The Power of Proofs
(Much) Beyond RCTs

Dirk Van Damme
Distinguished Senior Research Fellow, Centre for Curriculum Redesign

www.curriculumredesign.org

© Center for Curriculum Redesign – 2022 All Rights Reserved
### Table of Contents

Introduction ............................................................................................................................................... 3  
The sluggish growth of educational research ......................................................................................... 4  
Education research: improvement needed .............................................................................................. 5  
The one-dimensional view of scientific proof .......................................................................................... 8  
The evidence deficit in education ........................................................................................................... 11  
Teachers’ professional knowledge .......................................................................................................... 13  
A pluralist view of scientific proof ........................................................................................................ 15  
A novel approach to scientific proof and relevance ................................................................................ 16  
Profiles of the nine research designs or outputs ...................................................................................... 19  
  Descriptive research ............................................................................................................................ 19  
  Qualitative studies ............................................................................................................................... 21  
  Case studies ......................................................................................................................................... 22  
  Surveys and cross-sectional studies .................................................................................................... 23  
  Quasi-experimental and longitudinal studies ...................................................................................... 24  
  Randomized controlled trials .............................................................................................................. 25  
  Meta-analyses and reviews of experimental research ......................................................................... 27  
  Research synthesis papers ................................................................................................................... 28  
  Scholarly books ................................................................................................................................... 29  
Conclusions ............................................................................................................................................. 30

### List of Figures

- Figure 1. ESSA tiers of evidence as promoted by IES (USA) ................................................................. 9  
- Figure 2. Pyramid of evidence adapted by CCR .................................................................................... 10  
- Figure 3. Aleatoric versus epistemic uncertainty ................................................................................... 11  
- Figure 4. Knowledge profiles based on type of knowledge, thematic orientation and cognitive demand ............................................................................................................................................................... 14  
- Figure 5. Quality of teaching practices by domains of practice ............................................................. 14  
- Figure 6. Two-dimensional mapping of research designs or outputs .................................................... 19  
- Figure 7. Profile of Descriptive work .................................................................................................... 20  
- Figure 8. Profile of Qualitative studies .................................................................................................. 21  
- Figure 9. Profile of Case studies .......................................................................................................... 22
Introduction

Education systems are facing serious challenges. From the classroom up to the minister’s office, there are huge needs to rethink and redesign the existing education infrastructure and ways of working. But education reform can no longer rest solely on good intentions and noble ideas. Like most other modern public policy sectors, education needs the input of solid research evidence and reliable knowledge to bring about change. In order to fill that need, education research has grown tremendously both in quantity and quality. As a consequence, educational policy and practice are now increasingly becoming “evidence-based” and “data-driven.” As Andreas Schleicher at the OECD states in his signature: “Without data, you are just another person with an opinion.”

As a major independent organization aiming at improving education and making it more relevant, CCR is developing its work on the basis of the best possible evidence. In its 2015 paper “Theory of Change & Research Process,” 1 CCR describes its knowledge development process for developing and refining its frameworks. Its approach builds on three efforts to come to be knowledge foundations needed: synthesis, analysis, and organization of evidence into accurate and actionable knowledge. It uses five design principles to organize research evidence to develop frameworks that are comprehensive, compact, uncorrelated, appropriately levelled, and globally relevant.

This paper builds on the 2015 paper and expands and refines the argument. It argues that the conventional approach of translating and disseminating ever more rigorous research evidence towards policy and practice is failing the rapidly expanding knowledge needs of educational policy and reform. A reconsideration of what counts as scientific proof and relevant evidence is needed in order to expand the provision and usage of actionable knowledge to education policymakers and practitioners.

---

The sluggish growth of educational research

Education systems are increasingly looking for reliable sources of knowledge to inform and improve policy and practice, and they increasingly turn to scientifically generated research knowledge. Research evidence promises to provide the knowledge and information enabling complex systems to change in order to adapt to or anticipate changing circumstances. Over the past decades, the international movement towards evidence-based or evidence-informed educational policy and practice has become very influential. The supply of educational research knowledge has progressively increased. Based on a rough calculation by OECD researchers Vincent-Lancrin and Jacotin, the production of educational research papers has increased fivefold between 1996 and 2015.

Other large public policy systems, such as health care, have made the transition to evidence-informed policy and practice over the past decades. The input of scientific research knowledge into these systems has completely transformed them. The health sector provides a useful comparison to education, a public policy sector of similar size and comparable mission. Even a very superficial comparison will demonstrate that compared to the health sector, education still is very far behind in moving to an evidence-informed system. In its recent white paper, education.org has calculated that globally the health expenditure is 75% higher than education expenditure, but its knowledge synthesis production is no less than 2,600% higher than education’s. For every synthesis developed in education, health produces 26 times more syntheses. Whether the publication of research synthesis papers is a reliable metric can be disputed (although an extensive knowledge domain needs high-quality synthesis papers), but the magnitude of the gap certainly is significant. Education is still far remote from transitioning to a knowledge system based on research evidence comparable to the health sector.

Yet, although not at the same speed as health research, the growth of education research remains important. Several countries have set up specific institutions to support, fund, and manage educational research. Specialized journals, conferences, professional organizations, etc. have seen the light of day. Ministries have created specific knowledge departments. Knowledge brokerage institutions are being established that translate research findings into advice and recommendations for policy and practice.

