5

HOW DO WE KNOW EDUCATIONAL INTERVENTIONS WORK?

KEVIN D. HOCHARD

University of Chester, Chester, UK

ABSTRACT

The overall aim of this chapter is to focus on the process of, and issues warranting consideration for, the evaluation of educational interventions. In particular, to outline some key considerations for educators to follow when assessing the evidence-base for interventions they might be considering for use in their practice. Also, important considerations for those wishing to evaluate the effectiveness of an intervention they have initiated, as well as a useful checklist which summarises this all. Recognising that some readers of this chapter might be practitioners rather than researchers, it has been written with the practitioner in mind in, hopefully, a simple and practical way. There are, however, further opportunities for additional reading and resources signposted throughout for those who wish to read up on any of these areas more. In addition to those cited throughout and referenced in the Reference list at the end, there is also section that provides the author's Additional Recommended Readings and Resources to follow-up on. Readers might also want to refer to Chapter 3 in this book which discusses Single versus Multiple PPI approaches.

Keywords: Evaluation; PPI; positive–psychology intervention; educational intervention; robust; reliable

INTRODUCTION

Wasting resources, or worse – the harming of participants, can be the unintended consequence of failing to ensure any educational intervention can affect a desired change. Thus, any claim that an intervention 'works' must be made on solid foundations rooted in scientific evidence if we are to see consistent measurable improvements on a given behaviour or metric. Sadly, a large gap exists between what we think might work and what ultimately does work.

CONTEMPLATING STUDY DESIGN

At its most basic, determining that an intervention 'works' is a straightforward matter of observing a change in a selected outcome following the administering of the intervention. For example, a change in observable behaviours, or an increase in wellbeing as reflected by a significant increase in the scores on a scale. This can be done through a variety of means. However, all these methods share a common element. That is, one must assess if the outcome of interest has changed relative to either a baseline (i.e. the level of the outcome variable measured prior to the intervention), or a control condition (e.g. a group not receiving the intervention).

Although this explanation is an oversimplification, and throughout this chapter, we shall touch on a variety of nuances worth consideration when assessing the utility of an intervention, this basic principle remains. The intervention must change the outcome, either:

a) from Time 1 (before the intervention) to Time 2 (after the intervention); or
b) to a differing extent in the intervention group compared to a control; or, better yet,
c) from Time 1 to Time 2 for the intervention group, relative to any changes across time-points observed in the control group.

Fig. 4(a)– (c) provide a representation of these various scenarios.

Let's Consider

For example, imagine an intervention which aims to increase the wellbeing in children in a Year 11 class.

Under scenario (a), we would measure the children's wellbeing pre- and post-intervention. Assuming our intervention works, we should be able to see,

