WORKING
GROUP 04

CHAPTER

1

The EBE3 framework: extending evidence based education from causal ascriptions and effectiveness generalizations to relative effectiveness generalizations and local effectiveness predictions

DOI: https://doi.org/10.56383/QHEN7211

This chapter should be cited as:

Mercier, J. and Bourgault Bouthillier, I. (2022). 'The EBE3 framework: E] extending evidence based education from causal ascriptions and effectiveness generalizations to relative effectiveness generalizations and local effectiveness predictions' in Duraiappah, A.K., van Atteveldt, N.M., Borst, G., Bugden, S., Ergas, O., Gilead, T., Gupta, L., Mercier J., Pugh, K., Singh, N.C. and Vickers, E.A. (eds.) Reimagining Education: The International Science and Evidence Based Assessment. New Delhi: UNESCO MGIEP.

To more reliably achieve educational goals based on values and policies, quantitative and qualitative traditions should complement each other to strengthen the quality and impact of empirical research, under a broad banner of evidence based education (EBE). A different approach to EBE can help to solve questions relating to 'what works', by extending this question to 'what is working best generally' and 'will a given intervention work here and now?'. This chapter proposes a more complete framework for EBE by delineating the information and reasoning needed to address a cascade of questions that jointly determine the best course of action for obtaining the best educational outcomes. The traditional levels of evidence are revised and complemented by a proposal for levels of contextual fitting, grounded in both theory building and theory testing. Implications for conducting future applied research, for policy-making and for improving educational practice are discussed.

On average, the temperature in Alaska is above freezing, but if I am planning a trip and hope to avoid the snow I must figure out when it is above freezing and when it is below. Similarly, the greater the impact varies among sites and students, the less we learn from an average treatment effect, even if it is accurate for the broad population (Joyce, 2019).

Coordinating Lead Authors

Julien Mercier

Lead Author

Iris Bourgault Bouthillier

WORKING GROUP 04

CHAPTER 1

1.1 Introduction: 'what works' is not enough

The goals of education are based on values and policies **(Brighouse et al., 2018)**. This public policy-making is a political process that requires conflict, negotiation, the use of power, bargaining and compromise **(Anderson, 2011)**. When it comes to the means to achieve those goals, the relative benefits of some approaches over others are assessed through empirical research, where quantitative and qualitative traditions have complementary roles **(National Research Council, 2002; Karrigan and Turner-Johnson, 2019)**. This research, which is usually conducted with samples of learners, involves, among other things, the observation of gains on target outcomes and processes. Since human learning and

From the perspective of evidence based education, decisions about which practices to use in a given learning context should ideally be based on evidence.

development are the cornerstones of educational goals (albeit reformulated through reforms), the domains contributing to educational research rest essentially on a vast number of fields in the learning sciences and cognitive science (psychology and neuroscience (behaviour and brain processes), computer science (computer based learning systems, learning analytics), and economics and social sciences (the learner and their broader context).

1.1 .1

PROBLEM: THERE IS A NEED TO APPLY A HIGH MINIMUM STANDARD FOR WHAT COUNTS AS EVIDENCE OF IMPROVED LEARNING

Pertinent research relies on a variety of methods, which, in essence, focus on different aspects of theory building and validation. From the perspective of evidence-based education (EBE), decisions about which practices to use in a given learning context should ideally be based on evidence **(Slavin, 2020)**. Evidence starts with a demonstration of the effect of some treatment on a defined outcome **(Connolly, Keenan and Urbanska, 2018)** and, more broadly, of empirical support that a policy works generally or in a specific context **(Joyce and Cartwright, 2020)**. This essential foundation means that we should expect a higher standard: that is, to know whether an intervention works better than what we were already doing, compared to a control group and after eliminating as many possible sources of bias. To know this, a level of confidence in the inferences made from the empirical investigations need to be considered. The study of 'what works' is limited to causal ascriptions, that is, the estimated causal effect of an intervention on the targeted outcomes. These arise from a comparison of an experimental group with a control group. Causal ascriptions,

1

CHAPTER

In a logic of cumulative generalization and abstraction of claims of effectiveness, the best evidence is available when every possible intervention for a specific goal and target population - including the context of that population - has been tested with equally valid studies (ideally replicated) and then rank-ordered with respect to its established effect.

combined with assumptions about generalizability and replicated across a few studies, lead to general effectiveness claims. These are the inferences that results obtained with samples will apply to the corresponding population(s) and context(s). Thus, the demonstration simply indicates that a given intervention is better than the normal practice (which has proven difficult to define) **(see Kornell, Rabelo and Klein (2012) for an example)**. Even state-of-the-art experiments carried out at the cluster level (e.g. forty to fifty schools or classrooms) as advocated by Slavin **(2020)**, use designs limited to testing causal ascriptions, just like traditional comparisons between experimental and control groups. These complex experiments often use hierarchical linear modelling to take into account the similarity of the participants within a school or classroom. This only improves what Shadish, Cook and Campbell **(2002)** call statistical conclusion validity (by getting the standard errors right) but not internal validity (the potential to establish the unbiased effect of an intervention) or external validity (notably the potential for generalization). The results of a collection of high-quality studies (unbiased sampling, randomized treatment/group assignment, well-defined intervention, valid and reliable measures, statistical analyses with power, effect size and significance tests) comparing an experimental group given a target intervention with a control group has been the cornerstone of EBE for decades under the label 'what works'. It is the main, but not sufficient, building block of EBE, because such studies provide relatively isolated indications of the effectiveness of interventions, which remain to be further compared and rank-ordered empirically. Thus, there is a need for a higher minimum standard for what counts as evidence of improved learning. This chapter proposes an evolution of previous efforts and capitalizes on the EBE building block 'what works' to develop further rationales for establishing the efficacy of interventions.

In a majority of cases, the evidence is scattered, emerging and incomplete, or based on a multiplicity of research designs, methods and conceptual frameworks.

In a logic of cumulative generalization and abstraction of claims of effectiveness, the best evidence is available when every possible intervention for a specific goal and target population – including the context of that population – has been tested with equally valid studies (ideally replicated) and then rank-ordered with respect to its established effect. In such cases, choosing the best intervention and which to try first, second or third in terms of specific outcomes is straightforward, at least in terms of efficacy **(Goldacre, 2013)**. Unfortunately, educational issues tested this way are scarce but have been increasing during the last decade **(Connolly, Keenan and Urbanska, 2018)**. In a majority of cases, the evidence is scattered, emerging and incomplete, or based on a multiplicity of research designs, methods and conceptual frameworks. The common denominator is the level of trust in the inferences made from empirical investigations. It is important to consider that when alternatives exist, effectiveness of available interventions is always relative to the effectiveness of some other intervention(s). We call these inferences 'general relative effectiveness claims', because they stem directly from the comparison of effectiveness generalizations.

1.1 .2

PROBLEM: BEFORE THE NEED FOR ADDITIONAL EVIDENCE IN THE FORM OF NEW TESTS OF INTERVENTIONS, THERE IS A NEED FOR 'RELATIVE EVIDENCE'

We define relative evidence as the result of thorough comparisons of extant interventions, under the assumption **(see the Australian Society for Evidence Based Teaching)** that combined results coming from meta-analyses or systematic reviews are much more

Relative evidence arises from the combined results of multiple studies, using meta-analysis and made possible by thorough comparisons of effect sizes of multiple extant interventions.

informative than single – albeit excellent – studies when necessary precautions are taken **(Simpson, 2018)**. These necessary precautions consider that effect size (the indication of the impact of a given intervention) is due not only to the intervention, but also may be part of the whole study (e.g. sample size, test characteristics and comparison treatment). Relative evidence arises from the combined results of multiple studies, using meta-analysis and made possible by thorough comparisons of effect sizes of multiple extant interventions. The consistency or variability of effect sizes across studies of similar interventions is critical to support assertions regarding their general effectiveness. In addition, the consistency of effect sizes across studies is critical to empirically support assertions about what we have termed relative effectiveness generalizations, that is, claims that the relative effectiveness of interventions, tested with samples, will apply to the corresponding populations and contexts. However, there is a lack of relative evidence in extant literature regarding most educational issues: new interventions are tested against a control group (business as usual) and well-documented interventions rarely get rank-ordered through a proper meta-analytic approach.

Aside from scientific challenges, the lack of relative evidence may unfortunately be explained, at least in part, by policies governing research. Indeed, the neoliberal model underlying the funding of research and educational institutions 'has forced academic researchers to dismiss methodological limitations of social science research … and overestimate the impact of their research in order to obtain highly competitive, and scarce, research money … fueling a replication controversy in published research' **(Karrigan and Turner-Johnson, 2019, p. 290)**. Moreover, Chubb and Watermeyer **(2017)** synthesize a drift from traditional and still desirable norms in academia including communism, universalism, disinterestedness and organized scepticism; and the defence of critical, objective

truth. This drift pulls academics toward professional pragmatism and sponsorism as a survival response in the face of demands and directives of academic capitalism and 'managerial' governmentality, seen as hegemonic and inescapable. In the end, these forces rewarding short-term and shallow productivity do not encourage the undertaking of thorough synthesis work.

An effectiveness prediction is the prediction that a given intervention, abstracted through causal ascriptions, effectiveness claims and relative effectiveness generalizations will work concretely within the specific constellation of variables of a given application context.

