Towards Research-based Education: integrating policy, research and practice.

Hugh Burkhardt[1]

Executive summary

The challenge. The goals for STEM education are largely agreed, nationally and internationally. Work over the last 30 years in the research and development community has shown how to develop tools and processes for teaching, assessment and professional development that enable typical teachers to teach much better mathematics and science much more effectively. Why is this not reflected in most classrooms, and what could be done about it?

The diagnosis. This situation indicates that the fundamental problem lies in the systemic processes of improvement - in particular, when compared with research-based fields such as medicine, the lack of coherent collaboration between the policy, research and practice communities in the development and implementation of research-based 'treatments'.

The way forward lies in integrating the very different timescales of policy decisions and systematic R&D through a program to develop structures, like those in medicine:

A. **to support and evaluate innovation** by funding, in areas needing improvement, a vigorous program of research-based design, iterative development and refinement of effective treatments – notably, to support well-aligned teaching, assessment and professional development in schools;

B. **to gradually strengthen the research base** of policy and practice, by funding insight-focused research of direct relevance, such as 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;

C. **to evaluate potential policy moves and advise government** on their cost-effectiveness in the light of the evaluation evidence on their strengths, weaknesses and costs – so that government is not involved in the design of treatments but makes choices for implementation based on solid evidence.

A parallel purpose is to develop institutional memory and human capital in these areas across the policy, research and practice communities.

Evaluation. This program should build medium and long term benchmarks for success, while producing enough signs of progress (teacher learning, changes in classroom practice, etc.) year by year to justify expenditures until it is reasonable to expect more large-scale results.

---

[1] Shell Centre, CRME, University of Nottingham. Contact: Hugh.Burkhardt@nottingham.ac.uk see also https://www.mathunion.org/icmi/awards/past-receipients/2016-icmi-award-winners

Outline of the argument

Education policy-making is an area of government that is always active, with a regular flow of initiatives that aim to improve pupil learning. Yet often the outcomes are far from the intentions. How could the system do better? There has been a great deal of analysis of the various needs for educational change (e.g. Cockcroft Report 1992) and the change process itself (see e.g. Fullan 2016). This paper takes a different approach, treating educational improvement as a design and development research problem. It looks at the problem, the players and the processes, and suggests ways forward. It describes factors that contribute to the pattern of shortcomings and suggests ways to address these factors. In doing so it draws on experience in other parts of the world, and in other fields of importance to people's lives where research plays a much bigger role.

The path towards research-based practice. Engineering and medicine[2], to take two examples, began as and remain areas of professional practice, tackling problems of societal importance in a systematic way enshrined in a community of professionals sharing craft knowledge. In these fields, a scientific approach gradually developed, a research community emerged and, as time passed, increasingly informed practice. This research-based progress was recognized by policy makers and supported as an effective, and cost-effective element in meeting societal needs. Engineering is now largely research-based; though opportunities for creative design remain, they are increasingly constrained by research-based understanding – exemplified, for example, by the current similarity of shape of cars. Medicine is less far along the road, as anyone with back pain knows, but its practice has become increasingly research-based, with enormous strides in both insights and treatments coming from the growing research effort over the last century. Most doctors no longer design the treatments they use (and politicians never did!). Education is a long way behind medicine and remains largely craft-based. How far has it moved along the path to becoming more research-based, and how might it move forward more rapidly? That is our concern[3].

Is there a problem? In Section 1 we look at the nature of the challenge in more detail. We describe some examples of the gross mismatch between policy intentions and outcomes in practice. We point to some of the factors that made this likely, often inevitable[4] – notably the absence of expert design and iterative refinement in their development. (Examples of designs that tackled these and other challenges successfully are left to the Appendix; this paper is focused on structures and processes.)

In Sections 2, 3 and 4 we look at the participant communities – policy-makers, education researchers, practitioners – describing the different structures in which they work, the different pressures they face, and how these frustrate any coherent program of research-based improvement.

The policy makers' world. Ministers lives are characterized by *pressures*: time, diverse political input, government procedures for policy making, 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 education, while coping with week-by-week media-driven 'events' across the education system. These difficulties are real and need to be taken into account in the design of a more effective research-based improvement process.

---

[2] The comparisons with medicine that follow are focused on the relationship between research and development, policy and practice. Other differences, including the organizational structures, are not addressed here.

[3] The paper focuses on STEM education - the reasons are explained below - but most of the analysis is more general.

[4] Unintended consequences of the design of measures for accountability feature prominently.

The education research world functions effectively for its own purposes: producing dissertations and articles for academic journals that inform decisions on appointments and promotions and career reputations. Some of these analytic studies provide useful *diagnoses* of challenges at various levels; these sometimes inform policy. But in contrast with medicine, research focused on the development of new *treatments* that improve what happens in classrooms has low prestige in the academic value system.

The world of educational practice faces pressures at every level, flowing down from government to the individual teacher. Many of the pressures are in the name of accountability. These often distort teachers' core task, which is challenging enough: to help and guide some 30 children to become well-educated citizens.

Aspects of particular concern that emerge include:

- **Poor communication** between politicians and policy makers and the education communities.
- **Trying to 'fix the problem'** – a political tendency that fails to take into account system complexity, and that real improvement involving changes of well-grooved professional practice is inevitably gradual and complex.
- **Technical naïveté** – the tendency of politicians and policy makers to prescribe aspects of teaching and assessment at *a level of technical detail* that they would not dream of trying in, say, medicine or engineering – thus discounting the expertise of the education professions.
- **Pressures for uniformity** – the limited opportunities and shortage of support for pilot projects that can, after evaluation, 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.
- **No generally accepted body of results** – 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'.
- **The lack of authoritative structures** that integrate evidence from research and practice in a form that policy-makers respect and can use, a consequence of the above.

How could the system work better? Section 5 sets out and explains proposals for structures and processes that would make each of the three communities work more effectively, and together move forward towards society's goals for education. Section 6 returns to these concerns.

Key in the strategic design (Burkhardt 2009) is to separate activities with different timescales into complementary structures, so the gradual process of research and development of new treatments that work well only becomes an issue for policy makers through initial funding decisions and, much later, when fully developed and evaluated at pilot scale.

This requires, as in medicine and other applied fields, support for a vigorous program of research and development of robust improved 'treatments' for use in classrooms, professional development and examinations. Complementing this, education research needs to build a core of agreed results, with well-established boundaries of proven validity, on which future developments can rely.

Policy makers will, of course, have overall control of this program but it should *operate* largely independently. Cost-effectiveness analysis suggests that a serious program of this kind with a decade timescale for substantial improvement would require more funding than at present, but growing slowly to a level still well below 1% of the running costs of school education.

The suggested structures, which have approximate parallels in Medicine, are (*working titles*):

A. **National Institute for Educational Development** to fund, with others, research-based development of 'treatments' in areas of recognized need and/or policy priority;
B. **National Institute for Educational Research** to gradually strengthen the research basis of policy and practice by funding research that is directly relevant to these;
C. **National Institute for Educational Excellence** to review evaluation evidence on available 'treatments' and to advise government on their cost-effectiveness.

Over time this approach will produce, as in medicine, high-quality impact-focused research and development that proceeds on an ongoing multi-year basis *and* that also delivers, year-by-year, well-proven results that can form the basis of policy initiatives that are likely to prove successful, individually and cumulatively – and will thus deliver political capital. This is the path to research-based education.

1. Is there a problem?

"If it ain't broke, don't fix it."

Is there a serious problem with the design of education policy? A fair question. I propose to answer it through examples but, first, a general point. Design is about optimizing within constraints. While all aspects of design affect the quality of the outcomes, here we should focus on *strategic design,* i.e. those aspects of design that concern the interaction of the initiative with the system it aims to support – in particular, how to use constructively the predictable 'gaming' responses of the groups affected.