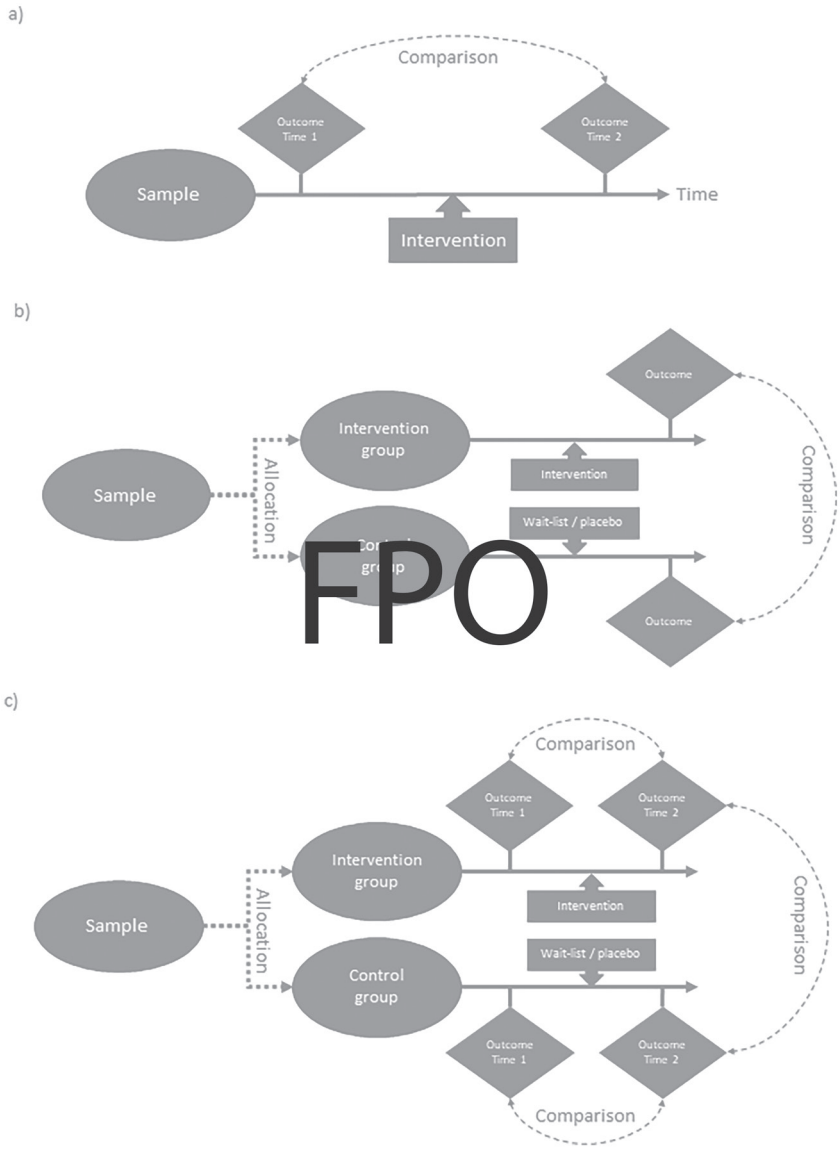

Fig. 4. Possible points of comparisons in experimental studies AQ1

via statistical comparison, that wellbeing is greater post-intervention relative to wellbeing pre-intervention.

Under scenario (b), our sample of Year 11's would be allocated to, either, the intervention or a control condition. This could be a *wait-list control*,[1] a *placebo/sham*[2] intervention, or an *alternative* intervention. The intervention group would receive the intervention, whilst the control group would receive

the alternative or simply wait, as is implied by the wait-list. The outcome would then be measured for comparison. If the intervention works, we should see the intervention group display significantly greater wellbeing than the controls on statistical comparison. Statistical significance is important in scientific research because it allows researchers to hold a degree of confidence that their findings are real, reliable, and not due to chance.

Under scenario (c), our sample of Year 11's would, again, be allocated to, either, intervention or control conditions, but this time measurements of wellbeing would be both pre- and post-intervention for each group. This would allow us to see if changes in wellbeing are significantly different as assessed by our statistical analysis, relative to the control condition. We should, also, look to see that the change in wellbeing for the intervention group from pre- to post- is greater than any change which may have been observed in the control condition.

When assessing the evidence that an intervention works, it is recommended that educational practitioners give more weight to research studies which provide the greatest level of comparison. In this case, studies that compare interventions to a control group, and where there is consideration of baseline levels of the outcome of interest (scenario 'c'). Should studies with these features not be available, as can often be the case, caution is recommended for reasons that will become clear throughout this chapter.

HIERARCHIES OF EVIDENCE

There are several ways to assess if an intervention brings about the desired outcome. However, as not all studies are designed in a way that will yield suitable evidence to draw this conclusion, not all studies are equal when assessing the quality of the evidence produced. Fig. 5 displays a hierarchy of evidence quality for study designs.

In addition to differences in the quality of the evidence, some designs are simply more suited to assessing how well an intervention works than others. For example, a *cross-sectional observational* study measures variables of interest at a singular time-point and, as such, this makes comparisons between time-points impossible to do. Further, as with all observational studies, with a cross-sectional observational study, no intervention would be delivered making it impossible to determine the causal influence on fluctuations in the level of a given outcome. Cross-sectional observational studies, therefore, are better suited to determine if variables are associated with each other and can yield information on the prevalence of a given behaviour.

As shown in Fig. 5, studies higher up on the hierarchy provide higher quality evidence due to the increasing control they place on extraneous[3] variables.