1.1 .3

PROBLEM: IF RELATIVE EVIDENCE IS AVAILABLE AND GENERAL RELATIVE EFFECTIVENESS CLAIMS ARE SUPPORTED BY APPROPRIATE EVIDENCE, THERE IS A NEED FOR STRONG ASSERTIONS ABOUT HOW THE LOCAL CONTEXT IN WHICH THE EVIDENCE IS TO BE APPLIED OUGHT TO AFFECT OUR EXPECTATIONS OF IMPACT

An effectiveness prediction **(Joyce and Cartwright, 2020)** is the prediction that a given intervention, abstracted through causal ascriptions, effectiveness claims and relative effectiveness generalizations will work concretely within the specific constellation of variables of a given application context. Such comparisons can enable assessment of the effectiveness against specific outcomes of all pertinent interventions, allowing practitioners to answer the question: given all the possible interventions available to me, which is most likely to succeed in my specific context? Ultimately, this context concerns a specific

teacher and a specific classroom at a specific moment in time, this specificity being the opposite of the potential for generalization sought by quantitative research. In other words, evidence is needed to support the prediction that a given intervention abstracted across causal ascriptions and general effectiveness claims will work concretely within the specific constellation of variables of a given context of application. These assertions are what Joyce and Cartwright **(2020)** have termed 'local effectiveness predictions'.

These local effectiveness predictions have proven elusive in the traditional view of EBE. Reasoning about how causal claims related to a given intervention will yield documented outcomes in a target concrete and specific context (a given school for example), evidenced by the right information, has not been clear or available. Consequently, EBE at this step has consisted of merely applying research-based practices, that is, causal ascriptions and general effectiveness claims.

This applicationist stance is accompanied by concerns about teacher training, teachers as technicians rather than professionals, educational leadership, accountability and scaling up of interventions. Local effectiveness predictions are generally either absent from implementation efforts, or tackled through biased, non-scientific reasoning, such as beliefs, peer pressure, marketing, and so on. It would be possible, in education, to be a lot more efficient in implementing best practices by applying a rationale increasingly used in other fields **(Pawson et al., 2005; Pawson, 2006)** that explicitly concerns how contextual elements facilitate the release of active ingredients in interventions documented as the most effective.

Local effectiveness predictions are generally either absent from implementation efforts, or tackled through biased, non-scientific reasoning, such as beliefs, peer pressure, marketing, and so on.

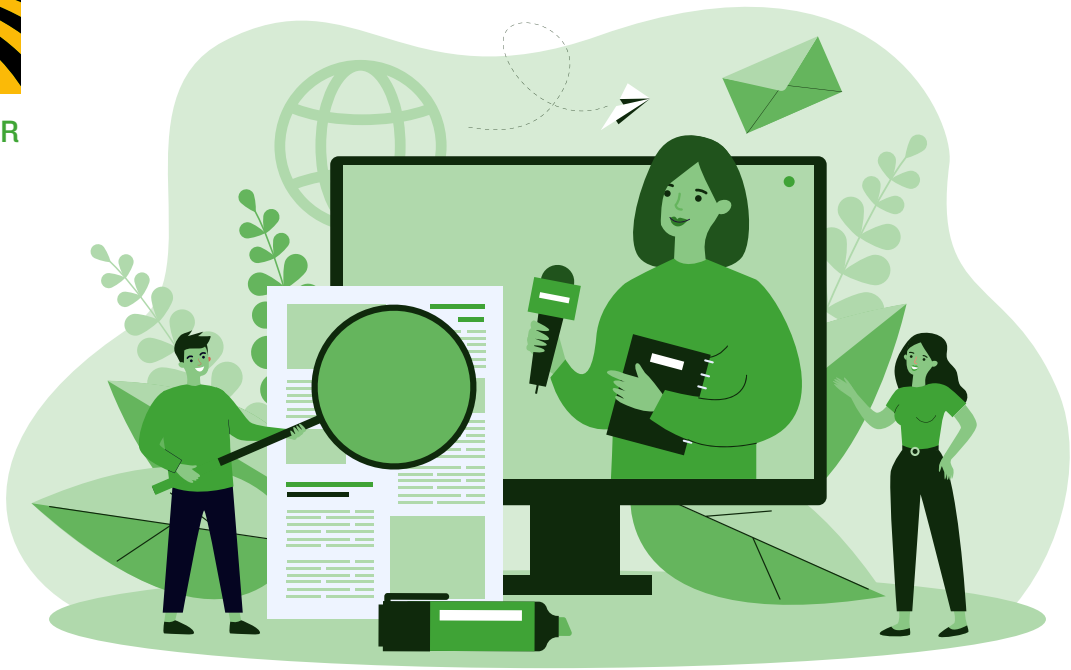
What constitutes a fully operational EBE has not yet been framed as a coherent cascade of questions related to the decision-making involved in implementing the best interventions and driving the production/consideration of the necessary information. Nor have these questions been

This chapter aims to provide an overview of the nature of scientific evidence in education and to suggest a framework that encompasses all current types of efforts related to the development of educational knowledge, and posits the overall progress of educational research as a compromise between theory building and validation.

operationalized in terms of required evidence paired with the necessary empirical work. This leaves the vast majority of educational research, synthesis work and application endeavours subject to gaps that need to be satisfactorily resolved in a specific sequence. Globally, in line with Joyce and Cartwright, **(2020)**, we are concerned with the information and reasoning needed to address a cascade of questions that jointly determine the best course of action for obtaining the best educational outcomes: what works? What is working best generally? Will it work here (tomorrow, in my classroom)? This chapter aims to provide an overview of the nature of scientific evidence in education and to suggest a framework that, firstly, encompasses all current types of efforts related to the development of educational knowledge, and, secondly, posits the overall progress of educational research as a compromise between theory building and validation. It is expected that this integrated framework is both practical and useful for stakeholders (researchers, policy-makers and practitioners) in educational systems. Hence, the first section of this chapter discusses the importance of theory building and theory testing in educational research. The second section discusses the levels of evidence, their usefulness and their limits. The third section presents an original framework aiming at the application of evidence in specific contexts, which to date has been underspecified. Finally, the usefulness of this new framework for stakeholders is discussed. An appendix outlines a procedure for obtaining the necessary information and making the necessary inferences from it to answer key questions in a process of EBE: after determining the most important educational goals, identifying the means to attain these goals by using or fostering necessary results from pertinent empirical work. The application of this procedure can ultimately be used as a practical tool for conducting literature reviews and implementation work as well as policy-making.

WORKING GROUP 04

CHAPTER 1



1.2 Theory building and theory testing in educational research: divide, compromise or synergy?

Besides the emphasis on empirical developments in EBE, another essential aspect of educational research is the development of theory. The National Research Council (NRC) **(2002)** briefly defines theory as follows: scientific theories are conceptual models used to explain phenomena. In the social sciences and humanities

Besides the emphasis on empirical developments in EBE, another essential aspect of educational research is the development of theory.

(including education) the nature of theories has been largely discussed. The NRC recognizes a continuum between 'grand' theories, that aim at generalizing theoretical understanding, and research that seeks to achieve deep understanding of particular events or circumstances. In between these two extremes are mid-range theories attempting to account for social aspects and particular elements of situations. All theories, wherever they are located on this continuum, consist of representations or abstractions of some aspect of reality that can only be approximated by such models. We place limited emphasis on the 'grand' theories that aim at generalizing theoretical understanding and focus on mid-range theories attempting to account for social aspects and particularities of situations. Mid-range theories consist of representations or abstractions of aspects of reality that can be approximated by conceptual models, which can be subjected to empirical tests. According to Maciver et al. **(2019 pp. 13 14)**:

The term "middle range" theory refers to the level of abstraction at which useful theory for realist work is written: detailed enough and "close enough to the data" that testable hypotheses can be derived from it, but abstracted enough to apply to other situations as well ... Middle range theorization is useful because it offers an analytical approach to linking findings from different situations.

According to the NRC, one of the main principles of scientific inquiry is to link empirical research to relevant theory. Empirical research can be linked in many ways to theory. Depending on the underlying epistemology and the advancement of knowledge in the field, theory can either be what guides a study or what emerges from it. In many cases, theory can be linked to research in both ways when a study is based on theory and at the same time enriches it. In short, theory is what 'drives the research question, the use of methods, and the interpretation of results' **(National Research**

Council, 2002). Thus, theory has an undeniable importance in applied science.

In the learning sciences, theory is notably what allows researchers, decision-makers and practitioners to support the use of interventions in specific contexts and understand the underlying mechanisms **(Joyce, 2019)**. When reviewing the scientific literature about a topic, stakeholders in education should therefore be able to determine the contribution of a study or group of studies to the advancement of theory. Research can contribute to theory in two main ways: theory building and theory testing (validation). These two types of contribution are not mutually exclusive. Research has shown in some fields that the more an article contributes to theory in one or both ways, the more it will be cited **(Colquitt and Zapata-Phelan, 2007)**. While the citation rate is not the only way to measure the importance of a scientific publication **(Sugimoto and Larivière, 2018)**, it can be considered a general indicator of the impact of research. Hence, in some fields, the more an article is contributing to theory, whether by building it, testing it, or both, the more impactful this piece of research tends to be for the scientific community, as reflected by its citation rate.

The next paragraphs describe a taxonomy created by Colquitt and Zapata-Phelan **(2007)** that can be used to capture many facets of the theoretical contributions of an empirical study. Although their article is focused on the field of management, it can easily be transferred to the field of education, given these fields share many similarities. For example, they are both social sciences based on values and policies and the research methods and nature of theories used in both fields are mostly the same. The taxonomy is built on two orthogonal axes, theory building and theory testing, which are both divided into five ordinal levels. A given empirical study is situated on both axes. Qualifying a corpus of studies in a given field this way may help assess the maturity of the research on a given educational issue

In the learning sciences, theory is notably what allows researchers, decision-makers and practitioners to support the use of interventions in specific contexts and understand the underlying mechanisms

and may help in extracting the information needed to address the main questions of the framework proposed.

