—were involved and others emerged from the development process.

In education, it is rare for a single research result to form the basis for a change in practice. Where it has been tried it has often proved counterproductive. For example, treating the behaviorist approach to learning as a comprehensive theory, rather than as one *effect* (Burkhardt 1988) among the many that operate in learning situations, proved a simplistic dead end. More recently, the power of phonics in decoding text into sounds has driven a simplistic view of teaching young students to read. Both effects are valid and important but, as theories, far from complete. Simple ideas are almost always too simple for complex educational contexts and systems.

While exciting new research ideas may stimulate a search for change, turning such findings into effective improvements in practice is a complex process involving people with quite different complementary skills and insights—particularly in design—and a systematic learning process

✉ Hugh Burkhardt
Hugh.Burkhardt@nottingham.ac.uk

¹ University of Nottingham, Nottingham, UK

² University of California, Berkeley, USA

often called development that also uses research methods. A whole body of established research must be taken into account—much that is most useful coming from studies-in-depth of innovative practice, whether those studies are formalized as research or not. The essential attributes of this process are the theme of this paper, which builds on earlier work (Schoenfeld 2002; Burkhardt and Schoenfeld 2003; Burkhardt 2006, 2009). We will not attempt to review most of the other work in various places on more-or-less similar lines. Going back to the 1960s in STEM education, such systematic development has rarely been encouraged in academic circles, which have viewed its focus on improving products and processes as ‘not really research’. We are aware, of course, of advances in design research and design experiments, a tradition dating back to papers such as Brown (1992) and Collins (1992). We note that the vast majority of studies in that tradition, while appropriately iterative, have not been trialed in an increasingly broad range of circumstances that approximate “real world” implementation.

Where does that leave *implementability*? Can one predict whether a research result can be turned into a valuable contribution to teaching and learning? We return to that question in the Conclusions. Meanwhile, in Sect. 2 we look in some detail at successful implementations and key features they share. Section 3 looks at this ‘engineering research’ methodology while Sect. 4 looks at the various roles that theory can and should play in advancing implementability. Section 5 looks at the challenges of achieving significant impact on practice at system level, leading into the Conclusions.

2 What implementation models seem to work well?

It is useful to examine some improvement initiatives that are widely recognized for the quality and large-scale impact of the outcomes they have produced, with an eye toward lessons learned. We shall look at some examples, two in detail, that suggest some principles for analysis in the following sections as we look for patterns to guide a research program that takes implementability seriously. The two examples we feature were the first products to be awarded the Prize for Excellence in Educational Design of the International Society for Design and Development in Education (ISDDE).

2.1 Aspects of implementability

It may be helpful to point out in advance the key features of strategic design (Burkhardt 2009) that are common to these examples, and to many other successful materials. They include:

- a coherent long-term research and development program, with insights flowing in both directions
- design creativity in turning learning principles into stimulating learning activities—for students and for teachers
- research-in-depth on the design, and the theoretical principles that underlie it, established across examples of use in diverse situations
- constructive engagement with as many of the stakeholder groups as possible including subject teachers, subject specialists, the research communities, external assessment providers, school leadership, and school system policy makers
- iterative development across all the key implementation variables: students, teachers, school support, etc.—in particular, explicit support with the pedagogical challenges that the innovation presents
- retaining ‘design control’ to ensure integrity over the long term.

These factors are standard in many fields of practical importance that are research-based, such as engineering or medicine, but are still the exception in education. Missing out on any of them and the empirical evidence they provide through the research-design-development-research cycle is likely to reduce the effectiveness of the implementation. The reader may like to note their roles in the examples that follow.

2.2 Connected Mathematics

This program set out to design a comprehensive set of teaching materials for US middle school students aged 11–14. It was one of thirteen projects funded by the US National Science Foundation to help teachers implement the *Standards* set out by the National Council of Teachers of Mathematics (National Council of Teachers of Mathematics 1989) which had been developed over the previous decade, integrating a combination of research results and “best practice”. This and several related documents represented a unique time as the US mathematics education community enthusiastically embraced the *Standards* and the vision it represented for moving the field forward—particularly the focus on mathematical processes such as problem solving, reasoning, making mathematical connections and communicating with mathematics.

The funding of these projects, at around \$1,000,000 for each grade, was generous by the prevailing standards of curriculum development funding. Viewed strategically, however, this required teams of a dozen or so to develop each lesson in less than 2 days work in total. This is a considerable challenge for which the *Connected Mathematics Project* (CMP) was unusually well-prepared. CMP was able to build on twenty years of prior R&D at Michigan State, led by

Glenda Lappan, Elizabeth Phillips and Bill Fitzgerald. This earlier work in the *Middle Grades Mathematics Program* included the development of teaching materials along with a multi-year professional development program, widely seen to exemplify high-quality teaching as reflected in the *Standards*. Simultaneously the team conducted research around the learning and teaching of algebra and functions. From their ongoing research and development work CMP articulated “a guiding philosophy and commitment” for materials design:

- Love for mathematics; looking for the big ideas, what it means to understand these ideas, finding ways to embed these in a sequence of problem-solving activities; building on and connecting to other big ideas
- Passion and commitment to making a difference, sustained over the long term
- Focus on curriculum design, research, revision, evaluation, research, revision, evaluation, ...
- Continued focus on research and development on teaching and learning of mathematics and other areas that affect curriculum design and implementation
- Close contact with middle school classrooms including the team’s experience in teaching these ideas in middle school and pre-service education
- Simultaneously attending to both student and teacher learning
- Providing support for teachers in conferences, workshops and website, always making the support for teachers respectful, and listening to feedback
- Willingness to balance idealism and practicality
- Never losing sight of the ideals (like inquiry, openness) through changing pressures from publishers, standards, etc.

CMP’s track record explains the NSF support for *Connected Mathematics*. In (Lappan and Phillips 2009) the lead designers give an in-depth description of this program. Here we have space only to note and exemplify some key features, the first of which is the substantial research and development program over many years on which it was based. Equally important, the team had built up a large community of teachers who were partners in the design and development process. This established the base for iterative cycles of R&D, in instructional contexts that reflected the target population for the materials.

The next core feature is that CMP adopted a context-based approach, relating mathematics to explicit situations in the real world and within the subject—the *connections* of the program’s title. For example, well-chosen real world situations with two variables make it relatively easy to ask: How are the variables related? How can the relationships be represented in tables, graphs, and symbols? If you know

the value of one variable, how can you find the value of the other? These ideas are central to CMP, in which algebra is a major strand.

Consider a familiar topic: an introduction to quadratic functions (see Lappan and Phillips 2009). CMP uses the relations between area and perimeter for rectangles. It starts from the context of “staking claims” in the California gold rush, asking students to investigate the ‘sheep pens’ problem—how big a rectangle you can rope off with a given length of boundary fence. Students are asked to try some examples with different side lengths, investigating systematically how the problem can be represented in tables, then graphs using specific numerical examples viewed in multiple ways, fixing one quantity then another, leading up to the general statements in algebraic form. This reflects research that documents the value of spending time to look in depth at one rich problem from different perspectives, as opposed to working through a sequence of closely similar exercises. The unit also brings out the differences from linear and exponential relationships, setting quadratics in the broader context of functions.