---


4 Unpublished communication.


Has this growth of the education knowledge base resulted in visible improvements in the sector? In other words, what is the macro-efficiency of education research? Education doesn’t seem to have benefitted a lot from the increase in educational research. Since the 1970s, tens of billions of dollars have been spent in the US on educational research, yet student performance in reading and math, as measured by the NAEP, has not improved. “After decades and billions of dollars of expenditure, the education research establishment, by and large, does not know how to reliably identify the factors that account for success in large-scale education systems” is the verdict of former NCEE President Marc Tucker. 7

The harsh verdict remains valid even at the international level and when different measures of effectiveness and efficiency are used. Compared with the evolution of the major systemic risks facing education today, the track record of educational research looks very weak: research has not prevented educational systems to become more costly (per student funding increased)8 while average overall quality has deteriorated in many countries (learning outcomes as measured by international learning assessments such as PISA).9 Education fads abound and come and go, with no clear development and maturation in pedagogical practices. Educational research has not been able to fundamentally address the enduring equity issue, nor has it made education systems more innovation-friendly.

More research is definitely needed to fill the knowledge needs of the education sector looking for better evidence to strengthen policy and practice. Research that continues to improve in quality and rigor. But more of the same will not be enough. This paper argues that some fundamental changes in the production of education research will be necessary.

**Education research: improvement needed**

Together with the expansion of educational research funding, activity and output, the methodological quality of educational research has improved. Even if one can still find a lot of contemplative papers in educational journals, the “empirical turn” in educational research has clearly changed the research output in the direction of quantitative, methodologically sound research. Research methodologies have become more sophisticated and research designs more complex. This development has resulted in thousands of really valuable and interesting research papers in specialized outlets. It also has led to growing optimism that educational research could potentially resolve some of the most difficult issues in educational policy and practice.

However, concerns about the quality of education research are dampening this optimism. To some extent, these quality issues are not typical for education research but affect social sciences in general. A very serious challenge in social science research is the so-called “replication crisis,” which indicates that the results of many scientific studies are difficult or impossible to reproduce.10 A recent study found huge differences in the interest in replicability of research

---

7 https://ncee.org/tucker-writing/education-research/
findings across disciplines, with psychology in a far better position than sociology. The replication crisis in social sciences has hit educational research very hard. Already in 2014, an analysis of the top 100 education journals revealed that only 0.13% of the published papers were replications, far less than in any other field of social research. It looks difficult for educational research to adequately respond to the replication crisis. Moreover, it does not appear that replication failure much reduced the influence of nonreplicated findings.

Education and social sciences suffer from troubling trends in scientific research more generally. For example, there are some serious flaws in the standard production and publication processes in sciences, such as peer-reviewed journals publishing hundreds of non-sensical papers. A recent paper has shown that the prolific publication trends are working towards the canonization of already well-established findings rather than advancing novel ideas. Evaluation of research proposals often suffers from conservatism bias. The result is stagnation instead of advancement through the confrontation between old and new ideas. Negative citations are extremely rare, suggesting that citation patterns mainly aim at confirmation. Social sciences seem to be particularly vulnerable to confirmation and publication bias. For instance, a recent study found that after correcting for publication bias the research evidence for the popular concept of “nudging” was poor to non-existent.

The excessive publication trend also leads to the publication of what one author provocatively calls “scholarly bullshit” papers, contributing little to none to the scholarly body of knowledge. For example, in journals covering the field of sustainability and transition studies, up to 50% of

published papers could be classified as “bullshit.” 21 Publication pressure is also the cause of outright fraud, of which the much-publicized cases are probably only the tip of the iceberg. The retraction rate of scientific journals is only about 0.04% and it ought to be much higher.

As a relatively young field in the social sciences, education research is particularly vulnerable to those methodological issues. Although much progress has been made in recent years, methodological flaws still harass education research publications, for example confounding statistical correlations for proof of causality. Methodological conflicts have absorbed enormous research power and energy. Philosophical and normative debates on methodology still occupy a large part of publication output.22 Equally, education research is easily affected by various kinds of bias. According to one study, the “insider bias” in educational research is very important.23 “Insider bias” refers to the widespread practice that actors or organizations designing and implementing an educational intervention also are the ones evaluating its effects and impact. External, independent evaluation studies are quite rare in the educational field.

Much like other social research domains,24 also educational research seems to be vulnerable to ideological or political bias. Many educational researchers are strongly motivated by compelling feelings of sympathy for certain political causes, which is implicitly or explicitly influencing their research objectives, design, and findings. This also gives way to in-group favoritism and group thinking. The field of education research is quite strongly divided into specific “camps,” with high risks of confirmation and publication bias within each one.

Related to the replication bias, is the issue of overreliance on statistical significance testing (by a simple \( p \leq 0.05 \) decision rule) and the misinterpretation or even exaggeration of effect sizes, often by lumping together effect sizes of real and non-existent effects. It is well known that low-powered randomized controlled trials (RCTs), for example, tend to systematically exaggerate effect sizes. A retrospective study of promising RCTs found that the estimated effect sizes are exaggerated by an average of 52% or more.25 Misinterpretation or lack of comparability between effect sizes has been the major criticism of one of the most popular books in education research, John Hattie’s best-selling *Visible Learning*.26 In one of his excellent blog posts, titled “John Hattie is wrong,” the late Bob Slavin criticized Hattie for simplifying vast amounts of research findings and introducing many errors in his attempt.27 Moody et al. rightly conclude that significance tests should be part of a decision about substantive significance rather than the only

21 https://www.science.org/content/article/what-massive-database-retracted-papers-reveals-about-science-publishing-s-death-penalty
27 https://robertslavinsblog.wordpress.com/2018/06/21/john-hattie-is-wrong/
determining factor. Effect sizes and confidence intervals should always be presented in a way that highlights substantive importance.\textsuperscript{28}

A lot more could be said about the conceptual and methodological challenges of education research. It still is a young and maturing field. But its challenges and flaws jeopardize its capacity to meet the knowledge needs of the rapidly expanding education system. Providing better and more strategic funding for education research, establishing education as an academic discipline, improving self-regulation and evaluation, and accepting limitations seem to be the necessary steps to move education research to more rigor and maturity.\textsuperscript{29} But will the problem be solved then?