One axis presents five levels of theory building. The first two levels of theory building on the axis are considered low-level contributions. The first level represents attempts to replicate results that already support existing theories. Replication studies are very important to science because they offer substantial protection to the quality and credibility of empirical scientific work; specifically, issues linked to false positives results, null results and questionable research practices **(Frias-Navarro et al., 2020)**. Despite their importance, they are considered the lowest level in terms of contributing to building new theories. Level 2 attempts to examine effects that have already been the subject of prior theorization. Level 3 includes studies that introduce new variables (e.g. mediators or moderators) to existing theories on relationships or processes. Level 4 studies explore new relationships or processes. Finally, level 5 includes studies that propose entirely new theories, models or concepts, or that significantly reconceptualize existing ones.

The other axis illustrates five levels of theory testing. Studies from the first level are either inductive or ground their predictions within logical speculation. In this level, one may find exploratory studies that are not necessarily based on prior theory or concepts. Level two studies ground their predictions with references to past findings. This means that the results are put in relation to other findings but are not explicitly based on prior theory or concepts. Level three includes studies that ground their predictions with existing conceptual arguments, while level four studies' predictions are grounded within existing models, diagrams or figures. Finally, level five studies explicitly ground their predictions on existing theory.

The interaction between the two axes enables us to distinguish five discrete article types in terms of

their theoretical contribution: the reporters, the testers, the qualifiers, the builders, and the expanders. For specific examples of articles that fit into each of these categories, see the article by Colquitt and Zapata-Phelan **(2007)**.

The reporters category includes empirical articles that score low on both axes. For example, an article that aims at replicating a previous study **(level 1 of theory building)** with hypotheses based on findings of several prior other studies on the topic **(level 2 of theory testing)** would be classified in this category. Even when studies are considered to be low on both axes, it is important to stress that they can still be constructive and useful for science. Testers includes articles that show high levels of theory testing and low levels of theory building. This category includes articles that aim primarily at testing existing theories empirically without incorporating new constructs or variables. The qualifiers category is composed of articles that contain moderate levels on both axes. They can be articles that push previously demonstrated relationships a little further. For example, articles in this category can be based on previously demonstrated

relationships between concepts and try to add a new mediator to qualify this relationship. Builders are articles that score high on the theory building axis and low on the theory testing axis. This category includes, amongst others, inductive studies that elaborate new constructs, relationships or processes. Finally, the expanders are articles that are high on both theory testing and theory building axes. Like builders, they focus on new constructs, relationships and processes that have not already been theorized, but they do it while also testing existing theory.

While the taxonomy of theoretical contributions for empirical articles that allows classification of articles according to their level of theory building and theory testing contribution can be very informative, it only depicts empirical studies intended theoretical development, not how well it is done

While the taxonomy of theoretical contributions for empirical articles that allows classification of articles according to their level of theory building and theory testing contribution can be very informative, it only depicts empirical studies intended theoretical development, not how well it is done **(Colquitt and Zapata-Phelan, 2007)**. As the authors themselves argue, many other important underlying factors could be added to their taxonomy: how interesting is a new construct, how much a new relationship adds to the relevant literature, how rigorously a theory is tested, and so on. This taxonomy conveys a profound message: theory is at the heart of the advancement of science and the value of empirical observations is contingent on their contribution to theory building and theory testing. As will be discussed in the next sections, theory is central to progress in the hierarchies of the EBE3 framework. To answer the question of what is working best generally, theory defines and isolates the active ingredients in interventions. This is critical for classification of interventions in meta-analytic work so that the comparisons are warranted and interpretable. To answer the question about replicating the efficacy of a given intervention in a specific context, pertinent theory defines experimental and observational elements to take into account and mechanisms and processes not to take into account for the purposes of predicting efficacy **(Pearl and Bareinboim, 2014)**.

WORKING GROUP 04

CHAPTER 1

1.3 What is working best generally: levels of evidence

Along with Joyce **(2019)**, we consider causal ascriptions, on which the so-called 'what works' approach hinges, to be extremely limited in informing the implementation of interventions in EBE. Consequently, we begin our discussion of the necessary ingredients of an empirical demonstration of effectiveness with the notion of general effectiveness claims. General

In light of the cumulative nature of empirical evidence, the levels of evidence are operationalized domain by domain, from a gradation of internal and external validity of the available evidence.

effectiveness claims build upon causal ascriptions and consist of a further empirical demonstration of: (1) the relative effectiveness of available intervention; and (2) the variations in effect across studies, contexts and populations. This empirical demonstration requires a meta-analytic approach, conducted with state-of-the-art procedures to avoid common, published mistakes **(Borenstein, 2019)**.

Insofar as applied research improves professional practices in education, and given the impact of these practices on learners, it seems desirable to be able to judge the relative value of available research results relevant to practice, following a set of considerations pioneered by Cochrane **(1972)**. For each aspect of the role of the teacher or professional, it must be possible to determine either an absence of research, the presence of poor-quality research, the presence of quality research and possibly the accumulation of relevant and converging research. From an interventionist perspective that follows a basic premise, namely that the best information for practice is of an applied and causal nature **(Joyce, 2019)**, it is necessary to formulate unambiguous inferences between an intervention and its effect on the learner. In this regard, consensual criteria on which these causal inferences can be established, taken up across a majority of applied fields emanating from the human sciences, are brought together through the notion of levels of evidence.

In light of the cumulative nature of empirical evidence, the levels of evidence are operationalized domain by domain, from a gradation of internal and external validity of the available evidence. Also, considering a standard benchmark of effectiveness, the most common being effect size, is essential in merging evidence about relative effectiveness across increasingly broad educational areas of intervention in order to prioritize intervention in these areas.

CHAPTER 1

In areas related to learning, different types of research questions are needed to design and document the effectiveness of practices empirically. These types of questions are accompanied by different methods: manipulation of experimental groups, correlational studies, single-case designs and qualitative methods. Several authors have suggested hierarchies allowing classification of scientific evidence according to the level of confidence that can be attributed to the inferences drawn from them (see the literature review on the subject) **(Nutley, Powell and Davies, 2013)**. Most of these classifications are generally similar to one another in content. In Table 1, we propose such a classification of the pseudoscientific and scientific evidence applied to educational research. This proposal of criteria for the efficacy of intervention seeks to extend prevalent hierarchies of evidence to encompass the various types of evidence created and disseminated, including inadequate, pseudoscientific evidence, **(e.g. Evans, 2003; Burns, Rohrich and Chung, 2011)**. It allows the distinction of: (1) information of pseudoscientific or non-scientific nature; 2) the results emanating from a scientific approach; and 3) probative evidence concerning the relative convergence and divergence of the integrality of available research results. The terms probative, scientific and pseudo-scientific/non-scientific are used for clarity in relationship with the common language of researchers, practitioners and policy-makers in education. They are used to provide clear benchmarks to classify sources of evidence and should not be seen as exclusive or unrelated. Hansson **(2009)** defines a pseudoscientific assertion using three criteria: (1) it pertains to an issue within the domains of science (in the wide sense); (2) it is not epistemically warranted; (3) it is part of a doctrine creating the impression that it is epistemically warranted. Scientific, in the context of applied educational research, is meant to provide limited empirical indications about the efficacy of a given intervention. Probative is understood as the ability of

The terms probative, scientific and pseudo-scientific/non-scientific are used for clarity in relationship with the common language of researchers, practitioners and policy-makers in education

evidence to make an assertion true, in this case the assertion pertaining to 'effectiveness'. The pseudo-scientific category comes from belief, biased observation, and so on. The scientific category, on the other hand, comes from rigorous research answering valid research questions. The probative nature of research results refers to the best level of confidence that can be placed in the results of scientific studies aimed at establishing the effectiveness of interventions.

Each level of evidence is described, in descending order of potential to empirically answer the question of what is working best generally. Contrary to Goldacre's **(2013)** claim that students are 'similar enough that research can find out which interventions will work best overall' **(p. 7)**, it is essential to stress the importance of carefully analysing the circumstances of practice that we want to support scientifically **(Joyce, 2019)**. Thus, the learning object, the learner's particularities, as well as the context of intervention are among the elements to be considered to establish the correspondence between the educational act and the available scientific literature. Any discrepancy between the circumstances of 'real' practice and the circumstances of practice as studied in the scientific literature decreases the level of scientific evidence. It can be suggested that the 'real' practice circumstances prevail, and that this will establish the level of scientific evidence that applies, rather than implementing practices supported by the best scientific evidence that would prove unrelated to the current practical needs. Although this is tangential to this chapter, it should be noted that proper training and expertise of the educational professional are necessary for the analysis outlined above.

The only probative sources of evidence are grouped at level 1. Probative qualifies evidence that fully proves a given assertion about the relative effectiveness of interventions. Levels 2, 3, 4 5 and 6 constitute the scientific range because they support causal

TABLE 1 LEVELS OF EVIDENCE APPLIED TO EDUCATION RESEARCH TOWARD EFFECTIVENESS GENERALIZATIONS

RANGE	LEVEL	SOURCES OF EVIDENCE	MAIN LIMITATIONS
Probative: provide effectiveness generalizations	1	Mega-analysis, meta-analysis, narrative literature review, evidence-based review	Abstracted, decontextualized recommendations
Scientific: provide causal ascriptions	2	Experimental studies	Do not provide relative effectiveness generalizations
	3	Quasi-experimental studies	Internal validity
	4	Correlational studies, quantitative case studies	Impossible to verify causality
	5	Experts committees, clinical experience from experts (teamwork reports)	Opinions subject to political or personal influences
	6	Qualitative research, single case protocols	Lack of generalizability
Pseudo-scientific and non-scientific: beliefs not related to solid observation or reasoning	7	Bad quality research (qualitative or quantitative)	Improper methodology
	8	Absence of research, practice reports, trends	Lack of systematic empirical observations

inference, generalizability and replication to varying degrees. The pseudo/non-scientific range is included last, with levels 7 and 8 as red flags, because practitioners in education are frequently exposed to information pertaining to these levels. Levels 1 and 2 are discussed in more detail below, level 1 because although it represents the best sources of general effectiveness claims, it is not exempt from issues in improving educational intervention, and level 2 because it has been seen as the gold standard for EBE for decades despite significant strengths and limitations. Solutions to the limitations of level 1 are suggested later in this chapter.