The third key feature is attention to the pedagogical challenges that a “thinking curriculum” poses. Addressing pedagogical support for the teacher, Elizabeth Phillips explained to us:

“As we wrote CMP, we did so with students sitting on one shoulder and teachers on the other. Many times, decisions about what would go in the student book were based on teacher needs. For the better part of the 1990s we met with other NSF curriculum projects; our stance on teacher needs was unique. Many curriculum developers write teacher support materials after the student books are finished. We wrote extensive teacher support and developed professional development activities side by side with the student materials. Creating an instructional model—*Launch, Explore, and Summarize*—focusing on the embedded mathematics in the problems was a critical part of the success of our professional development.”

With this instructional model in mind, CMP teacher support materials include suggestions on how to plan for instruction. These are not meant to be algorithmic, but to provide teachers with ways of thinking about their planning and enactment of lessons. The design team imagined themselves as sitting on the shoulder of a teacher and having a conversation about a set of possible ways to engage the students in a lesson. Each lesson provides, as well as with detailed instructions for the students in the Explore phase, examples of questions that can be used to elicit student thinking, to push students’ thinking toward deeper mathematical insights, and questions that can be used at


<p>Each participant in the walkathon must find sponsors to pledge a certain amount of money for each kilometer the participant walks. The students in Ms. Chang’s class are trying to estimate how much money they might be able to raise. Several questions come up in their discussions:</p> <ul style="list-style-type: none"> • What variables can affect the amount of money that is collected? • How can you use these variables to estimate the amount of money each student will collect? • Will the amount of money collected be the same for each walker? Explain. <hr/> <p>Each student found sponsors who are willing to pledge the following amounts.</p> <ul style="list-style-type: none"> • Leanne’s sponsors will pay \$10 regardless of how far she walks. • Gilberto’s sponsors will pay \$2 per kilometer (km). • Alana’s sponsors will make a \$5 donation plus 50¢ per kilometer. <p>The class refers to these as <i>pledge plans</i>.</p> 	<p>What variables can affect the amount of money collected?</p> <p>How can you estimate how much money each student will collect?</p> <p>Will the amount be the same for each walker? Explain.</p> <p><i>then, after the detailed work in the Exploring phase</i></p> <p>How is the cost per kilometer similar to a person’s walking rate?</p> <p>How can you recognize the patterns are the same in a table graph or equation?</p> <p>Describe another pledge plan (and give its equation) whose graph is a horizontal line.</p>
<p>Raising Money - the task</p>	<p>Some suggested questions</p>

Fig. 1 Connected Mathematics instructional model (from Lappan and Phillips 2009)

the end of a class as a quick formative assessment to help in planning for the next lesson.

In any new comprehensive curriculum there is a delicate strategic design choice of ‘how big a step to take’—too big and few will follow, too small and why bother? The designers of Connected Mathematics found a successful balance. Figure 1 illustrates this from a lesson on linear relationships that asks students to compare ‘pledge models’ for a sponsored walk. The questions show how, in the Launch and Summarize phases, designers encourage the teachers to help students take a higher-level view of the mathematics they have been exploring in detail (and with close guidance). The materials include detailed answers to each question, some of which are mathematically challenging. It should be evident that when teachers are being asked to move into new territory, as here, they will need more support than when on familiar ground. This instructional model is widely used in other curricula. Smith’s Five Practices (Smith & Stein 2011) and Herbel-Eisenmann’s discourse moves (Herbel-Eisenmann & Breyfogle 2005) evolved from studying CMP classrooms.

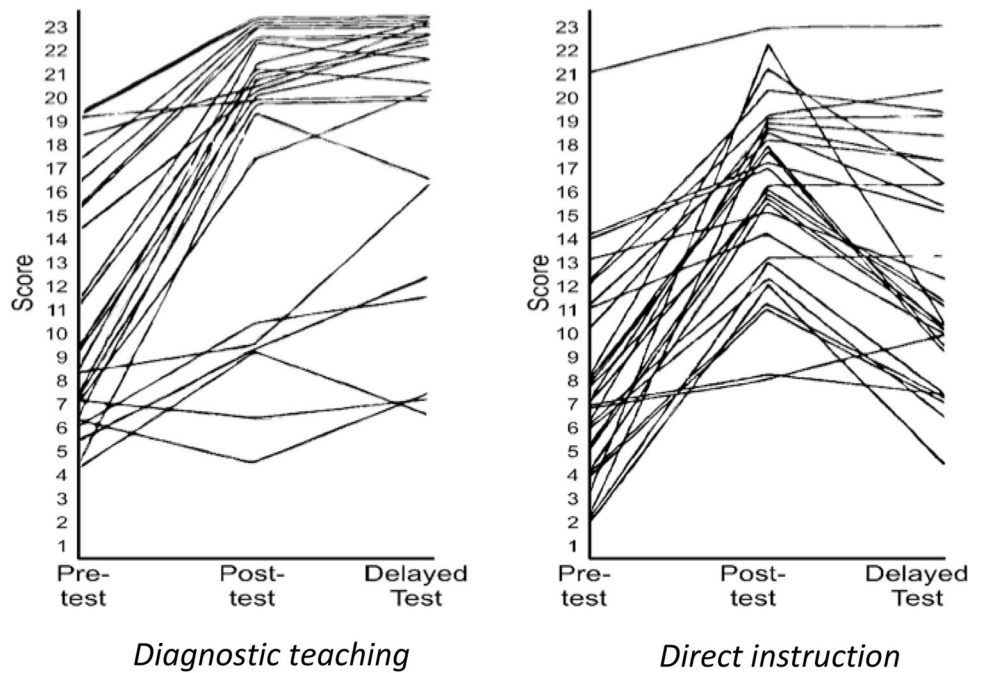
Connected Mathematics has been widely used and admired. Though commercial publishers do not release sales figures, the more than 300 research articles by researchers

who have chosen to study CMP classrooms (see <https://connectedmath.msu.edu/research/completed-research-and-evaluation> for details) is evidence of its influence and importance. Basing research on well-engineered materials that are widely used in this way gives in-depth insight along with greater confidence that the treatment is at-least-roughly stabilised, providing a framework for exploring generalizability across studies and helping the ongoing implementability of the new research (Burkhardt 2016).

Finally, the authors have maintained intellectual control over the content including format, presentation, etc. This is fairly uncommon in the education world, where publishers’ rather different priorities often lead to unintended damage to the design. It may have its downside in marketing terms but it is essential for maintaining the integrity of a long-term design and development program.

Brown and Campione (1996) noted that all curricula undergo mutations in practice. The challenge, they said, is to avoid “lethal mutations,” maintaining some fidelity to the designer’s intentions. The R&D program discussed here both identified key aspects of the CMP work that needed to be maintained in implementation and provided consistency and support in those directions. We think it is a strong factor in the continuing success of CMP.

Fig. 2 Mean scores on pre-, post- and delayed post-tests (from Birks 1987)



2.3 The Shell Centre program

Here we describe a similar multi-decade trajectory of research-based design, development and refinement through successive projects leading to the *Mathematics Assessment Project* (MAP), for which the authors were principal investigators. Although the work was carried out differently, it displays all the key features noted above—long-term coherence around a developing research basis, in-depth student work on rich tasks as the core cognitive demand, close attention to the pedagogical challenges and building a core constituency of users.