The analysis is focused on STEM subjects, with the examples drawn mainly from mathematics education – and not only because this is an area where I have some expertise. Why? First, the gulf between official goals and widespread practice is particularly wide in mathematics. Equally, the issues of teaching STEM subjects are very different from those in the humanities. There, most teachers have established genres of lesson types into which they insert appropriately chosen texts from the literature; in STEM the technical demands of the literature make it inaccessible to pupils so the specific design of each learning or assessment activity presents subject-specific technical challenges. Each is a 'treatment' that can be more or less effective, depending on the quality of the 'engineering' – the research-basis and subsequent design and development.

The following outlines exemplify the gross mismatches that often arise between worthy intentions of policy initiatives and the outcomes in practice, primarily through failure to anticipate users' responses.

"Teaching to the Test." A system of examinations where the results have serious consequences for the lives of pupils, teachers and their institutions means that, despite urging to the contrary, "teaching to the test" dominates the learning activities in most classrooms. Yet most current tests actually assess only a small subset of the official performance goals. This leads to a sharp narrowing of the *enacted curriculum* in comparison with *intended curriculum* in mathematics, leaving out, for example, the substantial chains of reasoning that pupils will later need for problem solving in life and work.

Defining levels of performance in mathematics in terms of detailed lists of skills expected at each achievement level actually *drives down* standards. Why is this inevitable? Given the importance attached to test results, *fairness* requires that pupils be given the opportunity to reach as high a level as they can. That opportunity lies in the easiest tasks involving that skill: short, specific items with no other factors increasing the difficulty. In particular, tasks are avoided if they involve any extended chains of autonomous reasoning like those involved in non-routine problem solving, a key learning goal.

Competition between examination providers in England is another prime example. Introduced to encourage competition, schools naturally use the freedom of choice to try to select the easiest, most predictable exams, leading to a "race to the bottom" between providers – *not* the "rigorous, high-quality examinations" that the government intended. This is a classic *client-customer mismatch*.

Ensuring comparability of standards. The government was then driven to find ways to counter this market pressure – basic fairness requires parallel exams to be comparable in difficulty. Among the many methods available it chose to micromanage the design of the tests, again narrowing the range

of performances assessed – and eliminating the variety that competition, let alone innovation, might have produced.

Assessing higher-level skills is well-recognized as a crucial assessment design challenge, reflecting the societal importance of such skills in the modern world where straightforward technical skills are absorbed by technology. International research and development has produced many robust tests of this kind; these have not yet penetrated high-stakes assessment in England which, for a variety of non-educational reasons, continues to focus on short, separate elements of performance.

"Coursework", sometimes called portfolio assessment, is the assessment of extended pieces of pupil work as part of a high-stakes assessment system. While it is accepted that such work is closer to the kind of performance needed outside school, and rigorous trials have shown it can be assessed rigorously and reliably, politicians don't trust it - for plausible but incorrect reasons – and teachers are happy to avoid the extra work involved. Thus, again, a narrower curriculum results.

Professional development programs often "will the ends but not the means". Governments recognize that continuing professional development is important for improving teaching and learning. However, additional resources are not realistically matched to the objectives. For example, there is no change in teaching loads to make room for the professional development time that would be needed. Even when time is set aside, the opportunity is often lost for lack of a well-designed structure of activities - so teachers welcome the chance to catch up on other pressing needs.

Computer-based testing has an obvious political attraction. It promises lower costs and seems more objective than written examinations with complex responses that require human scoring. Unfortunately, despite claims from some providers, the range of responses that can be computer-marked does not include extended autonomous reasoning or substantial problems (Black et al 2012, Section 4).

It is notable, and no accident, that assessment or testing features so often in these examples of system failure. "What you test is what you get" (*WYTIWYG)* (Burkhardt, Fraser & Ridgway 1990). The powerful influence of high-stakes tests (Black et al 2012) raises an inherent design conflict. To optimize such tests politically, the priorities are simplicity, low cost, and 'reliable' results that do not provoke criticism or appeals. But there are easier ways to raise marks than teaching for understanding. To optimize pupil learning, the full range of performance goals needs to be assessed, requiring more complex assessments of the kind recommended for the National Curriculum by the expert group (TGAT 1987) – but rejected for STEM subjects[5] by ministers who had priorities of the first kind.

There are working examples, outlined in the Appendix, which show that most of these failures could be, and have been, avoided by better research-based design and development. Why are policy making and practice not built on such well-engineered solutions? The next three sections aim to throw light on this question.

[5] If English tests had been restricted in the same way as for Mathematics, there would be no extended writing, only spelling and grammar exercises. Politicians understand language but dismiss mathematical thinking ("I don't use any maths") because it does not match what they had at school – an indictment of that curriculum.

2. The policy makers' world

"If I want to talk to Education, who should I call?"

Channeling Henry Kissinger's famous question about Europe, this is the policy maker's dilemma. The problem does not arise in medicine; as a Permanent Secretary who had been in both departments pointed out, if there is a problem in medicine, you call the President of the relevant Royal College who will point you to the leading researchers in the field. Why is education so different? First and most obvious, there are no institutions like the Royal Colleges that, while not immune from criticism, are accepted as authorities in their field. But this is just a symptom; there are deeper causes, related to the nature of education research and practice and the working environment in which policy is made.

In this section I try to set out what I have learned about that critical part of the education system: policy and how it is developed. This determines the frameworks and the resources within which education practitioners operate. Reducing the gulf between what happens in typical classrooms and what we know those teachers could achieve depends on our first understanding the dynamics of this very different world, then developing ways to inform and influence it much more effectively than in the past. I know of no well-established solutions; the observations in this paper are offered as a contribution to their development.

A politician's life: pressures and constraints.

The development of policy involves, like all design challenges, optimizing within constraints. What are the constraints for policy makers? What are the pressures in their lives?

- **Time.** Politicians in government lead very busy lives, working within a complex system that imposes a great variety of demands on their time – from their civil servants, their colleagues in government, and in their party, the media and 'keeping in touch' with the public.
- **Pressure.** Politicians work within a political context that generates constant inputs from the media, fellow ministers, party members with an interest in education, and a great variety of lobbyists with political, commercial or professional arguments to 'sell'. Many ideas for initiatives emerge from this rather random set of biased inputs, in which the educational research community rarely figures large! At the same time, when national statistics identify, for example, poor comparative results in a school or region, we must acknowledge that ministers are left with no choice but to 'do something' and do it quickly.
- **Procedures**. Governments are complex organizations that need to have well-defined practices and procedures for the generation, implementation and monitoring of policy. From a policy maker's point of view, these may often feel like constraints – indeed, they are meant to be[6].
- **Money.** Funding is a major issue in any government department, with constant pressure to keep budgets down. Any initiative will involve the addition or diversion of funding, either of which will meet resistance; so the initiative needs to be seen to be both effective and cost-effective in terms of advancing the minister's goals.

[6] "The Green Book", http://www.hm-treasury.gov.uk/data_greenbook_index.htm, lays down procedures to be followed by all parts of central government.

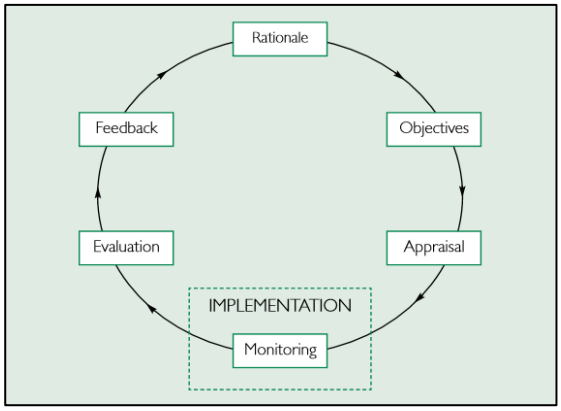
- **Timescales.** Last but perhaps most important, these pressures have much shorter timescales than the decade-long process of real improvement in education. Tomorrow's headlines and the 'events'[7] that inspire them require a swift response. Policy discussions happen week-by-week. Elections come every few years. Ministerial appointments are often shorter than this, leaving each minister with little time 'to make their mark', and thus advance their careers. And they will have moved on long before the outcomes emerge and are evaluated.

