Received: 8 October 2023 | Accepted: 19 November 2023

DOI: 10.1002/rev3.3448

SHORT COMMUNICATION

Review of Education BERA

Judging the relative trustworthiness of research results: How to do it and why it matters

Stephen Gorard

Durham University Evidence Centre for Education, Durham, UK

Correspondence
Stephen Gorard, Durham University Evidence Centre for Education, Durham, UK.
Email: s.a.c.gorard@durham.ac.uk

Abstract

This paper describes, and lays out an argument for, the use of a procedure to help groups of reviewers to judge the quality of prior research reports. It argues why such a procedure is needed, and how other existing approaches are only relevant to some kinds of research, meaning that a review or synthesis cannot successfully combine quality judgements of different types of research. The proposed procedure is based on four main factors: the fit between the research question(s) for any study and its design(s); the size of the smallest group of cases used in the headline analyses; the amount and skewness of missing data; and the quality of the data collected. This simple procedure is now relatively widely used, and has been found to lead to widespread agreement between reviewers. It can fundamentally change the findings of a review of evidence, compared to the conclusions that would emerge from a more traditional review that did not include genuine quality rating of prior evidence. And powerfully, because it is not technical, it permits users to help judge research findings. This is important as there is a growing demand for evidence-led approaches in areas of social science such as education, wherein summaries of evidence must be as trustworthy as possible.

KEYWORDS

Evaluations, meta-analysis, systematic review

Context and implications

Rationale for this study

Currently, much research which is mostly paid for by public taxation or charitable donations is wasted. It is either ignored because users do not understand it, or used to draw unwarranted conclusions. The number and range of robust research results have improved over 20 years, but robust research is still in a minority. This means that in any meta-analysis, systematic review or synthesis the weak research is usually bundled together with and given equal emphasis to the most trustworthy wok. This yields misleading conclusions. But the problem can be addressed by the sieve approach described in this paper.

Why the new findings matter

The sieve approach provides a simple and usable approach for judging the relative rigour of different research studies on any topic. It means that researchers can weight their syntheses of prior evidence appropriately, and that teachers and policy-makers can assess the trustworthiness of any research relevant to their real-life decisions. This should improve the use of evidence, much of which is otherwise either wasted or given unwarranted attention, and so processes and outcomes for learners in education.

Implications for practice

Users now have a simple tool to assist them when assessing the quality of research evidence. This, or something like it, should be used both to assess the trustworthiness of individual studies, and to help calibrate the weight to give to each study in any future syntheses.

INTRODUCTION

This paper is concerned with how to judge the quality or trustworthiness of any piece of empirical research, and what the key characteristics of high quality research are. Here high quality work is research from which the evidence can be trusted (as far as possible in real life) and is judged to be robust and secure against bias and errors. The paper starts by rehearsing the need for such judgements, considers a number of alternative approaches, and then suggests four key factors to consider. These factors appear to be generic (relevant to all empirical research). The paper describes a process of 'sieving' that uses these four factors to provide a composite judgement of research quality, and ends with an illustration, a discussion of the process in use, and possible modifications.

THE NEED FOR QUALITY JUDGEMENTS

In education and social policy more generally there is an increasing number of syntheses of existing evidence intended to summarise what is known on any topic in a way that helps to promote evidence use, or identify evidence gaps for future research (Polanin et al., 2022). These syntheses include meta-analyses, hyper-analyses, and systematic, structured, rapid, and scoping reviews (Cirkony et al., 2022). Such syntheses matter because there is also a growing demand for real-life users to base policy and practice decisions on 'what the evidence says' (Flynn, 2019; Nutley et al., 2019).

These syntheses may be conducted by academics, organisations and government departments. Traditionally the emphases have been on scrupulous searching to try and collate all relevant evidence, and on transparency so that the findings can be checked or even replicated. These are ideals. No review can expect to find everything, and most reviews would never actually be replicated.

More recently, there has been an increased emphasis on judging the quality of the individual studies synthesised in any review. In the past, the issue of the quality of the underlying studies has often been ignored (e.g. Hattie, 1992) or it has been recommended that studies below a certain scale are simply excluded (e.g. Torgerson, 2003). Both approaches lead to clear bias.

There has been considerable criticism of meta-analyses like that of Hattie (1992) or the Education Endowment Foundation (EEF) Toolkit in England, saying that these analysts have not engaged with the quality of the underlying studies (De Vrieze, 2019), and/or have engaged in unsuitable aggregation (Bergeron, 2022). And many other syntheses of evidence, even those published by official What Works Clearing House, show considerable inconsistency when evaluating the same interventions (Wadhwa et al., 2023).

It has been demonstrated that giving equal weight in a synthesis to studies of different quality yields invalid and misleading results (Gorard & Chen, 2023). Poor quality studies tend to produce more extreme 'effect' sizes than better-designed ones, and so will dominate the results when all are aggregated. However, ignoring 10 studies each of 10 cases that show different results to one study of 100 cases, as Torgerson (2003) and others suggest, is equally poor practice. The idea of a synthesis is to aggregate studies. Therefore, while the number of cases is one factor that should be used to judge the quality of each study, and therefore its weighting in the synthesis, smaller studies should not just be routinely ignored.

