Judgement in Measurement and Analysis

Stephen Gorard

Durham University, UK

DOI: https://doi.org/10.47772/IJRISS.2026.10100109

Received: 04 January 2026; Accepted: 09 January 2026; Published: 23 January 2026

ABSTRACT

True measuring scales behave in the same way as the real-life things that they are measuring. They also permit an estimate of the level of error in making a measurement, through calibration. Where these two characteristics are not present, then a purported measurement is not a true one. Numbering is not the same as measuring. Errors in measurement also propagate in calculations, but without a true measure we cannot tell how. There is no technical or statistical solution to this. When judging the trustworthiness of the findings from a piece of research, the quality of the measurement used is an important criterion. Without this knowledge, research cannot be trusted. Therefore, analyses and the trust placed in them must be based on appropriate judgement.

Keywords: Data quality, error propagation, fake measures, inferential statistics, judgement, judging research quality, latent measures, measurement calibration, reliability, representational errors

Key points

- Measurement is more than numbering
- Some “measures” common in social science are not trustworthy
- There is no technical way of overcoming these limitations
- Only transparent judgement can be used to assess the quality of education research using measurement

SUMMARY

This paper considers the kinds of measurements used in social science research. Not everything that is numbered is, or can be, a measurement. A measurement scale must behave in a very similar fashion to the thing being measured. If it does not, or it is not clear that it does, then it is not a true measurement. Where calibration of the measure with the thing being measured is not possible, because the thing being measured is not directly accessible, then there is no true measurement. The fact that different attempts to measure something are consistent with each other is of no help in establishing the validity of a purported measure. Without a true measure we cannot know what the range of error is in measurement, or even whether there is anything to measure at all.

All measurements will have errors, even if only in operation, and studies tend to have missing values or cases. These initial errors propagate when used in analyses, usually in unnoticed ways, often making the results meaningless. For these reasons, any analyses based on conditional probabilities are inappropriate. There is no technical solution to these issues. Analyses and the trust placed in them must be based on appropriate judgement. The paper discusses why and how.

When judging the trustworthiness of the findings from a piece of research, the quality of the data/measurement used is an important criterion. If we are unable to judge the quality of data, then we are unable to judge the quality of any research using that data. In that situation, the research must be deemed untrustworthy.

Introduction to measurement in social science

The practice of measurement, and the analysis of measured variables, are both relatively common in social science research. Repeated reviews and surveys have shown that work not involving numbers, sometimes misleadingly referred to as "qualitative", dominates social science output (Smith et al. 2025). Nevertheless, work referred to as "quantitative" forms a substantial subset of research reports. In fact, these divisions or classifications of research are of no help and may even be harmful (Gorard 2025). Most studies involve some form of numeric assessment (even if just saying that something is "more common" than something else). And all studies involve judgement of qualities.

Part of the damage created by the false dualism of quantitative and qualitative work is that commentators might envisage all "quantitative" work as involving measurement, rather than just using numbers for categories or words. Measuring involves numbers, by definition, but not all numbers used in research are measurements. This paper looks in more detail at what this means, and the possible implications for judging the strength of research claims. It starts with a consideration of what measurement is.

Measurement

Despite its prevalence in research, the idea of measurement itself is seldom discussed (except perhaps in creating educational assessments). There is plenty of writing and commentary on how to collect numeric data, how to assess it, analyse and present it. Standard texts tend to focus on supposed levels of measurement, or on a specific form of reliability, and on inferential statistical analysis, or complex modelling. These issues are generally and incorrectly dealt with technically (rather than in terms of the judgements and compromises needed). There tends to be little about the construction, meaning and validity of the measurements involved. These issues are somewhat taken for granted by many commentators. The process of measurement is just assumed to be understood by all.

