WORKING GROUP 04

CHAPTER

2

How well does this intervention work? Statistical significance, uncertainty and some concepts to interpret the findings of evaluations of educational interventions[1]

This chapter should be cited as:

Rodriguez-Guzman, G. (2022) 'How well does this intervention work? Statistical significance, uncertainty and some concepts to interpret the findings of evaluations of educational interventions' in Mercier, J. (ed.) Data and Evidence in Reimagining Education: The International Science and Evidence based Education Assessment [Duraiappah, A.K., Atteveldt, N.M. van et al. (eds.)]. New Delhi: UNESCO MGIEP. In Press.

Coordinating Lead Authors

Guillermo Rodriguez-Guzman

[1] This document discusses the role of uncertainty in the interpretation of empirical evaluations in education. This is an expanded and revised version of a statement on statistical significance first published by the Education Endowment Foundation (EFE). The EEF is an independent charity based in England, dedicated to breaking the link between family income and educational achievement.

The author and the EEF are grateful to Ronald Wasserstein, Nicole Lazar, Allen Schirm, and Steve Higgins for the invaluable comments and suggestions made to the original statement published by the EEF.

The authors are also grateful to Julien Mercier and Drew Bailey for outstanding recommendations to multiple versions of this paper; and to the three anonymous peer reviewers for other thoughtful suggestions.

WORKING GROUP 04

2

CHAPTER

2.1 Introduction: the importance of understanding and interpreting uncertainty

For anyone interested in how evidence can support more effective decision-making in education, the term ‘statistical significance’ will be a familiar one – and yet one probably shrouded in confusion. Despite the claims one might hear circulating in the media, policy circles and from different pundits, no study will give the ultimate and unquestionable truth about whether a programme or intervention will achieve a specific impact.

Policy decisions and prescriptions for action are often made on the basis of incomplete and imperfect information and the uncertainty around quantitative results is one of the key factors at play. As the eventual implementation of interventions may have positive

Despite the claims one might hear circulating in the media, policy circles and from different pundits, no study will give the ultimate and unquestionable truth about whether a programme or intervention will achieve a specific impact.

or negative impact on learners, understanding uncertainty of impact estimates is integral to educational practice and policy-making. In principle, not considering this uncertainty means that policies and changes in practice, despite being based on research evidence, overlook relevant scenarios. This can lead to overly cautious decision-making in some cases or risk detrimental effects to learners in others.

Reflecting this complexity and uncertainty, researchers have been using 'statistical significance' to attempt to deal with uncertain, incomplete answers. But the use of statistical significance divides the research community in a range of disciplines, from statistics to social policy, including education. Some consider statistical significance an essential part of impact evaluation, just one aspect of a broader picture, while others regard it as a meaningless and misleading concept that should be abolished altogether **(Shrout, 1997; Ziliak and McCloskey, 2008; Trafimov and Marks, 2015; Gorard, 2016; Hubbard, 2016; Wasserstein and Lazar, 2016; Amrhein, Greenland and McShane, 2019; McShane et al., 2019; Wasserstein, Schirm and Lazar, 2019).**

For the average classroom teacher, school leader or policy-maker, this lack of consensus among educational researchers is highly problematic, making it difficult to answer the very reasonable question: 'How well does this intervention work?'

This chapter outlines some key concepts underpinning notions of uncertainty, and proposes a way forward, which is then adopted in the subsequent chapter that presents estimates of impact, costs and certainty for a range of common education interventions and approaches. The key proposal is that impacts should be reported as effect sizes, and interpreted alongside internal validity and uncertainty when making a decision about a programme. We summarise relevant scholarship in this topic, which proposes moving away from a dichotomous interpretation of p-values and significance testing as the means to gauge the effectiveness of a programme.

2.2 How well did this intervention work? Some building blocks and an example

2.2 .1

KEY CONCEPT 1 – EFFECT SIZE

An effect size is a number that conveys the strength of the relationship between two variable factors. This number is obtained, for any given dependent variable, by scaling the difference between group means by the dispersion of the observations (the standard deviation).

In education, factors manipulated experimentally usually are subject to a specific intervention to measure the outcomes achieved by learners (e.g. educational attainment). In an experimental setting, this would usually compare the average in the intervention group and the average in the control group, scaled by how dispersed the results are

..when communicating evidence of impact, it can be helpful to translate outcomes into other more meaningful measures while trying to introduce them into the common parlance of decision-makers.

(i.e. the standard deviation). The larger the effect size, the larger the difference between the two groups and the stronger the relationship between the intervention and the outcomes being measured.

Effect sizes are an important and useful metric because they enable us to move away from the simplistic question of whether something works or not (further complicated by the reliance on a dichotomous interpretation of statistical significance – more on this below). Instead, effect sizes help to answer the more relevant question 'How well did this work?' **(Coe, 2002; Major and Higgins, 2019; Higgins, 2021)**. Effect sizes are also useful as they provide a common metric to compare the relative effectiveness (see chapter 1) of different interventions, which is more meaningful for decision-makers choosing between competing alternatives.

A key challenge regarding the use of effect sizes is that they describe differences in terms of standard deviations rather than measures that are more readily understood by the very audience who should be able to make the most of research results: policy-makers and teachers.

This is why, when communicating evidence of impact, it can be helpful to translate outcomes into other more meaningful measures while trying to introduce them into the common parlance of decision-makers.

2.2 .2

KEY CONCEPT 2 – MONTHS OF (STANDARD) PROGRESS AS A PRACTICE-ORIENTED TRANSFORMATION OF EFFECT SIZE

To overcome this communication challenge, the Education Endowment Foundation's (EEF) toolkit **(Major and Higgins, 2019; Higgins, 2021)** transforms effect size into a single scale of school progress: months of progress.