**The one-dimensional view of scientific proof**

In the previous overview of methodological challenges to education research, we deliberately saved the most important one for a more detailed discussion, namely the issue of what counts as evidence. The growth of education research, its progressing methodological quality and the increasing sophistication of research tools have led to growing optimism that educational research could potentially resolve some of the most difficult issues in educational policy and practice. The expansion of experimental design and especially of randomized controlled trials (RCTs), often proclaimed to be the gold standard of educational research design, brought many people to believe that it was possible to develop a catalog of “what works” interventions, that could be readily implemented. In the US, this belief was clearly visible in the 2002 No Child Left Behind Act, which used the phrase “scientifically-based research” more than 50 times, and in the 2015 Every Student Succeeds Act (ESSA), which calls schools to implement “evidence-based interventions.”


The “what works” idea became a real movement, fueled by the Institute of Education Sciences (IES) and its funding mechanisms, and also spread to other countries and continents, most notably in the English-speaking world. The success of the “What works” movement led to the reformulation of research agendas, focusing attention on intervention and effectiveness research.30

Through an interesting initiative to support and guide grant-seekers for support in the ESSA framework, the IES introduced a model of four “tiers of evidence” (see Figure 1), going from “strong evidence” at Tier 1 to “demonstrates a rationale” in Tier 4.31 Each tier consists of specific values on five criteria: study design, results of the study, findings from related studies, sample size & setting, and match. The idea is that a grant application for a specific intervention to improve student learning under ESSA has to demonstrate the quality of the evidence base.

The idea that not all evidence counted equally, first developed in medical research, gained strong ground over the past decades. The ESSA tiers of evidence in some way mirrors the hierarchy of levels of evidence in the well-known pyramid of levels of evidence, originally developed in the

---

field of medical research. Various versions of the pyramid have been developed, with weaker study designs such as case studies and qualitative studies at the bottom, and randomized controlled trials (RCTs) and meta-analyses at the top. In the middle, surveys, cross-sectional studies, and quasi-experimental and longitudinal studies are represented.

Figure 2 presents a slightly adapted version of the well-known pyramid. Under the levels of scientific evidence, the pyramid also includes pre-scientific forms of knowledge which take a strong role in educational policy and practice, such as expert opinion, professional knowledge of teachers, experiential knowledge, and educated opinion.

Indeed, the main shortcoming of the standard one-dimensional, hierarchical view of levels of evidence is the trade-off between the push towards methodological rigor by pushing educational research to ever higher levels of the pyramid on the one hand, and the decreasing relevance for educational policy and practice on the other. In reality, the “gold standard” of methodologically sound research in education, symbolized by RCTs, proves very hard to reach and attracts trenchant criticism in the educational community. Educators doubt whether experimentation is ethically acceptable in education. Others argue that the experimental design implies an exaggerated simplification of the often-messy reality in schools, ignoring context and

---

32 There are many sources for the pyramid of evidence, see for example https://amedd.libguides.com/c.php?g=476751&p=3259492#:~:text=The%20evidence%20pyramid%20represents%20the%20strongest%20evidence.

experience. Yet the most potent criticism is that the outcomes of RCTs are very difficult to generalize to other settings and conditions.

## The evidence deficit in education

To better understand the issue at stake here, we can introduce the concept of “epistemic uncertainty.” Originally coined in modeling, machine learning and artificial intelligence, the concept of “epistemic uncertainty” refers to things we could in principle know but do not know in practice because of incomplete knowledge of the system, because of a lack of data, or because of measurement errors. It has found its way into various disciplines, such as medicine. It can easily be applied as well to applied intervention research in education.

![Figure 3. Aleatoric versus epistemic uncertainty](image)

Epistemic uncertainty is to be distinguished from aleatoric uncertainty, which is the uncertainty caused by probabilistic variations in a random event, or natural variation or inherent randomness.

---


An interesting visualization of the difference between the two forms of uncertainty is presented in Figure 3.37

Translated to the social sciences, research sophistication is primarily oriented towards minimizing aleatoric uncertainty, at the expense of increasing epistemic uncertainty. By knowing more about an ever-smaller part of reality, we finally know less about reality.

Education research suffers from the problem that the more sophisticated the research design is, the higher its methodology ranks on the pyramid, the farther it goes from the messiness of educational practice, and the more difficult it becomes to replicate its results. Thus, paradoxically, research at the higher levels of the pyramid of evidence is confronted with higher risks to forsake the evidence needs of the system. In education research, a lot of variables come into play. It requires enormous methodological sophistication and scrutiny to control for all possibly relevant variables. And the more variables the research design tries to control for, the higher the distance with the messiness of educational practice and the less relevant the research results risk to become. It is extremely difficult for education research to escape this trade-off.

This is one of the main reasons, next to the quantitative and qualitative deficiency of education research to meet the abundant and diversified knowledge needs of education systems, why other knowledge sources still occupy such an important place in education. Teachers and policymakers have to call on other knowledge resources than scientific research to meet their knowledge needs and to address the evidence deficit in educational policy and practice created by epistemic uncertainty.