These pressures all influence ministerial choices as to what initiatives to pursue. There is a continuum in between between the 'fast thinking' based on gut reaction that has been important for survival and the reflective 'slow thinking' that we need for finding high-quality solutions to complex problems. In the pressured world in which ministers take decisions, it is not surprising that world of research and systematic development seems too slow to be useful, and not surprising that plausibility rather than proven effectiveness becomes the criterion for choosing initiatives. Yet our experience, and the examples in the Appendix, shows that "slow design" pays off. Section 5 discusses how this timescale mismatch can be reconciled and used positively.

The policy-making process.

The function of the UK Civil Service is to help ministers implement their (officially the government's) policies. So, the process is fundamentally inward-looking, driven by the priorities that 'The Minister' chooses. Note also that both ministers and civil servants normally move between departments every few years; so they are not specialists in education – rather they draw on whatever expert advice they think they need.

How does the policy-making process work? The UK government has a process, mirrored by the research cycle shown. The need for evidence and analysis to underpin good policy is recognized – but what kinds of evidence?



There are several sets of key players involved. *Ministers* set agendas and want to drive change, to make their mark. They do not have *time* to engage with the knowledge base in detail. *Policy makers* are responsible for identifying how to enact ministerial wishes, including options and models of implementation. They draw on *Analysis teams* – statisticians, operational researchers, economists, social researchers, supported by external experts and advisers. Only some of the last group are likely to come from education.

What the policy makers are looking for includes 'problem scoping' – as to size, shape, boundaries, and dynamics – leading to option generation that combines rational analysis and political pragmatism. This is informed by data collected through monitoring, including impact assessment and equalities assessment. They use it for modelling, forecasting and evaluation. These professions and their uses of data are sometimes not as seamlessly integrated as one might hope; that can depend on whether policy makers recognize their need for advice – and ministerial priorities do change!

[7] Harold Macmillan, when asked what ministers fear most, replied "Events, dear boy, events."

In this process, the pressure 'to do something' about a recognized problem, means that initiatives are implemented before they are thoroughly developed through piloting, let alone evaluated. Given the political capital involved, it is equally unsurprising that there is little enthusiasm for commissioning subsequent evaluation, which might prove negative.

Other political input - select committees, think tanks and the media.

Under the British constitution, governments are responsible to Parliament. However, the government largely controls the agenda and scheduling of parliamentary activities so, provided the party in power has a 'working majority', members of parliament can ask questions and raise issues but rarely divert the government from its program. The most effective instruments for raising questions are probably the 'Select Committees' of MPs. These normally work on a cross-party basis, aiming for consensus in reports and recommendations. A select committee has the power to call witnesses including ministers, expert advisers, and those outside government whose actions seem to impinge on policy and practice. Operating under much less pressure than ministers they often take a more holistic view of their field.

This has been true of the Education Select Committee which has asked "what are we trying to do" on various fronts, such as work and changing employment patterns, the purpose of education (it isn't just assessment), life skills and using skills to solve problems. Select committees reach out to stakeholders of all kinds, more often 'think tanks' like the Education Policy Institute than academic research groups. They look out for evidence and unintended consequences of policy decisions. Despite their limited resources, their work is valuable in raising issues with government and the public - they get media coverage - but less often in changing policy.

Think tanks contribute diagnostic analysis to departments across government. Some, like the Institute for Public Policy Research or the Resolution Foundation have a broad brief and, often, a political slant. Others, like the Institute for Fiscal Studies or the Education Policy Institute, focus on a specific area. Most aim to be non-partisan. In addition to their analysis, they often make policy suggestions but do not see the detailed design and development of policy initiatives as their role.

Probably the strongest influence on policy making (perhaps, but only perhaps, apart from government policy as set out in its election manifesto) is the constant input from the media, particularly newspapers that support the party in power. Inevitably this is amplified and distorted by social media. It is not the education correspondents' pieces that tend to be of concern to ministerial meetings; it is the dramatic reporting of 'events' that can be presented as 'scandalous', so that 'something must be done', often 'to ensure that this never happens again'[8]. This follows from and reinforces the unquestioning belief in 'accountability', with little regard for its cost or consequences. Ministers in the firing line look for 'who is to blame'.

Having outlined my understanding of pressures and processes through which policy-makers work, let us turn to things specific to education.

[8] While this is an international phenomenon, the British seem to gain particular satisfaction from outrage.

On goals for education and how to get there.

Goals for education.

I believe that most politicians, and the civil servants who support them, want to improve the learning of all pupils. Indeed, they say it is their highest educational priority. They try to promote improvement within the constraints and pressures they face, day-by-day and on longer timescales. Further, within STEM education there is broad agreement on goals. In mathematics almost everyone agrees that: pupils should acquire skills in and understanding of procedures and their underlying concepts in number, algebra, geometry and (more recently) data analysis; they should be able to apply these to both standard and non-routine problems within mathematics, in other school subjects, and in the world outside. More broadly, they should develop powers of reasoning, not merely answer-getting, that enable them to think about the world in a mathematically literate way. Policy makers know that achieving this costs money - in England the education budget is about £50 billion a year. That's the good news.

On how to get there.

What is not agreed, indeed is often controversial, is an entanglement of two issues: how to move towards these goals, and their relative priority. The first is, or should be, a technical issue for the research and development community to resolve – as in medicine, determining what 'treatments' are effective in what circumstances, and what support practitioners need to implement them well. The issue of priorities is societal, and therefore fundamentally political but even here technical considerations of feasibility and cost are central – as, for example, in deciding on size of typical classes, the single most important variable in the cost of running the system.

Let us look at a key example. In mathematics, a politician's view that "*You can't solve problems until you have a solid basis of skills*" has regularly been used to defer teaching useful problem solving skills 'until later'. As a result, many pupils never get there, and are left with the common view that mathematics is pointless as well as punitive. The quoted statement, which sounds technical, is basically wrong. You can usefully reason about real world problems with minimal mathematics – even if you can only *count* and *distinguish objects*, there's a rich world of important problems that can be better understood using mathematics. Does the statement actually reveal priorities – that problem solving is not really seen as important? Perhaps because "I didn't do it at school". Much more than in medicine, the interface between technical and political issues is blurred in this way.

Technical naïvete, or arrogance?

Less happily, many politicians feel free to believe that they know how best to achieve the improvements in learning they choose to address, that they understand teaching and learning at a technical level. For example, they maintain a firm grip on the details of national standards for curriculum, with mixed results. They decide, against the evidence, that " practice in arithmetic is what you need for understanding maths" or that, in reading, decoding "phonics" should come *before* understanding the meaning of text. The balance of research suggests that there is long term benefit when these things are *developed together*. Policy makers *do* take advice from educational professionals but, typically, only on the details – and they choose to consult those whose advice will not conflict with their convictions. (Everybody shows 'confirmation bias'.)

Their choices tend to include totemic beliefs – for example, in the importance of fluency in 'long division', a vivid example of a skill in pencil-and-paper arithmetic where fluency is long outmoded,

except in classrooms. It is probably no accident that these beliefs reflect their own school experience, 13 years of which seem to convey a feeling of expertise.

The contrast with medicine is stark – it is inconceivable that a minister of health would seek to specify medical treatments: "The Chinese have used acupuncture for thousands of years, and it has worked well. We're going to change the health service over to acupuncture." It just wouldn't happen, but similarly naïve technical decisions are common in education policy.

The 'expertise' confusion.

How and why is education so different? Many factors have contributed to this disregard for research results and professional expertise, sometimes even a denial that it exists. We shall discuss them in more detail in the next two sections noting, for example, the absence of broadly accepted research-based authority to guide policy decisions. So whom should policy-makers consult? For consult they do - indeed they like to believe that "my door is always open". But if that were really so they would be deluged with opinion, more or less well founded, from many who work in education and from the even-more-numerous others who have strong opinions about it. "Cocktails from a fire hose" indeed. In practice, politicians' choose to consult like-minded 'experts' and people in their own party and 'circle'. So establishing a better collecting and filtering mechanism is important.

To summarize.

The policy maker's challenge is to match the educational reality of the teacher with the political realities. We offer a definition:

Well-designed policy makes solving educational problems politically practical.