This transformation is done by dividing effect size, which is a measure of progress in terms of standard deviations, by the progress that could be expected in a school year for a given group of learners, also measured in standard deviations. The result is the amount of progress that would have been made in comparison to the average progress made in a year. That is, a standardised benchmark that allows drawing comparisons between multiple interventions in a metric that is easier to understand for teachers and decision-makers.

The average progress in a year is estimated to be around one standard deviation; and while this is likely to be a conservative estimate which may vary for different ages and types of tests, a crude measure is preferred to ensure findings remained more accessible and meaningful **(Major and Higgins, 2019; Higgins, 2021)**.

Other transformations and metrics have been proposed and reviewed by **(Bloom et al., 2008; Lipsey et al., 2012; Baird and Pane, 2019; Evans and Yuan, 2019)**. These include months of progress measures that account for differences across tests and the speed at which pupils learn over time, as well as alternatives like percentile ranges. These alternatives have their merits, as they address some of the methodological shortcomings of the simpler months of progress measure used by the EEF. However, this can also result in more complex interpretation, which is the problem these alternatives are trying to address. Stakeholders may decide to use one or several of these transformations, depending on the levels of literacy and exposure of the decision-makers they are seeking to inform or influence. For example, using months of progress as a metric, researchers can explain that an intervention that had an impact of 0.3 standard deviations could be represented as achieving the equivalent of 3 months' progress – a measure that is likely to be easily understood by practitioners and decision-makers.

Stakeholders may decide to use one or several of these transformations, depending on the levels of literacy and exposure of the decision-makers they are seeking to inform or influence.

A confidence interval is a range that is often used to measure uncertainty around an estimated value, such as an effect size or the mean of a distribution.

In addition to the 'mean' effect identified by an evaluation, quantitative researchers need to clearly express the uncertainty around those results – that is, other results that would be plausible under the statistical model being used and considering characteristics of the data.

2.2 .3

KEY CONCEPT 3 - 'CONFIDENCE' INTERVALS (OR 'COMPATIBILITY' INTERVALS)

A confidence interval is a range that is often used to measure uncertainty around an estimated value, such as an effect size or the mean of a distribution. This range of values is bounded above and below the statistic's mean. A 95% 'confidence interval' includes a range of values for which 95% of the confidence intervals computed from many hypothetical studies would contain the unknown population parameter if all the conditions under which the intervals are built hold. The interpretation of confidence intervals can be challenging and has been extensively criticised **(Greenland et al., 2016; Morey et al., 2016)** for reasons akin to the problems with p-values (see below).

2.2 .4

KEY CONCEPT 4 - P-VALUES AND STATISTICAL SIGNIFICANCE

Another standard way of assessing this uncertainty is using a p-value. These are measures of the compatibility between the observed data and a particular model of the data and are closely related to the idea of a 'confidence interval'. Both concepts are probabilities computed for many hypothetical studies under a set of conditions. We define these terms

A result is deemed 'statistically significant' if the 95% confidence interval does not include zero or if a p-value is below a given threshold, often 0.05, which is symmetrical to the 95% confidence interval.

in greater detail in the section 'The problems with statistical significance'.

P-values are difficult to interpret for researchers and practitioners alike and have been widely criticised for misleading decision-making and biasing the literature, particularly given the tendency to interpret them in a dichotomous way due to a reliance on the idea of 'statistical significance' **(Greenland et al., 2016; Wasserstein and Lazar, 2016; Amrhein et al., 2019; Wasserstein, Schirm and Lazar, 2019).**

A result is deemed 'statistically significant' if the 95% confidence interval does not include zero or if a p-value is below a given threshold, often 0.05, which is symmetrical to the 95% confidence interval. When a result is 'statistically significant' it is often interpreted as meaning that the intervention 'had an effect'. As explained in the section 'the problems with statistical significance', this is not true. This dichotomous interpretation is at the heart of the problems with p-values, confidence intervals and significance testing.

Nonetheless, the interpretation of p-values could be seen as more heinous than confidence intervals because a range of values is more likely to be interpreted with caution. A range of values is more plausible than imprinting a false sense of certainty for decision-makers who observe a result that is 'statistically significant' and believe it to be the 'true' effect. This has been reflected in the preference of a growing number of journals to report confidence intervals instead of p-values **(Greenland et al., 2016).**

2.2 .5

KEY CONCEPT 5 – INTERNAL VALIDITY

To evaluate the impact of a programme or intervention, researchers would like to compare the 'treatment' outcomes those without the 'treatment' or intervention. This scenario is called the counterfactual. Clearly, it is not possible to observe both scenarios in the real world, which requires researchers to

Most EEF-funded evaluations use a randomised controlled trial (RCT) design to estimate the impact of a programme; this is one of the most robust ways to identify a valid counterfactual.

compare the results of the group that was treated with those of a group identified as a suitable comparison (that is, a valid counterfactual). The differences in outcomes between the treatment and the comparison groups, considering the mean outcome and its variability in each group, is interpreted as the estimate of impact and measured as an 'effect size'.

Most EEF-funded evaluations use a randomised controlled trial (RCT) design to estimate the impact of a programme; this is one of the most robust ways to identify a valid counterfactual. The evaluation design, in this case an RCT, is one of the crucial factors defining how confident we can be that the findings are a good representation of the impact of the intervention. However, to make this assessment, it is also important to consider other dimensions including:

- the overall size of the study[2];

- whether the relevant information from participants is present, and, if not, understanding why (outcome attrition);

- whether appropriate and reliable outcome measures were used to track progress;

- whether those in the control group received the intervention being tested or experienced any other changes that could affect their behaviour and progress, such as non-compliance or experimental effects, among others.