We know from several iterations of OECD’s Teaching and Learning International Survey (TALIS) that teachers feel confronted with severe gaps in their knowledge base.38 They report lacking the necessary knowledge, for example, to work with special needs students, to integrate technology in their classroom practice, to teach in highly diverse classrooms or in mixed ability settings, etc. Yet, teachers don’t view research evidence as a useful resource for filling these knowledge gaps, especially not if it contrasts with their own experiential knowledge or the transmitted professional knowledge. An EEF/NFER survey in 2017 found that academic research was having only a small to moderate influence on decision-making by teachers relative to other sources.39 Important barriers related to skills and time to access scientific evidence prevent the use of research evidence by teachers.40 A quote from a superintendent in the US, mentioned by the late Bob Slavin in one of his many interesting blog posts on the relationship between educational research and practice, summarizes the issue quite eloquently: “GOOD research is

---


that which confirms what I have always believed. BAD research is that which disagrees with what I have always believed.”

In a recent article, Daniel Willingham and David Daniel list the main reasons why most teachers seem to ignore research evidence. “First, teachers may view research as somewhat removed from the classroom, with further translation needed for the practice to be ready to implement in a live setting. Second, teachers may judge a practice to be classroom-ready in general but delay implementation because their particular students and setting seem significantly different from the research context. Third, teachers may resist trying something new for reasons unrelated to its effectiveness—because it seems excessively demanding, for example, or because it conflicts with deeply held values or beliefs about what works in the classroom. Finally, teachers may be unaware of the latest research because they only rarely read it.”

Teachers’ professional knowledge

Instead, they heavily rely on professional knowledge, the accumulated – often over generations of teachers – experiential knowledge base about “what works” in teaching and learning. A recent project of OECD’s Centre for Educational Research and Innovation (CERI) tried to assess the state of teachers’ pedagogical knowledge base as a critically important component of teacher professionalism. In order to stimulate the process of professionalization, teachers would need to continuously update and improve their core professional knowledge. A pilot study assessed teachers’, teacher candidates’ and teacher trainers’ pedagogical knowledge in five countries with regard to the domains of assessment, instructional and learning processes. The pilot found that there were important gaps in teachers’ pedagogical knowledge, but also that there were different profiles of teachers’ knowledge in different countries. In general, knowledge about assessment scored stronger than knowledge of instruction and learning. This suggests that the reception and integration of scientific knowledge into the professional knowledge base worked better for those aspects of professional practice which are highly valued by the educational system. And the poor results with regard to learning and development (brain functioning, memory, metacognition, etc.) shows that recent scientific findings have a very hard time being integrated into the professional knowledge base.

As shown in Figure 4, all three samples (teachers, teacher candidates and teacher educators) exhibited a strongly practice-based knowledge, while theoretical-scientific components in professional knowledge are of less significance. When in-service, teachers rely even less on scientific knowledge than when they were in teacher training.

---

45 Ibid., Figure 4.12
This relates to an observation in another recent OECD study, the Video Study of Teaching, where trained scorers rated teaching quality in lesson videos across eight jurisdictions.\textsuperscript{46} The rated quality of instruction was lower than that of classroom management or even social-

emotional support. Overall, teachers excelled at managing the classroom, gave students average social-emotional support, but provided them with instruction of subpar quality. This suggests that either usable but research-based knowledge about instruction is not readily available to teachers or that teachers don’t integrate that knowledge into their professional knowledge base.

While there are strong doubts whether teachers’ professional knowledge qualifies as evidence- or research-based knowledge, the reality is that teachers’ professional knowledge is a very important and powerful component of the knowledge infrastructure in education. And most likely, it will remain so, even in the context of growing education research and increasing efforts of translating and disseminating research findings into policy and practice.

That’s why it is legitimate to include teachers’ professional knowledge in the pyramid represented in Figure 2, together with other forms of legitimate knowledge, such as expert opinion or even lay knowledge. For example, we rightly consider parents as partially behaving as rational actors in educational decision-making, so their opinion and even what is called “common sense” have a role in the education knowledge infrastructure. The image of a pyramid also suggests that the knowledge systems at the bottom occupy a much larger space of the knowledge needs of the system than the more sophisticated and powerful levels of evidence at the top.

A pluralist view of scientific proof

So, let’s return to the domain of scientific research. The argument developed so far is that education systems’ knowledge needs are enormous and only very partially met by education research, especially the kinds of research meeting the requirements at the higher levels of the pyramid of evidence. This creates a problem of evidence deficit and epistemic uncertainty, as well as a trade-off between research excellence and relevance for policy and practice. In order to solve this problem, we need to drive the argument further. We need more evidence developed through rigorous research such as RCT’s, but not every piece of evidence should come from RCT’s. We need a more pluralist view of what constitutes scientific proof.

In the philosophy of science, the concept of epistemic or epistemological pluralism has gained a lot of ground. In the Wikipedia definition, “epistemological pluralism is a term used in philosophy, economics, and virtually any field of study to refer to different ways of knowing things, different epistemological methodologies for attaining a fuller description of a particular field.”47 Ganeri defines epistemic pluralism as “the view that there are different but equally valid ways of knowing the world.”48 Narrowing the concept to research methodologies, the concept of methodological pluralism, as defined by the American Psychological Association (APA), is “the belief that various approaches to conducting research, qualitative and quantitative, each have their respective strengths and weaknesses such that no one method is inherently superior to any other and no single method is best overall.”49 These definitions stand in contrast to the one-dimensional hierarchical view in the pyramid of evidence.