Where the quality of the underlying studies *is* used, it is important that all partners in the process of synthesis have a shared understanding of how to judge quality, and can (largely) agree on the weighting to give each piece. This applies both to those in a team doing an evidence synthesis, and to the eventual users of evidence who might also need to assess the quality of individual studies. Over the past 20 years a relatively large number of systems for judging and classifying the quality of evidence have appeared. They may differ in terms of the fields they appear in (sociology, policy, economics, psychology, and so on), and in terms of the types of research they relate to.

A summary of different approaches

A key prerequisite for judging the quality of a research report is that it must be reported well. The reporting must be honest, clear and include all of the factors that might be used to judge the quality of the research being reported. Sadly, this is rare. There is widespread evidence of dishonesty in research publishing, from the existence of 'paper mills' to individuals making up or suppressing data, and in all fields apparently (Ioannidis, 2021; Sabel et al., 2023). There is also considerable publication bias, with larger and more positive outcomes apparently being published more often or more easily. Attempts have been made to reduce some of these problems by pre-registration of studies, and pre-publication of analysis protocols. But these are not really effective (Brodeur et al., 2022; Florez et al., 2023; Sarafoglou et al., 2023), and they only concern some of the issues of trustworthiness (such as publication bias, and misuse of statistics).

Even worse, the quality and comprehensiveness of most research reporting in social science is very poor. Much reporting makes it just about impossible to judge the quality of the study being reported (Gorard, 2021). Therefore, it make sense that there are some checklists for what a study has reported or not, and what has been reported well. If authors do not report well then it is hard to trust their research, and their research should play no part in a

synthesis and generally have no real-life use. However, such checklists are only a precursor to judging quality, and must not be mistaken for anything else (Logullo et al., 2020). Paper mills often produce short, neat papers with very traditional subheadings about materials and methods but the research is weak as well as often being totally made up (I did see one paper in a Russian journal that was an evaluation I had published years earlier with just the authors and place names altered).

One of the earliest and perhaps best-known systems to aid in the judgement and grading of research quality is the Maryland Scientific Methods Scale (Farrington et al., 2002). This scale has five grades of quality ranging from comparisons between naturally occurring groups (lowest quality) to full randomisation (highest quality), such as in randomised trials with low attrition where no statistical manipulation is required (What Works for Local Economic Growth, 2023). The factors suggested in this scale, such as the research design, are clearly relevant to judging research quality, and the Maryland Scale is relatively easy to use. It has two main drawbacks for the purposes described in this paper.

First, by having levels and descriptors for complete studies it assumes that the characteristics of better or worse studies are tied together in a way that is not realistic in practice. In a review, we might want to find a large randomised control trial (RCT) with low attrition and high quality data, but there are many small trials, trials with very high attrition, or with poor or compromised data. Being an RCT in itself does not necessarily mean that other aspects of the study are desirable (Ginsburg & Smith, 2016; Nevill, 2016). Having low attrition does not mean that a study is designed well or collects useful data. And collecting good data does not mean that the study is an RCT. Such factors are independent of each other in a way that the Maryland Scale does not cater for.

Second, the scale was created for a specific subset of research studies – evaluations of interventions. This means that the emphasis is on causation, which is why randomised designs are highlighted (Gorard, 2013). But much research is quite properly descriptive, comparative, or assessing a trend, for example. A good comparative study (about the difficulties faced by children with a disability compared to their peers in school perhaps) would be rated as low level on the Maryland scale simply because the cases were not randomised to be in the disabled group or not. What is urgently needed is a process for judging the quality of all and any empirical research, whether causal or not.

What has happened instead is the creation of a whole set of different criteria for judging the quality of different kinds of research. Some are attempts to help users understand biases and errors in research, such as That's a Claim (https://thatsaclaim.org/). Some are not for judging single studies but assessing larger bodies of work consisting of many individual studies (Gough, 2007), such as systematic reviews of evidence (Madaleno & Waights, n.d.). These are not relevant to this paper, although they include some of the same factors that are discussed below.

For single studies, there is the Cochrane ROBINS-I approach to assessing the risk of bias in non-randomised studies (Sterne et al., 2016). This is still largely about judging causal evidence. There is a Newcastle-Ottawa scale for judging the quality of non-randomised comparison studies (Ottawa Hospital Research Institute(ohri.ca)), which has separate procedures for judging the quality of cohort, cross-sectional and case control studies (Deeks et al., 2003). Each of these makes sense, in the same way as the Maryland Scale, but each has the same two limitations. Using these approaches it is not feasible to synthesise evidence from studies of different designs because some of the factors judged are specific to each design.