Taken together, these may be understood as the internal validity of a study. EEF-funded studies are assigned a 'padlock rating' using the EEF's classification of

[2] Sample sizes are intrinsically linked to the level of 'uncertainty' in a study, but they are also related to its internal validity. While one can obtain an unbiased (yet imprecise) treatment impact estimate from a small study, a larger study is less likely to suffer internal validity problems such as randomisation failure whereby the two groups are substantially different. The effectiveness of randomisation relies on the law of large numbers and the central limit theorem, which are compromised in smaller samples.

The EEF's classification system for single studies summarises relevant aspects of the internal validity of findings and considers the professional judgement of the peer reviewers assigning them.

the security of the findings. This systematically summarises the characteristics that define the internal validity and whether these make an estimate of impact from a given study more or less credible.

These dimensions cumulatively affect how much credence we give to a study. For instance, a study that succeeds to capture information on every participant would be more credible than one where only 60% sat the relevant exam (all else being equal). Failing to include every learner in the follow up (called outcome attrition) can be a problem because those who did not sit the exam could have been different from those who did in a way that is related to the intervention.

The EEF's classification system for single studies summarises relevant aspects of the internal validity of findings and considers the professional judgement of the peer reviewers assigning them. These ratings should not be understood in a definite manner either, but as providing useful information to interpret findings. However, there are many other tools and resources used to gauge the robustness of a single study: from relatively simple approaches focusing on study design such as the Maryland Scientific Methods Scale **(Farrington et al., 2002)**, to others that consider multiple sources of bias and external validity problems depending on the type of design being considered **(Higgins et al., 2016; Sterne et al., 2017)**.

2.2 .6

AN EXAMPLE

Now, using the key concepts described above, imagine you have three studies in the same domain, each with the goal of establishing the impact of an intervention:

- the evaluation of programme A was well-designed and well-conducted and found an effect size (ES) of 0.10; compatibility interval (CI): −0.10, 0.3; not statistically significant;

- the evaluation of programme B, also well-designed and well-conducted, found an ES of 0.10; CI: −0.01, 0.21; not statistically significant;

- the evaluation of programme Z1 was fraught with problems of internal validity that reduced its credibility; it found an ES of 0.20; CI: −0.20, 0.4; not statistically significant.

- the evaluation of programme Z2 was fraught with problems of internal validity that reduced its credibility; it found an ES of 0.20; CI: 0.10, 0.3; statistically significant.

The evaluation of programmes Z1 and Z2 suffered from important internal validity limitations[3] and thus the results are more likely to be called into question. One additional difficulty is that these problems with the design and implementation of a study are not always measurable and might be operating in different directions. This means that we might be overstating or underestimating the impact of an intervention, but the magnitude and direction in which this is happening is both difficult to ascertain and quantify. On these grounds, researchers are unlikely to recommend the use of Z as the evidence is not credible enough to claim that Z might be effective at improving outcomes. The findings could be understood as tentative at best and additional evidence of the effectiveness of Z would be necessary, by means of a better study.

The findings could be understood as tentative at best and additional evidence of the effectiveness of Z would be necessary, by means of a better study.

Studies for programmes A and B were well-conducted and methodologically robust[4] and had

[3] Using the EEF's classification system for single studies, these studies would be awarded a very low rating – probably one or two padlocks. For example, this could be an observational study designed to compare outcomes before and after without a control group. As it would not be possible to distinguish the effects of the intervention and the natural progress of pupils, we are unable to confidently conclude the intervention can improve pupil outcomes.

[4] Using the EEF's classification system for single studies, these studies would be awarded the maximum of five padlocks.

Quantitative studies in education and other applied domains provide a range of possible answers that need to be analysed, considering multiple sources of uncertainty.

Quantitative studies in education and other applied domains provide a range of possible answers that need to be analysed, considering multiple sources of uncertainty.

the same estimate of impact: an ES of 0.10, which is equivalent to +2 months' additional progress[5]. However, as we stated above, studies do not give a single, unequivocal and definitive answer. The CI associated with both studies indicates that the data for programme B were also compatible with a range of effects from no impact to moderate impact, whereas the data for programme A was also compatible with a range of effects from a small negative impact to high impact.

Using statistical significance as the only criteria, researchers would have concluded that programme Z2 had statistically significant results (which is often understood as 'having an impact') while both programmes A and B had non-significant results (which is often understood as 'not having an impact').

This dichotomous interpretation of statistical significance is at the core of its problems and the source of contention around its use **(Wasserstein and Lazar, 2016).**The advancement and use of scientific knowledge in the quantitative approach is not as simple as concluding that something works, and something does not. This example illustrates how the exclusive reliance on statistical significance could be very misleading as it obscures a much more nuanced picture: one where we are interested in understanding how well something works and which are the plausible scenarios that we can expect – that is, the uncertainty around the results.

Quantitative studies in education and other applied domains provide a range of possible answers that need to be analysed, considering multiple sources of uncertainty. Otherwise, decision-making is severely impeded. In the context of the example, no sound decision can be made exclusively on the basis of statistical significance because the uncertainty highlighted by coupling the effect size with confidence intervals

[5] The estimate of months of progress is based on EEF Guidance.