3. The education research world

"If you so smart, how come you ain't rich?"

Why is the education research community not at the centre of education policy design? Why, when policy makers see an educational problem, do they so rarely turn to the research community to develop a solution? In medicine, while the strategic organization and funding of health care systems is an essentially political domain, the modes of working of medical professionals are recognized as areas of expertise of the medical community, closely informed by medical research on health and, in particular, diseases – their nature and how to diagnose and treat them effectively and cost-effectively. We have seen that the situation is very different in education. In this section I look at the pattern of research in education, and suggest reasons why it has such a limited role in guiding policy and practice.

The academic value system

Education research is a vigorous field that serves its members well (Burkhardt 2016); it is less successful in improving the practice of education, let alone the formation of policy aimed at improvement. How does this manifest itself?

In most research fields, while there are disputes about new research results, there is a large body of knowledge that is broadly accepted as a foundation on which to build both current practice and future research and development; as we have noted, in education research there is no such body of accepted results – *nor any collective attempt to establish one*. Why?

First, the field flourishes on disputation, giving major credit for new ideas and 'theories', provided only that they are plausible. It gives minimal credit for replication studies that seek to discover the *boundaries of validity* of such theoretical claims, which are commonly based on small-scale studies or, more often, the author's untested 'observations'. There are reasons for this, which I discuss below, but it is a major obstacle to education becoming a research-based field. Replication is at the heart of the scientific method – in medicine new treatments are subject to testing and evaluation in depth. This in turn leads to new insights, and new products.

Secondly, the lively 'scientific' research that goes on in education is mostly diagnostic -– either survey-based or, in the absence of replication, too small-scale to reliably inform either design or policy. In contrast to more research-based fields like medicine, there is relatively little research with an 'engineering' approach – that is 'treatment-focused' research, development and evaluation.

These academic values as to what is "good research" partly reflect the constraints under which most researchers operate: shortage of time for research, limited resources and pressure to publish frequently. Whatever the causes, the value system might be summarized in a list of academic priorities, which favour:

- new ideas *over* results with a body of evidence that can be relied on
- disputation *over* consensus building
- new investigations *over* replication and extension
- first author *over* team member, hence
- personal research *over* team research, hence
- small studies *over* major programs
- papers in academic journal *over* products and processes that improve practice.

These priorities are the reverse of what is needed to provide a reliable basis for the design of tools and processes that will improve practice and inform policy-making. They encourage the reworking of familiar concepts in different terms, leading to a multiplicity of closely related theoretical viewpoints rather than convergence on agreed terms and statements. They reward new perspectives over the consolidation of a solid body of agreed results of the kind needed for design and development - and wider credibility. Other fields show that it doesn't have to be this way (Burkhardt and Schoenfeld 2003).

Styles of research in education

Where does this pattern of research come from? Educational research is a mixture of traditions with very different views of research and scholarship – essentially those of the humanities, sciences, and engineering. The focus of both the humanities and the science approaches is the search for improved insights – into learning, teaching, professional development, and the behaviour of education systems. The engineering research approach has a different priority: impact on practice.

The humanities approach is the oldest tradition, based on scholarly acquisition of knowledge and, in the light of prior work, critical analysis of it – but with no tradition of empirical testing of assertions made. The key product is critical commentary – just as for works of literature or art – so the quality of the writing is a core criterion. The ideas and analysis, based on the authors' reflections on their experience and observation, are often valuable. Importantly, without the requirement of empirical testing, a great deal of ground can be covered[9]. This is still the most influential approach, perhaps partly because it allows politicians to regard it as "opinion" and rely for policy formation on their own "common sense" beliefs – with mixed results like those outlined in section 1.

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

The science approach is also focused on better insights, on improved understanding of "how the world works" through the analysis of phenomena, and the building of models that help to explain them. The process involves exploring the system and generating insights, now called hypotheses, but with an additional requirement for testing them empirically. Testing hypotheses takes time and effort, which sharply narrows the range of what can be covered in a single study - rarely providing evidence of wider generalizability of the results (see e.g. Schoenfeld 2002, Burkhardt 2013). Even when such methodological issues are resolved, such research is fundamentally diagnostic. If well done it provides reliable insights, identifying problems, and suggesting possibilities. However, it does not itself generate robust practical solutions, even on a small scale; for that, it needs to be linked to the 'engineering' approach.

The engineering approach is directly concerned with supporting practice – not just understanding how the world works but helping it to work better (Burkhardt 2006, 2014). It does this by developing solutions to practical challenges 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. Again there is an essential requirement for empirical testing of the products and processes, both formatively in their development and in later evaluations in use. The key products

[9] ... as in this paper!

are new or improved tools and processes that work well for their intended uses and users; but the work also produces new theoretical insights that come from the feedback in the design and development process. With these elements, development is research. While there has always been some support for engineering R&D in education, such work is often undervalued in the academic community – in some places only 'insight' studies in the science tradition are regarded as "research". Medicine takes a very different view, valuing greatly research for the development of new treatments.

All three research traditions have contributions to make but currently the balance among them is far from optimal for translating insights into practical improvements in classrooms and school systems (see for example, Schoenfeld 2009). What balance, of effort and of "academic credit," would be most effective, and how does it differ from the current pattern? I believe that that there should be more 'engineering' research and that this needs, and in turn produces, reliable research insights to build on. A parallel implication for 'science' research in education is an increased focus on evaluation in depth of treatments in the field, to provide formative input to policy, and to further engineering development.

Scales of research and development

Finally, it useful to distinguish four different foci for education R&D: learning, teaching, teachers, and school systems. The very different scales needed are set out in Table 1. The differences may be summarized by the relevant research domains: a laboratory for L; a classroom for T; many classrooms for RT; and, for SC, whole school systems.

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: class-rooms, assessment, professional development, community relations

Currently, the great majority of research is confined to L and T, where some progress can be made by single researchers in a year or two of their available research time. But in system terms, there is a crucial difference between T, which is about teaching possibilities, and RT, which is about what can be achieved in practice by typical teachers with available levels of support. In engineering research

in education (Burkhardt 2006), the process of design research at T is continued through further rounds in more typical classrooms of trialing, observation and revision until the products work well for a well-defined target group of real users, RT[10].

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. Burkhardt and Schoenfeld (2003) list the elements that are key to a successful link between research and practice, as evidenced in other research-based fields of practice like engineering or medicine:

- Robust mechanisms for taking ideas from laboratory scale to widely used practice. Such mechanisms typically involve inputs from prior research, imaginative design of prototypes, systematic development, and marketing mechanisms that rely in part on respected third-party in-depth evaluations.
- 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.
- A reasonably stable theoretical base, with a minimum of faddishness and a clear view of the reliable range of each aspect of the theory.
- Stable design teams of adequate size to grapple with large tasks over the relatively long timescales required for sound work of major importance in both research and development.
- Sustained funding to support the process on realistic time scales.
- Individual and group accountability for ideas and products — do they work as claimed, in the range of circumstances claimed?

Within the insight-research community, a change in the balance of effort could greatly enhance the impact of their research (Burkhardt 2016). Key changes would be to build the collaborations needed for showing generalizability, to focus on evaluation-in-depth of specific well-engineered products and processes, observing what happens in the classroom as well as pupil performance outcomes. This is the way to build a solid body of reliable results, together with evidence of the range of their validity and applicability - and to identify (and publicize) successful initiatives.

All these elements require encouragement and support from government.

[10] The Education Endowment Foundation, funded by the Gatsby Foundation with government support, is pioneering this approach on a small scale.

4. The world of educational practice

"In a completely rational society, the best of us would aspire to be teachers and the rest of us would have to settle for something less, ..." Lee Iacocca

The key constituencies in the education system all have much the same broad aims, focused on improving pupil learning and developing good citizens. However, like policy makers, each profession faces day-to-day pressures that do not directly support those aims. In this section we look at what this means for practitioners, in particular for those most important guides, and gatekeepers, to pupil learning: teachers. What they create with their pupils in "the zone of instruction" (Elmore 2011) determines how far the system meets its aims.