It gets needlessly worse. Some commentators try to divide research studies into those that are termed ‘quantitative’ (involving numbers) and ‘qualitative’ (not involving numbers, but text, speech, pictures, sounds, or other sensory data). Sale et al. (2002) makes this split, claiming that only quantitative studies are about science, while qualitative studies are about multiple realities. Because RCTs and even cross-sectional studies are seen as being

somehow 'quantitative', systems have been set up to judge the quality of only research that does not involve numbers (e.g. Stenfors et al., 2020). And there are other systems such as that devised by the Joanna Briggs Institute (JBI) with explicitly different criteria for all 'qualitative' research lumped together regardless of design (© Joanna Briggs Institute 2017 Critical Appraisal Checklist for Qualitative Research (jbi.global)), and for specific designs separately (© Joanna Briggs Institute 2017 Critical Appraisal Checklist for Quasi-Experimental Studies (jbi.global)). Several of the criteria in these schemes are actually about the quality of reporting (see above) not the quality of the research. And they insist that the philosophical perspective or 'positionality' of the researcher must be included in the research report, in a way that the Maryland, Cochrane, Newcastle-Ottawa criteria sensibly do not.

There are three main problems with this attempted separation into qualitative and quantitative work. The split is not based on actual research behaviour, the idea of positionality is a distraction and, above all, it is again important to have scales that work for all research and not just a subset.

All designs, including those mentioned so far such as longitudinal, cross-sectional, or RCT, are independent of the nature of the data collected and analysed. A longitudinal study would be longitudinal whether it repeatedly collected numeric or non-numeric data, or both. Many of the factors considered in judging the quality of research in the Maryland Scale, for example, are applicable to all research – including the number of cases and the attrition level. And almost all research questions could involve collecting both numeric and non-numeric data. The purported split between qualitative and quantitative work cannot be defended – whether on the basis of data collection or analysis, or underlying philosophy such as supposed 'paradigms' (Gorard & Taylor, 2004; Symonds & Gorard, 2010). In fact, observation of researchers actually conducting research demonstrates that they do not approach research differently, whatever they may say about their 'paradigms' or methodology (Kuehn & Rohlfing, 2022; Postan, 1971; Rorty, 1999).

Positionality or reflexivity statements attached to social science research reports are growing and there is pressure from some commentators for all reports to contain them. They arise from the logic and limits of knowledge in the philosophy of science. However, adopting a position via-à-vis knowledge production is not a mere fashion choice, and should arise from a clear understanding of the philosophy of science. It is unreasonable to expect every empirical report to delve deeply into one of the most complex and important areas of philosophy, but it is also unreasonable to expect a few paragraphs at the start of a report to do much other than confuse the reader (and probably the author). Positionality statements therefore appear merely to encourage researchers to ignore the key ideal of impartiality. Impartiality is not helped by adopting a position. Positionality statements are anyway impossible to maintain because they are, presumably, bound by a positionality as well (Savolainen et al., 2023). They lead instead to an infinite regress (Boghossian, 2007).

There are of course multiple perspectives for any research site – we could look at a school in terms of its architecture, heating, lighting, funding, background of students, average attainment, and so on. These perspectives would not be incompatible, any more than would arise from an economist and a psychologist choosing to examine different aspects of a problem. Each account can be accurate without contradicting the other. There can be many 'true' descriptions of a finite set of events – as long as they are consistent with each other. This is the principle of relativity – urging us to examine phenomena from different viewpoints in an attempt to provide a way of expressing any research findings so that they include all of these viewpoints, and would therefore be understandable from each (Turner, 2002). Perhaps most famously, Einstein (1920) produced theories of special and general relativity in physics which can demonstrate both the importance of observer standpoints, *and* how the phenomenon under investigation can be understood/resolved at a meta-level for all standpoints. This is very different from something like relativism – which might assert that

because there are many perspectives then all are inevitably equal, and anything can be or must be true (Gorard & Tan, 2022).

The key point for research synthesis is that weighting each study in terms of its quality or trustworthiness is a key part of assessing the overall picture of the research literature on any topic. We cannot simply ignore research because it is from a different perspective. That would lead to bias. But we cannot assemble the full body of evidence with appropriate weighting if that weighting is substantially determined not by what the research was but by which position was chosen by the author (or reader). We would not then be able to compare purportedly different kinds of research directly. If each perspective uses different criteria to judge the quality of any piece of research then we will not be able to have wide inter-rater reliability or agreement in judgements. This means we would not be able to synthesise all evidence properly, to benefit the public who pay for it and whose lives may be affected by its use. This would not be an ethical position.

In fact, we *can* synthesise all empirical studies. We can judge quality based on the integrity of researchers (see above), methodological transparency and the rigour of the study. Positionality is not needed to judge the quality of research. Issues like clear bias by the researcher, attempts to deceive, and conflicts of interest, might all be important (see below), but these are all to do with the conduct of research *not* the nature of knowledge itself.

The foundation of the process

Given all of the above, there is a need for a process for judging the trustworthiness of individual research studies before synthesis (or real-life use). The process needs to cover all (or almost all) types of social science research, insofar as they are empirical studies. And while recognising the importance of conducting research ethically, ethics is not a major component of judging how good a study is. In fact, it is more appropriate to judge how ethical any study is by judging its robustness (Gorard, 2002).

Judging the quality of a piece of research, or the trustworthiness of its findings, is almost exclusively about its robustness or internal validity. The issue of whether its findings would be more generally true of a larger number of cases that are not in the study is a secondary issue, dependent on internal validity. If a study is not trustworthy, then the issue of whether it applies more generally is moot. In a review of studies the issue of generality is partly addressed by the entire body of evidence assembled, and only after each study has been judged for its quality.