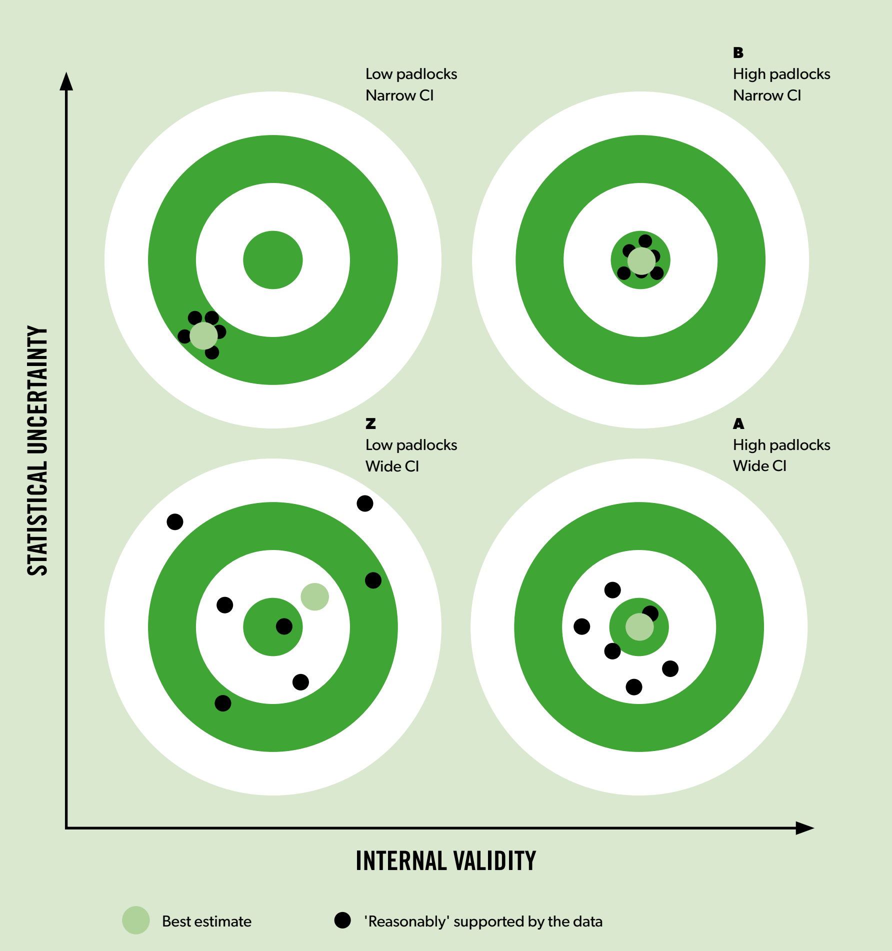

Figure 1. Schematic representation of internal validity and uncertainty

(an aspect commonly neglected) means that the findings in A are also compatible with a negative impact (−0.1) or a larger positive impact (0.3) while those of B are compatible with an educationally-very-small negative effect (−0.01) or a larger impact (0.3). Note that these are not the only values that are compatible with the data because confidence intervals should not be interpreted in a dichotomous way either, see **section 2.5**.

For a teacher or policy-maker deciding which of two similar programmes to invest in, both pieces of information are important and are represented conceptually in **Figure 1**.

Comparing Z with A or B would be like a vertical comparison in

Figure 1: between not-so-well-designed, tentative studies, and well-conducted, more credible studies. This comparison could be interpreted as the internal validity of the finding.

However, to discern between programmes A and B it is also relevant to consider other aspects. Even if both have the same estimate of impact (effect size), the findings of programme A are compatible with more variability (confidence intervals): from negative effects to larger positive impacts included in the intervals. In contrast, the findings of programme B show less variability being only compatible with a very small negative effect or a larger positive effect. This compares the uncertainty of the findings.

Making this distinction—between internal validity and uncertainty—accessible to decision-makers is fundamental: while the best estimate of A suggests a positive impact, the variability around it suggests more caution as the model of the data is compatible with the programme being harmful; however, the best estimate of B found the same positive impact, but at worst the model of the data was less compatible with the programme being harmful. Thus, with this information, a decision-maker may be more confident to implement B.

Decision-makers also need to consider a series of aspects when deciding which programme to implement; these include costs and resources, for example, which is why each EEF evaluation report provides an estimate of the required investment. For more information see EEF Cost Evaluation Guidance. Other aspects include the programme's acceptability, its relevance to the problems faced by a particular school and the quality of programme implementation, among others. EEF evaluations strive to cover such topics as part of the Implementation and Process Evaluation component of all EEF-funded studies. For more information, see EEF IPE Guidance.

Decision-makers also need to consider a series of aspects when deciding which programme to implement; these include costs and resources, for example, which is why each EEF evaluation report provides an estimate of the required investment. For more information see EEF Cost Evaluation Guidance.

2.3 Where does uncertainty come from?

There are multiple sources of uncertainty; but in the context of evaluations, two types are particularly relevant: sampling uncertainty and allocation uncertainty.

Even in a well-designed and well-conducted study with good internal validity, there are at least two steps in an RCT that introduce uncertainty.

1. When a group of schools or pupils is selected to take part in a study, random sampling leads to sampling uncertainty. This uncertainty is accepted because it is not practically feasible or economically viable to include

every school in every single study. Even if a random sample from the population is selected, such schools or pupils might be different from the population at large for reasons we might not be able to identify. Note that in most cases, samples of participants taking part in a RCT are not drawn at random from the population.

2. When these schools or pupils are subsequently randomly allocated to the intervention or control group, random assignment leads to allocation uncertainty. Even if these are randomly assigned, there might be differences between the two groups for reasons we might not be able to identify.

These two processes thus introduce sampling uncertainty and allocation uncertainty, respectively.

Even if the same experiment is repeated a large number of times, these sources of uncertainty imply that the observed differences between groups could differ under each of these identical hypothetical experiments. These types of uncertainty are closely linked with the heterogeneity between units in the population and the sample.