Level 1 shows the relative effectiveness and variability in outcomes of all the interventions tested experimentally. Mega-

Any discrepancy between the circumstances of 'real' practice and the circumstances of practice as studied in the scientific literature decreases the level of scientific evidence.

analyses (the meta-analysis of meta-analyses, also called meta-meta-analysis) and meta-analyses are preferred because they provide relatively unbiased empirical results. Narrative literature reviews (a discussion of important topics on a theoretical point of view **(Jahan et al., 2016)** and evidence-based reviews (also called systematic reviews) also qualify as probative because they concern available interventions and their relative effectiveness, although it must be noted that they are much weaker than the meta-analytic approach; they are more subjective and may lack the sensitivity to extremes and combination of factors that is characteristic of meta-analyses. The major limitation of this level is that it provides abstracted, decontextualized recommendations. Indeed, the increasing level of aggregation of results needed for probative evidence implies a gradual dissociation with the contexts of the experiments. It is important to point out that evidence at this level is absolutely necessary to qualify research results as probative for any given issue, but the quality of evidence at this level depends on the quality of the primary studies in the scientific range, which get aggregated in the probative range. Also, the demonstration in this chapter that there is no substitute for properly aggregated results at the probative level indicates that interventions implemented should be properly documented at the probative level. If educational goals in policy-making involve means not documented at the probative level, then the implementation of these means in practice should be deferred until the necessary evidence is available. In fact, these goals should drive the production of this evidence.

Level 2 contains the best experimental evidence to support causal ascriptions and effectiveness generalizations. Experimental studies, the gold standard being randomized-controlled trials, allows adaptation of the design to specific target populations and the intervention context. As stated earlier, the more an experimental design is closely related to the

Because of the complexity of educational issues, a conservative position seems warranted and it appears that, all things considered, opinions remain weaker than scientific observations.

real context of practice, the more confidence one can have in the interpretations drawn from the evidence in their own context of practice. The main caveat, as Joyce **(2019)** describes, is the difficulty of determining which characteristics of the populations and the intervention contexts must be considered salient for educational decision-making. Selected characteristics are used as evidence of the representativeness of sampling without supporting their relevance for educational outcomes with evidence. Applying them indiscriminately will not help educators find studies that are appropriately representative and may even lead them astray.

Level 3 shows that quasi-experimental studies have documented the effect of an intervention. However, sampling and assigning to different conditions does not guarantee the equivalence of groups. Also, the internal validity is compromised and does not unequivocally link a difference between the groups with the effect of the intervention tested.

Level 4 indicates the presence of correlational studies or quantitative case studies that do not allow establishing the causality between an intervention and its effect. As intervention involves causal reasoning supported by indications showing that such intervention produces such results ('if I do this, then the student should progress'), studies that do not show a directional link explaining the learning gains contribute very little to the orientation of the interventions. Note that in cases where variables of interest cannot be manipulated, such as gender for example, correlational studies are entirely adequate or even decisive.

Level 5 refers to various reports, think tanks and recommendations from the judgement of expert researchers or clinicians on a predefined question presumably in the absence of higher-level scientific evidence. Because of the complexity of educational issues, a conservative position seems warranted and it appears that, all things considered, opinions

By nature, a qualitative study does not aim to generalize results, but instead explain a specific situation in its context.

remain weaker than scientific observations. It should be noted at the outset that relying on experts' good reputation does not overcome inherent weaknesses in this type of consensus exercise **(DellaVigna and Pope, 2018)**. In addition, DellaVigna and Pope **(2018)** show that groups always perform better at predicting a rank-ordering of the efficacy of treatments than single individuals, even when these individuals are recognized experts. The judgements formulated by groups of experts take the form of projections, hypotheses and extensions of the available data, which are subject to a large number of biases, including, to begin with, the choice of experts consulted. However, groups of experts may be used very productively to answer another type of questions. DellaVigna, Pope and Vivalt **(2019)** propose a methodology to use expert judgement in novel ways in the conduct and dissemination of research results that may improve the use of evidence at higher levels.

Level 6 refers to the exclusive reliance on qualitative studies. Their interpretative nature shows what is possible but not necessarily probable in terms of the effect of given interventions. By nature, a qualitative study does not aim to generalize results, but instead explain a specific situation in its context. This can lead to the identification of pertinent variables to study experimentally **(Slavin, 2020)**. Alternatively, single-case designs are available, which demonstrate the effect of an intervention experimentally and clearly, but are not generalizable, unless a large number of single-case studies are available to submit to a meta-analytical approach, in which case they will lack representativeness.

Level 7 implies the availability of relevant empirical observations that can serve to instigate future research but with problematic methodological origins. It should be noted that the levels of scientific evidence beyond this level apply only to well-conducted scientific studies.

1

CHAPTER

Level 8 indicates the absence of systematic empirical observations. Thus, this level includes professional success stories, principled positions, trending and hot topics, or media attention to strategically selected research results.

The applied and professional fields, which rely on scientific knowledge, can view research results according to the levels of scientific evidence presented. Levels of scientific evidence help establish a level of confidence in research results, which seems essential given that many

professionals in education (including special education teachers) work with vulnerable populations of learners. It should be noted that a professional stance based on knowing and applying research-based practices reinforces, rather than diminishes, the importance of professional judgement. Professionals become responsible for knowing the aspects of their role that can be oriented by evidence and those that cannot. They also become responsible for applying evidence in their practice, an application that requires great expertise to match interventions with given needs, rank them according to their likely effect and contextualize the best intervention without threatening its active ingredients. The levels of evidence are also useful in helping researchers classify evidence that supports their own research processes and results. Finally, they are instrumental in guiding policy-makers in their decision process regarding educational practice and the appropriateness of interventions.

THEORY BUILDING AND THEORY TESTING, AND THE NEED TO MOVE UP ACROSS LEVELS OF SCIENTIFIC EVIDENCE IN EDUCATIONAL RESEARCH.

Going back to the need to know the likely effect of an intervention and, importantly, its mechanism as a condition for its implementation (i.e. general effectiveness claims), it is therefore possible to conclude that potential best practices based on evidence will initially be drawn from cumulative and converging evidence originating from experimental research (participants randomly assigned between groups), quasi-experimental research, and single-case studies on more or less proximal target outcomes. Such evidence is currently expressed in terms of effect size, a notion originally used to design better replications of a study in terms of statistical power **(Cohen, 1962)**, and recuperated following the need to establish

Although the evidence-based trend is widespread in education, its application by practitioners has been the subject of widespread criticism targeting in turn internal and external validity.

practical significance **(Kirk, 1996)**. Technically, an effect size is the mean difference (standardized or in natural units) in outcome scores between a study's intervention and comparison groups **(Simpson, 2018)**. It is generally considered the best estimation of the effectiveness of an intervention, which can be compared across studies and interventions. However, Simpson argues that an effect size is mostly a measure of the clarity of the results of a study because it is also influenced by the psychometric characteristics of the outcome measures and characteristics of the samples, in addition to the effect of the intervention. Therefore, the effect size is the best solution to date, but technical improvements are warranted. It should be noted that the publication process likely inflates effect sizes because of a scarcity of reporting confidence interval and statistical power in the context of dichotomous statistical decision-making **(Fritz, Scherndl and Kuhberger, 2012)**. In other words, if statistically significant findings tend to get published more, then conditional on being published, effect sizes will be larger than they are in all of the studies (or, more importantly, statistical tests) undertaken. While traditional thinking underscored that confidence in research relied on it being of a high-quality standard (e.g. correct and faithful implementation) with solid psychometric measures and with little or no subject attrition, the present reasoning implies a major reconsideration of the veracity of research findings. Ioannidis **(2005)** has boldly demonstrated that, in principle, more than 50 per cent or research findings are very likely to be false as a result of bias such as research design, nature of the data, analysis strategy and reporting. Consequently, he concludes that confidence in research should arise from larger samples, larger effect sizes, more uniformity in research designs, definitions, outcome measures and analytical strategies. Because of the difficulty of conducting experimental studies in a school environment, we take a realistic stance to insist on the accumulation of quasi-experimental studies. Thus, within these constraints, the convincing

Often seen as a hierarchy of scientific methodological quality because of its grounding in internal validity, the applicability of EBE according to a policy (decision-maker) perspective is frequently overlooked.

nature of an intervention will typically be demonstrated by a large number of study results that demonstrate significant results or few studies that demonstrate mixed effects, with many studies demonstrating positive effects and no or few studies demonstrating negative effects.

Although the evidence-based trend is widespread in education, its application by practitioners has been the subject of widespread criticism targeting in turn internal and external validity. Internal validity is the extent to which an empirical study establishes and univocally explains a relationship between an intervention and its outcome; external validity refers to the possibility of applying the conclusions of an empirical study outside the context of the study.

Even in the case of the higher levels of evidence, the construct validity of studies regarding a given issue may be less than ideal: the definition of a given tested intervention may vary significantly across studies **(Davis, 2018; Simpson, 2018)** even if they stem from the same theoretical background. Thus, the cumulative evidence of desirable effects may be misleading in failing to capture the active ingredients in the approach as implemented in studies, departing from the apparently homogeneous theoretical definitions and further confounding the variability of impact across populations and contexts.