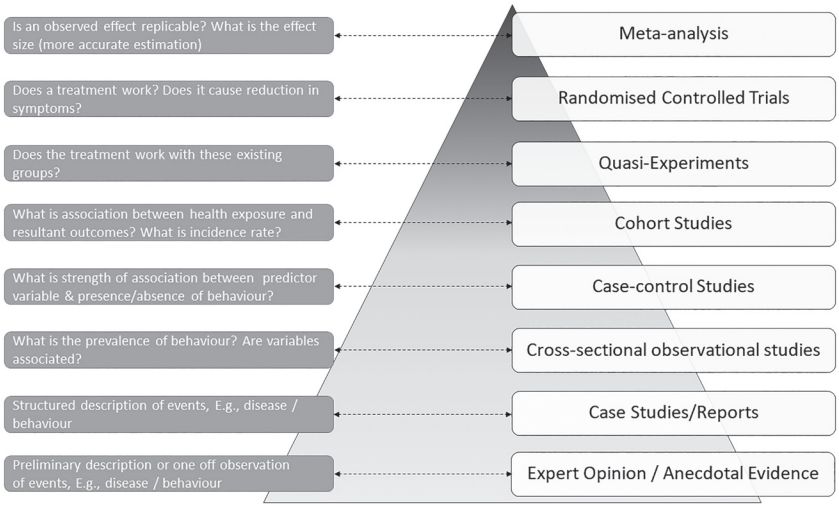

Fig. 5. Hierarchy of Evidence From Study Designs and the Type of Research Questions They Address.

For example, studies where one variable is manipulated under controlled conditions, known as experiments, are particularly well suited to inform on whether or not interventions work. *Quasi-experiments* and *Randomised Controlled Trials* (RCT) are two examples of types of experimental designs. In these studies, the manipulation lays in the delivery of an educational intervention compared to not delivering it. Likening it to flicking a light switch on and off, an experiment with an active intervention and a control group would aim to test if the intervention produces the light-like behaviour of interest relative to the darkness of the control group. Assuming that the intervention 'works', the lack of any effect reported by the control group is of great importance as this demonstrates that the intervention is responsible for changes in the desired behaviour, and not other environmental factors.

A PLAUSIBLE CONTROL

A hallmark of good experimental studies is their focus on eliminating sources of bias and alternative explanations for an observed change in a given outcome. This can be achieved by including several design features to the study. One such feature is the inclusion of a *plausible control* condition. Trials with weaker, or less plausible, controls tend to show interventions as having a greater impact than those with a good control, such as an alternative active control.

Let's Consider

For example, consider an intervention study aiming to test if one can increase levels of physical activity in primary school Physical Education classes by providing wrist-worn activity monitoring devices to students. With these devices, visual representations of activity levels appear in the form of a circle being filled by a bright colour. We might assume that the mechanism of effect is the positively reinforcing nature of the bright-coloured circles representing the number of steps taken.

Let us further assume that the children are allocated to, either:

1. an *active intervention* group, who are provided with wrist-worn actigraphy devices which they are asked to wear for a certain period of time each day; or

2. a *control* group, placed on a wait-list and receiving no intervention.

After a few weeks, the researchers find that the children in the active intervention group are more physically active than those in the control group. One may surmise, therefore, that the intervention has been successful. However, it is also possible that some of the increase in activity levels observed in the active intervention group may be due to the *placebo effect*. For example, children in the active intervention group may feel more motivated to be active because they have received and are wearing the devices; they believe that they are part of a special group that is supposed to be more active. This belief may lead them to being more active even if the devices themselves do not directly cause increased activity levels.

To account for this possibility, researchers could use a placebo condition in which the control group is given non-functional wrist-worn devices that look identical to those given to the active intervention group. This can help to control for the placebo effect and ensure that any differences in activity levels between the groups are truly due to the intervention, and not just to the participants' beliefs or expectations. Evidence from a study with stronger controls can help rule out alternative explanations of the intervention effect and increase the credibility of the intervention, due to the comparison being more stringent.

REGRESSION TO THE MEAN

Experimental studies also incorporate control groups to account for a phenomenon known as *regression to the mean*. This occurs when extreme or unusual values in a dataset, caused by random or temporary factors, become

less extreme or more 'normal'. This can happen when measured repeatedly, or under different conditions, as temporary factors dissipate, or random factors are not present in the new context.

Let's Consider

For example, refer back to the aforementioned physical activity intervention but this time when no control group is used. It would make common sense if the class selected for the intervention due to low activity levels were measured to assess their baseline levels of activity prior to the start of the intervention. Physical activity levels may appear to increase after our intervention, but this could be due to regression to the mean. This is where low initial activity levels rise naturally over time, instead of due to the intervention itself.

Regression to the mean can be caused by temporary factors, such as illness, that may have affected the class' baseline activity levels. The addition of a control group, assuming similarly low activity levels at baseline for both intervention and control participants, would account for this bias.

