Statement on statistical significance and uncertainty of impact estimates for EEF evaluations

February 2020

Education Endowment Foundation

This document discusses the role of uncertainty in the interpretation of EEF-funded evaluations and explains why the EEF does not use statistical significance to discuss findings.

It builds on, and supersedes, the EEF's note, Statistical uncertainty in randomised controlled trials, published in August 2018.

This document has been developed in association with the EEF's Evaluation Advisory Group, other experts, and some members of the EEF's panel of evaluators. The EEF is grateful for all the feedback received.

Internal validity and uncertainty of impact estimates: an example

Imagine you have three studies:

- the evaluation of programme **A** was well-designed and well-conducted and found an effect size (ES) of 0.10; confidence interval (CI): -0.10, 0.3;
- the evaluation of programme **B**, also well-designed and well-conducted, found an ES of 0.10; CI: -0.01, 0.21; and, finally
- the evaluation of programme **Z** was fraught with problems that reduced its credibility; it found an ES of 0.20; CI: 0.10, 0.3.

The evaluation of programme Z suffered from important limitations and was likely to be biased.[1] On these grounds, evaluators 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.

Studies for programmes A and B were well-conducted[2] and had the same estimate of impact: an ES of 0.10, which is equivalent to +2 months' additional progress.[3] However, studies do not give a single, unequivocal, and definitive answer. The advancement of scientific knowledge is not as simple as that. Instead, studies provide a range of possible answers that need to consider multiple sources of uncertainty.[4] This uncertainty 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).[5]

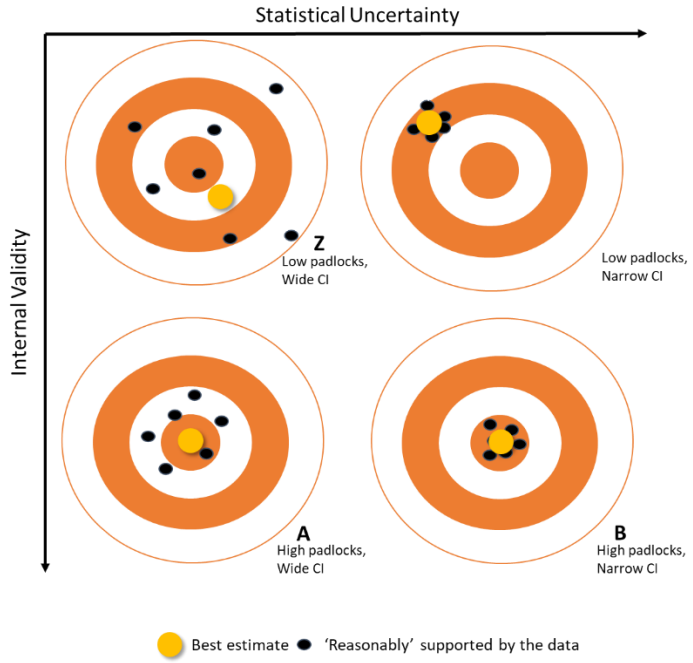
For a teacher deciding which of two similar programmes to invest in,[6] both pieces of information are important and are represented conceptually in Figure 1.

[1] The potential direction and magnitude of these biases are very difficult to predict.
[2] Using the EEF's classification system for single studies, these studies would be awarded the maximum of five padlocks.
[3] The estimate of months of progress is made on the basis of Table 2.
[4] See section: Where does uncertainty come from?
[5] Note that these are not the only values that are compatible with the data, see section: Recommendations to discuss uncertainty in findings.
[6] Decision-makers 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.

Comparing Z with A or B would be like a vertical comparison in Figure 1: between not-so-well-designed, tentative studies (at the top),[7] and well-conducted, more credible studies (at the bottom). 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, the findings of programme A (left) are compatible with more variability: from negative effects to larger positive impacts included in the intervals. In contrast, the findings of programme B (right) show less variability being only compatible with a very small negative effect or a larger positive effect. This compares the **uncertainty** of the findings.

***Figure 1**: Schematic representation of internal validity and uncertainty*



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 it may also be harmful; however, the best estimate of B found the same positive impact, but at worst the programme was not harmful. Thus, with this information a decision-maker may be more confident to implement B.