Starting in the 1980s Alan Bell, Malcolm Swan and their students explored, in a coherent sequence of small-scale studies, the validity of an approach they called *diagnostic teaching*—a specific approach to formative assessment based on eliciting student thinking in a way that surfaced misconceptions, resolving them through structured discussion, first in small groups and then across the class. The team studied this approach across three key variables: students, mathematical topics, and different designers of the experimental teaching material. Ongoing research also focused on teachers, showing that collaborative discussion materials can be effective when used appropriately, even with typical teachers and low attaining students. This research program (Swan 2006) also offered insights into the ways in which teachers’ beliefs (about mathematics, teaching and learning) affect the ways in which they use teaching materials and, conversely, the ways in which the materials affect beliefs and practices.

The studies showed that diagnostic teaching, when compared with the standard direct instruction approach of the

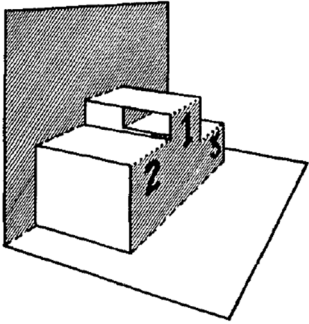
time, showed a common pattern of much improved long-term learning, illustrated by Fig. 2. Note in the right hand graph the subsequent loss, so familiar to teachers, of most of the gains made during the teaching of the unit; this does not occur with the diagnostic teaching approach. This key result was stable across the various parallel studies, providing evidence of generalizability of the design principles.

In addition to sound theory-based design principles, detailed design calls for a mixture of insight and creativity. Two examples, for which Malcolm Swan was the lead designer, must suffice. *Be a Paper Engineer* (Swan et al. 1987–89) encourages students to investigate the geometrical principles used in making pop-up cards and gift boxes, going on to use those principles to make new designs themselves.

The first task in Fig. 3 uses an investigative approach to parallelogram geometry. It sets students the challenge of making a pop-up card, and discovering in the process that there are principles for the positioning of fold and cut lines so that the card does not crease in the wrong place or protrude when the card is closed. This process enables a broad range of students to conjecture and justify the parallelogram theorems that they usually just learn. Later the students explore mathematically more challenging possibilities, where the folds are not all parallel. *The Language of Functions and Graphs* (Swan et al. 1985) was one of the first examples of mathematics curricular material that focused on graphs of real-life situations. The graphing task stimulates rich discussions, revealing that many students misunderstand graphs as pictures, thinking the motion goes up and down because the graph does, perhaps also not noticing it is a speed-time graph. The pedagogy was made explicit

Task from: Be a Paper Engineer

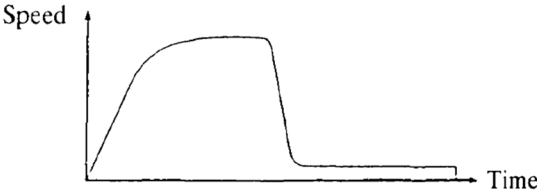
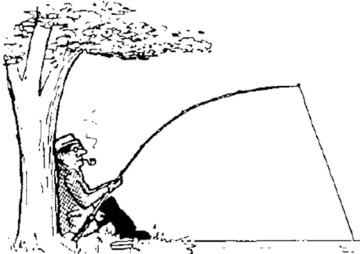
Make this Winners' podium from a single sheet, without using any glue.



From:
<http://www.mathshell.com/materials.php?item=paperengineer&series=numeracy>

Task from: The Language of Functions and Graphs

Which sport will produce a graph like this?

- Fishing
- Pole Vaulting
- 100 metre Sprint
- Sky Diving
- Golf
- Archery
- Javelin Throwing
- High Jumping
- High Diving
- Snooker
- Drag Racing
- Water Skiing

From:
<http://www.mathshell.com/materials.php?item=lfg&series=tss>

Fig. 3 Diagnostic teaching tasks that probe misconceptions

in the teaching materials through explanation and through detailed guidance in the lesson activities—again, a key feature. Absent such guidance, many teachers default to “demonstrate and practice” pedagogy.

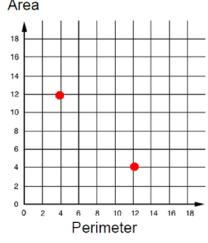
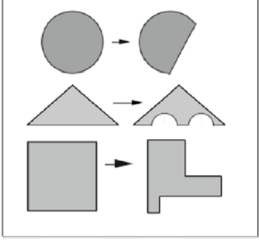
The *Testing Strategic Skills* project (TSS) in the mid 1980s was the first large-scale implementation of this work. Strategically it initiated an examination-driven approach based on *gradual change*. In collaboration with a large examination provider, one new task-type was introduced each year to the high-stakes examination for age 16. The Shell Centre team provided curriculum and professional development materials to support teachers tackling the new tasks. These modules became known as the Blue Box (*Problems with Patterns and Numbers* Swan et al. 1984,) and the Red Box (*The Language of Functions and Graphs*, Swan et al. 1985). Each module included five examples of the new task-type (five to show the variety to be expected), teaching materials for the three weeks of new teaching involved, and materials including video and software to guide do-it-yourself professional development activities in school. The first phase of development focused on a handful of volunteer schools, allowing close observation; the second phase involved only written and interview feedback, but from classes in about 30 schools. The social importance of the examination ensured

the creation of a large user community, which provided feedback for later refinement of the materials. These modules were popular with teachers and students but the gradual improvement process was halted by a government change to the examination system.

The subsequent *Numeracy through Problem Solving* project developed five modules (Swan et al. 1987–89) that pioneered 3 week small-group modelling projects, again with external examinations to assess transfer. *Be a Paper Engineer* was one of these.

While these modules focused on problem solving and modelling, rich task types were also devised for concept development. One design principle is that all practice should be embedded in rich tasks that are extendable, generalizable and make connections. The example on the left hand side of Fig. 4 makes connections between areas, perimeters and functions. The right hand side illustrates the ‘justify and prove’ task genre. Students are given a number of statements and are asked to either justify why each is *always* or *never* true or to identify all the conditions under which it is true. The card below the statement is provided only to students whose ‘productive struggle’ ceases to be productive. This *differentiation through support* is another research-based design principle that proved powerful.

Fig. 4 Task types from the *Improving Learning in Mathematics: Challenges and Strategies*

<p>Possible and impossible shapes.</p> <p>Plot points on the grid that represent squares (then other classes of shapes). Find a shape that would be plotted at (4, 12), then (12, 4). Find which points on the grid represent possible shapes and which do not – and why.</p>	<p>Always, sometimes or never true?</p> <p>Is this statement always, sometimes or never true? Justify your answer, giving examples and counterexamples.</p> <p>“When you cut a piece off a shape, you reduce its area and perimeter.”</p>
	

These materials were developed through a sequence of projects to support both teaching and professional development. A multimedia package of professional development support materials, *Improving Learning in Mathematics: Challenges and Strategies* (Swan 2005) was built around four such task genres. It was distributed by the British government to all secondary schools and colleges (and prisons!).