RANDOMISED CONTROLLED TRIALS (RCTs)

Most fields of knowledge hold RCTs as the gold standard design when assessing the effects of an intervention (e.g. Worral, 2010). With RCTs, this design randomly allocates participants to, either, an intervention or control group to ensure that any differences in outcomes between the groups can be attributed to the treatment being tested, rather than any variance in factors between groups.

Let's Consider

For example, referring back to the aforementioned physical activity study, this could implement random allocation of pupils to the active intervention or control condition. Doing this would reduce biases in allocation and avoid systematic differences between the groups. It also accounts for confounding variables and ensures that the groups being compared are similar in all relevant characteristics. For instance, pupils with higher athletic abilities would not be disproportionately allocated to one group or another following a randomisation procedure, thus avoiding this characteristic from skewing the results of the intervention. RCT designs aim to maximise the internal validity[4] of

the findings, ensuring that the efficacy of the intervention is established for a specified population.

CHALLENGES WITH THE HIERARCHY OF EVIDENCE

Whilst it is true that study designs which appear higher on the hierarchy of evidence are more appropriate for assessing interventions, it is not always feasible for such study designs to be conducted. For example, in the early stages of evidence gathering for a newly introduced intervention, it may be challenging to persuade ethics committees or gatekeeper organisations to implement an intervention with a limited track record of success. Additionally, study designs higher in the hierarchy require significantly more resources to conduct, and securing funds may be difficult without firstly demonstrating some promising, yet basic, results with lower quality studies. For example, it may be necessary to demonstrate that an intervention brings about a desired change from baseline to post-intervention in as many pupils as possible for the effective statistical evaluation of a study's findings. This is especially important if the study has a small participant pool, or if gatekeeper organisations are reluctant to allow more pupils to participate, making it difficult to include control groups in the analysis.

Consequently, practitioners interested in brand new interventions may encounter a plethora of published studies ranked lower in the evidence hierarchy due to pragmatic limitations for the researcher developing the intervention. It is, therefore, recommended that practitioners exercise caution when considering the application of novel interventions, albeit they should avoid dismissing them outright. Only through continued research can confidence be gained in how well an intervention does, or does not, work. Evidence from a mix of observational and non-controlled experimental studies can yield compelling evidence when RCTs are difficult, or unethical, to perform. For an excellent example, see the classic article from White (1990), who details how epidemiological studies and in-lab cellular experiments were instrumental in establishing smoking as a causal factor for lung cancer.

USING ESTABLISHED INTERVENTIONS

More established interventions will generally benefit from a number of higher quality studies, such as RCTs, or a *meta-analysis* of RCT studies. Meta-analyses, as the name suggests, is a study of existing studies. These assess the sum of available evidence on a given topic by pooling data from all studies on that topic which have been carried out. For example, in a meta-analysis, all studies

aiming to replicate the effects of an intervention would be searched for, and screened for their appropriateness and should they meet a strict inclusion criterion, the key findings – as well as information relating to the sample in each study, would be extracted.

Assuming all the studies included in the meta-analysis have been conducted reliably and robustly, and without bias when selecting the studies to be included, they generally yield higher quality evidence than a singular study. A selection criterion to control for this is often part of the process. This higher quality is achieved by the following:

1. increases in the sample size, by pooling together the number of participants from all the included studies, leading to greater statistical power;
2. the dilution of the influence of random error in one study across multiple studies;
3. the increase in generalisability of findings from data obtained in varying samples and contexts; and
4. highlighting the subtle variations from one study to the next giving an insight into why findings may be different from one study to the next, and ultimately the resolving conflicting findings across several studies of equivalent quality such as RCTs.

If performed correctly, a meta-analysis could also provide a robust empirical assessment of an intervention's performance and quantify the magnitude of change in a given outcome.

CAUTION

It is important to note, however, that meta-analyses are only as good as the study from which they are derived. As the saying goes, 'garbage in, garbage out'. As such, practitioners reading meta-analyses should consider the design of the studies included. After all, a meta-analysis of observational cross-sectional studies cannot provide insight as to how well an intervention works, given that individual observational cross-sectional studies lack the design features to evaluate this aspect.

In addition to benefiting from an evidence-base comprising higher quality studies, other benefits of using established interventions are that the designs will often present examples of useful evaluation tools that a practitioner may also decide to use. Further, they will benefit from tried and tested procedures, which can eliminate the need of running a pilot study.