When individuals in the population are very different from each other, it is more likely that a random sample would end up with a group with very different characteristics for which the estimate of impact could be different from the 'true' population effect (1). Likewise, even within a given sample, the random allocation might lead to a treatment group with very different characteristics for which the estimate of impact could also be different from the impact estimate that would be obtained with a different random configuration of the treatment and control groups (2).

When these schools or pupils are subsequently randomly allocated to the intervention or control group, random assignment leads to allocation uncertainty.

[6] It is not possible to know the 'true average effect size' as that would require pre-test and post-test outcomes for each member of the sample/population both with and without the intervention, which is not possible.

This means that it is always possible that the true effect size[6] observed in an RCT will differ from the true average effect size in the sample because, even for two identical experiments, the observed effect size is likely to differ a bit, and will occasionally differ a lot, as a result of this statistical uncertainty.

Likewise, the observed effect size may also be different from that on the population. In addition to the problems related to inferences in a sample, to make broader claims around the external validity of the findings to a population it is necessary to consider many other aspects beyond statistical uncertainty, which are more likely to influence whether the results observed in a sample can be expected to be replicated for the population **(Deaton and Cartwright, 2018)**.

However, these are not the only sources of statistical uncertainty. For instance, to focus on one of the most common, the accuracy and reliability of an outcome test may also introduce measurement uncertainty from the selected instruments. This relates to the margin of doubt that exists for the result of any measurement that could be due both to the instrument being used (e.g. a test, a timer) and how this translates the relevant behaviour into a quantitative value (e.g. a score). This can also be affected by the construct being measured (e.g. algebra, self-efficacy). Hence every measurement differs from the 'true' value that it is trying to capture. This difference is the error; while measurement uncertainty is the quantification of those expected errors and is often expressed as a confidence interval around a measurement. The measurement uncertainty introduced by using a specific outcome measure could be considered an internal validity problem but it also adds to the variability of the results observed.

This means that it is not possible to isolate the multiple sources of uncertainty from some aspects of internal validity.

2.4 The problems with statistical significance

To assess uncertainty, many researchers consider a hypothetical situation where:

1. a (random) sample is drawn from the population of interest[7] (which would be related to sampling uncertainty);

[7] RCTs are hardly ever a random sample from the population. EEF-funded studies are not random samples. This means that the interpretation of the p-values should not be considered as making claims about the external validity of the study (inferences on the impact on the population) but only as relating to the sample at hand (inferences on the internal validity of the study on the sample).

One of the reasons for misinterpretation is that p-values give the right answer to the wrong question

2. the same experiment is conducted a large number of times on samples drawn from the same population (which would be related to allocation uncertainty, and other sources of uncertainty related to the internal validity of the study); and

3. the intervention has no true impact on the population (that is, the real impact of the intervention is zero).

Then, researchers estimate how likely it would be, in this hypothetical situation, to observe a difference at least as big as the difference they observed due to the statistical uncertainty.

This probability to observe a difference at least as big as the difference they observed is called the p-value.

This statistic has been strongly criticised because frequent misuse and misinterpretation lead to distortions in scientific enquiry **(Wasserstein and Lazar, 2016; Amrhein, Greenland and McShane, 2019; Wasserstein et al., 2019)**. One of the reasons for misinterpretation is that p-values give the right answer to the wrong question. In practice, the question we want to answer is, 'Does this intervention work?' Instead, p-values explain, 'How rare would these results be in a world where the intervention had no effect? (i.e. the hypothetical situation, which also requires fulfilling the other assumptions mentioned above)'. For example, imagine you want to identify whether a programme improves pupil outcomes and you found a difference equivalent to three months of progress. The question we want to answer is: Given that we observed a difference of three months of progress, how likely is it that this programme had no effect? This is not what a p-value tells us. The p-value shows the probability that you would observe a difference of three months or more given that the intervention had no impact (the hypothetical situation, which also includes the other relevant assumptions described above).

P-values neither give an indication of the likelihood that the intervention had an effect nor give the probability that the observed result was produced by random chance alone **(Greenland et al., 2016; Wasserstein and Lazar, 2016; Amrhein et al., 2019; Wasserstein, Schirm and Lazar, 2019)**. P-values give a very indirect answer to the question we are truly interested in. The smaller the p-value, the more unusual the results if all the assumptions under the hypothetical situation are true. However, a very small p-value does not tell us which of the assumptions might be incorrect even if we are only truly interested in the question of whether this intervention worked – closely related with the third assumption above **(Greenland et al., 2016)**.

...the most salient problem with p-values (and also similar statistics such as confidence intervals, discussed below) is the convention to treat them in a dichotomous way around a 0.05 threshold - a 'bright-line'

However, the most salient problem with p-values (and also similar statistics such as confidence intervals, discussed below) is the convention to treat them in a dichotomous way around a 0.05 threshold – a 'bright-line' where on one side an impact is inferred to exist, while on the other, the possibility of an impact is entirely disregarded as inconsistent with the data.

This simplification is a caricature of the necessary complexity to make inferences to advance scientific knowledge and violates the spirit of how p-values are supposed to be interpreted. Originally, the 0.05 threshold was chosen as a way to limit the risk of false positives. It means that if you were to repeat the experiment 100 times under the hypothetical situation (that is, the programme has no effect), in five of them, you would see results as extreme or more extreme than yours. The original proponent of the p-value, Ronald Fisher, argued that a statistically significant finding was worthy of further investigation. Alas, in a gross misrepresentation of that spirit, this threshold became the value to consider a finding 'true', which is not true **(Wasserstein and Lazar, 2016)**.