The ultimate product of this 30-years program of research-based design and development is the 100 formative assessment lessons of the US-based *Mathematics Assessment Project* (MAP), designed and developed with the Berkeley team to support teachers and students in the implementation of the US Common Core State Standards (Burkhardt and Swan 2014). The project set out to explore how far well-engineered teaching materials of this kind can enable typical teachers to acquire the adaptive expertise needed to handle formative assessment for learning in the classroom (Black and Wiliam 1998). The 20 lessons for each grade, 6 through 10/11, combined the concept development approach from the diagnostic teaching strand of work with a complementary strand of research and development on the teaching of modelling skills. There have been over 8,000,000 lesson downloads from map.mathshell.org. Elsewhere (Burkhardt and Schoenfeld 2019a) we analyse a few of these lessons, showing how they exemplify the important aspects of successful implementation. Remarkably positive formal evaluation results (Herman et al. 2014; Research for Action 2015) showed that, as well as strongly advancing student learning, using these lessons led to teachers broadening their expertise in ways that carried over to their teaching in other lessons.

2.4 Other examples

We mention just a few of the other successful projects that share the key features noted at the beginning of this section.

Each adds a new dimension to the above examples as outstanding examples of ‘implementation’.

Realistic Mathematics Education (RME) was built on the ideas of Hans Freudenthal. A team of design-researchers at the University of Utrecht led by Jan de Lange developed an approach to learning through a process of increasing abstraction of the mathematics used in modelling concrete real world situations. (Focused on the mathematics, this approach to mathematical modelling complements the more usual view of modelling, exemplified in the MAP/Shell Centre lessons, as a way to solve real world problems.) The RME work was developed over decades through a sequence of projects in the Netherlands and abroad, notably *Mathematics in Context* in the US (developed in parallel with *Connected Mathematics*)—a rare example of international transfer of a curriculum. Here too there was a theory-based program of research and development, with materials refined in a series of context-based trials.

Everyday Mathematics, another NSF-supported project, has been widely used in US elementary schools. Part of the University of Chicago School Mathematics Project, it too grew out of a long-term program of research and development by Max and Jean Bell, which also viewed mathematics as a tool for modelling real world situations. Again, an iterative sequence of revisions led to the current influential version. The other NSF-supported curricula developed in the 1990s follow these design and development principles more or less closely.

VCE Mathematics, the school leaving examination in the Australian state of Victoria, introduced in the late 1980s a broad-spectrum examination including an unusually wide variety of types of non-routine problems. These included 2-days “take home” investigations, with an ingenious monitoring system to ensure that the student had indeed produced the solution. Subsequent classroom research (Barnes, Clarke and Stephens 2000) found work of this kind happening at

all grade levels. This initiative had, arguably, the highest mathematical validity of any broadly implemented high-stakes examination to date. As with TSS, social and political pressures in the high-stakes assessment system limited its longevity.

Beyond mathematics, *Nuffield A-level Physics* pioneered the introduction of active scientific experience by students into high school physics, which had long been based entirely on learned knowledge. It influenced the syllabuses of other providers and, developed for the digital age, remains in use after many decades. Outstanding here was the time and effort the designers spent initially in building a consensus that this change was needed across all the stakeholder communities: physics researchers, science educators, a group of interested teachers, school leaders and, crucially, an examination board. The key designers, Paul Black and Jon Ogborn, were a university physicist active in education and an innovative teacher in whose classroom the initial development took place. Black (2008) describes the project and analyses the decisions about the design and development process that contribute to its implementability and impact at scale.

3 The engineering research approach

Research in education reflects three traditions (Burkhardt and Schoenfeld 2003), from the *humanities*, *science* and *engineering*. Most research on STEM subjects follows the *science tradition* of empirical exploration, analysis and the generation of hypotheses that are tested empirically—a slow and demanding process that makes it difficult to cover a lot of ground in a single study. In contrast the *humanities tradition* is based on producing critical commentary that, in the absence of empirical testing, is judged by that community on its plausibility and originality. (Most educational materials are written and published in this tradition, with no empirical evidence on what happens in the hands of typical users. As a result, less-than-optimal untested approaches are likely to be adopted.) Despite being fundamentally speculative, this approach remains highly influential for at least two reasons. A lot of ground can be covered in a single piece. Also policy makers can, and do, contribute at a technical level they would not contemplate in, say, medicine—and they often find their own views most plausible! The key research product in both traditions is articles in academic journals; typically practical implementation is left to others. As implementability is not usually a concern, this special issue is, indeed, breaking new ground.

Implementation implies that the primary outputs are products and processes for use by practitioners. If they are to be judged according to how well they work in practice in forwarding the educational goals, we believe this

requires research in the *engineering tradition* (Burkhardt 2006). As in engineering and medicine, research insights are both a key input and a second kind of output from the central research-based design and development process. This process is well known from these and other applied fields (see, e.g., Archer 1974; Morris 2009; Norman 2013). It involves:

1. a specific improvement goal, grounded in robust aspects of theory from prior research
2. design tactics that combine these core ideas with the design team's creativity, always providing close support with the pedagogical challenges
3. flexibility in the draft materials that affords adaptation to the range of contexts across the intended user community
4. rapid prototyping followed by iterative refinement cycles in increasingly realistic circumstances, with
5. feedback from each round of trials that is rich and detailed enough to ensure the robustness and adaptability of the final reproducible materials, and
6. continued refinement on the basis of post-implementation feedback 'from the field'—this also informs thinking for future developments.

Most educational materials are produced without 4, 5 and 6. They are usually written, revised by the author in the light of comments from wise persons, and published. In the absence of empirical evidence as to what will happen in their use by others in diverse classrooms, they lack research validity. One would not accept a medicine or any other critical product on this basis. Should education accept a lower standard?

Rigorous qualitative research methods are involved in one to six above, not just for the input phase but throughout the process: in the selection and the training of designers; the choice of sample classrooms or other contexts for trials at each stage of development; the design of protocols for capturing the relevant information in the most cost-effective way, for example for classroom observation; the form of presentation of the materials to optimise communication, which always involves choices that balance information with usability—a typical design trade-off. This combination of analytical and creative design skills is unusually demanding and outstanding exponents are rare; yet there is a huge difference in students' learning experience between the best and the standard "perfectly good" materials that are so widespread. Examples of evidence for this can be found in Senk and Thompson (2003) for the projects that were supported by the NSF, including Connected Mathematics; for the Mathematics Assessment Project (MAP) materials the independent evaluations are summarised by Burkhardt and Schoenfeld (2019a).

The process is exemplified by the lesson development in discussed in Sect. 2 for MAP. It might be described as a further implementation of the diagnostic teaching research (a robust theoretical research base, point 1 above) but, as usual, with many other inputs. The goal was to see how far typical mathematics teachers in supportive school environments (a specific user group, 3 and 4 above) could be enabled to implement high-quality formative assessment for learning (a specific improvement goal guided by research, see Black and Wiliam 1998) in their own classrooms, when primarily guided by published lesson materials designed for this purpose. The design team was led by Malcolm Swan who had, over 30 years, built a track record of outstanding design achievements (point 2 above). Flexibility (3) was addressed through advice on ways to time these supplementary lessons so as to best enhance whatever mathematics curriculum the school is using. Feedback of sufficient quality for guiding revisions (5 and 4) was based on direct observation of a handful of lessons at each stage by experienced classroom observers using a structured protocol. This emphasised what actually happened in the lesson and its relation to the design intentions. The report also included interviews with the teacher. The 700 lesson reports together also provided research insights to guide future work (1); in this it was supplemented by independent evaluations and informal feedback from users (6). The work is described in (Burkhardt and Swan 2014) and analysed as formative assessment in (Burkhardt and Schoenfeld 2019a), which reviews the evaluative evidence showing the power of the lessons as teacher professional development as well in student learning.