In fact, when we say we are measuring something we tend to assume a number of things before we get to the measure itself (Berka 1983). For example, the thing that we are measuring must exist. Measuring something that does not exist makes no sense. Second, it must be observable, or at least its existence must be observable from the way that it influences the environment. It would be fake "measurement" to measure something that could not be observed, directly or indirectly, in some way, because we could not then associate the measurement scale with these variations. Not everything can be measured (and not being able to measure something does not make it unimportant). Third, there must be a regularity or pattern to the characteristics of the thing we are measuring. Measuring involves mapping this pattern of observed characteristics or behaviour to a set of numeric values. So, we need detailed information about the characteristics or behaviour of the thing we are measuring. Fourth, we also need to create a measurement scale, with steps in it that match the observed behaviour of the thing we are measuring. The behaviour of the thing we are measuring and the scale used to measure it must be analogous. The scale must be standard, so that measurements taken at different times or in different places are comparable. It should have units that can be directly related to the properties and behaviour of the thing being measured. In this way, the measurement can be calibrated, and we can estimate how accurate our measurements are, and how large the error component is. It is important, if obvious, for what follows that measurement involves the identification of two distinct things - the object of measurement and the measuring scale for it.

With well-established and simple measures (like counts of objects) it may be reasonable to ignore these pre-technical steps of creating a measure before using it. But it is still salutary to look back on them from time to time. For new measures and those that are complex or difficult to calibrate, it is essential to consider these first steps. The danger in ignoring the process of measurement is that researchers can then mistakenly imagine that the mere allocation of numbers to something constitutes a measurement (Gorard 2010a). If one of the elements of a purported measurement is lacking then the measurement could be deceptive and misleading. Numbers can quite properly be used as names for objects or categories, such as the serial numbers on car engines, but this does not imply the existence of any quantity that can be measured. Not all number systems are actually measurements.

Some examples of measurements

Counts might be the simplest examples of measurements, measuring the number of things in a certain category. These could be frequencies or percentages, for example. The quality of a count depends largely on the certainty with which we can identify cases to count, and the accuracy of the counting process. Counting a small number of people in a room should yield an accurate answer. People are generally easy to distinguish from anything else, and we have a shared understanding of what a person is. We can sense when a room is more or less crowded, and associate this observation with a simple numeric scale that others will understand. The count can vary easily as people enter or leave the room, or we can close the door. And we can get several observers to repeat the count as a check. There are no instruments and no estimates involved. This would be a true numeric measuring scale.

Counting the number of people in a busy metropolitan railway station would be much harder, not because the cases are necessarily harder to identify but because the turmoil creates a danger of missing cases or counting some twice as they move around. Counting the number of objects in a small room is also difficult, because it is hard to define what an object is. But our simple measuring scale itself is still valid, even if harder to use in some circumstances. In social science research both frequencies and percentages are useful for description and comparison. The measuring scales involved are clearly robust. Most of any difficulty comes from identifying cases accurately (such as when looking at frequencies of people being in specific occupational groups, or having a particular ethnic origin).

Measuring the heights of the people in a small room is also feasible. Length is a conceptual measure not a count. It is based on observations that things protrude more or less than others. Standing in a room, the tops of some peoples' heads protrude higher than others. From this we could rank them in order, from shortest to tallest. We can capture the amount of variation with a standard ruler, and so associate the amount of protrusion with a measurement. Our standard could be a fixed case, such as the shortest person, or we could create/use a standard unit like a metre. We can measure people more than once as a check, using this standard, and we can get several people to make the measurements. Of course, this scale is harder to use than a count, and so will yield more errors on average. Peoples' height varies with their stance, time of day, their shoes, and so on. However, once we have agreed on a standard, we can use it to assess those variations and so, improve our recorded measurements. Again, we have a real measuring scale, that can be calibrated by comparing the measurement scale and the proximity of peoples' heads to the floor or ceiling.

Explicitness is key to forming and using a good measure – and one that is reliable in the true sense that it can reach the same result when used properly in the same situation by more than one researcher. If there is a dispute about the number of people in a room, it is possible to settle the matter by reference to the identifiable and separate things being measured. We have the measure (such as seven people) and we have bodies in a room that we can line up, and count repeatedly, until a consensus view is arrived at or a judgement becomes more general. Of course, there can still be errors of measurement, but in principle these can be resolved by direct comparison of the figures and manifestation of the quality being measured. Resolving a dispute about the heights of people is more complex than one about how many people there are, but still possible. We can get others to compare the ruler (standard of height) with the lengths of the people. In each example, we have explicit, stable and easily re-counted evidence of a phenomenon and we have a separate system of measures. We can compare the two.