Rather than a 'bright-line' where effectiveness can be decided, p-values provide a continuum of how compatible the data is with the hypothetical situation. Values at either side of the threshold should not be treated as definitive answers but as different tonalities of grey – data that is more or less compatible with the estimate of impact. Even if actionable recommendations may require an affirmative answer, making inferences on the basis of an arbitrary threshold is incorrect and has distorted decision-making **(Wasserstein, Schirm and Lazar, 2019).**

A finding might be of educational/practical significance (represented as a large effect size) even if it is not deemed 'statistically significant' by reaching the arbitrary 0.05 cut-off point. A finding might be of educational/practical significance (represented as a large effect size) even if it is not deemed 'statistically significant' by reaching the arbitrary 0.05 cut-off point.

This dichotomy at each side of the threshold also conflates practical and statistical relevance. A finding might be of educational/practical significance (represented as a large effect size) even if it is not deemed 'statistically significant' by reaching the arbitrary 0.05 cut-off point. This problem is particularly heinous because when a study is large, even small violations of the assumptions can lead to a 'statistically significant' result that affects how decisions are made.[8] Contrariwise, even an educationally relevant difference could fail to be 'statistically significant' if the sample is not large enough. Sometimes a statistically significant result simply means that a very large sample was used.[9]

The most common alternative is to report confidence intervals or compatibility intervals (CI). As is the case with p-values, confidence intervals are also prone to misinterpretation **(Greenland et al., 2016; Morey et al., 2016).** These estimate that if the same experiment were conducted a large number of times and interval estimates are made on each

[8] For example, (Sullivan and Feinn, 2012) mention an example of a study for aspirin. In the study, more than 22,000 subjects used aspirin over 5 years and the authors identified a statistically significant reduction in heart disease even if the reduction in risk was very small – and clinically negligible – for most patients. However, aspirin was recommended for general prevention for years. More recent studies confirm aspirin should be taken only for those who have suffered heart disease or a stroke and medical guidelines have been adapted accordingly.

[9] This also highlights the importance of relying on bodies of evidence, instead of single studies. By combining the information from multiple studies, systematic reviews (and statistical methods such as meta-analysis that combine different findings into a single metric) help to use information across all observations, which can help mitigate some of the problems related to single studies relying on statistical significance. However, it is important that the interpretation of these analyses is not subject to the same dichotomous interpretation of statistical significance.

P-values and CI are calculated based on similar hypothetical situations, and suffer from similar problems; including the erroneous dichotmous interpretation

occasion, the resulting intervals would bracket the true population parameter in approximately 95 % of the cases if the hypothetical situation is true.

P-values and CI are calculated based on similar hypothetical situations, and suffer from similar problems; including the erroneous dichotmous interpretation. CI are often interpreted as 'not crossing zero' to suggest that a result is 'statistically significant' and thus, 'true'. This is untrue. Symmetrically to p-values, a CI can only help to conclude how compatible the results are with a given statistical model. Just because a value lies outside of the specific CI, it does not mean that this value can be refuted or excluded from the data – just that it is less compatible with the assumptions used.

However, as argued above, using CI is seen as superior to p-values because presenting a range of values that is consistent with a given model of the data is more likely to be interpreted with caution rather than a single value that is often understood as evidence that an effect 'exists' or not **(Greenland et al., 2016)**.

In short, the issue around the interpretation and use of p-values, CI and statistical significance has less to do with the assumptions upon which they are constructed than with the obsession with a clear decision rule (i.e. a threshold) to conclude whether something is 'true' or not. This shows a naïve interpretation of the statistical assumptions underpinning these concepts but, more importantly, it steers decision-makers and practitioners away from key pieces of information needed to formulate new policies and introduce changes.

2.5 The way forward: bringing together internal validity and uncertainty to make the best use of evidence in educational decision-making

Effect sizes provide a better indication of the magnitude of impact and thus should be reported for all estimates

Internal validity and uncertainty should be considered in tandem when making a decision about a programme, as illustrated in the discussion above. Internal validity measures the suitability of the design of the study to produce estimates close to the true estimate of impact, that is, how close one is to the bull's eye or the bias of the estimate. Uncertainty measures how likely it is that the same experiment, repeated under the same conditions, would find a similar effect, that is, how close are different estimates of impact to each other or to the spread of the estimate. This was represented conceptually in **Figure 1**.

Ideally, a study should be well-designed and well-implemented (good internal validity) and likely to find a similar effect if replicated under the same conditions (low uncertainty). However, studies are hardly ever definitive and both aspects need to be factored into any interpretation of the results.

To address the criticisms above we propose that findings should be discussed in terms of effect sizes, with a thorough description of their internal validity using well-regarded tools; and importantly, emphasising the role that uncertainty plays in decision-making and moving away from a dichotomous interpretation of statistical significance. Commissioners and researchers may also consider translating these measures into other, more readily understood, measures such as months of progress.

To aid the effective communication of findings for educational interventions, we propose the following principles, which distill work by Wasserstein and Lazar **(2016)**, Wasserstein, Schirm and Lazar **(2019)**, and Amrhein, Greenland and McShane **(2019)**.

1. Use effect sizes to focus on the practical/scientific significance of a finding rather than relying on whether the finding was statistically significant.

The arbitrary 0.05 cut-off conflates practical and statistical relevance. However, statistical significance does not explain whether a finding is practically/scientifically/educationally interesting. Effect sizes provide a better indication of the magnitude of impact and thus should be reported for all estimates. These may be considered alongside other transformations to aid interpretation such as measures of months of progress that might be more accessible for decision-makers.