The cost of such a process of research, design and development is much higher than the cost of the more typical “authorship” model, largely because of the iterative learning process of refinement through the trials of successive versions. For the MAP Formative Assessment Lessons there were at least three rounds of materials. It is essential to note the design decision to obtain *rich, detailed feedback from a small number of lessons*, enough to distinguish the generic from the idiosyncratic. In contrast the other end of this trade-off, sparse feedback from large samples, is little help for revision when a qualitative change is the goal, as with MAP. Coherent structured observation is expensive but invaluable in the depth of feedback it allows. Each MAP lesson cost around \$30,000 *in toto*, excluding the earlier developments on which some lessons were based. This is much more than the <\$10,000 per lesson we noted for the NSF curriculum projects. Nonetheless, if widely used, this cost is highly cost-effective when related to the running cost of the education system (Burkhardt 2006). But it is still unfortunately rare that the funding of such a thorough engineering research project allows this quality of process.

Where does ‘design research’ (Brown 1992) fit into this engineering framework? Typically it addresses at most items

1 and 2 above, along with the first part of 4. There is rarely concern for a user community beyond the teachers working with the research team in whose classrooms the new learning activities are developed—teacher variability and the support that different teachers may need is not explored. Teaching materials are rarely developed into a form that others can use and evaluate—or published at all. This reflects the primary goal of design research: academic journal articles. Implementability, let alone implementation, is not an issue. These limitations distinguish design research from engineering research in education (Burkhardt 2006).

4 Robust theory as a guide and constraint

What are the roles of insight-focused research in supporting implementability? What kinds of contribution can the results make to implementation? What aspects of theory need to be strong for this? That is our focus here. We shall distinguish four kinds of research contributions to implementation: *inspiration, guidance, constraint, and moving onward*.

4.1 Research as a source of inspiration

A new and surprising research result may suggest a new approach to, say, teaching (or assessment or professional development). This result may then inspire the design of teaching materials that seek to realise that approach in classrooms in an effective and, if well done, more motivating form. Success may motivate funders to support the rest of the engineering process needed to develop products that are robust in the wider range of circumstances across the intended user community. Research on non-routine problem solving (Pólya 1945; Schoenfeld 1985), diagnostic teaching (Bell 1993; Swan 2006), and cognitively guided instruction (Carpenter et al. 2014) are examples of such research.

What are the essential characteristics for research to be worth the effort that good engineering involves? Schoenfeld (2002) identified three dimensions along which to characterize a research study: Importance, Trustworthiness, Generalizability.

- *Importance*. Does the result address learning goals that are socially or intellectually important and appropriate to students at a specific age or stage? For example, the recognition that reliable procedural skills in arithmetic were no longer an adequate basis for employment made the research on problem solving socially important. The changed circumstances of the second half of the twentieth century required the ability to respond flexibly and effectively to a range of new types of problem, so we needed to learn how to teach the skills involved.

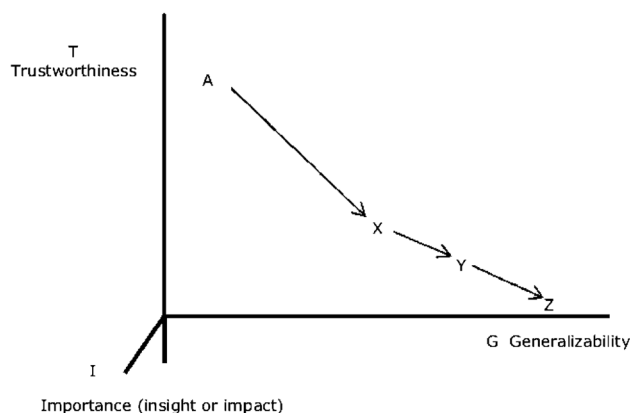


Fig. 5 The trajectory of a typical research report

- *Trustworthiness*. Do the design of the study, the data collected, and the logic of the analysis presented justify the claims made? Academic standards for research rightly give priority to trustworthiness; without positive answers to these questions, the research will not be published in a respectable journal—the prime currency in academic success.
- *Generalizability*. Is there evidence on the validity of a research result across the range of relevant circumstances of use? This is essential in establishing its value for implementation. The sequence of diagnostic teaching studies mentioned above, designed on the same principles by different designers on different mathematical topics with different students, was evidence that these were indeed reasonably robust principles. Any single study was merely ‘promising’.

In practice there is a trade-off among these desirable qualities in the design of any research study. There is no shortage of topics that are widely recognised as *important* though, as the examples in Sect. 2 show, the scale of any study is often constrained by the resources available to academic researchers. *Trustworthiness* usually requires a focus on specific research questions in a range of circumstances limited by the available resources of time and funding. This limits the evidence on *generalizability* that can be collected, addressing a limited subset of the circumstances that important issues generally concern. In practice, research reports usually contains assertions in different parts of the three dimensional space illustrated in Fig. 5, focused on the latter 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 and tried out in a few classrooms, collecting and analyzing the teacher and student responses to the intervention. If well done, the results are high on T but, because of the limited range of the variables explored, low

on G; this is shown as the zone A on the graph. However, the conclusions section of a typical paper goes on to discuss the “implications” of the study. These, whether explicitly claimed or more subtly implied, are usually much more wide ranging but *with little or no empirical evidence to support the generalizations 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 represent 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.

For these reasons, exciting research results are typically as much inspiration as guide, with the generalizability issue left to the engineers in the design and development process. For example, the study of problem solving was initiated by the reflections of Polya (1945), who identified a set of problem strategies. This inspired one of us to undertake a series of empirical studies (Schoenfeld 1985) focused on undergraduate students working on non-routine geometric problems. This detailed analyses of real student behavior added further theoretical insights, particularly on movements between levels of thinking (from “control” (monitoring and self-regulation) to “strategic” to “tactical” to “technical” and back) and the influence of teacher beliefs about the nature of “doing mathematics.” When the Shell Centre TSS team, having persuaded the examination board to add a problem solving task to the examination, set out to develop teaching materials, it had to discover how to transfer these ideas to the different age range and, necessarily, a different problem area. *Problems with Patterns and Numbers* (Swan et al. 1984) teaches explicit strategies (look for simple cases, organize and represent the information you find, look for patterns, generalize), while making clear that their meaning depends on the type of problem, and emphasizes the value of reflecting on your own problem solving process as it proceeds. The meaning of these things was refined through the trialing process and their value through the levels of student responses, before and after the teaching. This work also added to the body of research.

4.2 Guidance: research specifically designed to inform implementation

Although research is often “use-inspired” in that the questions addressed arise in or from practice, it is much less common for bodies of theoretically oriented research to be deliberately framed to guide or shape implementation. One such example is Black and Wiliam’s work on formative assessment (e.g. Black and Wiliam 1998). The work was undertaken to document the impact of formative assessment,

clarify what it is and is not (e.g., it is *not* frequent testing), document its impact (e.g., student work improved when students received comments but not grades, and did not improve when they received grades as well as comments), and point to directions for further development. As noted above, such work was part of the foundation for the Shell Centre’s work. Similarly, the body of research on summative assessment (Mathematics Assessment Project 2016) led to a new kind of test aimed at assessing a much broader spectrum of mathematical understandings than is typical of high stakes assessments (see, e.g., Smarter Balanced Assessment Consortium 2012).