How long each person has been in the room is more complex again. We either have to monitor entries and exist scrupulously, or ask people when they entered if monitoring is not possible. This will inevitably be less accurate than a count but because of the procedure involved, not the measure itself. The actual measure of elapsed time is, like height, well-agreed, reliable and useful.

All three examples so far are relatively simple and robust. All of these scales have a clear zero value, are equal interval, and ratio (2 metres is exactly twice as much as 1 metre). And all can be calibrated with the quality of presence, length or duration that is observable separately to the measure of these things. There is isomorphism between the physical characteristics and the measuring scales. These examples were described in order to help provide a contrast with the more complex scales commonly used in social science. Perhaps an extreme example of a purported measurement with less security would be a psychometric or personality scale. As soon as researchers go beyond relatively simple measurements, such as the observable number of pupils in a school, to latent and unobservable variables then their measurement problems and concerns will multiply.

Less clear measures

The problem with trying to measure less explicit phenomena such as attitudes, self-esteem, motivation, metacognition, growth mindset, and so on, is that we usually have only the "measurement" from a test devised for each phenomena. These same phenomena such as state of mind can have completely different tests. Unlike measures like length there are no standards, such as a metre. There is not even a scale really. It is not possible to check such measurements with the observable thing being measured, for validity, calibration, or measurement error. What has developed instead, because researchers cannot compare their purported latent measures with anything observable, is an obsession with a kind of internal coherence. The coherence sought is within groups of questions that are intended to become measures of a latent variable when responses to all of them are put together.

Supposed reliability of latent measures

For example, Nunally (1975) suggests that best way to develop an instrument like a questionnaire that correlates (calibrates) highly with the thing being measured is to make the items in the questionnaire homogeneous. "One should select those items that correlate well with the test as a whole and throw out those items that do not…" (p.9). This kind of bold claim actually confuses reliability for validity. It proposes making all of the items in a questionnaire used to measure a latent variable the same as, or very similar to each other, without regard to whether they are actually measuring the thing they are meant to. Using this logic, a good measure of health, for example, can be obtained by asking a range of very similar questions about trust, or wealth, or education, or whatever. As long as the multiple questions yield similar responses then this completely irrelevant measure should be deemed a good one, according to Nunally (1975) and many other commentators.

However, this consistency or reliability does not even make sense on its own terms. With a real measure like length or a count, we can assess the level of measurement error by repeating the measurement using a specified process, and with different operators or at different times. This yields an idea of how accurate the measurement is, and the importance of this is discussed in a later section. This kind of test-retest reliability for a measure is important. In the measurement of latent variables, a much weaker kind of reliability is used (Gorard 2021). This is simply an estimate of the extent to which all survey items used to measure a latent variable are actually asking the same question over and over again. It is not clear why this is desirable. It is not a test-rest situation, and yields no calibration or estimate of measurement error. This is largely because each item is not the measure. The measure is meant to be the aggregate of all items.

It is not clear that asking the same thing many times in slightly different ways, not to calibrate a measure but to create the measure in the first place, will improve accuracy. If the responses to each item used to create a latent variable are biased, or incorrect, or the respondent does not know the answer or what they want the answer to be, then having several versions does not help. The idea only make sense if errors in the responses are random, so that repeated questions will somehow overcome these random errors. The idea also only works when the same respondent gives different or contradictory responses to two or more questions. Nothing is gained if each respondent gives the same or consistent responses to each item – here one item would have been sufficient. If the majority of respondents give a correct response to the first item, then random errors will mean that the majority of incorrect responses to the second item will be from people with correct responses to the first item. Only a small number of correct responses will come from those who gave an incorrect response the first time. Correct here means as they would, on reflection or after checking, wish to respond. Asking a question more than once therefore reduces rather than increases the accuracy of evidence (see worked example in Gorard 2010a). It also has an opportunity cost meaning that other items cannot be included, and leads to boredom for the respondent.