THE BASIS FOR QUALITY JUDGEMENTS

So, what are the general characteristics of a high quality or trustworthy research study, other than clarity and integrity? As illustrated in this section, these characteristics have to be applicable to all (or nearly all) studies, they are largely relative to other studies, and will be based on assuming that all other characteristics are equal (or at least equivalent).

So, for example, it would not make sense to ask as a general factor whether the interviews were conducted well, or the 'effect' sizes calculated appropriately, in any study because issues like these and many others would only apply to a subset of studies. It does not make sense, in this context, to ask whether collecting data online or face to face was better, if the achieved samples and response rates were not equivalent in two studies. If, as may be true, the interviews held online were more efficient and cheaper but interviews held face to face allowed better communication and fieldnotes, then the method of data collection is associated with other differences. Hence, the format of data collection cannot be considered

a general factor ceteris paribus. Note that if everything else, such as time taken, quality of data collected and response rate, was equivalent then just whether data was collected face to face or online does not then appear to make a clear difference to the quality of a study.

Compare this to the issue of scale. If two studies are equivalent in all other respects, such as design, response rate, quality of data and so on, then a study with more cases is considered of higher quality (or the results more trustworthy) than a smaller study. Imagine a survey of 100 randomly selected participants, and another survey of 100,000 randomly selected participants. Both have the same response rate and coverage, and they use the same instruments and methods of analysis. Any findings would be much stronger from the second survey than the first.

One conclusion from that study might be that older survey participants answered differently, on average, to younger participants. Because the number of cases used in the comparison would be so much higher in the second survey the finding about any different responses by age would be much stronger. Whatever else we know about research, and whatever type of research it is, if other factors remain the same then the scale of the study matters. This is not so obviously true for more minor issues such as how the data was collected, and is actually not true for most differences between specific research studies.

Another illustration could be two studies of the views of school headteachers about the curriculum in their schools, as gathered by in-depth interviews. Again, assume that the response rate, quality of transcripts, and other factors are equivalent between the studies. If one study included 60 headteachers and the other included only 6, then the first study would be considered stronger. Any claims made on the basis of 60 interviews would be more generally trustworthy than claims made about 6 interviews. And any claims about sub-groups, such as the headteachers of specific types of schools or from different regions, would be stronger also.

Whenever a comparison is made between groups in a research study the key issue is not the total number of cases, but the number of cases per group. A study that compared the views of 30 headteachers of one kind of school with the views of 30 other headteachers would be stronger than a study comparing the views of three headteachers with 57 others. This is true, even though both studies have an overall scale of 60 participants.

A second general factor in judging the quality of research would be missing data. Each of the studies used for illustration so far would be weaker if the achieved number of participants actually came from a much larger set of participants, many of whom had refused to take part. Or if the participants taking part had refused, or were unable, to answer some key questions. The same situation would occur in a longitudinal study, looking at the same cases over time, if cases dropped out and refused to cooperate after the study had started. Or in an experiment where participants dropped out once they had been allocated to a treatment group. In all of these examples, and many others, the missing data creates the potential for considerable bias in the study results (Gorard, 2020). In summary, if all other factors including the number of cases remain the same, then a study with less missing data will usually be more trustworthy.

The situation is not quite as simple as that, because the source of the missing data also matters. If a survey has a number of missing cases, or cases missing values for key variables, and these occur randomly within the data then the damage is less than if all of the missing data is of the same kind (e.g., from low attaining students) or in the same group in an RCT. So, the judgement has to be about the amount of missing data – because all of it can create bias – and the nature of the missing data – any indication of pattern or skew in the missing data.

Another important characteristic in judging the trustworthiness of research studies would be the quality of the data collected. This is not always under the control of the researcher. In general, simple counts such as the number of pupils attending in a class are more accurate than measurements such as how tall each pupil in a class is (Gorard, 2010). Further,

measurements of height are generally more accurate than tests of attainment (e.g. maths test scores), which are in turn more accurate than measurements of or stories about attitudes (or self-reports of enjoyment or self-esteem, for example). This ranking is partly related to how easy it is to check or calibrate each kind of data. Some kinds of data, such as recollections and attitudes, may not even have anything to check or calibrate with, and so their level of accuracy ('measurement' error) is unknown.

However, researchers do not select research topics on the basis of what is easiest to measure. Therefore, in the areas that they research, the best quality of data that they can collect is not up to them. Nevertheless, this must be a factor in how trustworthy (believable) any study is. And there are other issues to look for, such as whether the instruments used to collect the data, and even the data collection process itself, are independent of the researcher.