Teachers are recognized as the key to pupils' progress, in learning and in character development, as doctors are in treating our ailments. Yet their lives are very different. Teachers' salaries are much lower than doctors' or those in many other professions that teachers could have chosen. The hours they are expected to work are long, filling their evenings in term time – though they *do* have long holidays. They face pressures, day-by-day and in the longer term, many in the name of 'accountability'. Good teachers assess each child as a core part of teaching but the system requires them also to produce summative 'data' that, supposedly more 'reliable', takes teacher time without adding anything to pupils' learning[11]. Further, this data is a major element in their evaluation as professionals in a way that is not so true for doctors. In practice, such tests are far from balanced across teachers' supposed goals, encouraging the narrowing of the curriculum focus onto the easy-to-measure elements that we pointed out in Section 1. This is symptomatic of a lack of trust in teachers as professionals; justified or not, such attitudes undermine teachers' confidence and self-worth – and thus performance.

Given the demands of the job, and the specific demands of STEM subjects noted above, it is unreasonable to expect teachers to design their own lessons – any more than doctors design their own treatments. The best outcomes are likely to come from providing practitioners with the (large) spectrum of well-engineered 'treatments' they need to support their clients – pupils or patients.

School leadership is intended to provide inspiration, guidance and support to their teaching staff – and many principals do, insofar as they can. But, like middle-management everywhere, they are squeezed between pressures from above and below. They are the channel through which government constraints and pressures are fed into the school. They, too, are judged by aggregations of many kinds of data, financial and demographic, as well as the dominant measure: tests scores. Their school may be inspected at any time by Ofsted, whose reports have consequences for the school and for individual teachers. So leadership transmits these pressures onto the teachers, who spend time and emotional energy responding to them – for example, 'teaching to the test' and always being prepared for an inspection.

Government, quite apart from the policy development that is the focus of this paper, plays various ongoing roles in the operation of the system from month-to-month and, if a problem hits the media, from day-to-day. The Department for Education monitors the system in many ways – accountability

[11] Teachers are not the most extreme sufferers from this focus of contemporary society. Social workers, for example, spend more time creating a 'paper trail' than they do with the clients they serve. Policy makers rarely account for the cost of accountability, let alone its cost-effectiveness, and there is little public discussion. Practitioners know that any mistake, or simply a difficult case, could threaten their careers when media-driven politicians know that "something must be done so that this never happens again" - though, inevitably, it usually will.

again. It collects data from schools and local school systems around the country on test results and a variety of politically sensitive demographic issues – for example, the proportion of pupils on free school meals. Ministers and their civil servants feel political pressure to respond to such evidence. If schools in one part of the country are doing relatively poorly on tests, what should ministers do? Improvement needs a balanced combination of pressure and support. Test results and inspections provide pressure; the support provided is rarely developed so it works as intended, matching outcomes to intentions; what is needed in each case is *a coherent program that has been shown to enable the teachers in question to meet the challenge*. This can probably be achieved with a well-engineered integrated program of continuing professional development for teachers, tutorial help for struggling pupils, and perhaps breakfast to ensure that learning can take place. Effective support like this costs more than pressure – and needs much more design and development effort than just testing. The real cost of accountability, including diversion of learning time and undermining of teachers' professional standing, is rarely considered.

Publishers have long played an important role in STEM education, where the teaching materials they offer are used by most teachers and thus largely determine the classroom activities through which pupils learn. There is a long history of excellent textbooks playing a key part in educational improvement projects supported by government or foundations, notably the Nuffield Foundation and the School Mathematics Project – usually with associated examinations. This approach has been greatly suppressed since the 1989 introduction of the National Curriculum, which set out what pupils were to learn. While in the humanities, these Attainment Targets were expressed in broad terms (Reading, Writing, Speaking and Listening in English) they were much more detailed in the STEM subjects, reducing Mathematics in particular to lists of individual skills – reflecting politicians memory of the subject. The picture has been further complicated in the last decade by two factors: the emergence of free material on the web along with a movement for schools to develop their own 'schemes of work'. Neither of these is likely to help schools deliver better learning without support in the challenging task of designing a rich and coherent curriculum.

Assessment is the obvious powerful link between all these four constituencies. Government, which repeatedly says it doesn't want to tell teachers *how* to teach, regards examinations as the measure of school effectiveness in delivering *what* the National Curriculum says they should teach. In the name of accountability, it maintains tight control of the examinations at various ages. After long resisting the responsibility that flows from the empirical fact that, when test scores have important consequences, teachers will 'teach to the test', there is now acceptance that "What You Test Is What You Get". However, other pressures on the design and delivery of examinations outweigh the responsibility to have 'tests worth teaching to', balanced across the performance goals in the intended curriculum. This leads on to a fifth constituency.

Assessment providers in England are 'examination boards' that design the tests in each subject, deliver them to schools, collect and mark the pupils' responses, and aggregate these scores into grades for each pupil. There is an oversight body that, on behalf of government, approves each board's syllabus and monitors the exam papers and results. These are then integrated across classes and schools, with published "league tables" that rank order schools. Boards have to reconcile conflicting factors: a wish to make their exams educationally sound; government directives; commercial pressures to make them attractive to schools that are looking for the best results for their pupils; the design capabilities of their teams; no time to refine their tests through trialing; and concern that their tests will be subject to challenges from anxious parents and schools. Together these make it very difficult to create the "tests worth teaching to" that would form "pressure that drives improvement".

5. How could the system work better?

"... if you don't have no scheme, how you going to have your scheme come true."

Malcolm Swan, after Oscar Hammerstein

Having outlined the worlds in which the key players in the education system operate, let us return to the challenges set out in Section 1. The aim now is to identify changes that will increase the probability that the outcomes of policy initiatives are much closer to policy makers' intentions than at present, avoiding the failures and unintended consequences that so often arise. Where existing institutions seem to represent a step in the right direction, I shall point to them.

As a result of research and development over the last 40 years, we now know a lot about how children learn. We have developed effective treatments, teaching approaches and materials that embody this knowledge. These enable typical teachers to help their pupils perform across the widely-accepted range of learning goals. The fact that this doesn't happen in most classrooms is the main driver for this paper.

The underlying challenge is to reconcile the mismatch of timescales, for improvement in education and for political decision-making, to the advantage of both. This will need a system in which R&D processes get ahead of policy, offering ministers alternatives for implementation and minimizing the need to base policy initiatives on plausible ideas that lack rigorous testing.

This requires, as in medicine, that high-quality impact-focused research and development proceeds on an ongoing multi-year basis *and* that it also delivers, year-by-year, well-proven results that can form the basis of policy initiatives that are likely to prove successful, individually and cumulatively – and deliver political capital.

This will, in turn, require some new structures within both government and the education communities. These could take a variety of forms; however, the necessary functions are clear:

A. To identify on an on-going basis, areas for improvement that are recognized as important for moving towards educational goals, both broad and specific.
B. To identify, filter and classify on the basis of research evidence, past and on-going work that shows potential to contribute to improvement in these areas.
C. To support through appropriate funding structures a substantial program of research-based design, development and evaluation in depth, focused on these challenges.
D. To advise government on proven products and processes that merit consideration for large-scale initiatives.

Short term steps

To this end, there are changes that policy makers can make in the short term on a trial basis within existing structures, as a step towards establishing the above structures (linked as shown).

- Recognize that every policy initiative they consider needs to be treated as a design and development problem.
- In order to cover a broad range of options, include from early in the process of developing the policy, a range of advisers with diverse expertise in research-based design and development across the various elements of the policy (A, B above).

- Focus on strategic design, matching the scope and pace of the changes involved to the resources available for effective support of those who will need to change (A).
- Give explicit focus of attention in the design of initiatives to alternative models for the change process involved (A, B).
- Commission a number of candidate designs from groups with relevant, successful track records to offer ministers a variety of solutions along with some evaluation of their strengths and weaknesses (C).
- If there are novel areas involved where there are no groups with proven expertise, establish a preliminary research and development project to explore ways in which one might move forward - or not (C).
- Consult with all the main constituencies facing change and modify the design to take into account how their behaviour is likely to change in response to the initiative (D).