This tradition means that social science research has adopted almost unquestioningly a large number of apparently quantified things that have not been demonstrated via calibration to be good and accurate measures of anything. They cannot be calibrated with anything except themselves. If the thing purportedly being measured can only be seen via the measure then it may not actually exist. Or it may not behave in the way that the measure is designed to do, analogously. Perhaps the most famous example is IQ meant to measure intelligence, and the tautologous definition that "intelligence is what IQ tests measure" (Prandy 2002). If this definition were true, then IQ cannot be a real quantity for the reasons given so far. Further examples might include attitude measures, or constructs like the self-concept. These are also questionable in terms of what they measure, and whether the scales used do actually behave in the way that the underlying thing does.

Clearly, these ideas for measures can be correlated with more overt observations. Intelligence can be correlated with success in other cognitive tasks or assessments, attitudes with can be correlated with habits, motivation with revealed action, and anxiety with physiological indicators. But this generally yields three things to compare and consider, not just two. These could be a “measure” of attitude, for example, the unseen attitude being measured, and the habits associated with the unseen attitude. So, then the question arises why the attitude concept is needed. If it depends only on validation from a pattern seen in observable habits, then measures of habits could be used instead of measures of attitudes. This does not happen with true measures like height, based on observance of protrusion or length (Gorard 2002), and a measure that behaves isomorphically with it. No third correlated measure is needed to try and validate the first measure.

Additionally, in many real-life examples there is no substantive correlation between latent measures and their observable correlates. Students’ reports of attitudes to learning are weakly but negatively related to actual decisions to study science in the future (Gorard and See 2009). Aspirations to attend university are not good predictors of actual attendance at university, or even of application to university, for example.

Implications of poor measurements

When a measurement has an unknown level of measurement error this has serious consequences for its use in research in a variety of ways, dealt with below.

The behaviour of errors

Measurement error is almost inevitable, but does not mean that a measuring scale is not analogous to the thing being measured. Put another way, it is not the presence or absence of error that makes a measuring scale a good one. However, it is crucial to know whether we have the correct number when measuring, or how far from the correct number any measurement is likely to be. If this is not possible then considerable problems ensue.

Errors involved in using a good measure can come from copying, misreading, miscommunication, or missing data. They occur simply in converting number bases, as when entering denary numbers into a computer or calculator (e.g. the simple denary fraction 0.1 cannot be represented exactly in binary). Then, whatever the initial error is it will tend to propagate when used in any calculation (Gorard 2010b). Adding two numbers also adds the errors in both of them, and so on. The more complex a calculation is the worse this situation gets, and the harder it is to track. Some measures will end up consisting almost entirely of error, in what are termed “ill-conditioned” calculations. It is good practice to check whether this is happening.

One simple example is a difference-in-difference calculation commonly used in social science. This could involve calculating a change over time in each of two groups. The change in each group will be smaller than the numbers that generated it, but will contain the error component of both numbers used to compute the change. The error relative to the answer will have grown considerably – a larger absolute error in a much small answer. The next step in the difference-in-difference calculation is to find the difference in the changes between the two groups. This will tend to lead to an even smaller final answer, which now includes the errors from both change figures. At this stage the error can be larger than the purported answer. For the sake of research quality, it is important to check.

Similar problems arise in even simpler cases. For example, in a simple trial, a researcher finds a mean score for the intervention group of 80, and a mean score for the control group of 70. They want to conclude that the intervention group has done better. If both means are considered 90% accurate, then the real figure for the intervention group is between 72 and 88, and for the control between 63 and 77. The difference between the answers could therefore be +25 or -5. Each result is equally likely. The range of possible answers is 30, which is three times the size of the achieved answer (10). In spite of measurements that are 90% accurate, a reasonable figure for social science or assessment, and an apparent difference between groups, it is not clear whether there is actually a difference or whether it is positive or negative. The measurement error in the original measures has been propagated by the ensuing simple calculation. The same would happen when looking at trends, or other patterns in data. The situation is far worse in more complex analyses, which is one reason why simple calculations and models should generally be preferred (Gorard 2013).

