#

June 2017 • Standards of Evidence

**Understanding the impact dimension**



# The impact dimension of the Standards of Evidence provides information about the extent of measured improvement.

# Understanding the impact of initiatives helps focus our attention on those initiatives which are most effective.

# Accurate measurement of impact may be influenced by the quality of the initiative’s design.

# The impact dimension includes five levels to indicate the strength of the evidence demonstrating impact. (See diagram to the right)

#

**Strength of Evidence**

# There are many different ways to measure impact and the way chosen may be influenced by practical realities such as resourcing or data collection constraints.



**Different ways to measure impact**

|  |  |  |
| --- | --- | --- |
| **Time-based comparison***Comparing pre-test scores with post-test scores* | **Group-based comparison***Comparing Group A test scores with Group B test scores* | **Effect size** |
| * Compares performance before and after, through gathering pre-initiative (Time 1) and post-initiative (Time 2) performance data. Ideally performance is measured using the same scale.
* The simplest way to measure change in performance.
* A time-based comparison will not give information about whether improvements in performance would have occurred in the absence of the initiative. That is, time-based comparisons, by themselves, will not support attribution (as per the design dimension) or provide a sufficient measure of impact (as per the impact dimension).
 | * Compares the performance of one group who receives the initiative (target group) against the performance of another group who does not receive the initiative (comparison group).
* Comparing impact across groups may support attribution as per the design dimension.
* A comparison group helps to account for improvement that might have occurred without the initiative.
* It may be challenging to find two separate comparable groups. Other possible options include use of historical comparison groups or published norms relevant to the target group.
 | * An effect size indicates the strength of difference in SD units, enabling comparison between initiatives assessed on different scales.
* There are different ways of calculating effect size, depending on the comparison group and the nature of the data (e.g. percentages versus numbers).
* When interpreting an effect size, reference should first be made to the existing literature.
* The department has developed effect size thresholds for each of the five levels within the impact dimension based on a commonly used ‘rule of thumb’. These are intended as a guide.
 |
| **Things to consider:*** Data gathered before an initiative is implemented (Time 1) is known as baseline data. Baseline data is a pre-requisite for measuring impact.
* An aggregate of the data collected at each time point should be used to summarise the performance of a group (for example, the average, median or mode).It may be informative to look at aggregate data separately for different sub-groups of students as this could show group specific changes in performance.
* A direct comparison of time-based group averages can serve as a useful indicator of change when Time 2 is measured on the same scale. For example, when using a common writing marking guide across two writing tasks, the change in group average between Time 1 and Time 2 can meaningfully describe the measured change in performance of the group.
* Standard deviation (SD) is a measure of the spread of all individual scores relative to the average score of the group. Examining differences in SD between Time 1 and Time 2 could provide information about the effectiveness of an initiative for students performing at different levels. For example, a reduction in the SD indicates a smaller distance between the lower and higher performing students. Together with other data items this could indicate the initiative has lifted the lower end of the performance spectrum, or conversely, has failed to lift the performance of the top performing students.
* When Time 2 performance is assessed on a different measurement scale, the two scales need to be standardised or converted to a common scale before assessing a change in performance.
* Larger groups deliver more reliable information regarding the impact of initiatives. Data drawn from groups of 20 or less should be interpreted with caution.
 | **Things to consider:*** The target and comparison groups should:
* be roughly equal in size
* be as similar as possible in important characteristics (e.g. demographic composition, academic ability)
* have similar Time 1 performance on the key measure of interest. This is known as baseline equivalence. Establishing baseline equivalence supports attribution as per the design dimension.
* Scores for both the target and comparison groups should be normally distributed and not contain any large outliers.
* A standardised performance metric can be calculated (that is, an effect size). In two-group designs, the comparison group average at a given point of time (that is, Time 1 or Time 2) is usually subtracted from the target group average and then divided by a SD. Attention to the way aggregate data is visually presented can strengthen an initiative’s claim to impact.
* Larger groups deliver more reliable information regarding the impact of initiatives. Data drawn from groups of 20 or less should be interpreted with caution.
 | **Effect size threshold** | **Strength of impact** |
| **>0.6** | **Very high** |
|  | *Very large measured improvement* |
| **0.4 to 0.59** | **High** |
|  | *Large measured improvement* |
| **0.2 to 0.39** | **Moderate** |
|  | *Medium measured improvement* |
| **<0.2** | **Low** |
|  | *Small measured improvement that can be reasonably linked to the initiative* |
| **Not applicable** | **Impact unknown** |
|  | *Impact cannot be measured or unintended impact is identified* |
|  |  |