A third example comes from the arena of classroom studies and professional development. There is no shortage of individual research studies saying that particular classroom moves, or norms, or techniques, help students learn. That, in fact, is the problem: what is a teacher, coach, or administrator to do with a long list of important classroom techniques? If a teacher has to attend to even a dozen things at once, the task is impossible.

For that reason, the Teaching for Robust Understanding project (2018) made it an explicit theoretical aim to develop a theory of powerful classrooms that could guide implementation efforts. Specifically, the theoretical goal was to identify a small number of dimensions that were necessary and sufficient for powerful instruction—small in number so that reflection and professional development are feasible, necessary in the sense that if things go poorly along any of those dimensions at least some students will be ill-served, and sufficient in that if things go well along all of the dimensions the students will emerge as powerful thinkers (Schoenfeld 2014; Schoenfeld, Floden, and the Algebra Teaching Study and Mathematics Assessment Projects 2018; Teaching for Robust Understanding Project 2018).

A multi-year R&D effort distilled the literature into five key dimensions of practice: *the discipline* (Are students engaged with and in important content and practices?), *cognitive demand* (Do activities involve students in *productive struggle*?), *access* (Are all the students actively involved in each phase of learning activities?), *agency, ownership and identity* (Does each student feel that they can contribute and that their mathematical reasoning is recognized as “belonging to them” and their fellow students?), and *formative assessment* (Is instruction structured to consistently reveal student thinking and provide formative feedback?).

This list is short enough to keep in mind. Teachers, coaches, designers and administrators internalize it quickly and it becomes a frame for thinking about teaching. Each of the dimensions is “actionable”—a teacher, teacher and coach, or teacher learning community can focus on improving practice in one dimension at a time, a manageable task. Thus the framework is “implementation friendly” by design. It is generative, in that the dimensions

support tool development (see Teaching for Robust Understanding Project 2018) and support an implementation-focused research and development agenda. Evidence of the framework’s impact is given in the papers cited above.

4.3 Constraints: research as a source of design constraints

Fine design is creativity within constraints. As well as a source of ideas for implementation through engineering, high-quality research can provide a guide to constraints of at least two kinds that will improve the design process.

Type A. Checklists that embody research-based principles against which to check designs throughout the engineering process

For brevity, we present the following examples largely in the form of questions that a design team should ask, and verify by observation during the development process.

- *The TRU Framework* sets forth principles that powerful environments should satisfy. Thus, any design and development process should aim to ensure that its products effectively support teacher and students in each of these five dimensions. Burkhardt and Schoenfeld (2019a) includes an analysis of MAP lessons from this perspective, showing how the lesson structures lead teachers to handle all five dimensions. In a form of lesson study called TRU-lesson study (Schoenfeld et al. 2019) TRU is used to generate and review lesson plans, to structure observations, and to guide the lesson reflection and commentary process. More broadly, the framework is used to *problematize* instructional design at every stage of design and implementation, along the five dimensions listed above.
- *Roles analysis.* This approach monitors the roles that teacher and students play at various stages in the learning activities—and thus how far essential *role shifting* from traditional roles is achieved. One system (Phillips et al. 1988) distinguishes the *directive* roles that teachers traditionally play (*manager, explainer, task setter*) from the *facilitative* roles (*counsellor, fellow student, resource*) that develop student autonomy and agency. For example, are teachers’ questions designed to elicit explanations, or merely short answers? What opportunities do students have to explain their reasoning? Do students formulate mathematical questions for investigation? For how long will students typically work on a single problem without teacher intervention?
- *Technical constraints* of various kinds. These often arise from accumulated experience through the engineering process. Two examples are enough to make the point:

1. *The types and amounts of guidance offered in the materials.* Is this sufficient for a teacher in the intended user group to plan and realise the lesson as intended? Is it brief enough for them to be likely to read it?
2. *The use of visual aids to support or replace text.* Would a graphic at this point help more students understand the problem? How could this task be ‘brought to life’ with a picture?

Of course, Type A research can also play a role in inspiring new thinking by designers. Role shifting has long been an explicit design principle of Shell Centre work, and TRU generated a comprehensive set of tools for professional development (Teaching for Robust Understanding Project 2018).

Type B. Invaluable guidance on what *not* to try. Examples include:

- Teaching mathematics through ‘demonstrate and practice’ alone. While some degree of automaticity is of course useful, skills drop precipitously with time (e.g. the notorious “summer slump”). More effective is to teach skills combined with conceptual understanding so that students can regenerate knowledge and skills when they need them.
- Assuming that understanding will transfer across trivial changes, e.g. of notation. It doesn’t without hard thinking. Testing this informally, one of us asked a colleague “What is $d/dn(n^x)$?” “Something to do with the Gamma Function, isn’t it?” he replied. We all learn $d/dx(x^n) = nx^{n-1}$; exchange the symbols and we need a chain of careful reasoning to sort it out: xn^{x-1} , of course. Using variable names that link to physical quantities (time t) adds meaning but can trigger this problem for students who are fixated on x and y ; both practice and reflection are needed to broaden the learner’s thinking to pay attention to *structural* forms.

4.4 Moving onward: evaluative research to advance the state of practice

Evaluation has always employed traditional research skills but has not typically focused on providing information from a formative point of view. Too often summative studies have been confined to gathering superficial data—usually on student performance on tests designed for accountability purposes. While such data may satisfy funding agencies that wish to determine whether or not their money has been well spent, they are totally inadequate as a guide to improvement. How can this kind of research contribution be improved?

The engineering research approach of Sect. 3 ensures that the design and development team has a great deal of

evaluative information on the product in action, and has used it formatively. An independent research team will in addition bring a somewhat different perspective, detached from the details thinking of the creators. However, to be useful as formative input for the next stage of improvement, we need (see Burkhardt 2016) studies that:

- are in at least as much depth as in the development process, but with emphasis on analysis
- cover the range of important variables, notably students, teachers, levels and kinds of support and pressure
- focus on well-engineered treatments that are likely to be reasonably consistent in implementation throughout the study.

To achieve this it is hard to see a credible alternative to large-scale research collaborations, conducted in a collegial spirit. These do occur but are rare in the work of the education research community at large. Such collaborations need to:

- understand in depth, and accept, the goals of the designers of the treatment
- agree, across the collaboration and with the design team, a well-defined protocol of research methods, including the collection and integration of in-depth data and co-ordination of the analysis

while

- being free to investigate other aspects of the implementation that they collectively decide may be important.

Such an approach, with large collaborating teams from different institutions and joint funding, is common for tackling complex challenging problems in other fields. The Large Hadron Collider in physics and the Human Genome project in bioscience are examples where many researchers work to a common purpose. These are often referred to as Big Science; we see a need for Big Education.

5 Achieving systemic impact

Finally we move on to a more ambitious definition of implementability, going beyond building research results into tools, products and processes that can be used by *some* teachers. While such ‘existence proofs’ are valuable, implementability should surely aim to mean implementation on a significant scale.