A fourth, and perhaps the most important, feature of the quality of any study is the strength or appropriateness of the study design in addressing the research question. This is also the most difficult to judge as there will be a very large number of designs and combinations of designs, some of which are not yet invented (Gorard, 2013). But in the same way that any empirical study needs data, a comparative empirical study with a comparative research question needs data from two or more groups in order to make a comparison. This sounds obvious, but in social science such a basic idea is routinely ignored. The vast majority of social science research appears to draw comparisons or highlight the exceptional nature of one group, but without any direct comparison of two or more groups. A study of underachieving school pupils would typically only include underachievers, providing no direct evidence on whether other pupils have the same characteristics or experiences. A study of criminals would only be of criminals, or a study of homeless people would only include the homeless. Despite this lack of comparator such studies do not limit themselves to descriptive questions or conclusions, but will clearly state or assume that the data collected was somehow specifically true only of the group in focus. Such studies can be largely ignored when assembling evidence of a comparative nature.

Similarly, if a synthesis of evidence concerns comparisons over time, then at least two sets of data are required, each taken at a different time. These collections of data may be longitudinal, collected from the same cases repeatedly, or a time series, collected from successive cohorts, but they must be equivalent in order to make a strong comparison. Whether comparative over space, characteristics, or time, the cases being compared need to be as similar as possible except in terms of the key variable(s). In this way any differences in outcomes or responses can be linked more safely to the key variable. If, for example, a study was comparing the progress in attainment of girls and boys at school over time it would be preferable if the two groups had equivalent attainment at the outset.

If a review of evidence concerns a causal question then designs that explicitly address causation and impact are to be preferred in the individual studies. These might include, but are not limited to, RCTs, regression discontinuity designs, quasi-experiments and matched comparisons. This list appears in something like descending order of trustworthiness for a causal study (RCTs, all other things being equal, are more convincing than natural comparisons). But there will be gradations within each design. An individually randomised RCT is preferred, ceteris paribus, to a cluster randomised trial, for example. And there will be intermediary variations, such as difference-in-difference and instrumental variable approaches (Gorard, 2013).

One reason why judging the suitability and strength of a design is hard is that it is not like the other three factors in being a clear more-or-less issue. Given that all else is equivalent a study that is larger, has better quality data, or less (or less skewed) attrition, is more believable. But no design is intrinsically better than all others. It would be a waste of resource and unethical to conduct an RCT to answer a comparative question. It would be inappropriate to use a longitudinal design to answer a cross-sectional question, but

essential if the question were about change over time. Note that this is also often not an issue about the design used to answer the research question in an individual study, but about how appropriate the individual study design is to answer the new synthesis question. For example, if the synthesis of evidence were about which of several interventions was most successful for a desired outcome then a descriptive study would be of little use, and so would be rated poorly for this purpose. But the study might be intentionally and appropriately descriptive. And it might be rated highly for answering a different review question, of a descriptive nature.

These four appear to be the main factors that are common to all research and have a key influence on how robust any study appears to be – design, scale, missing data and data quality. Other less suitable candidates are discussed briefly below.

CREATING A COMPOSITE JUDGEMENT

How can these four elements be combined to make a composite judgement about any research study in a way that is comparable between studies, reviews and individual reviewers?

A suggested procedure is summarised in Table 1. A version of this 'sieve' first appeared in Gorard (2014). The idea is that it represents a sieve in which a study will sink to its lowest level based on the four factors. The reader starts with the first column, reading down the design descriptions until the research they are reading is at least as good as the descriptor in that row. In this row, they should move to the next column and read down the descriptions, if needed, until the study is at least as good as the descriptor in that row. If the study is better than the descriptor in any subsequent column, the reader stays in that row (not moving up). The reader repeats this process for each column. The final column in the table gives the estimated security rating for that study (between 0 and 4 padlocks). A much fuller description of this process appears in Gorard (2021).

For any column, if it is not possible to discern the quality of the study from the available report(s) then the rating must be placed in a low category (or the study simply ignored).

The overall rating suggests a research finding whose trustworthiness is at least at the level of the descriptions in that row. So, 4🔒 suggests a study that is as secure as could reasonably be expected (because no research will be perfect), and 0🔒 represents a study that is so insecure that it adds nothing safe to our knowledge (the situation for much actual research in practice).

A real example of how the final padlock ratings from several studies could be combined in a simple way is presented in Table 2. This review concerns the existing evidence on the use of financial intervention such as conditional cash transfers to help improve school attendance in developing countries, and comes from a larger project. In the report of that project, the weakest (0 rated) studies are simply ignored as they are judged not to contribute anything to knowledge (Gorard et al., 2022). The review studies are classified in terms of their quality (rows) and whether their results showed a benefit for attendance, no benefit or even harm, or a mixed set of results (columns). This table, unusually for reviews, includes some good studies and most of these are positive. More commonly it has been noted that the weaker studies are more positive.

Of course, this is only one way in which the scores from the sieve could be marshalled. But it is simple and reasonably clear for readers. The next step would be to summarise all studies and describe the key ones (like the 4-rated ones) in more detail. Given this pattern of results, the more definitive claims answering the research questions would be drawn from the 3 and 4🔒 studies.

TABLE 1 A 'sieve' to assist in the estimation of trustworthiness of any research study.