Often seen as a hierarchy of scientific methodological quality because of its grounding in internal validity, the applicability of EBE according to a policy (decision-maker) perspective is frequently overlooked **(Parkhurst and Abeysinghe, 2016)**. Indeed, evidence-based practice and evidence-based policy do not face the same challenges. Regarding evidence for policy-making, one may prefer to use the term evidence-informed because not only higher-level evidence is useful in the policy-making process. Higher-level evidence may be very useful to determine the effects of an intervention at the practical level **(Slavin, 2020)**, but evidence of a different nature

is needed from a policy-making perspective depending on the context. Particularly in a field like education, where practice is based on policies, aspects such as popular opinion of practices, social determinants of target groups and other contextual variables are important to take into account **(Parkhurst and Abeysinghe, 2016)**. These aspects may therefore also hinge on high-quality evidence, but with respect to a different criterion corresponding to a different type of assertions. As will be discussed in the next section, assertions related to a particular context have to be seen as complementing previous levels of evidence that support relative effectiveness generalizations. Doing so will contribute to developing the educational policies on which practices are ultimately based.

Another limitation of the hierarchy of scientific evidence is the external validity of the evidence **(Joyce, 2019)**. Higher-level evidence aims at increasing the internal validity of studies to better demonstrate the effect of an intervention, but the external validity of these studies remains limited **(Orr, 2015)**. In the biomedical field, for example, there is an expectation that one entity will be similar to another (e.g. one human body is similar to another). This allows extrapolation of the results obtained in the laboratory to other contexts. In psychosocial fields (e.g. education), these similarities between entities are harder to demonstrate. Hence, interventions are more likely to produce different results in different groups, contexts, and so on. In such cases, results from experimental studies are not always isomorphically transposable or transportable to 'real-life' contexts **(Schmuckler, 2001)**. Even meta-analyses are susceptible to introducing biases regarding the external validity of a body of research since they pool studies conducted in several contexts that are not necessarily comparable **(Parkhurst and Abeysinghe, 2016)**.

Another limitation of the hierarchy of scientific evidence is the external validity of the evidence.

Another aspect that can affect the external validity of meta-analyses is the publication bias from the articles they include **(Gage, Cook and Reichow, 2017)**. Publication

Thus, cumulative evidence of desirable effects may be misleading by not capturing the active ingredients in a given approach as implemented in studies that deviate from seemingly homogeneous theoretical definitions, thereby further confusing the variability of the impact between populations and contexts.

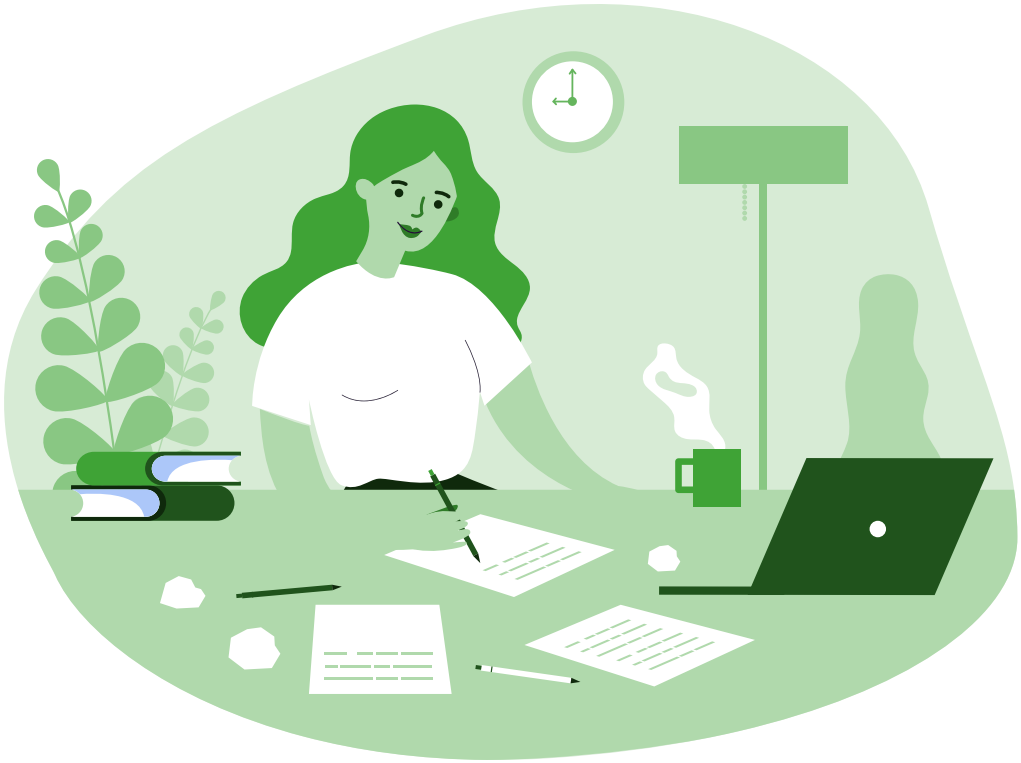
bias is defined as the fact that articles with greater effect sizes or statistical significance are more likely to be published, with articles with mixed results or statistically in significant results less likely to be published. Although both scientific and probative levels of evidence are affected by publication bias, the meta-analytic process can be particularly affected by it because, without rigorous pre-specification and inclusion of grey literature, it can carry this bias by selecting articles from among an already biased pool of published articles. By doing so, meta-analytic results can boost the effect size tainted by the publication bias **(Fritz, Scherndl and Kuhberger, 2012)**.

Thus, cumulative evidence of desirable effects may be misleading by not capturing the active ingredients in a given approach as implemented in studies that deviate from seemingly homogeneous theoretical definitions, thereby further confusing the variability of the impact between populations and contexts. All the previous caveats can, in principle, be alleviated by recourse to relevant theory. Indeed, these caveats stem at least in part from definitional issues related to critical aspects of empirical work, such as population characteristics, interventions, outcomes, control variables and contexts.

In sum, the first aspect of next-generation EBE is the provision of general relative effectiveness claims (which takes the form of a new, more stringent, probative level in the framework), indicating that an intervention has a stable causal capacity relative to all other comparable interventions. This is a significant improvement over traditional EBE based on 'what works', which culminated with a miscellaneous collection of interventions essentially shown to be better than nothing. What is needed to complement these general relative effectiveness claims are credible assertions about how a local context affords a causal pathway through which the most effective intervention can make a positive contribution.

WORKING GROUP 04

CHAPTER



1.4 Will it work here? How the local context affords a causal pathway through which the intervention can make a positive contribution

Ultimately, we don't just want to know if an intervention works, we want to know if it will work in the specific context in which it is intended to be used.

Ultimately, we don't just want to know if an intervention works, we want to know if it will work in the specific context in which it is intended to be used. This question implies a shift toward a context-focused approach to EBE **(Joyce and Cartwright, 2020)**, which, in our proposed framework, is the necessary complement to the general relative effectiveness claims discussed earlier. Answering the question 'will it work here and now?' amounts to demonstrating, by means of empirical data or literature, how the local context affords a causal pathway through which an intervention documented as effective can make a positive contribution. The inferences made through this reasoning have been termed local effectiveness predictions by Joyce and Cartwright **(2020)**. While local effectiveness predictions will never be certain, incorporating this information in the reasoning supporting the implementation of evidence-based practices can improve them **(Joyce and Cartwright, 2020)**.

Proponents of EBE generally attribute the gap between research and practice results to shortcomings in the way tasks are performed in either knowledge production or knowledge use in practice **(Joyce and Cartwright, 2020)**. However, we argue that a major part of the necessary reasoning in EBE, formulating local effectiveness predictions, has been overlooked. With this in mind, qualitative research, which appears to be lower-level evidence in the context of establishing what works best **(see Table 1)** becomes mandatory in our proposed framework to attain higher levels of evidence in the context of establishing a fit with local context **(see Table 2)**. For example, ethnographic approaches or local surveys are also needed in order to assemble a body of evidence supporting the utility of an intervention in a specific context **(Parkhurst and Abeysinghe, 2016)**.

What kind of reasons can support projectability and transportability of extant research in educational contexts? Results from a sample representing a given population permits generalizing results to that population, but not transporting

… the argument theory of evidence specifies that 'a fact counts as evidence for a specified claim when it speaks to the truth of that claim'…

results to specific targets within it **(Pearl and Bareinboim, 2014)**. To this end, and because a progression to higher levels of evidence does not provide effectiveness predictions (transportability is a causal, not statistical notion) **(see Pearl and Bareinboim, 2014)**, a complementary, mostly inductive rationale is needed. As discussed by Joyce and Cartwright **(2020)**, the argument theory of evidence specifies that 'a fact counts as evidence for a specified claim when it speaks to the truth of that claim' **(p. 1051)**. Additionally, the material theory of induction underscores the importance of empirical work; observations are encoded in substantive claims that connect the evidence with the hypothesis **(Norton, 2003)**. Considered in this light, a research result is evidence relative to a target hypothesis and to a set of additional claims describing material facts about the world **(Joyce and Cartwright, 2020)**. In considering local effectiveness predictions, the hypothesis to be evidenced is: the outcomes specified in claims about relative effectiveness generalizations will occur within a local context.

As illustrated next within the discussion of the realist approach, the evidence needed to test this hypothesis may come from empirical research, observations and credible theory. A formal graph-based procedure may also be used to logically encode and analyse differences between contexts **(Pearl and Bareinboim, 2014)**. Given the state of the research in education, in which mechanisms and processes are generally not sufficiently understood, this procedure may be best used for the moment to foster the necessary types of research, rather than to warrant the transportability of results across contexts.