See the checklist at the end of this chapter for what to look for when selecting interventions and assessing their evidence-base.

EVALUATING NEW INTERVENTIONS

Whilst there are several benefits to using established interventions, using new interventions is important and may be necessary too. Trialling new interventions can help us advance our understanding of what does, or doesn't work. They may also be more appropriate to change an outcome not previously investigated, or has been tailored for a specific context.

When evaluating new interventions, it is recommended that practitioners carefully consider the design of studies. Specifically, if the design is aligned to the aims of the intervention. They should also consider the quality of the control group against which the intervention is being judged. Interventions that demonstrate robust effects when compared to other active interventions will have, in essence, survived a harsher test compared to interventions judged against a waiting-list control.

Educational practitioners must also make use of their own knowledge of the field to make a reasoned assessment of evidence quality. Practitioners will be well placed to know pragmatic limitations, for instance, the possibility of randomising students to an intervention or control, or how feasible it would be to run a control group in the first place. You are also an expert on the students that you teach and, as part of this, what might, or might not, work for them. There is no one size fits all.

Practitioners should critically evaluate the strength of the evidence for any intervention in light of pragmatic limitations, as no applied research is ever perfect. It may also be necessary to conduct multiple rounds of testing in order to accurately determine the effectiveness of the intervention. Only through the repeated evaluation of an intervention can one truly gain a degree of confidence in the utility of said intervention. This replication process is a hallmark of the scientific process (Open Science Collaboration, 2015). As such, practitioners are strongly advised to seek out interventions with a good replication track record.

In the event that practitioners are interested in creating their own novel intervention, tools exist to help guide this process. For example, the 6SQuID (Wight et al., 2016) provides a six-step framework for designing effective interventions. This framework pinpoints the problem and its roots, allowing for the identification of modifiable factors and the primary beneficiaries of the intervention. By selecting the right mechanisms for change and strategising as to the best method of delivery, the approach allows for the novel intervention to be tested to gather compelling evidence to validate its effectiveness before a thorough evaluation.

EFFICACY VERSUS EFFECTIVENESS: EVIDENCE MEETS THE REAL-WORLD

The effectiveness of an intervention varies based on the specific context in which it is applied, and adjustments or adaptations may be needed to achieve the intended outcomes. As described previously, whilst many fields regard the RCT as the gold standard for evaluating an intervention, the high internal validity of these studies can also make it challenging to apply the findings to real-world situations. This is because the results from highly controlled environments, such as those established for RCTs, to create standardised conditions and eliminate biases or alternative explanations for potential effects, may not be universally applicable. This could be due to subtle contextual influences and the range of practical considerations an educator implementing these interventions might face on any given day. Therefore, the credibility of results indicating that an intervention can modify an outcome under optimal conditions (i.e. the efficacy of an intervention derived from an explanatory trial) may not align with the intervention's capacity to induce change in real-world settings (i.e. the effectiveness of an intervention derived from a pragmatic trial).

Pragmatic trials aim to test the effectiveness of an intervention in real-world settings within a participant sample that is typical of the intended recipients of that intervention. As such, the sample recruited to the trial would have a less restrictive inclusion criteria. With this broader eligibility criteria, the sample would be more representative making findings more generalisable. Pragmatic trials themselves are designed to mimic the real-world delivery of an intervention, with pragmatic constraints incorporated. This allows for findings from the trial to have direct applicability to educational contexts.

Additionally, delivering a trial under real-world settings allows for the use of the existing practice as the control group, rather than comparing against an artificial placebo control or wait-list. These trials also tend to make use of routinely collected data as outcome measures, allowing us to assess on outcomes of inherent importance to educational practitioners and stakeholders.

Let's Consider

For example, to highlight the distinction between explanatory and pragmatic trials, consider an educational intervention aimed at evaluating the impact of homework on student performance in standardised testing. In our explanatory trial, we would randomly assign students to either an intervention group, which receives a specific amount and type of homework, or a control group, which continues with the usual homework practices. Students would be

selected for inclusion into the trial based on a set of standardised criteria, ensuring that the intervention group and control group are matched (e.g. by age and baseline ability demonstrated in prior standardised testing).

Rather than use existing homework, and ideally working alongside the educator with this, the homework for the intervention group might be designed by the researchers themselves to ensure it aligns perfectly with the learning objectives of the standardised test. For example, that a test of recall is linked to a piece of homework designed to improve recall. Performance on the standardised test would, then, be compared between the two groups, with the effectiveness of the intervention evidenced by higher recall scores in the tests completed by the intervention group.