Design	Scale	Missing data	Measurement quality	Rating
Strong design for research question	Large number of cases (per comparison group)	Minimal missing data, no impact on findings	Standardised, independent, accurate	4
Adequate design for research question	Adequate number of cases (per comparison group)	Some missing data, possible impact on findings	Standardised, independent, some errors	3
Weak design for research question	Small number of cases (per comparison group)	Moderate missing data, likely impact on findings	Not standardised/independent, errors	2
Very weak design for research question	Very small number of cases (per group)	High level of missing data, clear impact on findings	Weak measures, high level of error, or many outcomes	1
No consideration of design	A trivial scale of study	Huge amount of missing data, or not reported	Very weak measures	0

TABLE 2 Strength of evidence and impact for studies to improve school attendance via finance.

Strength of evidence	Positive	Unclear/mixed	Negative/neutral
4🔒	2	1	–
3🔒	11	2	–
2🔒	16	7	2
1🔒	11	1	–

DISCUSSION

There are other factors that could be used with the sieve, but it is not clear that any of these are both generic and of considerable importance in the way that design and scale are. For example, in experimental designs the issue of whether participants are 'blinded' as to the purpose of the study is a relevant issue. But participants actually being aware of the study could be important for another design, and make no difference in another. Whether the researcher had a vested interest in the outcomes of a study (a conflict) is another issue that might affect its trustworthiness (Macnamara & Burgoyne, 2023). But assuming that the reporting of the research is honest (see above) then conflicts of interest are not as crucially important as using the right design in many research studies. It is not a good generic factor. Of course, anything spotted when reading a report could be important and should be taken into account. Idiosyncratic factors in any study can affect its trustworthiness and might be used to adjust the basic rating from the sieve. However, once the first four factors are decided on then the other factors either generally fall into line as well, or at least will not make the judgement worse. The sieve is a tough test.

Some researchers have queried why the descriptors in Table 1 are not more precise or prescriptive. For example, one cell in the table says a 'large number of cases' and new users might ask how many that is. The point is that, like the Maryland Scale and others with no numeric thresholds, the sieve is meant to be used with judgement. If a study wanted to know, for example, whether there was a tendency for middle-class parents to use different criteria to working-class parents when selecting a school for their child, consider how many cases you would feel happy with in order to trust the results. If there were two cases in the study – one middle- and one working-class parent – clearly no one should trust the results. If there were 2000 cases, 1000 of each, then this could clearly be an adequate number of cases (assuming that all other factors like data quality are also adequate).

Therefore, any thresholds between a very trivial scale and a good or adequate number of cases lie somewhere between 2 and 1000. A good number of cases surely represents hundreds of cases per comparison group. And perhaps 20+ per group represents a 'very small' scale rather than a 'trivial' scale. As already stated, precise agreement on every column is not essential. Perhaps it is unclear whether 100 cases is adequate or small. The next column might reveal 50% attrition or missing data, in which case that study must be rated as no more than 2🔒 (if that), and the borderline issue about the scale is solved. There is no magic number for any threshold of judgement. It would not be reasonable to claim that 400 cases was a good size, but that 399 was not. Note, some schemes use power calculations or similar to help assess sample sizes for studies. This approach is based on the obsolete idea of significance testing, that does not work in practice, and cannot be applied to most real-life samples, including population data, convenience samples and incomplete random samples (Gorard, 2021).

Generally, discussion within teams of reviewers can resolve any boundary queries. It is important that reviewers envisage genuinely caring whether the findings of a study are safe. They must try not to be biased by preconceptions, whether they agree with the findings, or

by the kind of research involved. The standard of evidence looked for must be at least as strong as reviewers would accept for established 'facts' in real life.

The sieve has now been used widely and formally by research teams to assist in their reviews of evidence since 2014 (e.g., El Soufi & See, 2019; Fan & See, 2022; Huang & Chalmers, 2023; Neelen & Kirschner, 2020; Owen et al., 2022; See et al., 2022; Siddiqui & Ventista, 2018), as well as by countless PhD students in their theses. It has been explained in videos for specific areas like evidence on second language learning (Chalmers, 2016), and discussed widely on social media in fields beyond education and social science (e.g., https://twitter.com/kmyersfilm/status/1635913895981268994?s=20). It was adopted by the Education Research Foundation in England as the basis for its security ratings (Classifying_the_security_of_EEF_findings_2019.pdf (d2tic4wvo1iusb.cloudfront.net)), although the EEF made unwarranted changes—such as dropping the requirement for high quality data, and permitting overly complex reports that their intended users are unable to read. And it was used by the Education sub-panel for its decisions in the Slovak Periodic Assessment of Research (REF) in 2022 (Periodic Assessment of Research, Development, Artistic and Other Creative Activities, Ministry of Education, Science, Research and Sport of the Slovak Republic (minedu.sk)).

Users have reported good agreement (over 90%) on initial ratings, never differing by more than one padlock/grade. Although the system is used to assess individual studies, the purpose is to assist with syntheses of overall bodies of evidence. Therefore, a difference of one grade in one study may not matter much when reporting the substantive findings from that overall evidence.

It is important that syntheses of evidence, including traditional literature reviews, adopt transparent quality judgement to decide how much weight to give each prior study, along with systematic searches for an unbiased set of studies. A generic process, without technical clutter, for judging the relative quality of individual studies before synthesis or use is clearly needed. This is to help reviewers, and also to help users not to be taken in by the nonsense of research involving neologisms or overly complex analyses. The sieve described here or something like it would be suitable based on its logic and the testing it has had in the field.