There is no level of measurement error that avoids this problem, and the propagated error is only partly related to the initial error amount. But the problem is so much worse if there is no way of estimating the initial error. This is one reason why not being able to calibrate a supposed measure is important. With no idea of the scale of the initial error in the measurement it is not possible to check the maximum error propagation, making any ensuing research results almost worthless. However, such issues are largely ignored in researcher training, and in methods resources.

Judging differences, trends and patterns

When answering a typical analytic question, such as ‘is there a real difference between two figures’, we need to take into account, in a way that traditional analysis simply ignores--the likely scale of any errors. Most texts and resources on data analysis do not consider either errors in measurement or error propagation. But all numeric analysis, from simply stating what a measured amount is to complex consideration of differences between groups, is affected by errors. Instead, methods resources tend to focus on random variation and how this might have affected the results. This is a mistake. Hardly any social science research involves randomisation or random error. A single measurement may be in error, but that error is unlikely to be random (Gorard 2010a). Multiple cases are seldom selected randomly – rather than being population data or convenience samples – and even where they are, dropout and missing values make them non-random again (Gorard 2020). This means that most of what researchers learn about analysis is mathematically incorrect, and therefore misleading. Randomisation is an absolute requirement for calculating standard errors, significance tests, or confidence intervals.

This requirement is clear because it is assumed in the calculation of inferential statistics. A significance test, like a t-test for example, is used to assess the probability of a null hypothesis usually of the form – any observed difference between two groups arises solely from their randomisation. If the probability from the test is low this hypothesis is rejected – it is decided that not all of observed differences between two groups are due to randomisation (Fielding and Gilbert 2000). This process would be pointless with non-random cases, because it is already clear that any difference cannot be due solely to the random nature of the groups. The groups are already known to be not random. This process is also misleading because some less astute readers will assume that something of value has been done, and that the results are somehow stronger or more scientific as a result. This is dangerous, for research progress, systematic reviews and so on.

However, that is not the end. Even with random cases or values, none of the inferential statistics apparatus yields a useful answer (Nickerson 2000). An analyst might want to know whether their results were created by chance, but statistics will only compute the probability of their results if they were created by chance. The analyst asks, given the data I collected what is the probability that it arose by chance? The statistical test replies, given that the results arose only by chance this is the probability of getting the data you collected. That is of no help. However, there is widespread dishonesty with commentators and researchers pretending that the probability of any data given that it arose by chance is the same as the probability of that data arising by chance. Again, it misleads the unwary into thinking that this peculiar process somehow validates the findings or makes then more scientific. In fact, it is the opposite of scientific. based on a fundamental misunderstanding of probability and inference (Rozeboom 1960).

It is not simply possible to convert the probability of the data given randomisation (what significance tests compute) into the probability of the results being due to randomisation given the data found (what analysts are looking for). The two probabilities are very different, and one can be large and the other small, or any combination. It is not possible to directly assess one from knowing the other. They are mathematically related by Bayes’ Theorem. However, to use this involves knowing the unconditional probability of the null hypothesis being true. If a researcher knew this then there would be no need to do the research. It is mathematical curio, of no relevance to empirical research.

All of this needs to be swept away. Instead, we need to build analyses based on judgement, about the nature of our numeric findings, and their robustness in light of initial and propagated measurement errors. This is simple, because it is not technical, but it is not easy or lazy, because there is no formulaic or button pressing solution.

The “new” statistics suggests using effect sizes with confidence intervals instead of significance tests (Cumming 2014). This idea has several fatal problems. Confidence intervals have the same flaws and are based on the same flimsy foundation as significance tests. Actually, they are worse because they are even less well understood (usually as a range within which a true value is likely to be). Effect sizes can be helpful, but most are seldom used – such as R squared or odds ratios. The most common form of a standardised difference between means (e.g. Cohen’s d) requires a number of assumptions that are rarely assessed, such as the distribution of the two sets to scores. Use of effect sizes in a range of situations, including in passive comparison designs, needs further development.

One promising approach is the use of sensitivity tests to help judge how robust any finding is (Gorard 2021). It is preferrable if the results of any sensitivity test can be directly compared to the level of measurement error, or the number of missing cases/values. One such is the number of counterfactual cases needed to disturb a finding (NNTD). It is simply a count of how many cases would need to be removed from one group, or changed into a counterfactual value, for whatever result is going to be reported to disappear. This requires no assumptions or distributions. It can be used with randomised, population, incomplete and convenience cases. It is also easy. For example, in a study using a standardised difference between means, NNTD would be this “effect” size multiplied by the number of cases in the smallest group being compared.