This inevitably introduces systemic issues. It may be tempting to ignore these as merely organisational but that

would be a mistake. Successful implementations of professional development indicate that, as a field:

We know how to enable typical teachers to teach much better mathematics, much more effectively than is common in current classrooms. BUT, we do not have established ways of getting school systems to make this happen.

Education systems are organised in very different ways in different countries so our comments here must be general. Some countries, in East Asia and elsewhere, have structures that work systematically to improve curriculum and support professional development—but often these are centralized, inherently conservative and slow-moving (not necessarily a bad thing). Anglophone countries’ systems are often somewhat decentralized, from the 7 Australian States and Territories to the US with its 50 states and 15,000 school districts.¹ In principle, this gives opportunities to explore different implementation models in similar cultural environments.

However, in all systems there are forces that frustrate implementation of improvements that research and development has proved possible. These arise from pressures on the various stakeholders that must move if change is to happen—notably teachers, school and system leadership, and policymakers. While all these would say that improving student learning is their top priority, the different pressures of their lives, day-by-day and month-by-month, push this strategic priority to the distant rhetorical horizon (Burkhardt 2019b).

5.1 Getting all the system stakeholders inside

Politicians and policy makers recognize education as important. But their lives are characterized by pressures of time, diverse political input, government procedures, budgetary limits and, perhaps most important, the clash of timescales. The decade timescale of significant improvement in education lacks urgency as ministers try to ‘make their mark’ in their year or two in charge of education, while coping with week-by-week media-driven ‘events’ across the education system. Often they have strong views on education and involve themselves at a detailed technical level they would not contemplate in, say, medicine or engineering. These difficulties are real and need to be taken into account in the design of a more effective research-based systemic improvement process.

¹ In England, tradition gave each school principal responsibility for the curriculum but, in practice, variety was sharply confined by the high-stakes examinations and textbooks. From 1989 the National Curriculum narrowed this further.

Similarly, the world of educational practice faces pressures at every level, flowing down from government through system and school leadership to the individual teacher. These distort teachers’ core task, to help and guide some 30 children to become well-educated citizens, which is challenging enough. Many of these pressures are justified in the name of “accountability”. Scores on tests that assess only some narrow aspects of doing mathematics acquire priority, distorting the implemented curriculum in most classrooms.

The education research community functions largely for its own purposes: producing dissertations and articles for academic journals that inform decisions on appointments and promotions and career reputations. But in contrast with medicine, research focused on the development of new *treatments* that improve what happens in classrooms has, at least until recently, had low prestige in the academic value system—implementability is rarely considered, let alone made a priority.

Systemic aspects of particular concern in at least some countries are: poor communication between these communities; the political tendency to ignore system complexity, that real improvement involving changes of well-grooved professional practice is inevitably gradual and complex; the tendency of policy makers to design aspects of teaching and assessment at a level of technical detail that discounts the expertise of the education professions; pressures for uniformity that limit opportunities and support for pilot projects that can grow into and improve the mainstream; imbalance in education research between the dominant *analytical-diagnostic* research traditions and *treatment-focused* research and development with an engineering approach; the failure of the education research community to develop, on the one hand, a solid body of agreed research results and, on the other, detailed evidence on the effects of specific ‘treatments’.

This leads to a lack of authoritative structures that integrate evidence from research and practice in a form that policy-makers respect and can use.

6 Conclusions

This paper has offered some examples that show large-scale impact of research-based change is achievable and analysed features that seem to be important for this. In this we have seen that shaping research results in ways that result in significant impact on educational practice is too complex a process to be characterized as “implementation.” Doing so requires many players with different skills building on not just a single research result but, in principle, the whole body of relevant prior research. Other research skills are needed in a design and development approach in the engineering research tradition to turn research results and creative design

ideas into robust products and processes that that work well in educational practice.

We have seen that research-based theory itself can play several roles:

- *in inspiring design* and the courage to tackle the challenges of implementation—for this it needs to be strong in importance and trustworthiness, with evidence on generalizability across a well-specified domain of validity
- *in providing “implementation friendly” theoretical framings* of key arenas, setting them up for productive R&D work
- *in providing guidance* on strategic and tactical issues
- *in using evaluation* in-depth of well-engineered treatments to suggest ways forward at the next stage.

For some of these roles, research needs Big Education—substantial cross-institution teams working coherently with agreed foci and protocols. These are the aspects of effective research that support implementability.

How can an education system support such an endeavor? We can learn a lot, as always, from the successes and limitations of the examples described. Strategically, the way forward seems to lie in recognizing the very different timescales of policy decisions, of systematic R&D, and of building a corpus of well-tested and widely agreed research results. This seems to require a support structure with three strands designed, like those in medicine:

- **to support and evaluate innovation in practice** by funding, in areas needing improvement, a vigorous program of research-based design, iterative development and refinement of effective treatments—for example, to support well-aligned teaching, assessment and professional development in schools;
- **to gradually strengthen the research base** of practice and policy by funding insight-focused research of direct relevance, for example: evaluation-in-depth of both current practices and new treatments, and building a body of well-validated research results that is broadly accepted across the field;
- **to evaluate potential policy moves and advise system governance** on their cost-effectiveness in the light of the evaluation evidence on their strengths, weaknesses and costs—so that policy makers are no longer involved in the design of treatments, but make choices for implementation based on solid evidence.

Big Education requires funding, of course; and in the present context, such funds are hard to find. Why? People make investments when they perceive there will be a return on them. As we pointed out in 2003, the US Federal government spent a total of \$300 Million in 1998 on educational

research, while the Pfizer Pharmaceutical Corporation said that “We lead the industry in research, spending over two hundred million dollars a year, looking for new treatments designed specifically for animals.” (Burkhardt & Schoenfeld 2003, p. 12) This is a statement of national priorities and about the incentive system—but priorities can change when the perceived value of the investment changes.

A parallel purpose is to develop institutional memory and human capital in these areas across the policy, research, design and development, and practice communities. These are among the challenges for system governance in the next phase.

Acknowledgements We are grateful to many colleagues for discussions over the years on approaches to using research in the improvement of practice, notably Phil Daro, Malcolm Swan and Mark St John. Our thanks to Elizabeth Phillips for her comments on the *Connected Mathematics* program.

Funding No funding for this publication.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Archer, B. (1974). *Design awareness and planned creativity in industry*. Toronto: Thorn Press Limited.
- Barnes, M., Clarke, D., & Stephens, M. (2000). Assessment: The engine of systemic curricular reform? *Journal of Curriculum Studies*, 32, 623–650.
- Bell, A. (1993). Some experiments in diagnostic teaching. *Educational Studies in Mathematics*. <https://doi.org/10.1007/BF01273297>.
- Birks, D. (1987). *Reflections: A diagnostic teaching experiment*. Nottingham: Shell Centre for Mathematical Education.
- Black, P.J. (2008) Strategic decisions: Ambitions, feasibility and context. *Educational Designer*, 1(1). <https://www.educationaledesigner.org/ed/volume1/issue1/article1/index.htm>. Accessed 1 November 2020
- Black, P. J., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education*, 5, 7–74.
- Brown, A. L. (1992). Design experiments: theoretical and methodological challenges in creating complex interventions in classroom settings. *Journal of the Learning Sciences*, 2(2), 141–178. https://doi.org/10.1207/s15327809jls0202_2.
- Brown, A. L., & Campione, J. C. (1996). Psychological theory and the design of innovative learning environments: On procedures, principles, and systems. In R. Glaser (Ed.), *Innovations in learning: New environments for education* (pp. 289–325). Mahwah: Erlbaum.