49 https://dictionary.apa.org/methodological-pluralism
We need to distinguish epistemic pluralism from epistemic relativism, which is a concept that states that every view is relative to the viewer’s position and counts equally. In its extremist definition, this concept invalidates the idea of truth, as every truth is relative to the individual, the position or purpose of the individual, or the conceptual scheme within which the truth was revealed.\(^{50}\) On the contrary, we should not reinforce the existing tendencies through which the epistemological crisis leads to fragmentation of truth and the decay of shared knowledge.\(^{51}\) Adopting a pluralist – not relativist – definition of knowledge production strengthens the concept of truth by recognizing the multiplicity of approaches to discovering the truth.

The case for epistemic pluralism in the social sciences has already been made in Michael Gibbons’ et. al. seminal book *The New Production of Knowledge*.\(^{52}\) They coined the concept of Mode 2 research, which is interdisciplinary, problem-solving oriented, applied, and contextually embedded, in contrast to Mode 1 research, which is disciplinary, characterized by theory building, and aiming at universal validity. The concept of Mode 2 research with its focus on application-driven, transdisciplinary, reflexive, and contextualized scientific knowledge production, has had a profound impact on educational research settings and has fostered a pluralist perspective on knowledge generation in education, especially in European countries.\(^{53}\) At the same time, however, it also intensified the debates and tensions between scientific quality and excellence on the one hand, and participatory and relevant research on the other.\(^{54}\)

In the literature, one may find many more arguments in favor of expanding the methodological scope of education research, in order to better meet the knowledge needs of the education systems. Implicitly and indirectly, such arguments boil down to loosening the methodological requirements of high-quality, rigorous research in order to be more relevant for educational policy and practice. The implicit idea is that scientific rigor is in a trade-off against practical utility. This is actually quite a simplistic approach. It doesn’t seem to be very productive to oppose different modes of scientific research against each other, which inevitably leads to polarization in the debate. It seems much more helpful to develop a more diversified perspective on how scientific proof can be assessed against relevance in different methodological approaches, in order to develop a landscape of research that really helps to reduce epistemic uncertainty in education. That’s what we attempt to do in the following sections.

### A novel approach to scientific proof and relevance

In the following sections, we attempt to develop a novel approach to assessing the value of research designs and research outputs through a two-dimensional framework combining their respective power of proof with relevance. We first distinguish nine research designs and outputs. The nine research designs or outputs we will assess are the following

---

\(^{50}\) https://en.wikipedia.org/wiki/Factual_relativism


• Descriptive work
• Qualitative studies
• Case studies
• Surveys & cross-sectional studies
• Quasi-experimental & longitudinal studies
• Randomized Controlled Trials (RCTs)
• Meta-analyses and reviews of experimental studies
• Research synthesis papers
• Scholarly books

Next, we assess the scientific power of proof of each research design or output on six criteria, of which four relate to the quality of hypotheses and two refer to the quality of the research results. The criteria identified for this are:

• Hypotheses are conceptualizable, i.e., there is a conceptual framework or even a theory behind the hypotheses.
• Hypotheses are testable, i.e., they lend themselves to rigorous testing, leading to discrete conclusions of confirmation or rejection.
• Hypotheses are modellable, i.e., given there are sufficient observations, they can be integrated in a model that is predicting empirical reality with a limited risk of error.
• Hypotheses are experimentable, i.e., they lend themselves to be tested in an experimental research design.
• Results are replicable, i.e., they are able to be replicated by other researchers.
• Results are generalizable, i.e., their validity extends to similar situations in different contexts.

We also assess the relevance of research findings for educational policy and practice on two criteria.

• The research findings have an acceptable position in the trade-off between depth and breadth, i.e., they are not too specific in order to lose relevance and at the same time they are not too broad to become meaningless.
• The research findings reduce epistemic uncertainty, i.e., they contribute to expanding and strengthening the actionable knowledge base of practitioners.

Scores on each of the criteria can vary between 0, 1, 2, and 3, in ascending order of strength. The scores are based on the author’s own expert opinion and do not involve any real measurement. As such, they certainly are disputable. The scores are given according to the typical quality of each research design or output. Obviously, there is a large variation even within the same category, so the score refers to a perceived median value.

Table 1 shows the scores of each research design or output on the criteria and dimensions. The scores for the six criteria of power of proof are added into one score and the scores for the two criteria of relevance are also added, so that we have total scores for each of the two dimensions. Figure 6 plots the scores on the two dimensions.
Table 1. Assessment of research designs or outputs on power of proof and relevance

<table>
<thead>
<tr>
<th>Research design or output</th>
<th>Hypotheses are:</th>
<th>Empirical results are:</th>
<th>Relevance</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Conceptualizable</td>
<td>Testable</td>
<td>Modellable</td>
</tr>
<tr>
<td>Scholarly books</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Research synthesis papers</td>
<td>3</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Meta-analyses and reviews of experimental studies</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Randomized Controlled Trials (RCTs)</td>
<td>3</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>Quasi-experimental &amp; longitudinal studies</td>
<td>3</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>Surveys &amp; cross-sectional studies</td>
<td>3</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Case studies</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Qualitative studies</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Descriptive work</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
This framework is superior to the one-dimensional pyramid of evidence presented in Figure 2 by integrating multiple criteria into the dimension of power of proof and by combining it with a dimension of relevance. What is clear is that scientific power of proof and relevance for policy and practice are two mutually independent dimensions. The Pearson correlation between the two dimensions over the nine research designs or outputs actually is very close to zero: -.05. So, in general, there seems to be a trade-off between scientific quality and practical relevance with the scientifically more rigorous designs or outputs being inferior to less rigorous ones in terms of relevance. However, the interesting observation is that at the level of each individual design or output the interaction of the two dimensions is quite unique. The trade-off doesn’t show up at the level of individual research designs or outputs.