A standard way of assessing this uncertainty of a finding is using a **p-value**.[8] However, these are difficult to interpret for researchers and practitioners alike and have been widely criticised for misleading decision-making and biasing the literature (Wasserstein and Lazar, 2016; Wasserstein, et al., 2019; Amrhein, et al., 2019).

The EEF's classification system for single studies discusses aspects of internal validity. This document expands that by introducing the EEF's position on how to discuss uncertainty.

[7] 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.

[8] This estimates a measure of the compatibility between the observed data and a particular model of the data (see section: Uncertainty and significance testing for a description of the p-value).

What is internal validity?

To evaluate the impact of a programme or intervention, researchers would like to compare the outcomes of those treated and the outcomes they would have had, had they not received the intervention. This scenario is called the *counterfactual*. Clearly, it is not possible to observe both scenarios in the real world, which requires researchers to compare the results of the group that was treated with those of a group that was identified as a suitable comparison (that is, a valid counterfactual). The differences in outcomes between the treatment and the comparison groups is interpreted as the *estimate* of impact and measured as an 'effect size'.[9]

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—sample size and Minimum Detectable Effect Size (MDES);
- 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 the security of the findings. This systematically summarises the characteristics that define the internal validity and makes an estimate of impact more or less credible.[10]

Where does uncertainty come from?

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 are selected to take part in a study, random sampling leads to **sampling uncertainty**. Even if a random sample from the population, such schools or pupils might be different from the population at large for reasons we might not be able to identify.[11]
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 them 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.

---

[9] An ES measures the difference between the means of two groups, scaled by a measure of how variable or disperse outcomes are: it divides the mean difference between groups by the standard deviation.

[10] Note that this security rating 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.

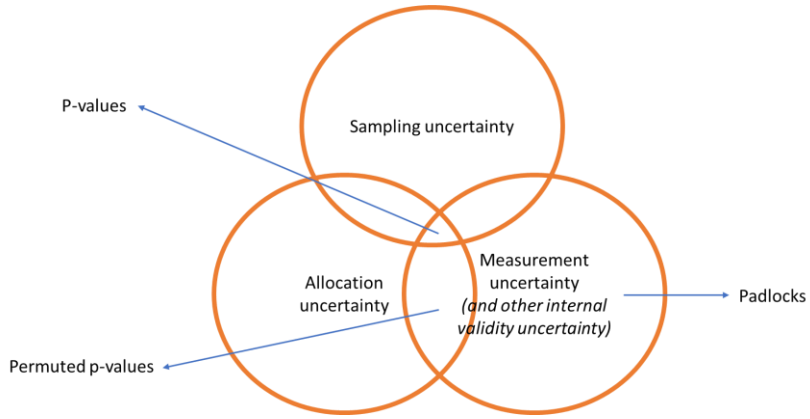
[11] Note that in most cases, samples of participants taking part in an RCT are not drawn at random from the population. However, **allocation uncertainty** would remain as one of the sources of uncertainty.

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

This means that it is always possible that the true effect size[12] observed in an RCT will differ from the true average effect size in the sample[13] or the population.[14] However, these are not the only sources of statistical uncertainty. For instance, the accuracy and reliability of an outcome test may also introduce **measurement uncertainty**. Therefore, it is not possible to completely isolate the multiple sources of uncertainty from some aspects of internal validity, as represented schematically in Figure 2.

Evaluators may choose to present statistics that rely on different assumptions and account for different aspects of uncertainty. For example, while **p-values** encompass information on sampling, allocation, and measurement uncertainties, alternatives like **permuted p-values** do not require the assumption of random sampling and therefore do not attempt to make generalisations beyond the sample. In any case, a transparent interpretation of these statistics requires a careful examination of the assumptions upon which they are constructed.

***Figure 2**: Schematic representation of sources of uncertainty*



Bringing together internal validity and uncertainty

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[15]—how close one is to the bull's

[12] You can never actually know the 'true average effect size' as you would need to know the pre-test and post-test outcomes for each member of the sample/population both with and without the intervention, which is not possible.

[13] 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.

[14] 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 & Cartwright, 2018).

[15] This relates to how close the results would be to the *true* estimate of impact in infinite hypothetical repetitions as is described in the section 'Uncertainty and significance testing'.

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—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 would 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.