- Burkhardt, H. (1988). The roles of theory in a “systems” approach to 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* (pp. 121–150). London: Routledge.
- Burkhardt, H. (2009). On strategic design. *Educational Designer*, 1(3). <http://www.educationaldesigner.org/ed/volume1/issue3/article9>. Accessed 1 November 2020
- Burkhardt, H. (2016). Mathematics education research: a strategic view. In L. English & D. Kirshner (Eds.), *Handbook of international research in mathematics education* (3rd ed., pp. 689–712). London: Taylor and Francis.
- 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., & Schoenfeld, A. H. (2019a). Formative assessment in mathematics. In R. Bennett, G. Cizek, & H. Andrade (Eds.), *Handbook of formative assessment in the disciplines*. New York: Routledge.
- Burkhardt, H. (2019b) *Improving Policy and Practice*, Educational Designer, 3(12). <http://www.educationaldesigner.org/ed/volume3/issue12/article46/>
- Burkhardt, H. & Swan, M. (2014) *Lesson design for formative assessment*, Educational Designer, 2(7). <http://www.educationaldesigner.org/ed/volume2/issue7/article24>. Accessed 1 November 2020
- Burkhardt, H., & Swan, M. (2017). Design and development for large-scale improvement Emma Castelnuovo Award lecture. In G. Kaiser (Ed.), *Proceedings of the 13th International Congress on Mathematical Education* (pp. 177–200). Cham: Springer International Publishing.
- Carpenter, T., Fennema, E., Franke, M., Levi, L., & Empson, S. (2014). *Children’s mathematics, cognitively guided instruction* (2nd ed.). Portsmouth: Heinemann.
- Collins, A. (1992). Toward a design science of education. In E. Scanlon & T. O’Shea (Eds.), *New directions in educational technology* (pp. 15–22). Berlin: Springer.
- Herbal-Eisenmann, B., & Breyfogle, M. (2005). Questioning our patterns of questioning. *Mathematics in the Middle School*, 10, 484–489.
- Herman, J., Epstein, S., Leon, S., La Torre Matrondola, D., Reber, S., & Choi, K. (2014). *Implementation and effects of LDC and MDC in Kentucky districts (CRESSST Policy Brief No. 13)*. Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESSST).
- Lappan, G., & Phillips, E. (2009) Challenges in US Mathematics Education through a Curriculum Developer Lens, *Educational Designer*, 1(11). <https://www.educationaldesigner.org/ed/volume1/issue3/article11/index.htm>
- Mathematics Assessment Project (2016). <https://www.map.mathshell.org/>. Accessed 1 November 2020.
- Morris, R. (2009). *The fundamentals of product design*. London: AVA Publishing.
- National Council of Teachers of Mathematics. (1989). *Curriculum and evaluation standards for school mathematics*. Reston: NCTM.
- Norman, D. (2013). *The design of everyday things: Revised and expanded edition*. New York: Basic Books.
- Phillips, R., Burkhardt, H., Fraser, R., Coupland, J., Pimm, D., & Ridgway, J. (1988). Learning activities and classroom roles with and without the microcomputer. *Journal of Mathematical Behavior*, 6, 305–338.
- Pólya, G. (1945). *How to solve it*. Princeton: Princeton University Press.
- Research for Action (2015). *MDC’s Influence on Teaching and Learning*. Philadelphia, PA: Author. <https://www.researchforaction.org/publications/mdcs-influence-on-teaching-and-learning/>. Accessed 1 November 2020
- Royal Society/British Academy. (2018) *Harnessing Educational Research*. <https://royalsociety.org/topics-policy/projects/royal-society-british-academy-educational-research/>. Accessed 1 November 2020.
- Schoenfeld, A. H. (1985). *Mathematical problem solving*. Orlando: Academic Press.
- Schoenfeld, A. H. (2002). Research methods in (mathematics) education. In L. English (Ed.), *Handbook of international research in mathematics education* (pp. 435–488). Mahwah: Erlbaum.
- Schoenfeld, A. H. (2014). What makes for powerful classrooms, and how can we support teachers in creating them? *Educational Researcher*, 43(8), 404–412. <https://doi.org/10.3102/0013189X1455>.
- Schoenfeld, A., Dosalmas, A., Fink, H., Sayavedra, A., Weltman, A., Zarkh, A., et al. (2019). Teaching for robust understanding with lesson study. In R. Huang, A. Takahashi, & J. P. Ponte (Eds.), *Theory and practices of lesson study in mathematics: An international perspective* (pp. 136–162). New York: Springer.
- Schoenfeld, A. H., Floden, R. B., & the algebra teaching study and mathematics assessment projects. (2018). On classroom observations. *Journal of STEM Education Research*. <https://doi.org/10.1007/s41979-018-0001-7>.
- Senk, S., & Thompson, D. (Eds.). (2003). *Standards-oriented school mathematics curricula: What does the research say about student outcomes?* Mahwah: Erlbaum.
- Smarter Balanced Assessment Consortium. (2015). Content specifications for the summative assessment of the common core state standards for mathematics (revised draft, July 2015). https://www.mathshell.com/papers.php#sbac_2012. Accessed 1 November 2020.
- Smith, M. S., & Stein, M. K. (2011). *5 Practices for orchestrating productive mathematics discussions*. Reston: National Council of Teachers of Mathematics.
- Swan, M. (2005). *Improving learning in mathematics: Challenges and strategies*. Sheffield: Department for Education and Skills Standards.
- Swan, M. (2006). *Collaborative Learning in Mathematics: A Challenge to our Beliefs and Practices*. London: National Institute for Advanced and Continuing Education (NIACE) for the National Research and Development Centre for Adult Literacy and Numeracy (NRDC).
- Swan, M., Binns, B., & Gillespie, J., Burkhardt, H, with the Shell Centre team. (1987–89). *Numeracy Through Problem Solving: five modules for teaching and assessment: Design a Board Game, Produce a Quiz Show, Plan a Trip, Be a Paper Engineer. Be a Shrewd Chooser*, Harlow, UK: Longman, and <http://www.mathshell.com>. Accessed 1 November 2020.
- Swan, M., Pitts, J., Fraser, R., Burkhardt, H, with the Shell Centre team. (1984). *Problems with Patterns and Numbers*. Manchester, U.K: Joint Matriculation Board and Shell Centre for Mathematical Education, and <http://www.mathshell.com>. Accessed 1 November 2020.
- Swan, M., Pitts, J., Fraser, R., Burkhardt, H, with the Shell Centre team (1985), *The Language of Functions and Graphs*, Manchester, U.K.: Joint Matriculation Board and Shell Centre for Mathematical Education, and <http://www.mathshell.com>. Accessed 1 November 2020.
- Teaching for Robust Understanding Project (2018). *Teaching for Robust Understanding*. <https://truframework.org/>. Accessed 1 November 2020.