This process in based on the norms in research-based fields, as described at the end of Section 3.

The Appendix describes examples from the past that show this can work well.

Medium term structures

New structures will be needed to institutionalize this approach to policy development and cover the functions A-D above. I see three distinct elements: engineering development, scientific research, and authoritative advice to government. There seem to be advantages in keeping these institutionally distinct with three bodies.

- **National Institute for Educational Development** to fund research-based development of treatments, both products and processes, in areas that government chooses for improvement.
- **National Institute for Educational Research** to gradually strengthen the research basis of policy and practice, by funding specific types of research including evaluation-in-depth of existing well-defined practices, and in the longer term, the building of a body of broadly-validated results, with well-defined boundaries of validity.
- **National Institute for Educational Excellence** to review proposals in the light of the evaluation evidence on the strengths and weaknesses of the education system, to evaluate available 'treatments' and to advise government on their cost-effectiveness.

Each of these bodies will need to be regarded as an experiment and evaluated on a 10-year timescale. A central role of all three is to develop institutional memory in its area.

It is *not* the aim of this paper to provide comprehensive specifications of roles and structure for these institutions; that is a task for the communities of Sections 2 to 4. But it is important here to say something more about my conception of them and show promising past and current developments.

A National Institute for Educational Development (NIED)

A **National Institute for Educational Development (NIED)** is designed to reflect a long history of funding for innovative projects in areas recognized as in need of improvement, some of it funded by government bodies from the Schools Council to the Qualifications and Curriculum Authority. The Education Endowment Foundation (EEF) currently plays a related role. Equally, many of the memorable contributions of the past have been funded by charitable foundations with either broad or specific briefs. The Nuffield Foundation had a particularly distinguished set of achievements that shaped the learning of science over the last half-century. For example, Nuffield A-Level Physics,

which continues to this day, changed the other board's syllabuses. The School Mathematics Project played a similar role in Mathematics in the 1970s and 80s.

The approach of the most successful initiatives had three key features. They started as small-scale enterprises, refining their products and processes over a period of years as they were taken up more widely. They took care to get all the key constituencies on board – teachers and their schools, the relevant scientific community, and an examination board – and to develop comprehensive support through examinations, teaching materials, and professional development offerings. The third crucial ingredient was government encouragement to explore improvement strategies. Although the 1988 Education Act, which introduced the National Curriculum, explicitly encouraged such projects, the pressures for uniformity in the name of accountability like those sketched in Section 1 choked off many promising initiatives at that time. This needs to change. Diverse channels of innovation are at the heart of progress in any field[12], as is the building of human capacity for the 'engineering' R&D that produces it.

A National Institute for Educational Research (NIER)

A **National Institute for Educational Research (NIER)** is designed to strengthen gradually the research basis of policy and practice by funding and otherwise encouraging types of insight-focused *research that directly support the improvement of practice* – thus complementing the engineering R&D support by NIED. These include evaluation-in-depth of existing well-defined treatments and practices and, in the longer term, the building of a body of carefully validated and broadly-accepted results as in research-based fields. The key shift needed for both of these functions is for researchers to work together in teams on well-defined programs. Only this will provide solid evidence – as they do when tackling complex problems in other subjects.

A National Institute for Educational Excellence (NIEE)

A **National Institute for Educational Excellence (NIEE)** is designed to provide an authoritative source of advice for government, evaluating available 'treatments' and advising government on their effectiveness and cost-effectiveness. As NICE does in medicine, it will review products and processes, well-established or innovative, in the light of the evaluation evidence on their strengths and weaknesses, and potential contributions to improving pupil learning across the spectrum of learning goals. The Education Policy Institute (EPI) has made a start in this direction. However, a larger base of well-engineered products and processes together with a much larger base of solid evaluative evidence on their effects in use is needed before this Institute can function properly. NIEE thus depends on the products of NIED and NIER.

Longer term monitoring and investment

Building this structure and the human capital to make it work well needs time. A 10-year timescale seems an appropriate target. The process can and should be tuned and evaluated for cost-effectiveness during this period.

The 1988 Education Act, which established the National Curriculum with all-party support, made specific provision for innovation of the kind suggested here, but without providing structures to support it. As with all such changes, the focus moved to implementation - to making the new system work, particularly the assessment system. This left little 'energy' for improving what happens in "the

[12] Comparability of standards in assessment can be achieved in other ways.

zone of instruction". There politicians adopted the naïve (or disingenuous) position: "We set the targets for attainment; we do not tell you how to teach to meet them." In fact, the tests, which were standardised, largely determine what happens in most classrooms because "What you test is what you get" (Burkhardt et al. 1990). As a result, the wave of well-evaluated innovation of the 1970s and '80s shrivelled in the UK. The approach described here will re-invigorate the improvement process.

Is good engineering cost effective?

The program outlined in this paper will cost money. While a lot of academic time goes into insight-focused research it is fragmentary; building teams and enabling them to work effectively together on specific problems needs some funding. The engineering research approach is more expensive than the craft-based "authoring", whether of teaching materials or of policies, on the basis of "experience" alone. For example, the Shell Centre over forty years has found that research-based design and development of tools to support improvement that work well costs roughly £20,000/$30,000 per class-hour, mainly because of the iterative development process. That is universal for new products in other fields.

Is this good value? It seems expensive; in system terms it isn't. To "do the maths", taking the English education system as an example, a comprehensive redevelopment over 10 years of materials for all 15,000 hours[13] of teaching in Years 1-13 would, at this price, cost roughly

£20,000 per hour*15,000 hours/10 years = £30,000,000 per year

which is *only 0.06% of the running costs of the system*[14]. Any field that takes research-based improvement seriously would regard this as a worryingly *low* level of investment in R&D. And this number is an overestimate because:

- Not all the curriculum needs re-engineering; there is already a lot that is well-proven.
- The human capacity needed to carry through high-quality R&D activity at this rate does not exist and will take a decade to build from its current small base.

How much is currently spent on R&D for improving school education? That depends on what you count – but it is an order of magnitude lower than this, and not well-focused on improving practice.

The impact ←→ funding catch

Improving education is a universally accepted social priority, yet R&D funding as a proportion of its cost is minuscule. Why? The reasons seem to be historical – certainly not rational. When you look at fields that get 'serious money', history suggests that the key to the door is an event with major impact that society recognizes. Until the Second World War, medical research was mainly done by academic doctors in teaching hospitals as part of their university work; then came antibiotics. The work and funding of the Medical Research Council expanded rapidly, complemented by private foundations like the Welcome Trust, leading to the current enormous collaborative-competitive research enterprise, epitomized by the Human Genome Project. Again, physics research had modest funding through universities until the late 1930s. It is reported that Rutherford's Cavendish Laboratory in Cambridge had a budget of £3,500 – even with inflation, not comparable to today's

---

[13] There are about 15,000 hours (Rutter 2009) of 'different' lessons taught in a school year – roughly 25 hours a week x 40 teaching weeks a year x 13 'years' from age 5 to 18 of schooling. This is increased by, for example, special needs provision.
[14] For comparison, reducing average class sizes by one pupil would cost around 3%. The evidence that this improves pupil learning outcomes is, to be generous, ambiguous - though it is obvious 'common sense'.

millions. Radar, Operations Research and The Bomb provided the vivid societal impact that has convinced governments ever since to fund major physics-based projects like Space Science or the Large Hadron Collider at CERN. Ongoing spin-offs, notably the World Wide Web, have sustained this faith in the societal value of such support.

The strategic challenge to education research and development is to produce improvements that are recognized as valuable sufficiently widely for governments to have faith that investment pays. Not an easy challenge; the suggestions in this paper aim to help its realization.

6. The initial concerns addressed?

I'll now look at each of the concerns set out at the beginning to see how the structures and programs set out in the previous section will address them.