CONFLICT OF INTEREST STATEMENT

The author is the developer of the approach described here.

FUNDING INFORMATION

No funding was received for this work.

DATA AVAILABILITY STATEMENT

Data sharing not applicable to this article as no datasets were generated or analysed during the current study.

ETHICS STATEMENT

No ethical approval was needed for this conceptual work.

ORCID

Stephen Gorard https://orcid.org/0000-0002-9381-5991

REFERENCES

Bergeron, P. (2022). How to engage in pseudoscience with real data. *McGill Journal of Education*, *51*, 2.

Boghossian, P. (2007). *Fear of knowledge*. Oxford University Press.

Brodeur, A., Cook, N., Hartley, J., & Heyes, A. (2022). *Do pre-registration and pre-analysis plans reduce p-hacking and publication bias?*SSRN.

Chalmers, H. (2016). *Assessing the trustworthiness of What Works research*, L3xiphile. (wordpress.com).
Cirkony, C., Rickinson, M., Walsh, L., Gleeson, J., Salisbury, M., & Cutler, B. (2022). Reflections on conducting rapid reviews of educational research. *Educational Research*, *64*(4), 371–390. https://doi.org/10.1080/00131881.2022.2120514
De Vrieze, J. (2019). *What science reporter should know about meta-analyses, medium*.
Deeks, J., Dinnes, J., D'Amico, R., Sowden, A., Sakarovitch, C., Song, F., Petticrew, M., Altman, D. G., & International Stroke Trial Collaborative Group; European Carotid Surgery Trial Collaborative Group. (2003). Evaluating non-randomised intervention studies, *technology. Health Technology Assessment*, *7*(27), 1–173, iii–x.
Einstein, A. (1920). *Relativity: The special and general theory*. Henry Holt and Company.
El Soufi, N., & See, B. H. (2019). Does explicit teaching of critical thinking improve critical thinking skills of English language learners in higher education? A critical review of causal evidence. *Studies in Educational Evaluation*, *60*, 140–162.
Fan, K., & See, B. H. (2022). How do Chinese students' critical thinking compare with other students': A structured review of the existing evidence. *Thinking Skills and Creativity*, *46*, 101145.
Farrington, D., Gottfredson, D., Sherman, L., & Walsh, B. (2002). *Evidence-based crime prevention*. Routledge.
Florez, M., Jaoude, J., Patel, R., Beck, E. J., Taniguchi, C. M., Minsky, B. D., Fuller, C. D., Lee, J. J., Kupferman, M., Raghav, K. P., Overman, M. J., Thomas, C. R., Jr., & Ludmir, E. B. (2023). Incidence of primary end point changes among active cancer phase 3 randomized clinical trials. *JAMA Network Open*, *6*(5), e2313819.
Flynn, N. (2019). Facilitating evidence-informed practice. *Teacher Development*, *23*(1), 64–82.
Ginsburg, A., & Smith, M. (2016). *Do randomized controlled trials meet the "gold standard"?* Do-randomized-controlled-trials-meet-the-gold-standard.pdf (carnegiefoundation.org).
Gorard, S. (2002). Ethics and equity: Pursuing the perspective of non-participants. *Social Research Update*, *39*, 1–4.
Gorard, S. (2010). Measuring is more than assigning numbers. In G. Walford, E. Tucker, & M. Viswanathan (Eds.), *Sage handbook of measurement* (pp. 389–408). SAGE.
Gorard, S. (2013). *Research design: Robust approaches for the social sciences*. SAGE.
Gorard, S. (2014). A proposal for judging the trustworthiness of research findings. *Radical Statistics*, *110*, 47–60.
Gorard, S. (2020). Handling missing data in numeric analyses. *International Journal of Social Research Methods*, *23*(6), 651–660.
Gorard, S. (2021). *How to make sense of statistics: Everything you need to know about using numbers in social science*. SAGE.
Gorard, S., & Chen, W. (2023). What is the evidence on research-informed education? In *BERA/SAGE Handbook on Research-Informed Education Policy and Practice, London BERA*. SAGE.
Gorard, S., See, B. H., & Siddiqui, N. (2022). *Making schools better for disadvantaged students*. Routledge.
Gorard, S., & Tan, Y. (2022). The difficulty of making claims to knowledge in social science. *Social Sciences Journal*, *28*, 170–202.
Gorard, S., & Taylor, C. (2004). *Combining methods in educational and social research*. Open University Press.
Gough, D. (2007). Weight of evidence: A framework for the appraisal of the quality and relevance of evidence. *Research Papers in Education*, *22*(2), 213–228.
Hattie, J. (1992). Measuring the effects of schooling. *Australian Journal of Education*, *36*(1), 5–13. https://doi.org/10.1177/000494419203600102
Huang, X., & Chalmers, H. (2023). Implementation and effects of pedagogical translanguaging in EFL classrooms: A systematic review. *Language*, *8*(3), 194.
Ioannidis, J. (2021). Hundreds of thousands of zombie randomised trials circulate among us. *Anasthesia*, *76*, 444–447.
Kuehn, D., & Rohlfing, I. (2022). Do quantitative and qualitative research reflect two distinct cultures? *Sociological Methods & Research*. https://doi.org/10.1177/00491241221082597. Online ahead of print.
Logullo, P., MacCarthy, A., Kirtley, S., & Collins, G. (2020). Reporting guideline checklists are not quality evaluation forms. *Health Science Reports*, *3*(2), e16.
Macnamara, B., & Burgoyne, A. (2023). Do growth mindset interventions impact students' academic achievement? A systematic review and meta-analysis with recommendations for best practices. *Psychological Bulletin*, *149*(3–4), 133–173.
Madaleno, M., & Waights, S. (n.d.). *Guide to scoring methods using the Maryland Scientific Methods Scale*. https://whatworksgrowth.org/public/files/Scoring-Guide.pdf
Neelen, M., & Kirschner, P. (2020). *Truth or truthiness? Analysing a VR study using Gorard's sieve, 3-star learning experiences*. (wordpress.com).