Profiles of the nine research designs or outputs

The following sections do not aim to provide a systematic introduction into research methods. There are many handbooks and online resources to introduce readers into research methods.55 The following sections mainly serve to discuss how the two dimensions, scientific power of proof and relevance, behave in specific research designs, in order to strengthen the argument for pluralism.

Descriptive research

Descriptive research is the lowest, entry-level research in the hierarchical approach to scientific research. In fact, it is not even accounted for in most variants of the pyramid of evidence. Descriptive research aims to accurately and systematically describe a phenomenon, a population

55 See for example the excellent overview in https://www.scribbr.com/methodology/
of a situation. Descriptive research is an appropriate choice when the research aim is to identify characteristics, frequencies, trends, and categories. It is useful when not much is known yet about the topic or problem, and when the first thing to do is to explore the research object. Explorative studies are a form of descriptive research. Before you can research why something happens, you need to understand how, when and where it happens.\(^{56}\)

As an example, a recent study to explore the use of mobile evidence-based practice app by students in health and social care programs can be mentioned.\(^{57}\) A descriptive approach was used to explore, categorize, and compare students’ experiences.

![Figure 7. Profile of Descriptive work](image)

As shown in Figure 7, descriptive research scores low on all criteria of scientific power of proof but can lead to conceptualization and to some degree of generalization. On the two criteria for relevance, it has some value. The basic significance of this kind of research design is that it is a stepping stone towards other research approaches. Without exploring a phenomenon, it is very difficult to design more powerful research and to ask the right kind of research questions.

---

\(^{56}\) [https://www.scribbr.com/methodology/descriptive-research/](https://www.scribbr.com/methodology/descriptive-research/)

Qualitative studies

There are many definitions and approaches to qualitative research. Proponents of qualitative research also have their own networks, journals, organizations and research infrastructure. In its simplest definitions, qualitative research refers to all research with non-numerical evidence. Qualitative research relies on data obtained by the researcher from first-hand observation, interviews, questionnaires (on which participants write descriptively), focus groups, participant-observation, recordings made in natural settings, documents, case studies, and artifacts. Qualitative methods include ethnography, grounded theory, discourse analysis, and interpretative phenomenological analysis. Qualitative research methods are often used in order to have a deep understanding of how people experience the world, to respect the meaning people attach to the phenomena studies and to avoid the reductionism of quantitative approaches.

A good, recent example is this study of sharing and reusing open education resources by educators in Dutch educational institutions. The study relied on in-depth interviews with 55 educators in order to provide recommendations to developing effective policies to raise awareness, organize adequate support and provide time to experiment with open educational resources.

Figure 8. Profile of Qualitative studies

58 https://en.wikipedia.org/wiki/Qualitative_research
As Figure 8 illustrates, qualitative studies can potentially lead to strong conceptualization, but generally score low on most other criteria for scientific power of proof. It also should be noted that too often the label ‘qualitative research’ is used to mask the lack of rigor in research design. On the other hand, qualitative research potentially can lead to very relevant findings, both with regard to reducing epistemic uncertainty on topics and issues where quantitative research fails to produce evidence, and with regard to the depth of understanding. Qualitative research is an important and indispensable part of the research landscape in education and should be recognized in its own right. When done well, it is certainly not inferior to quantitative or experimental research and it fills the knowledge deficit in education in areas where experimental research is not able or willing to go.

Case studies

A case study is an in-depth, detailed examination of a particular case (or cases) within a real-world context. A case study does not necessarily have to be one observation (N=1), but may include many observations (one or multiple individuals and entities across multiple time periods, all within the same case study). Case studies are often used in medicine, social care, business studies, but also as a pedagogical tool in various education programs. A case study is an appropriate research design when you want to gain concrete, contextual, in-depth knowledge about a specific real-world subject. It allows you to explore the key characteristics, meanings, and implications of the case. In most instances, case studies are a form of qualitative research, but some descriptive statistics can also be used.

---

60 https://en.wikipedia.org/wiki/Case_study
There are plenty of examples of good case studies in educational research. An example is a case study of how school leaders and educators view the integration of immigrant and refugee children in schools.61 Through semi-structured interviews, the research collects a good insight into how school leaders and practitioners see the challenges and is able to point to gaps and distortions in their views.

Figure 9 presents the profile of case studies over the eight criteria. Case studies show marked progress in scientific power of proof over descriptive and qualitative studies, by their potential to lead to conceptualization, by their capacity to lead to or to be led by testable hypotheses, and by their (still low) degree of replicability and potential generalization. As qualitative studies, they can – of done well – lead to important insights into certain phenomena, leading to less epistemic uncertainty. Incidental and anecdotal knowledge plays an important role in education. Every teacher knows his or her preferred anecdotes about what works and what doesn’t. Despite the methodological limitations, such knowledge “sticks” and plays an important role in professional experiential knowledge.

Surveys and cross-sectional studies

Surveys are a good way to collect data across large populations on characteristics, preferences, opinions or beliefs of people. In education, international large-scale assessment (ILSA) surveys, like PISA, TIMSS, PIRLS or PIAAC have become very popular. These surveys include assessments administered to large populations, either at school or at home. Surveys are mostly administered to a sample of the population. Various sampling methods are available, but representative sampling is most-widely used. Census-based surveys are sometimes administered for administrative data-collections and offer the opportunity to have data of entire populations.

Surveys lend themselves to cross-sectional analysis, where relationships between variables at a single point in time are analyzed. Cross-sectional analysis allows for relationships or correlations between variables, but do not allow to draw conclusions on causal inferences. Still, quite sophisticated analyses are possible with cross-sectional data collections.