- **Poor communication** between politicians and policy makers and the education communities. The three National Institutes are designed to address this directly, with NIED developing tools to support specific improvements, and NIEE filtering evidence from across the spectrum of research, development and policy to provide government with the best available advice on alternatives.
- **Trying to 'fix the problem'**. This political tendency fails to take into account system complexity, and that real improvement involving changes of practice is gradual.
  The Institutes will see the evidence that step-by-step specific changes are more effective at improving learning than the 'Big Bang' approaches and advise government of this.
- **Technical naïveté**. This is the tendency of politicians and policy makers to design aspects of teaching and assessment at a level of technical detail that they would not dream of trying in, say, medicine, engineering or agriculture – effectively discounting the expertise of the education professions. The new structures provide a solid base that should discourage this tendency – though it may well take time before politicians stop believing in their own educational expertise.
- **Imposition of uniformity**. It is the primary mission of NIED to encourage and fund high-quality research-based development, for pilot projects that may, after evaluation, grow into the mainstream. This will provide ministers with a range of options. NIEE needs to convince government that this is worth some reduction in uniformity and to adopt one of the many approaches to accountability assessment that accommodate such a change.
- **Imbalance in education research** between the dominant analytical-diagnostic research traditions and treatment-focused research and development with an engineering approach. NIEE will ensure that the balance will be improved gradually by funding appropriate kinds of research, and explicitly addressing the academic value system so as to recognize the importance of impact on practice and policy.
- **No generally accepted body of results**. The failure of the education research community to develop a solid body of agreed research results and evidence on the effects of specific 'treatments' undermines progress. There are reviews in the literature that attempt to remedy this. It is a prime role of NIER to collect and sift these, and to encourage further research that investigates the boundaries of validity of the assertions made, by funding research teams with this goal. In parallel with this, it will lead an ongoing process of consensus building so that denying the emergent body of well-established research findings will seem eccentric, rather like denying climate change.
- **The lack of authoritative structures** that integrate evidence from research and practice in a form that policy-makers respect and can use. NIEE, building on the work of the other two Institutes, is designed to meet this need.

There are, of course, no guarantees that the ambitious program outlined in this paper will eliminate these concerns, leading to a healthy education system that is steadily improving in the many specific ways that are needed. The concerns provide a set of headings for its evaluation over the decade after it is launched. By way of encouragement, the Appendix gives some examples of well-engineered changes that proved successful in the past or elsewhere, also addressing the specific problems listed in Section 1.

APPENDIX: What does good engineering look like?

To illustrate the fact that such success is possible, even when profound change is involved, we now describe and exemplify the strategic design[15] (Burkhardt 2009) of five initiatives where:

- these principles were applied
- the outcomes were close to the intentions
- the changes were widely welcomed by all the various groups involved.

We then return to the problems set out in Section 1, where the outcomes were far from the policy intentions. We sketch, in the spirit of a proposal, what a well-engineered solution would look like.

The support for each of the initiatives described below was produced by standard engineering methods. Their essential features are: input from prior research, imaginative design and systematic iterative development through successive rounds of classroom trials, with detailed feedback guiding each revision. The following examples were realizations of this engineering approach.

Integrated development

Integrated development (see e.g. Black 2008) has long proved the most powerful approach to improvement for teaching materials, an examination, and professional development for teachers. The key strategic design feature is development by an expert design team bringing in the academic subject community, teachers and their schools, an exam board and a national association. Though initially on a small scale, these initiatives survived and spread, also influencing other mainstream curricula and examinations in those subjects. This project-based approach was standard from the 1960s to the 80s, with a long list of memorable successes, including Nuffield A-level Physics (Black and Ogborn 1970, Ogborn 2003), School Mathematics Project (SMP 1970), and many others that changed thinking on education in their subjects.

Examination-led gradual improvement

This model is based on well-supported step-by-step improvement, driven by a high-stakes examination. One seriously new type of task was introduced each year into the examination for age 16, corresponding to 5% of the two-year syllabus and three weeks of teaching. Now often called "The Box Model" from the support materials (Shell Centre 1984, Swan et al. 1986), these 'replacement units' comprised exemplar exam tasks with commentary and solutions, materials for teaching the new curriculum component, and for in-school professional development. Most important, the approach was popular with teachers and pupils, who, as the exam results showed, acquired the new skills involved – non-routine problem solving in mathematics for the first module, translation skills for the second. Over half the board's schools bought the materials.

In the 1990s the Australian state of Victoria introduced a high-quality exam, emphasising extended problem solving (Victorian Board of Studies 1995). Though this was a school leaving examination, subsequent research (Clarke and Stephens 1996; Barnes, Clarke, and Stephens 2000) showed problem-solving activities in the enacted curriculum throughout the secondary schools. What you test is what you get– for better or, so often, for worse.

---

[15] In design, the tactical and detailed design are equally important – indeed the usual source of *excellence*; many *failures*, however, are due to bad strategic designs. (Burkhardt 2009)

Replacement units

In the 1990s, California set out to introduce a broader view of school mathematics that met international standards, including non-routine problem solving. It was recognized that teachers need support in the new aspects of teaching involved. An approach similar to the box model was adopted, with curriculum units designed to supplement existing curricula. However, the new state test that was developed was less well-aligned. So, as expected, the impact was more limited.

Focused supplementary support

The key strategic design feature here is to focus well-engineered support on a specific area of weakness in curriculum and assessment – for example, the mathematical practices within the Common Core State Standards for Mathematics in the U.S.A. In the Mathematics Assessment Project (MAP 2014) the proven potential classroom (Black and Wiliam 1998, 2001) of well-engineered formative assessment for improving pupil learning was the focus of the improvement program. A nationwide professional development effort was given both focus and power through supplementary teaching materials based on a diagnostic teaching approach that had been developed over 30 years of research and development (Swan 2006, Burkhardt and Swan 2017). This enabled enabled typical teachers to introduce this new and challenging aspect of teaching in their own classrooms. There were over 7,000,000 downloads of lessons by teachers and others. Independent evaluations showed large learning gains (see e.g. Herman et al 2014) in line with the findings of the Black and Wiliam.

Collaborative Assessment

Collaborative assessment systems in Scotland worked rather differently. A standing forum of researchers, civil servants and other experts helped to bridge the chasms between different expectations, leading to a more continuous dialogue on how well things were going and how they might be improved. This approach led to support being targeted on research that directly improved performance by teachers and pupils in the classroom.

Section 1 revisited: How could the problems have been solved?

Here we list a well-established solution or two for each of the problems outlined in Section 2. In each case there are, of course, other possibilities and variants that, with good engineering, can be made to work well. The following examples should be seen as 'proofs of concept'.

- **"Teaching to the Test"** is a problem if, as now, the test is not balanced across learning goals. Alternatively, it is an opportunity to influence teaching by commissioning tests that are balanced so that teaching to the test leads teachers to implement a balanced curriculum in their classrooms. The Box Model, B above, is an example of this.
- **Defining levels of performance** in mathematics in terms of detailed lists of skills, feeds teachers understandable craving for certainty with a false picture of mathematics[16]. A model that avoids this (Burkhardt 1990) has three elements: a description of mathematics at each stage in broad terms [17], giving equal weight to the topics and the processes or practices

[16] "The Devil is not *in* the details; The Devil *is* the details." Alan Schoenfeld 2010.

[17] The national curriculum specification in Denmark occupies 8 pages (see e.g. Niss 2003). In contrast, the Common Core State Standards gives 4 pages to "mathematical practices" and, despite the good intentions of the authors, 94 pages to the detailed lists of content that they were driven to provide. It is hardly surprising that content, familiar ground, commands most attention from practitioners.

of doing mathematics; a substantial set of task exemplars that show teachers the kinds of task their pupils should be able to tackle; a description of the various kinds of classroom activity that need to be part of the learning experience[18].