One reason why some commentators are resistant to such simple approaches is a mistaken belief that inferential statistics can tell us something about the generalisability of a finding. They cannot (see above). Drawing one marble randomly from a bag of many marbles with an equal number of two colours inside, it is possible to work out the probability of getting a marble of one colour. This is analogous to what a significance test does. Drawing a marble of a specific colour does not say anything about the colour of the marbles remaining in the bag if we do not already know the number and colour of all marbles in the bag. That would be generalisation, and it is not possible unless the answer is already known. This limitation applies to any number of marbles drawn (up to the total number).

Judging the generality of a research finding is anyway a secondary step. First, we need to know if the finding is trustworthy (Gorard 2006). Should we believe it? This is what the next section discusses. Once we are happy with the validity of a result/measurement then we may want to know if it applies in other contexts, or to other cases not involved in the research. It would be absurd to worry about whether a false or invalid result would also be true for other cases. Most studies in social science are of poor quality, and so the issue of how general their findings should never arise.

Judging the trustworthiness of research findings

A further implication of not knowing the level of error in a measurement is that any study using this measurement cannot be judged appropriately in terms of its quality. There has been considerable criticism of some systematic reviews and meta-analyses of existing research evidence, because they have not engaged with the quality of the underlying studies (de Vrieze 2019). It has been demonstrated that simply giving equal weight in a synthesis to studies of different quality yields invalid and misleading overall results (Gorard and Chen 2023).

However, many proposed schemes to judge research quality are specific to one design such as a randomised control trial – for example the Maryland Scale (Farrington et al. 2002). There is a Newcast;e-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). 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). For example, the Joanna Briggs Institute (JBI) has explicitly different criteria for all “qualitative” research lumped together regardless of their design (© Joanna Briggs Institute 2017 Critical Appraisal Checklist for Qualitative Research (jbi.global)). Several of the criteria in these schemes are actually about the quality of reporting not the quality of the research. But their biggest drawback is they are limited to only one kind of study.

Instead, Gorard (2024) proposed a simple template for assessing the quality of any piece of work, based on the four most important generic factors. These are the suitability of the study design for the question(s) being addressed, the scale of the study, the level of attrition or missing data, and the quality of the data used. These appear to be key to the quality of any study. An attempted comparison without a comparative design would be invalid. Whatever the study, all things being equal, the larger it is the stronger it will be. Any missing values can introduce bias into results and so reduce the validity of a study. And the better the data/measurement used the stronger the study is. Each of these factors is considered of equal weight, and the lowest score for any determines the overall score for the study.

This means that a study with a weak measure is necessarily a weak study overall. Having a good design, large scale and low attrition does not compensate for a result based on a weak measure. Where we cannot tell now accurate a measure is, the study involved will be given the lowest quality score (the same as if the scale or attrition were not reported).

Rethinking the use of measurement

Judgement is therefore key throughout the research and measurement cycle - from how accurate a single measure is, through substantive findings such as whether there is a noticeable difference between the measures for two groups, to how much reliance to put on the research findings. As this paper shows, the common technical approaches to analysis, such as significance tests, are of no use here, and will tend to mislead. Moving away from the technical to judgment is beneficial for many reasons. It makes the use of numbers less problematic and so encourages their use in research (Gorard 2021).

This paper illustrates the importance of true measurements, able to be calibrated with the thing being measured, and so provide an estimate of measurement error. Without this, it is not possible to tell how accurate our research numbers are, how the initial error in their measurement propagates, nor judge the quality of the research involved. The confusion between the measurement of observable events and the habit of assigning numbers to imagined events (including perceptions, attitudes, and intentions) presents real dangers. It can lead to a waste of research effort, corruption of new researchers, misleading of research users, and vanishing breakthroughs.