Cross-sectional analyses of survey data are particularly important in cross-national research, such as most reports on PISA data illustrate. A good example is a study of the relationship between family background, academic achievement in high school and access to high-status post-secondary institutions in three countries.62 Cross-sectional analysis allows to control variables in order to see what happens with the relationship between other variables.

Cross-sectional analysis of survey data is an important step towards better scientific proof, compared with qualitative research or case studies. As Figure 10 shows, research is highly conceptualizable, often driven by theory and clear hypotheses which are testable. Relationships between variables are to some degree modellable and the research is replicable with other survey data on other samples or populations. Findings are to some extent generalizable, if survey analyses are sufficiently robust. Such analyses also produce highly relevant findings with


sufficient depth and breadth. A lot of the evidence in contemporary educational policy and practice is based on cross-sectional analyses. Such analyses have contributed immensely to deeper understanding of, for example, the relationships between background variables and educational achievement and have strengthened the call for more equitable policies. The main handicap is that such analyses fall short in providing conclusive answers to questions about causality, which limits their power to inform policy interventions. Unfortunately, it is not rare to see cross-sectional studies where conclusions are formulated as if they were causal relationships.

Figure 10. Profile of Surveys and cross-sectional studies

**Quasi-experimental and longitudinal studies**

In longitudinal studies, researchers repeatedly collect data from the same individuals to detect any changes that might occur over a period of time. In contrast to cross-sectional survey data, longitudinal data collections can allow to make inferences about causal relationships between variables. In a quasi-experimental research design independent variables are manipulated without subjects being randomly assigned, in order to detect cause-and-effect relationships. Quasi-experimental research design studies are becoming increasingly frequent in educational research.63

A good example of a longitudinal study is the use of longitudinally linked PISA and administrative data to study the impact of inquiry-based teaching practices on students’ learning

---

outcomes. An example of a quasi-experimental research design is a study of the impact of educational tracking on interpersonal trust. The study used policy reforms as the manipulated independent variable to observe changes in the dependent variables.

As Figure 11 illustrates, the scientific power of proof of longitudinal and quasi-experimental research design again increases, compared to survey research. Hypotheses are now fully testable and to some extent modellable and even experimentable. Replicability and generalizability remain reasonably high. The relevance of such research is high. The main shortcoming of this kind of research is the scarcity of suitable data collections and the expensive nature of planning, designing and maintaining longitudinal data collections. But to the extent they are available, such research is potentially very powerful and relevant.

**Randomized controlled trials**

A randomized controlled trial (RCT) is a sophisticated experimental design where participants are randomly assigned to an experimental group and a control group, in order to control for as much variables as possible. RCTs are mainly used to study the effect of interventions. When

---


designed and implemented well, RCTs can control confounding variables and can lead to quite strong findings about treatment effects.66

A nice example is a study testing whether instructors’ use of different diversity approaches impacted the learning of new math and science content among students of color and white students.67 Students were randomly assigned to different treatment groups and control group. The study could demonstrate that the underlying diversity philosophy of teachers influenced the effectiveness of student learning among students of color. Another example is a study of how clicker-based technology in formative assessments among Dutch secondary school physics students reduce test anxiety and increase performance. Random assignment of students over control and treatment groups enabled the researchers to find causal relationships between the technological tools and the outcomes.68

---

**Figure 12. Profile of Randomized Controlled Trials (RCTs)**

Figure 12 shows that the scientific power of proof of RCTs is high. If developed well, hypotheses are conceptualizable, testable, modellable, experimentable and replicable. Findings are strong and can lend themselves easily to generalization. However, on the criteria of relevance, RCTs score much lower. Because of the requirements of the design, they often focus

---

on quite narrow problems. Breadth and depth of RCT studies are generally quite limited, so that they reduce epistemic uncertainty in the education sector only marginally.

However, as we have seen, RCTs are quite prominent in educational research and a lot of progress has been made over the years. Their growth is partly a result of policies of research funding agencies who see them as the gold standard of educational research. Researchers are strongly incentivized to develop RCTs to study the impact of specific educational interventions. The debate is far from closed, but questions can be asked about whether the added value of RCTs to increase and expand the evidence base in education is high enough to warrant their priority in research policies. The replicability of RCTs is high in theory, but in practice often limited.

Meta-analyses and reviews of experimental research

As more and more experimental design studies such as RCTs in education research are becoming available, also review studies are becoming more frequent. Reviews of experimental studies and meta-analyses, for example of effect sizes, are very useful in synthesizing the research literature and in evaluating the strength of the research evidence. Reviews often quantitatively synthesize the evidence using a meta-analysis, which is a statistical analysis tool to synthesize results from multiple studies. It’s a statistical analysis that combines the results of two or more studies, usually to estimate an effect size.

There are many excellent examples of review studies in education research. For example, a review study examined the research literature on the effectiveness of direct instruction curricula to find that all expected educational effects were positive with good effect sizes. Another study reviewed the research evidence on early non-cognitive skills development on educational and life outcomes and found that the quality of most research designs was too low to derive strong conclusions.

Meta-analyses of experimental studies are very powerful from a scientific point of view. That’s why, in Figure 13, they score very high on all criteria of power of proof. Hypotheses are conceptualizable, testable, modelable, experimentable, and replicable. By integrating all available research such studies arrive at generalizable conclusions. However, on criteria of relevance, such studies are much weaker. Many meta-analyses point to weaknesses in the research literature and come to conclusions which are often sobering. For example, they conclude that effect sizes are estimated too high or that there is conflicting evidence. From a scientific point of view, these are very valuable conclusions, but they decrease the power to reduce the epistemic uncertainty in the education system. Because such studies are often rather

---


narrow in scope, limited to a well-defined research question that is analyzed in a systematic way, they also lack the breadth needed to inform educational policy and practice.