- **Competition between examination providers**, leading to a "race to the bottom". The simple answer is to have a common set of papers across providers, so that all pupils are assessed on the same task set. Providers can then compete on price and service. This would allow the three stages of assessment - developing a collection of tasks, choosing a balanced sample, delivering the exam – which require different skills, to be individually optimized. (see Daro & Burkhardt, 2012)
- **Ensuring comparability of standards** between equivalent assessments without the need for the current elaborate and stultifying machinery can most simply be achieved by "double entry" – getting a sample of pupils to take both examinations, comparing results and adjusting grade boundaries to match. An alternative, to insist on a proportion of common tasks across all parallel exams, is only valid if the common tasks are either balanced across goals or randomly chosen from year to year.
- **Assessing higher-level skills.** International research and development over many years has produced many successful examples, including those outlined above. This is a solved problem – except, like so many listed in this paper, at system level.
- **"Coursework"** or "portfolio assessment", assesses pupils on types of work that reflect the actual performance goals in the subject. There is a long history in the UK of well-validated systems. Typically, teachers assess their own pupils' work on a mixture of 'set' and 'chosen' tasks, using well-defined marking schemes. This is followed by a process of 'moderation' in which teachers compare samples of their pupils' work, first within a school then, with a 'chief moderator' from the board, between schools. The professional development pay-off is clear. Such an approach was recommended for the National Curriculum by the expert group on assessment (TGAT 1987); despite the evidence, it was rejected by ministers – as it so often is.
- **Professional development programs**, to be effective in improving teacher's classroom strategies and skills, need the following: scheduled time as an integral part of the job; specific well-engineered activities[19]; formative feedback, based on what happens in the teacher's classroom. The goal is ongoing active participation in a professional learning community. The main cost is the time outside the classroom. (The phrase 'release time' gives a clue to how PD is currently viewed.)
- **Computer-based testing** – now that's a tough one, with no good solution except "Horses for courses". There are five phases in the use of an assessment task (Black et al, 2012, Section 4): presenting the problem; *pupil working on the problem and constructing a response*; capturing the work and the response; *evaluating the response*; recording and processing the data. While for a short right-or-wrong item all can be handled by computer, those in italics are beyond current AI for complex tasks requiring substantial chains of autonomous reasoning. So while video and interactive features can make the problem clearer than on

[18] The National Curriculum Working Group for Mathematics in England provided all three elements in its reports; the policy makers stripped it down to a list of "Statements of Attainment" (partly on grounds of what can be written into Statute - another thing to avoid) The result has impoverished mathematics education here for 30 years and counting.

[19] It is at last as challenging to design powerful professional development sessions as powerful lessons (see e.g. Bowland Maths professional development 2008). The tradition of the leader designing their own is as sub-optimal as for STEM lessons.

paper, the computer does not provide a natural working environment for doing mathematics[20], where making sketches and using symbols are central; these also make valid computer-scoring impossible. There are working systems but these rely on breaking up a complex task into steps - a very different kind of performance.

This last example makes a general point: as in medicine and engineering, not all attractive solutions will work.

[20] The situation is better for text-based subjects, though computer-scoring systems for essays rely on indirect measures, like readability, sentence length and vocabulary. The word "not", often crucial, is beyond them.

References

Barnes, M., Clarke, D., and Stephens, M. (2000). Assessment: The engine of systemic curricular reform? *Journal of Curriculum Studies,* 32(5), 623-650.

Black, P. (2008). Strategic Decisions: Ambitions, Feasibility and Context. *Educational Designer*, 1(1). http://www.educationaldesigner.org/ed/volume1/issue1/article1

Black, P., Burkhardt, H., Daro, P., Jones, I., Lappan, G., Pead, D., Stephens, M. (2012). High-stakes Examinations to Support Policy. *Educational Designer,* 2(5), Article 16. Retrieved from http://www.educationaldesigner.org/ed/volume2/issue5/article16

Black, P.J. & Ogborn, J.M. (1970) The Nuffield Advanced Physics course. *Physics Bulletin*. 21, 301-303.

Black, P. J., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education* 5, 7-74.

Black, P. J., & Wiliam, D. (2001). *Inside the black box: raising standards through classroom assessment*. London: King's College London School of Education.

Bowland Maths professional development (2008) downloaded from http://www.bowlandmaths.org.uk/pd/index.html

Burkhardt, H. (1990). Specifying a national curriculum. In I. Wirszup, & R.Streit (Eds.), *Developments in school mathematics around the world 2* 98-111. Reston, VA: National Council of Teachers of Mathematics.

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). Retrieved from: http://www.educationaldesigner.org/ed/volume1/issue3/article9

Burkhardt, H. (2013). Methodological issues in research and development. In Y. Li & J. N. Moschkovich (Eds.), *Proficiency and beliefs in learning and teaching mathematics - Learning from Alan Schoenfeld and Günter Törner*. Rotterdam: Sense Publishers.

Burkhardt, H. (2014). Curriculum design and systemic change. In Y. Li & G. Lappan (Eds.), Mathematics curriculum in school education. Heidelberg: Springer.

Burkhardt, H. (2016). Mathematics Education Research: a strategic view. In English, L. and Kirshner, D. (Eds.) *Handbook of International Research in Mathematics Education, 3rd Edn*. London: Taylor and Francis.

Burkhardt, H., Fraser, R., & Ridgway, J. (1990). The Dynamics of Curriculum Change. In I. Wirszup, & R.Streit (Eds.), *Developments in school mathematics around the world 2* 3-30. Reston, VA: National Council of Teachers of Mathematics.

Burkhardt, H., & Schoenfeld, A. H. (2003). Improving Educational Research: towards a more useful, more influential and better funded enterprise. Educational Researcher 32, 3-14.

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.

Clarke, D. J. & Stephens, W. M. (1996). The ripple effect: the instructional impact of the systemic introduction of performance assessment in mathematics. In M. Birenbaum and F. Dochy (eds), *Alternatives in Assessment of Achievements, Learning Processes and Prior Knowledge* (pp. 63-92). Dordrecht, The Netherlands: Kluwer.

Cockcroft Report (1982). *Mathematics Counts.* London: HMSO.

Daro, P. and Burkhardt, H. (2012) A population of assessment tasks*, J. Mathematics Education at Teachers College* 3, 19-25

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

Fullan, M. (2016) *The New Meaning of Educational Change,* 5th edition. Routledge: Abingdon

Herman, J., Epstein, S., Leon, S., La Torre Matrundola, D., Reber, S., & Choi, K. (2014). *Implementation and effects of LDC and MDC in Kentucky districts (CRESST Policy Brief No. 13).* Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

MAP (2014). *Mathematics Assessment Project*: introduction and materials downloaded from map.mathshell.org

Niss, M. (2003): The Danish KOM Project and possible consequences for teacher education. In R. Strässer, G. Brandell, B. Grevholm & O. Helenius (eds.). *Educating for the future: Proceedings of an international symposium on mathematics teacher education* (pp. 179-192). Gothenburg: Royal Swedish Academy of Science.

Ogborn, J. (2003) Advancing Physics evaluated. *Physics Education,* 38, 330-335.

Rutter, M. (2009) Fifteen thousand hours, Cambridge, MA: Harvard University 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, NJ: Erlbaum.

Schoenfeld, A. H. (2009). Instructional Research and the Improvement of Practice. In J. D. Bransford, D. J. Stipek, N. J. Vye, L. M. Gomez and D. Lam (Eds.), *Educational Improvement: What Makes It.*

SMP (1970). School Mathematics Project materials: downloaded from https://www.stem.org.uk/resources/collection/283319/school-mathematics-project

Shell Centre (1984). Swan, M., Pitt, J., Fraser, R. E., & Burkhardt, H., with the Shell Centre team*, Problems with Patterns and Numbers*. Joint Matriculation Board, Manchester, UK; downloaded from http://www.mathshell.com/scp/index.htm.

Swan, M. with the Shell Centre team: (1986). *The Language of Functions and Graphs*, Manchester, UK: Joint Matriculation Board, reprinted 2000, Nottingham, UK: Shell Centre Publications. downloaded from http://www.mathshell.com/scp/index.htm.

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). ISBN 1-86201-311-X.

TGAT (1987) Report of the *Task Group on Assessment and Testing*; downloaded from http://www.educationengland.org.uk/documents/pdfs/1988-TGAT-report.pdf

Victorian Board of Studies (1995). *VCE Mathematics: Specialist – Official Sample CATs*. Blackburn: HarperSchools.