A realist approach to the review and synthesis of evidence from the literature and to the evaluation of implementation of a given intervention seems particularly productive to answer the question 'will it work here?' The goal of a realist review is to explore the contexts that trigger certain mechanisms and the resultant 'outcomes of interventions' **(Defever and Jones, 2021, p. 9)**. Moreover, in light of the need for

evidence of contextual fitting in EBE, the realist review appears to be a mandatory analysis following systematic review and meta-analysis in our proposed framework. In that sense, coupling systematic reviews and meta-analyses with realist reviews is the only way to be fully probative in EBE. The approach underlying a realist review focuses on the same key aspect as the levels of evidence, that is, causality between interventions and outcomes. Indeed, mechanisms, in the realist approach, represent causal processes **(Caswell et al., 2020)** in the form of structure, culture and agency **(De Souza, 2016)**. According to De Souza **(2016)**, these pre-existing conditions establish boundaries that contribute to constraining or enabling the effectiveness of different aspects of a complex programme. To strengthen the impact of EBE, these conditions need to be reported as evidence in research findings. 'Gaining insights about the contexts within which programmes are implemented can point to the conditions needed to help trigger its potential successful workings. It also enables explanations about the conditions existing that might be hindering the intended integration, uptake, or outcome of the program.' **(De Souza, 2016, pp. 226-227)**. In our view, it is the process of looking beyond variables that are studied, compared or controlled in quantitative work.

A realist synthesis is a narrative summary focused on interpretive theory that applies a realist philosophy to the synthesis of primary study results that affect a single applied research question. Realist review and classic systematic reviews procedures are relatively similar. An essential difference, however, is an insistence on the notion that experimental results are always context-dependent and that interventions are never implemented in the same context **(Smets and Struyven, 2018)**. A realist review uses an interpretive inter-case comparison to understand and explain, how and why the observed results occurred in the studies included in a literature review **(Wong et al., 2012)**. Realist

evaluation provides a framework for understanding how the context and underlying mechanisms affect the outcomes of an intervention **(Ericson et al., 2017)**. In trying to understand why policy programmes are usually not implemented as designed, Verger, Bonal and Zancajo **(2016)** emphasize one aspect of the realist approach, the agency of actors. These authors insist on the notion that the application of policy programmes is mediated by the previous experiences, values and interests of the subjects, and by the ways in which they interpret the rules of the programme.

These methods were originally developed by Pawson and Tilley to evaluate complex intervention policies in health and social services **(Pawson and Tilley, 1997; ;**

... the success of an intervention depends on how participants interact with it in local contexts, and a realist approach should uncover these processes.

Pawson et al., 2005; Pawson, 2006). In a realist approach, data is collected and analyzed in order to determine context–mechanism–process effect configurations **(Haynes et al., 2017)**. An explanation and understanding of the interaction between the context, the mechanism and the impact of the intervention is then produced **(Wong et al., 2012)**. This joint focus on context, mechanism and process effect should overcome one crucial limitation of quantitative research: authors have argued that traditional study designs such as randomized controlled trials, and non-randomized and prospective cohort studies, although useful, depending on the objective of the evaluation, overlook a key element, namely being able to identify contextual information that is useful when replicating the results in another context **(Graham and McAleer, 2018)**.

In other words, the success of an intervention depends on how participants interact with it in local contexts **(Haynes et al. 2017)**, and a realist approach should uncover these processes.

The working hypothesis behind a realistic synthesis is that a particular intervention (or class of interventions) will trigger particular mechanisms somewhat differently in different contexts. In realism, it is the mechanisms that trigger change rather than the interventions themselves, and realistic reviews therefore focus on 'families of mechanisms' rather than 'families of interventions' **(Wong et al., 2012)**.

1.4 .1

LEVELS OF CONTEXTUAL FITTING APPLIED TO EDUCATIONAL RESEARCH TOWARD LOCAL EFFECTIVENESS PREDICTIONS

In the same way that levels of evidence establish the information

TABLE 2 LEVELS OF CONTEXTUAL FITTING APPLIED TO EDUCATIONAL RESEARCH

RANGE	LEVEL	EVIDENCE REQUIRED	MAIN LIMITATIONS
Probative	1	Realist review	
Scientific	2	Qualitative research during implementation work	Correspondence between studied population/context established for the target population, but without taking into account all contextualized elements from the literature
	3	Qualitative research during experimental work	Correspondence between studied population/context established only from the population studied
	4	Exclusive reliance on relative effectiveness generalizations	Correspondence between studied population/context unestablished
Pseudo-scientific/ non-scientific	5	Exclusive reliance on causal ascriptions and general effectiveness claims	Based on arbitrary[1] choices among 'what works'

[1]Arbitrary is meant to include, but is not restricted to epistemological biases, personal preferences, emphasizing the latest research or more globally acting without the required information.

needed to make relative effectiveness generalizations, **Table 2** proposes a classification of the contextual fitting of effective interventions based on scientific evidence. Akin to the previous levels of evidence, this proposal of criteria allows us to distinguish between: (1) information of pseudoscientific/non-scientific

The qualitative work in this level is very similar to that in level 3, with the important difference that the observations are conducted in the context of application.

nature; (2) the results emanating from a scientific approach; and (3) the probative level in which the relative convergence and divergence of results is uncovered based on a thorough literature review. The facts needed to improve the level of contextual fitting come from empirical research, observations and credible theory.

As shown in **Table 2**, level 5 is considered pseudo/non-scientific, whereas levels 2 to 4 are deemed scientific. The probative range is limited to level 1.

Level 1. This level is the only one to provide probative information necessary to test the hypothesis that the outcomes specified in claims about relative effectiveness generalizations will occur within a local context. The information is probative because it is based on a review of the literature, and can be considered the best way to identify, define and establish the salience of the variables involved in effectiveness predictions.

Level 2. The qualitative work in this level is very similar to that in Level 3, with the important difference that the observations are conducted in the context of application.

Level 3. Level 3 involves qualitative research during quantitative experimental work, a strategy underlying mixed-methods research. While the quantitative approach provides causal ascriptions, qualitative work establishes, inductively, a complementary model to explain the results. The limitation, especially in comparison with level 2, is that this explanation is part of an 'external' study, the results of which have to be transported to the context of application.

Level 4. In level 4, the reliance on relative effectiveness generalizations established from meta-analytic work and syntheses does not provide evidence of the transportability of a relatively effective intervention to a new context, beyond a collection of sampling variables that may not be salient in making effectiveness

predictions.

Level 5. Given the limitations of causal ascriptions and general effectiveness claims presented earlier, especially with respect to a lack of information about how a given intervention compares to others (and not just to business-as-usual teaching), level 5 posits that choosing an intervention to replace the one currently implemented in this context is so likely to be suboptimal that the status quo is probably better. As such, the hypothesis that the outcomes specified in claims about relative effectiveness generalizations will occur within a local context cannot even be tested.

As this proposal for levels of contextual fitting aims to demonstrate, for credible evidence-based policy or practice, the assumption that populations are alike must be supported.

As this proposal for levels of contextual fitting aims to demonstrate, for credible evidence-based policy or practice, the assumption that populations are alike must be supported **(Joyce, 2019)** by theory and other empirical results. Judging when generalized results from studies and specific applied settings are similar enough and in the right ways requires theory – lots of it and of very different kinds. Key aspects of the realist approach are linked to the use of theory in the form of context–mechanism–process effect configurations **(Haynes et al., 2017)**.

1.4 .2

THEORY BUILDING AND THEORY TESTING, AND THE NEED TO MOVE UP ACROSS LEVELS OF CONTEXTUAL FITTING IN EDUCATIONAL RESEARCH

Effectiveness predictions are obtained through the identification of contextual influences **(Joyce and Cartwright, 2020)**. Because we contend that contextual fitting necessarily occurs after obtaining the best level of evidence for relative

While local effectiveness predictions will never be certain, we propose that the sources of information used to formulate them can inform us about their accuracy and potential for transportability.

effectiveness generalizations, we specify the identification levels of contextual fitting as a process of disaggregation of contextual influences. This takes place through cumulative abstraction, in which relative effectiveness generalizations are 'reverse-engineered' once the target best intervention has been determined. The disaggregation of contextual influences through a realist review involves analyzing intervention characteristics that generate observed changes (i.e. mechanisms) and can inform the development or refinement of a conceptual framework **(Defever and Jones, 2021)**.

Also, we suggest that this process of disaggregation cumulatively leads to an increase in what we call levels of contextual fitting. Incorporating this information into the reasoning that supports the implementation of evidence-based practices will, in principle, improve the likelihood of replicating documented outcomes **(Joyce and Cartwright, 2020)**. While local effectiveness predictions will never be certain, we propose that the sources of information used to formulate them can inform us about their accuracy and potential for transportability. This increase in levels of contextual fitting hinges on theory building in the sense that identifying the causal mechanisms behind the effectiveness of an intervention constitutes the main asset for transporting (from one context to another by re-examining the variables, different from generalizing across contexts) effectiveness predictions. An increase in levels of contextual fitting signifies more reliable predictions about what might work in a given school or district, and with targeted students and, as Joyce and Cartwright **(2020)** insist, how it might work.

WORKING GROUP 04

CHAPTER 1

1.5 Conclusion: the framework and its implications

This chapter tackles the issue that while reviewing the scientific literature, it is sometimes difficult stakeholders such as policy-makers and practitioners to apply the evidence for the best possible effects in specific contexts, given the plethora of studies available. This chapter considered the importance of scientific theory in

... the question 'will it work here?' is now posited as absolutely necessary to complement the information and reasoning pertaining to 'what works best'.

explaining phenomena, and the contribution of empirical research to theory building and theory testing. It then examined the levels of scientific evidence and the need to accumulate appropriate evidence across these levels in order to support specific inferences in education. Finally, it discussed the need to fit general relative effectiveness claims to specific contexts of application.