Results should be accompanied by a thorough description of the different elements that affect the internal validity of the study. This could be reported either using standardised tools such as Robins I or Risk of Bias Assessments, or bespoke tools such as EEF's Padlocks Rating

2. Include assessments of internal validity.

Results should be accompanied by a thorough description of the different elements that affect the internal validity of the study. This could be reported either using standardised tools such as Robins I or Risk of Bias Assessments, or bespoke tools such as EEF's Padlocks Rating. Threats to internal validity should always be reported transparently, even if the magnitude and direction of biases are difficult to quantify.

3. Accept uncertainty in findings and always present a measure of this uncertainty.

Statistical modelling should not be interpreted as providing unique and definitive answers, or what Gelman **(2016)** calls 'a sort of alchemy that transmutes randomness into certainty'. Instead, it is paramount to understand that, in real-world situations, statistical modelling only attempts to identify 'signals' in noisy data with considerable variability. Therefore, we should acknowledge that statistical models only provide incomplete and uncertain – yet potentially useful – answers to scientific questions. Abandoning a dichotomous interpretation of p-values and other statistics, including 'compatibility intervals', advances in this direction– moving us away from the detrimental simplification of findings as 'true' or not. Thus, researchers must present a measure of the uncertainty around all effect sizes, recognising that uncertainty is an integral part of statistical modelling and scientific enquiry.

..in real-world situations, statistical modelling only attempts to identify 'signals' in noisy data with considerable variability.

4. Use precise language and clearly consider assumptions behind the statistics used to represent uncertainty.

P-values do not measure the probability that 'the studied hypothesis is true' nor the probability that the 'data were produced by random chance alone' **(Wasserstein and Lazar, 2016)**. Similar misinterpretations are common when describing confidence intervals **(Greenland et al., 2016; Morey et al, 2016)**. To a large extent, the problem with p-values is that they offer an answer to a question we are not necessarily seeking to answer – that of the hypothetical scenario. However, ignoring the assumptions upon which p-values are calculated goes a long way toward explaining why they have become contentious and potentially misleading. Thus, researchers must be accurate in the interpretation of p-values (or any other statistic used), what they are and what they are not, carefully considering the assumptions upon which these are constructed.

5. Report continuous p-values (or other measures of statistical uncertainty), interpreting them as varying degrees of statistical uncertainty and avoiding dichotomisation of decisions around the arbitrary cut-off of p = 0.05.

P-values are the probability, under a specified statistical model (the hypothetical scenario), that the mean difference between two groups would be equal or more extreme than the observed value in the study **(Wasserstein and Lazar, 2016)**. As a continuous probability, p-values are a measure of the degree of compatibility of the data with the hypothetical model imposed on that data. Claiming a finding as 'statistically significant' suggests a dichotomous interpretation that contravenes **Recommendation 1**. Therefore, abandon the dichotomous interpretation of p-values, recognising that different p-values suggest different levels of strength of the evidence and thus should be reported as a value and interpreted as a continuum. Findings should be interpreted

To report statistical uncertainty around the point estimate, discuss the educational/scientific relevance of the point estimate and also the extremes of the compatibility intervals. To report statistical uncertainty around the point estimate, discuss the educational/scientific relevance of the point estimate and also the extremes of the compatibility intervals.

neutrally, irrespective of whether results are 'positive' (positive effect size, not statistically significant) or not. Other statements that suggest a dichotomous interpretation around the 0.05 should also be shunned. For example, phrases such as 'no evidence of impact', 'there is no difference', and 'nearly statistically significant' should be discontinued entirely.

6. Discuss the practical relevance of 'compatibility intervals'.

Avoiding referring to 'confidence' intervals as the word confidence suggest ungranted certainty **(Amrhein, Greenland and McShane, 2019; Greenland, 2019; Wasserstein, Schirm and Lazar et al., 2019)**. To report statistical uncertainty around the point estimate, discuss the educational/scientific relevance of the point estimate and also the extremes of the compatibility intervals. Note that these compatibility intervals reflect other values, under the hypothetical statistical model used, that are also compatible with the data. Even if intervals are estimated based on a predetermined threshold – conventionally 95% aligned with a p of 0.05 – they should also not be interpreted in a dichotomous way as outlined in Recommendation 5: values closer to the point estimate (the best estimate of impact) are better supported by the data, while those farther away are less compatible with it. Values outside these intervals are less compatible with the data, not inconsistent with it.

7. Consider accompanying p-values and 'compatibility intervals' with other statistics.

Explore other statistics that could help interpretation, rather than interpreting them in a dichotomous way regardless of which statistic is chosen. Researchers may, for instance, consider permuted p-values that do not rely on the assumption of random sampling and thus do not intend to make generalisations beyond the sample, or other statistics like Bayesian Compatibility Intervals, which

rely on other assumptions. The American Statistical Association's (ASA) Special Issue, Statistical inference in the 21st century: A world beyond p<0.05, offers some suggestions. Researchers may also want to present alternatives to test the sensitivity of the statistical uncertainty captured by different models.

8. Discuss practical and scientific significance considering all relevant information.

Interpret the findings considering internal validity, statistical uncertainty, the strength of the existing evidence, the plausibility of the causal mechanism, the evidence of the quality

...we propose that findings should be discussed in terms of effect sizes, with a statement about the internal validity of the finding and representing the statistical uncertainty of the finding as a continuous p-value, 'compatibility intervals', and/or alternative statistics.

of the implementation, and considerations of the context. Also consider the process through which the statistics were obtained: For example, if the design and analysis were pre-registered, the effect size is more likely to approximate the true effect of interest than if the effect was observed only after exploring a range of subgroup, outcomes, and/or treatment variations and selected on the basis of its magnitude or associated p value. If a design and analysis are not pre-registered, or if the analytic process is not transparently described, a promising effect should be appropriately discounted.