Nevill, C. (2016). *Do EEF trials meet the new "Gold Standard"?* EEF Blog: (educationendowmentfoundation.org.uk).
Nutley, S., Boaz, A., Davies, H., & Fraser, A. (2019). New development: What works now? Continuity and change in the use of evidence to improve public policy and service delivery. *Public Money & Management*, *39*(4), 310–316.
Owen, K., Watkins, R., & Hughes, C. (2022). From evidence-informed to evidence-based: An evidence building framework for education. *Review of Education*, *10*, e3342.
Polanin, J., Zhang, Q., Taylor, J., Williams, R., Joshi, M., & Burr, L. (2022). Evidence gap maps in education research. *Journal of Research on Educational Effectiveness*, *16*(3), 532–552. https://doi.org/10.1080/19345747.2022.2139312
Postan, M. (1971). *Fact and relevance: Essays on historical method*. Cambridge University Press.
Rorty, R. (1999). *Phony science wars, review of hacking, I. The social construction of what?* Harvard University Press. The Atlantic Monthly online, November 1999.
Sabel, B., Knaack, E., Gigerenzer, G., & Bilc, M. (2023). *Fake publications in biomedical science: Red-flagging method indicates mass production*. (medrxiv.org).
Sale, J., Lohfeld, L., & Brazil, K. (2002). Revisiting the quantitative-qualitative debate: Implications for mixed-methods research. *Quality and Quantity*, *36*, 43–53.
Sarafoglou, A., Hoogeveen, S., & Wagenmakers, E.-J. (2023). Comparing analysis blinding with preregistration in the many-analysts religion project. *Advances in Methods and Practices in Psychological Science*, *6*(1), 251524592211283.
Savolainen, J., Casey, P., McBrayer, J., & Schwerdtle, P. (2023). Positionality and its oroblems: Questioning the value of reflexivity statements in researchm. *Perspectives on Psychological Science*, *18*(6), 1331–1338. https://doi.org/10.1177/17456916221144988
See, B. H., Munthe, E., Ross, S. A., Hitt, L., & El Soufi, N. (2022). Who becomes a teacher and why? *Review of Education*, *10*, 3.
Siddiqui, N., & Ventista, O. (2018). A review of school-based interventions for the improvement of social emotional skills and wider outcomes of education. *International Journal of Educational Research*, *90*, 117–132.
Stenfors, T., Kajamaa, A., & Bennett, D. (2020). How to… assess the quality of qualitative research. *The Clinical Teacher*, *17*(6), 596–599. https://doi.org/10.1111/tct.13242
Sterne, J., Hernan, M., Reeves, B., Savovia, J., Berkman, N., Viswanathan, M., Henry, D., Altman, D. G., Ansari, M. T., Boutron, I., Carpenter, J. R., Chan, A. W., Churchill, R., Deeks, J. J., Hróbjartsson, A., Kirkham, J., Jüni, P., Loke, Y. K., Pigott, T. D., … Higgins, J. P. (2016). ROBINS-I: A tool for assessing risk of bias in non-randomised studies of interventions. *BMJ*, *355*, i4919.
Symonds, J., & Gorard, S. (2010). The death of mixed methods?: Or the rebirth of research as craft. *Evaluation and Research in Education*, *23*(2), 121–136.
Torgerson, C. (2003). *Systematic Reviews*. Continuum.
Turner, D. (2002). *The class struggle: The place of theory in education?, inaugural lecture*. School of Humanities and Social Sciences, Glamorgan University.
Wadhwa, M., Zheng, J., & Cook, T. (2023). How consistent are meanings of "evidence-based"? A comparative review of 12 clearinghouses that rate the effectiveness of educational programs. *Review of Educational Research*. https://doi.org/10.3102/00346543231152262. Online ahead of print.
What Works for Local Economic Growth. (2023). *The Maryland scientific methods scale (SMS) - What Works Growth*. https://whatworksgrowth.org/resource-library/the-maryland-scientific-methods-scale-sms/

How to cite this article: Gorard, S. (2024). Judging the relative trustworthiness of research results: How to do it and why it matters. *Review of Education*, *12*, e3448. https://doi.org/10.1002/rev3.3448