CONCLUSION

It is important that measuring scales can be calibrated, shown to behave in the same way as the thing being measured, and are able to generate an estimate of measurement error. Without these a numbering scale is not a true measurement, and there will be many examples of these in social science. It is not possible to track the propagated error in such scales, nor to judge the trustworthiness of research using them. Traditional statistical analysis cannot help make these numbering scales any better. Judgement should be to the fore, and new researchers should be encouraged to be humbler about their numeric claims.

REFERENCES

1. Berka, K (1983) Measurement: its concepts, theories and problems, London: Reidel
2. Cumming, G. (2014) The new statistics: why and how, Psychological Science, 25, 1, 7-2, https://doi.org/10.1177/09567976135049669
3. De Vrieze, J. (2019) What science reporters should know about meta-analyses, What science reporters should know about meta-analyses before covering them | by Jop de Vrieze | Medium
4. Deeks J., Dinnes J., D'Amico R., Sowden A., Sakarovitch C., Song F. et al. (2003) Evaluating non-randomised intervention studies, Technology Assessment, 7, 27, iii–x, 1–173, doi: 10.3310/hta7270
5. Farrington, D., Gottfredson, D., Sherman, L. and Walsh, B. (2002) Evidence-based crime prevention, London: Routledge
6. Fielding, J. and Gilbert, N. (2000) Understanding social statistics, London: Sage Gorard, S. (2006) Towards a judgement-based statistical analysis, British Journal of Sociology of Education, 27, 1, 67-80, https://doi.org/10.1080/01425690500376663
7. Gorard, S. (2010a) Measuring is more than assigning numbers, pp.389-408 in Walford, G., Tucker, E. and Viswanathan, M. (Eds.) Sage Handbook of Measurement, Los Angeles: SAGE
8. Gorard, S. (2010b) Serious doubts about school effectiveness, British Educational Research Journal, 36, 5, 735-766, **https://doi.org/10.1080/01411920903144251**

9. Gorard, S. (2013) The propagation of errors in experimental data analysis: a comparison of pre- and post-test designs, International Journal of Research and Method in Education, 36, 4, 372-385, http://dx.doi.org/10.1080/1743727X.2012.741117
10. Gorard, S. (2020) Handling missing data in numeric analyses, International Journal of Social Research Methods, 23, 6, 651-660, https://www.tandfonline.com/doi/full/10.1080/13645579.2020.1729974
11. Gorard, S. (2021) How to make sense of statistics: Everything you need to know about using numbers in social science, London: SAGE
12. Gorard, S. (2024) Judging the relative trustworthiness of research results: how to do it and why it matters, Review of Education, 12, 1, https://doi.org/10.1002/rev3.3448
13. Gorard, S. (2025) Mixing methods is still wrong, in Morrison, K. and See, BH (Eds) Handbook of Mixed Methods, Sage
14. Gorard, S. and Chen, W. (2025) What is the evidence on research-informed education?, Chapter 2, pp.55-76 in Wyse, D., Baumfield, V., Mockler, N and Reardon, M. (Eds.) The BERA/SAGE Handbook of Research-Informed Education Practice and Policy
15. Gorard, S. and See, BH. (2009) The impact of SES on participation and attainment in science, Studies in Science Education, 45, 1, 93-129, https://doi.org/10.1080/03057260802681821
16. Nickerson, R. (2000) Null hypothesis significance testing: a review of an old and continuing controversy, Psychological Methods, 5, 2, 241-301, 10.1037/1082-989x.5.2.241
17. Nunnally. J. (1975) Psychometric theory 25 years ago and now, Educational Researcher, 4, 7, 7-21
18. Prandy, K. (2002) Measuring quantities: the qualitative foundation of quantity, Building Research Capacity, 2, 2-3
19. Rozeboom, W. (1960) The fallacy of the null hypothesis significance test, Psychological Bulletin, 57, 416-428
20. Smith, E., Gorard, S., Morris, R., Perry, T. and Pilgrim-Brown, J. (2025) Then and now: Twenty years of Education research methods use in the UK, British Educational Research Journal, 51, 1, 2347-2400, Does school matter for children's cognitive and non-cognitive learning? Findings from a natural experiment in Pakistan and India - Siddiqui - British Educational Research Journal - Wiley Online Library