Uncertainty and significance testing

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

1. a (random) sample is drawn from the population of interest;[16]
2. the *same* experiment is conducted a large number of times on samples drawn from the same population; 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 would it 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 is called the **p-value**, a statistic that has been strongly criticised for its proclivity to being misused and misinterpreted, leading to distortions in scientific enquiry (Wasserstein and Lazar, 2016; Wasserstein, et al., 2019; Amrhein, et al., 2019).

The reason for this common 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* (the hypothetical situation, which also requires fulfilling the other assumptions mentioned above)?'[17]

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 (Wasserstein and Lazar, 2016; Wasserstein, et al., 2019; Amrhein, et al., 2019). P-values give a very indirect answer to the question we are truly interested in.

One of the most salient problems with p-values 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 whilst, at the other side, 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 were supposed to be interpreted.[18] P-values provide a *continuum* of how compatible the data is with the hypothetical situation.

---

[16] 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). For this reason, in addition to computing p-values, the EEF encourages evaluators to use other statistics to represent uncertainty that are only related to the characteristics of the sample. For example, permuted p-values or Bayesian credibility intervals.

[17] Imagine you want to identify whether a CPD 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 says. In turn, 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).*

[18] 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

Values at either side of the threshold should not be treated as definitive answers but as different tonalities of grey effect—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, et al., 2019).

Recommendations to discuss uncertainty in findings

To address the criticisms above, the EEF recommend the following principles, which distil work by Wasserstein and Lazar (2016), Wasserstein, et al.(2019), and Amrhein, et al. (2019):

1. **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 understanding 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 advances in this direction, away from the detrimental simplification of findings as 'true' or not. *Evaluators must present a measure of the uncertainty around all ESs recognising that uncertainty is an integral part of statistical modelling and scientific enquiry.*

2. **Focus on practical/scientific significance**

   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 impact and should be discussed in all cases.* These may be considered alongside other sources of information to aid interpretation.

3. **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 of the implementation, and considerations of the context, among others. *Be thoughtful in describing how the finding shifts the evidence-base and existing priors.*

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

   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 sample mean between two groups would be equal or more extreme than the observed value in the study (Wasserstein & Lazar, 2016). As a *continuous* probability, p-values are a measure of the *degree of compatibility* of the data with the hypothetical model imposed to 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*

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 & Lazar, 2016).

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 neutrally, irrespective of whether results are 'positive' (positive ES, not statistically significant) or not. Other statements that suggest a dichotomous interpretation around the 0.05 should also be avoided. 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',[19] avoiding referring to 'confidence' intervals as the word confidence suggest ungranted certainty**

To report statistical uncertainty around the point estimate, discuss the practical 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, not interpreting them in a dichotomous way regardless of which statistic is chosen. Evaluators may 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 less stringent assumptions. The ASA's Special Issue, Statistical inference in the 21st century: A world beyond p<0.05, offers some suggestions. The EEF will consider commissioning a methodological piece exploring these alternatives.

The EEF encourages evaluators to present alternatives to test the sensitivity of the statistical uncertainty captured by different models.[20]

For the reasons outlined above, as from the publication of this statement, *the EEF does not describe findings as statistically significant or not.*

The EEF prefers presenting the point estimate as the best estimate of impact with a statement about the internal validity of the finding, which is captured by the EEF's classification of the security of the findings. Moreover, it is also important to represent the statistical uncertainty of the finding as a continuous p-value, 'compatibility intervals', and/or alternative statistics.

Acknowledgments

The EEF is grateful to Guillermo Rodriguez-Guzmán for leading on this statement, and Ronald Wasserstein, Nicole Lazar, Allen Schirm, and Steve Higgins for the invaluable comments and suggestions made to this statement. Any mistakes or omissions are the sole responsibility of the authors.

[19] 'Compatibility' intervals as referred by Amrhein, et al. (2019), Greenland, (2019), and Wasserstein, et al. (2019).
[20] For instance, when the CIs around a point estimate are very close to zero (for example, -0.01), the EEF encourages evaluators to test other analytical models as robustness checks to explore to what degree the statistical uncertainty is dependent on the analytical model chosen or how the uncertainty is measures (for example, p-values vs permuted p-values).

References

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

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

Gelman, A. (2016) 'The problems with p-values are not just with p-values', *The American Statistician, supplemental materials to ASA Statement on p-values and statistical significance,* 70, pp. 1–2.

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.

Kraft, M. A. (2018) 'Interpreting effect sizes of education Interventions', Brown University Working Paper.

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.