![Figure 13. Profile of Meta-analyses and reviews of experimental studies](image)

**Research synthesis papers**

Compared with meta-analyses and review studies, research synthesis papers have somehow more modest scientific ambitions. They mainly aspire to synthesize the available research literature. They do not necessarily aim to critically review and evaluate methods used. Still, a good research synthesis allows to test and validate hypotheses, understand key processes, and better design future research efforts. They are not only invaluable for the research community, but also serve to translate and transmit the available evidence to a wider audience in the education system. Good research synthesis papers, which are more than just summaries of literature, are rather rare in educational research. Apart from the obligatory literature review in papers and theses, few researchers devote time and energy to produce good synthesis papers. Research funding agencies are not sufficiently rewarding such papers. That’s the reason why research synthesis papers are often produced by intermediary organizations, brokerage institutions or sectoral agencies.
A good example is this synthesis of gender inequities in STEM in K-12 education in the US.\textsuperscript{73} Or the review of all available evidence on vocational education and training for secondary students in Australia.\textsuperscript{74}

Figure 14 shows that on criteria of scientific power of proof research synthesis papers often do not score as well as rigorous review studies and meta-analyses. Hypotheses are not always formulated in a systematic way that can lead to testable, modellable, and experimentable hypotheses. On the other hand, synthesis papers add enormously to criteria of relevance. They combine depth and breadth and lead to relevant insights that reduce epistemic uncertainty in the sector.

![Figure 14. Profile of Research synthesis papers](image)

**Scholarly books**

The last research output we will discuss is the scholarly book. Books written by scientific scholars often take a wide perspective and discuss the scientific evidence in relation to broad social and policy issues. They serve both to synthesize the evidence, but also to put the evidence into a specific orientation. They present research, analyze trends, and otherwise communicate with their peers in the same field, as well as with practitioners and the general public. As such,


books provide an invaluable contribution to the landscape of evidence and often stimulate debate.

Good examples of scholarly books in education really abound. Take for example this book by Audrey Watters on the history of personalized learning.⁷⁵ Or Stan Dehaene’s seminal book on learning science.⁷⁶

As Figure 15 demonstrates, the scientific power of proof of scholarly books is quite modest. Although most scholarly books written by scientists and researchers rely on abundant scientific research, the books themselves do not aspire to contribute to the research base. Their power is in using the research evidence to make a relevant contribution to the field, not only the scientific field but also the field of policy, practice, and the wider public debate. Such books help enormously to reduce epistemic uncertainty and they often have an excellent combination of breadth and depth.

![Figure 15. Profile of Scholarly books](image)

**Conclusions**

Education research has grown significantly over the past decades, both in quantity and quality. Yet, it still fails to provide the quantity and quality of evidence expected by the education system. The most immediate challenge for education research simply is to accelerate its quantitative growth and quality. As education systems expand and social expectations surge, teachers, school leaders and policymakers are confronted with ever more diverse and difficult

---


challenges which they cannot solve without more abundant and better research evidence. To meet the knowledge needs of education systems, higher quantity and better quality of education research is desperately needed.

Despite the growth of educational research and expanding intermediate systems of knowledge brokers and various kinds of institutions active with translating and transmitting research evidence into the system, the current situation is one of huge knowledge deficits and epistemic uncertainties. Education practitioners solve this deficit by tapping into alternative knowledge systems such as teachers’ professional and experiential knowledge. Although barely grounded on solid research evidence, the experiential nature of professional knowledge provides guidance to practitioners when research evidence is lacking. But this situation should not endure.

Policies trying to strengthen educational research have mostly aimed to prudently expand its quantity while arguing for more rigor. The quality issues in educational research, which are very real and pressing, are considered to be tackled by insisting on more rigorous research methods. In the US and other countries, this argument has taken the shape of a rather strong focus on randomized controlled trials as the gold standard of research at the top of the pyramid of evidence. A hierarchical view on quality of educational research has been promoted and widely adopted by policy makers and research funding agencies.

It certainly is the case that much educational research lacks sufficient rigor in design and methodology. But it would be a mistake to think that more rigor automatically means more research at higher levels of the pyramid of evidence, such as RCTs and experimental research. This approach risks to fall in the trap of a trade-off between quality and relevance. RCTs are often rather narrow in scope, precisely because of the methodological requirements. While they provide powerful scientific proof on specific issues, they are simply too limited to significantly address the knowledge deficit in education systems.

We need to move to a more pluralist perspective, where power of proof and relevance are seen as equally important dimensions of research quality. A pluralist approach, rather than a hierarchical, would value both criteria of scientific power of proof and criteria of relevance. Some research approaches and outputs produce very valuable knowledge for teachers, school leaders and policymakers, with maybe lower power of proof, but with better depth, breadth, and capacity to reduce epistemic uncertainty. A pluralist approach to educational research should not be seen as an argument against rigor and quality of research design and methodology. It is an argument to constantly evaluate the two dimensions of power of proof and relevance when making decisions about research.

In this perspective, education desperately needs more research outputs which score high on criteria of relevance such research synthesis papers and scholarly books. It also needs more research that combines high – if not top – scientific power of proof with high relevance, such as surveys, cross-sectional analysis, longitudinal studies, and quasi-experimental design studies. Such outputs contribute better to solve the knowledge needs of education systems than for example experimental studies and RCTs, which are very costly and difficult to implement.

A change in course of educational research is needed.