Articulating the two main ingredients of next-generation EBE posited in this paper – general effectiveness claims and effectiveness predictions – in an effort to go beyond 'what works' leads to a new articulation of applied empirical research within a given educational field, as seen in **Figure 1**. A few notable proposals emerge from the current work. Within the traditional view of levels of evidence, the probative level now concerns only relative effectiveness generalizations (i.e., a rank-ordering (generalizable to a population) of the effectiveness of all pertinent interventions), and not effectiveness generalizations (how a given intervention compares to a control). This places the meta-analytic approach as key to the provision of the required information to answer the most important question: what works best? Consequently, the gold standard of EBE, the randomized controlled trial, is no longer in the probative range. In addition, the conceptualization and operationalization of the levels of contextual fitting, in response to the need for local effectiveness predictions, can be seen as the most important contribution of the current work. Its most constructive implication is that the synergy between quantitative and qualitative approaches in applied research is more apparent. Also, the question 'will it work here?' is now posited as absolutely necessary to complement the information and reasoning pertaining to 'what works best'.

The proposed articulation of causal ascriptions, relative effectiveness generalizations and local effectiveness predictions generated by empirical research in education in the form of the EBE3 framework has implications

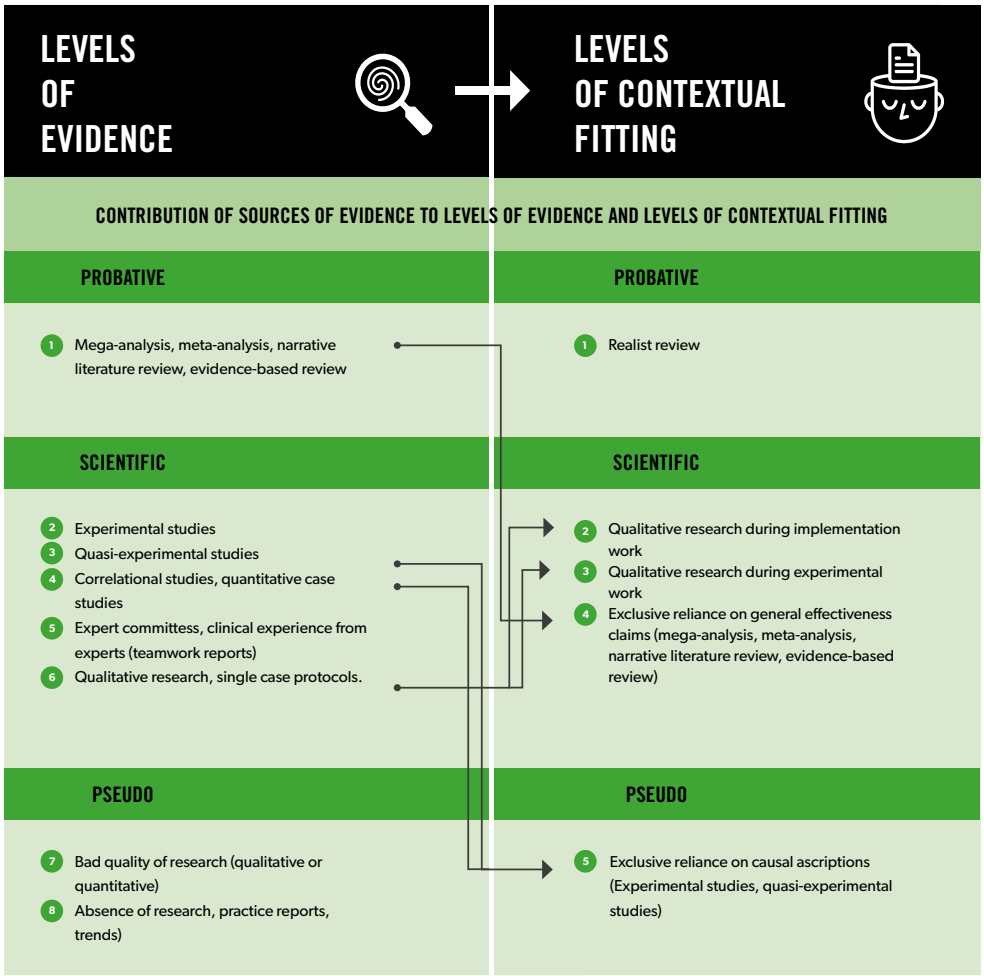


Figure 1

for conducting future research, for policy-making and for improving educational practice.

Concerning the orientation of applied scientific research, the framework in **Figure 1** may shed light on the need for specific

By reviewing and integrating the state of the art in EBE, it becomes clear that quantitative and qualitative research leverage each other in achieving the cumulative steps necessary for better intervention in a given domain.

kinds of quantitative studies, meta-analyses and synthesis of work, as well as qualitative implementation work. Thus, it helps in bridging the perceived divide between quantitative and qualitative research in education by suggesting a sound integration of quantitative and qualitative methodologies around a common applied goal: providing the necessary information for the improvement of educational intervention. By reviewing and integrating the state of the art in EBE, it becomes clear that quantitative and qualitative research leverage each other in achieving the cumulative steps necessary for better intervention in a given domain. As De Souza **(2016)** notes, methodologies for realist evaluation and review are still in development and are likely to make increasing contributions to the application of empirical research.

In light of the importance of meta-analyses and systematic reviews underlined when discussing the need for effectiveness generalizations, it should be noted that the realist review process presented as a method for establishing effectiveness predictions can be reused to facilitate the automation of meta-analyses and enable living reviews of evidence. The realist approach has provided a consistent rationale for synthesizing evidence across forms and types of interventions **(Pearson et al., 2015)**. Indeed, realist reviews can be key in standardizing coding frameworks for studies, with common coding of cohorts, intervention delivery mechanisms and core components. In addition, the framework presented in **Table 2** helps in focusing research efforts directly on a frequently overlooked issue, that is, how to build local effectiveness predictions. It outlines various kinds of information that can improve predictions and encourages using appropriate methods for acquiring that information.

With respect to policy-making, the framework presented in **Tables 1** and **2** may feed into the mechanisms identified by Langer, Tripney and Gough

(2016) as facilitating research use by policy-makers, beyond the preconditions regarding enhancing decision- makers' opportunity, capability and motivation to use evidence. By insisting on a more complete scientific demonstration of efficacy, from causal ascriptions to effectiveness generalizations and effectiveness predictions, the framework may provide the materials for interventions facilitating access to research evidence and for interventions building decision-makers' skills to access and make sense of evidence.

At the level of organizations and systems, this more complete scientific demonstration of efficacy outlined in **Table 2** may help identify the right information for the right people that can be used in the design of interventions that foster changes to decision-making structures and processes. Notably, an increased focus on core components, that is, mechanisms that represent active ingredients in interventions, can help policy-makers avoid biases toward scientific disciplines that may seem compelling but do not provide the best explanations about how interventions work and why. The consequences of evidence-based reform refined operationally in this paper could be profound. If educational policies begin to favour programmes with clear evidence, publishers, software developers, university researchers and entrepreneurs will have an incentive to engage in serious development and evaluation efforts. Governments, seeing the cumulative impact of such research and development, might provide substantially greater funding for these activities in education.

Finally, practice should be greatly improved by a widened view of the necessary evidence in the implementation of so-called best practices, especially regarding effectiveness predictions. Effectiveness predictions help frame practitioners' reasoning concerning the match between general, abstracted evidence and their own specific and idiosyncratic context around a specific kind of inference that is amenable to analysis and testing in

Finally, practice should be greatly improved by a widened view of the necessary evidence in the implementation of so called best practices, especially regarding effectiveness predictions.

In sum, the EBE3 framework presented in this paper may be one of the most integrative in terms of research traditions and with respect to the different roles (teachers, researchers, policy-makers) involved in EBE.

the context of day-to-day practice.

Evidence brokerage is also crucial to bridge the gaps between research and practice **(Langer, Tripney and Gough, 2016)**. Because the EBE3 framework identifies the reasoning and the supporting information for next-generation EBE, it could be used in information design, to enhance the structure of evidence repositories and other resources. Langer, Tripney and Gough **(2016)** also conclude that interaction among professionals can build a professional identity with common practices and standards of conduct fostering EBE. Using social influence and peer-to-peer interaction as catalysts, districts may be able to use support specialists (e.g. curriculum specialists, programme specialists) and schools may be able to use onsite personnel, including literacy facilitators or highly effective general or special education teachers (peers) as coaches. The focus could then be on those teachers who need follow-up support instead of providing the same support for all teachers across all professional development activities.

In sum, the EBE3 framework presented in this paper may be one of the most integrative in terms of research traditions and with respect to the different roles (teachers, researchers, policy-makers) involved in EBE. Future work should evaluate the implications of such an integration in terms of its conceptual, operational and organizational aspects.

CHAPTER 1

1.6 Key messages

The results of a collection of high-quality studies comparing an experimental group given a target intervention with a control group (usually receiving business-as-usual teaching) has been the cornerstone of EBE for decades under the label 'what works'. It is the main, but not sufficient, building block of EBE, and there is a need for a higher minimum standard for what counts as evidence of improved learning.

For a given educational issue, what is needed is a complete inventory of available interventions, rank-ordered in terms of relative efficacy to answer the question 'what works best generally?'.

An EBE initiative is not complete without solid indications that a specific application context will enable the 'working best in general' intervention to yield the expected benefits. This will answer the question 'will it work here'. Concretely, a realist review should be seen as complementary to a systematic review and meta-analysis and therefore should be conducted in tandem.

1.7 Key recommendations

The potential of the EBE3 framework to go beyond 'what works' will be fully realized by:

emphasizing effectiveness generalizations by the production of meta-analytic work as soon as there are enough published experimental studies on a given issue; and

emphasizing effectiveness predictions by undertaking qualitative work relating to effectiveness predictions in given contexts as soon as meta-analytic results are available.

The potential of the EBE3 framework to provide greater cohesion to applied empirical work on a given issue will be fulfilled by:

- ***focusing on theory building and theory testing*** in conducting empirical studies, despite the applied nature of educational research.

- ***aligning the goals/research questions of quantitative and qualitative research*** with the maturity of a field to optimize the outcomes when applied to educational interventions.

REFERENCES

Anderson, J.E. (2011) Public policymaking. Boston: Wadsworth Cengage Learning.