Furthermore, researchers should be thoughtful in describing how the finding shifts the evidence-base and existing priors. This is important because these statistics should be understood in the context of the processes that generated them, and thus, bringing additional information is crucial to decisionmaking.

In sum, we propose that findings should be discussed in terms of effect sizes, with a statement about the internal validity of the finding and representing the statistical uncertainty of the finding as a continuous p-value, 'compatibility intervals', and/or alternative statistics.

Advancing scientific knowledge in education is a complex endeavour. But it is also a laudable one - it has the potential to improve people's lives by fostering learners' strengths and, if needed, by providing scaffolding to move past difficulties. We hope that these principles will help researchers move closer to that goal by providing decision-makers the necessary information to make the right decisions about educational interventions grounded in evidence of what works, and eventually what works best **(WG4-ch1)**.

REFERENCES

Amrhein, V., Greenland, S. and McShane, B. (2019) 'Retire statistical significance', Nature, 567, pp. 305–307.

Baird, M. and Pane, J. (2019) 'Translating standardized effects of education programs into more interpretable metrics', Educational Researcher, 48(4), pp. 217–228.

Bloom, H.S., Hill, C.J., Black, A.B., and Lipsey, M.W. (2008) 'Performance trajectories and performance gaps as achievement effect-size benchmarks for educational interventions', Journal of Research on Educational Effectiveness, 1(4), pp. 289-328.

Coe, R. (2002) 'It's the effect size, stupid. What effect size is and why it is important', British Educational Research Association Annual Conference, Exeter.

Deaton, A. and Cartwright, N. (2018) 'Understanding and misunderstanding randomized controlled trials', Social Science & Medicine, 210, pp. 2–21.

Evans, D. and Yuan, F. (2019) Equivalent Years of Schooling. A Metric to Communicate Learning Gains in Concrete Terms. Washington DC: World Bank Policy Research Working Paper 8752.

Farrington, D.P., Gottfredson, D.C., Sherman, L.W. and Welsh, B.C. (2002) The Maryland Scientific Methods Scale. Milton Park: Routledge.

Gelman, A. (2016) 'The problems with p-values are not just with p-values', The American Statistician, 70, pp. 1–2.

Gorard, S. (2016) 'Damaging real lives through obstinacy: re-emphasising why significance testing is wrong', Sociological Research Online, 21(1), pp. 102–115.

Greenland, S. (2019) 'Valid p-values behave exactly as they should: some misleading criticisms of p-values and their resolution with s-values', The American Statistician, 73(Sup1), pp. 106–114.

Greenland, S., Senn, S.J., Rothman, K.J., Poole, C., Goodman, S.N. and Altman, D.G. (2016) 'Statistical tests, P values, confidence intervals and power: a guide to misinterpretations', European Journal of Epidemiology, 31, pp. 337–350.

Higgins, J., Savovic, J., Page, M.J. and Sterne, J.A. (2016) Revised Cochane Risk of Bias Tool for Randomized Trials (RoB 2.0).

Higgins, S. (2021) Improving learning. Meta-analysis of intervention research in education. Cambridge: Cambridge University Press.

Hubbard, R. (2016) Corrupt research: the case for reconceptualizing empirical management and social science. London: SAGE Publications.

Lipsey, M., Puzio, K., Yun, C., Hebert, M., Steinka-Fry, K., Cole, M., . . . Busick, M. (2012) Translating the statistical representation of the effects of education interventions into more readily interpretable forms. Washington DC: National Center for Special Education Research.

Major, L.E. and Higgins, S. (2019) What works? Research and evidence for successful teaching. London: Bloomsbury.

McShane, B.B., Gal, D., Gelman, A., Robert, C. and Tackett, J.L. (2019) 'Abandon statistical significance', The American Statistician, 73(1), pp. 235–245.

Morey, R., Hoekstra, R., Rouder, J., Lee, M. and Wagenmakers, E. (2016) 'The fallacy of placing confidence in confidence intervals', Psychonomic Bulletin & Review, 23(1), pp. 103–123.

Shrout, P.E. (1997) 'Should significance tests be banned? Introduction to a special section exploring the pros and cons', Psychological Science, 8(1), pp. 1–2.

Sterne, J., Hernán, M.A., Reeves, B.C., Savović, J., Berkman, N.D., 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., Ramsay, C.R., Regidor, D., Rothstein, H.R., Sandhu, L., Santaguida, P.L., Schünemann, H.J., Shea, B., Shrier, I., Tugwell, P., Turner, L., Valentine, J.C., Waddington, H., Waters, E., Wells, G.A., Whiting, P.F. and Higgins, J.P.T. (2017) 'ROBINS-I: a tool for assessing risk of bias in non-randomised studies if interventions', British Medical Journal, 335, p. i4919.

Sullivan, G.M. and Feinn, R. (2012) 'Using effect size-or why the P value is not enough', Journal of Graduate Medical Education, 4(3), pp. 279–282.

Trafimov, D. and Marks, M. (2015) 'Editorial', Basic and Applied Social Psychology, 37(1), pp. 1–2.

Wasserstein, R.L. and Lazar, N.A. (2016) 'The ASA Statement on p-values: context, process and purpose', The American Statistician, 70(2), pp. 129–133.

Wasserstein, R.L., Schirm, A.L. and Lazar, N.A. (2019) 'Moving to a world beyond "p<0.05"', The American Statistician, 73(Sup1), pp. 1–19.

Ziliak, S. and McCloskey, D.N. (2008) The cult of statistical significance: how the standard error is costing us jobs, justice and lives. Ann Arbor: University of Michigan Press.