Borenstein, M. (2019) Common mistakes in meta-analysis and how to avoid them. Englewood: Biostat.

Brighouse, H., Ladd, H.F., Loeb, S. and Swift, A. (2018) Educational goods: values, evidence, and decision-making. Chicago: University of Chicago Press.

Burns, P.B., Rohrich, R.J. and Chung, K.C. (2011) 'The levels of evidence and their role in evidence-based medicine', Plastic Reconstruction Surgery, 128(1). https://doi.org/:10.1097/PRS.0b013e318219c171.

Caswell, R.J., Maidment, I., Ross, J.D.C. and Bradbury-Jones, C. (2020) 'How, why, for whom and in what context, do sexual health clinics provide an environment for safe and supported disclosure of sexual violence: protocol for a realist review', BMJ Open, 10. https://doi.org/10.1136/bmjopen-2020-037599.

Chubb, J. and Watermeyer, R. (2017) 'Artifice or integrity in the marketization of research impact? Investigating the moral economy of (pathways to) impact statements within research funding proposals in the UK and Australia', Studies in Higher Education, 42. https://doi.org/10.1080/03075079.2016.1144182.

Cochrane, A.L. (1972) Effectiveness and efficiency: random reflections on health services. London: Nuffield Trust.

Cohen, J. (1962) 'The statistical power of abnormal-social psychological research: a review', Journal of Abnormal and Social Psychology, 65(3), pp. 145–153.

Colquitt, J.A. and Zapata-Phelan, C.P. (2007) 'Trends in theory building and theory testing : a five-decade study of the Academy of Management Journal', Academy of Management Journal, 50(6). https://doi.org/10.5465/amj.2007.28165855.

Connolly, P., Keenan, C. and Urbanska, K. (2018) 'The trials of evidence-based practice in education: a systematic review of randomised controlled trials in education research 1980–2016', Educational Research, 60(3). https://doi.org/10.1080/00131881.2018.1493353.

Davis, A. (2018) 'Evidence-based approaches to education: direct instruction, anyone?', Management in Education, 32(3), pp. 135–138.

De Souza, D. (2016) 'Critical realism and realist review: analyzing complexity in educational restructuring and the limits of generalizing program theories across borders', American Journal of Evaluation, 37(2), pp. 216–237.

Defever, E. and Jones, M. (2021) 'Rapid realist review of school-based physical activity interventions in 7 - to 11 - year-old children', Children, 8. https://doi.org/10.3390/children8010052.

DellaVigna, S. and Pope, D. (2018) 'Predicting experimental results: who knows what?', Journal of Political Economy, 126(6). https://doi.org/10.1086/699976.

DellaVigna, S., Pope, D. and Vivalt, E. (2019) 'Predict science to improve science', Science, 366(6464). https://doi.org/10.1126/science.aaz1704.

Ericson, A., Löfgren, S., Bolinder, G., Reeves, S., Kitto, S. and Masiello, I. (2017) 'Interprofessional education in a student-led emergency department: a realist evaluation', Journal of Interprofessional Care, 31(2). https://doi.org/10.1080/13561820.2016.1250726.

Evans, D. (2003) 'Hierarchy of evidence: a framework for ranking evidence evaluating healthcare interventions', Journal of Clinical Nursing, 12, pp. 77–84.

Frias-Navarro, D., Llobell, J., Pascual-Soler, M., Perez-Gonzalez, J. and Berrios-Riquelme, J. (2020) 'Replication crisis or an opportunity to improve scientific production?', European Journal of Education, 55. https://doi.org/10.1111/ejed.12417.

Fritz, A., Scherndl, T. and Kühberger, A. (2012) 'A comprehensive review of reporting practices in psychological journals: are effect sizes really enough?', Theory & Psychology, 23(1), pp. 98–122.

Gage, N.A., Cook, B.G. and Reichow, B. (2017) 'Publication bias in special education meta-analyses', Exceptional Children, 83(4), pp. 428–445.

Goldacre, B. (2013) 'Building evidence into education'. Available at: https://www.gov.uk/government/news/building-evidence-into-education (Accessed: 30 June 2021).

Graham, A.C. and McAleer, S. (2018) 'An overview of realist evaluation for simulation-based education', Advances in Simulation, 3(13). https://doi.org/10.1186/s41077-018-0073-6.

Hansson, S.O. (2009) 'Cutting the gordian knot of demarcation', International Studies in the Philosophy of Science, 23(3). https://doi.org/10.1080/02698590903196007.

Haynes, A., Brennan, S., Redman, S., Williamson, A., Makkar, S.R., ... and Butow, P. (2017) 'Policymakers' experience of a capacity-building intervention designed to increase their use of research: a realist process evaluation', Health Research Policy and Systems, 15(99). https://doi.org/10.1186/s12961-017-0234-4.

Ioannidis, J.P.A. (2005) 'Why most published research findings are false', PLOS Med, 2(8). https://doi.org/10.1371/journal.pmed.0020124.

Jahan, N., Naveed, S., Zeshan, M. and Tahir, M.A. (2016) 'How to conduct a systematic review: a narrative literature review', Cureus, 8(11). https://doi.org/10.7759/cureus.864.

Joyce, K. and Cartwright, N. (2020) 'Bridging the gap between research and practice: predicting what will work locally', American Educational Research Journal, 57(3). https://doi.org/10.3102/0002831219866687.

Joyce, K.E. (2019) 'The key role of representativeness in evidence-based education', Educational Research and Evaluation, 25(1–2). https://doi.org/10.1080/13803611.2019.1617989.

Karrigan, M.R. and Turner-Johnson, A. (2019) 'Qualitative approaches to policy research in education: contesting the evidence-based, neoliberal regime', American Behavioral Scientist, 63(3). https://doi.org/10.1177/0002764218819693.

Kirk, R.E. (1996) 'Practical significance: a concept whose time has come', Educational Psychological Measurement, 56(5), pp. 746–759.

Kornell, N., Rabelo, V. C. and Klein, P. J. (2012) 'Tests enhance learning: compared to what?', Journal of Applied Research in Memory & Cognition, 1(4), pp. 257–259.

Langer, L., Tripney, J. and Gough, D. (2016) The science of using science: researching the use of research evidence in decision-making. London: EPPI-Centre, Social Science Research Unit, UCL Institute of Education, University College London.

Maciver, D., Rutherford, M., Arakelyan, S., Kramer, J.M., Richmond, J. and Todorova, L. (2019) 'Participation of children with disabilities in school: a realist systematic review of psychosocial and environmental factors', PLOS ONE, 14(1). https://doi.org/10.1371/journal.pone.0210511.

National Research Council (2002) Scientific research in education. Washington, DC: National Academies Press.

Norton, J.D. (2003) 'A material theory of induction', Philosophy of Science, 70. https://doi.org/10.1086/378858.

Nutley, S., Powell, A. and Davies, H. (2013) What counts as good evidence? Provocation paper for the alliance for useful evidence. Available at: https://www.alliance4usefulevidence.org/assets/What-Counts-as-Good-Evidence-WEB.pdf (Accessed: 30 June 2021).

REFERENCES

Orr, L.L. (2015) '2014 Rossi Award Lecture: beyond internal validity', Evaluation Review, 39(2). https://doi.org/10.1177/0193841X15573659.

Parkhurst, J.O. and Abeysinghe, S. (2016) 'What constitutes "good" evidence for public health and social policy-making? From hierarchies to appropriateness', Social Epistemology, 30(5-6). https://doi.org/10.1080/02691728.2016.1172365.

Pawson, R. (2006) Evidence-based policy: a realist perspective. London: Sage.

Pawson, R. and Tilley, N. (1997) Realistic evaluation. London: Sage.

Pawson, R., Greenhalgh, T., Harvey, G. and Walshe, K. (2005) 'Realist review: a new method of systematic review designed for complex policy interventions', Journal of Health Services Research & Policy, 10, pp. 21–34.

Pearl, J. and Bareinboim, E. (2014) 'External validity: from do-calculus to transportability across populations', Statistical Science, 29(4). https://doi.org/10.1214/14-STS486.

Pearson, M., Chilton, R., Wyatt, K., Abraham, C., Ford, T., Woods, H.B. and Anderson, R. (2015) 'Implementing health promotion programmes in schools: a realist systematic review of research and experience in the United Kingdom', Implementation Science, 10(1). https://doi.org/10.1186/s13012-015-0338-6.

Schmuckler, M.A. (2001) 'What is ecological validity? A dimensional analysis', Infancy, 2(4), pp. 419–436.

Shadish, W.R., Cook, T.D. and Campbell, D.T. (2002) Experimental and quasi-experimental designs for generalized causal inference. Boston: Houghton Mifflin.

Simpson, A. (2018) 'Princesses are bigger than elephants: effect size as a category error in evidence-based education', British Educational Research Journal, 44(5). https://doi.org/10.1002/berj.3474.

Slavin, R.E. (2020) 'How evidence-based reform will transform research and practice in education', Educational Psychologist, 55(1). https://doi.org/10.1080/00461520.2019.1611432.

Smets, W. and Struyven, K. (2018) 'Realist review of literature on catering for different instructional needs', Educational Science, 8. https://doi.org/10.3390/edusci8030113.

Sugimoto, C. R. and Larivière, V. (2018) Measuring research: what everyone needs to know. Oxford: Oxford University Press.

Verger, A., Bonal, X. and Zancajo, A. (2016) 'What are the role and impact of public-private partnerships in education? A realist evaluation of the Chilean education quasi-market', Comparative Education Review, 60(2), pp. 223–248.

Wong, G., Greenhalgh, T., Westhorp, G. and Pawson, R. (2012) 'Realist methods in medical education research: what are they and what can they contribute?', Medical Education, 46. https://doi.org/10.1111/j.1365-2923.2011.04045.x.

A procedure for next generation evidence based education