This explanatory trial, however, is designed to test the efficacy of the homework intervention under ideal conditions. It yields strong evidence that any observed improvements in performance are due to the intervention itself, rather than other factors.

Conversely, our pragmatic trial aims to test the effectiveness of the homework intervention in real-world conditions. Therefore, this trial would be designed so that the intervention can be implemented across various classrooms (e.g. of varying ages and not controlling for ability), by regular teachers, who may adjust the intervention homework to suit their teaching style, or the specific needs of their students. The students' performance might also be assessed using a variety of measures, including standardised tests, teacher assessments, and student self-reports. Moreover, the results might be compared, not only between the intervention and control groups, but also across different classrooms, teachers, and schools.

BENEFITS OF TRIALS

Each trial type provides valuable information on the homework intervention. The explanatory trials provide us with an indication of the efficacy of the intervention under controlled (ideal) conditions, and insights into whether or not the intervention caused the changes in standardised test scores. Meanwhilst, the pragmatic trial provides insight into contextual factors and the feasibility of implementing the intervention on a larger scale. Whilst the above serves as an example of these differing trial types, however, readers should note that trials are never entirely explanatory or pragmatic; rather they fall on a continuum (Gartlehner et al., 2006; Patsopoulos, 2011).

For example, through a randomised controlled trial design Hochard et al. (2021) compared the efficacy of a brief values and acceptance exercise (common to Acceptance and Commitment Therapy) compared to cognitive restructuring

techniques (common to Cognitive Behaviour Therapy) to improve social resilience in university students. However, the trial was designed to mimic pragmatic limitations of university support services such as the likelihood that interventions would be delivered by non-experts in psychotherapy. This allowed the findings to inform on the feasibility of delivering the intervention in real-world settings whilst maintaining sufficiently rigorous experimental control to indicate which intervention was more efficacious.

Strictly speaking, if a practitioner is interested in whether the intervention itself yields the effect, then explanatory trials should be given particular attention. However, the generalisability of the literature's findings is key to ensure the intervention produces the desired effects. To avoid disappointment and wasting resources, practitioners are recommended to consider where and how they intend to implement the intervention of interest, and if the evidence-base for the intervention has applicability to those settings. The more the intervention has been shown to work in real-world settings, the better.

SELECTING EVALUATION MEASURES

Given that observing changes across time, or relative to a control condition, is essential to determine if an intervention works, selecting a tool to capture that change on the outcome of interest must be carefully considered and specified on a measurable criterion. Whilst there is also value in obtaining qualitative data, the quantifying of the outcome allows for interventions to be evaluated in an objective and reliable fashion. Further, quantifying the outcomes allows us to apply statistical inference to help rule out that changes in our outcome could have been due to chance, for example, because of variations in scores, or frequencies of the outcome in our sample. Well-constructed measures also reduce *random error*, providing more precise estimates of the impact of an intervention.

Let's Consider

Successful interventions change outcomes. For example, an intervention designed to increase wellbeing over a school term will be deemed successful if the wellbeing of participating pupils has improved over the course of that term. To check that this has occurred, wellbeing would need to be measured, and this can only be achieved if we have clearly defined what was meant by wellbeing and operationalised this definition. That is, to create a precise description of how the variable will be measured. See Chapter 7 for more about the 'complexity' of wellbeing.

In this case, we might operationally define wellbeing as:

> *a state of positive physical, social, and emotional functioning characterised by blood pressure in normal range, positive affect, low stress, and the absence of depression and anxiety.*

Based on such a definition, we might expect a study to make use of well-validated instruments designed to measure each of these variables. So, we could measure blood pressure via a mercury sphygmomanometer, whilst affect, depression, anxiety, and stress might be measured by a self-report tool, such as the Positive and Negative Affect Scale (PANAS; Watson et al., 1988) and the Depression Anxiety Stress Scale (DASS-21; Henry & Crawford, 2005), respectively. One of the additional benefits of using tried and tested interventions is that the research will often recommend a reliable scale. The same research may also provide comparable scores (e.g. means), which can also be useful too.

Clear operational definitions help ensure that researchers are all measuring the same characteristic in a consistent manner, making their results comparable and replicable. It also helps to avoid ambiguity or confusion about what is being studied, and the impact of interventions to be tested empirically by assessing changes in the intervention's targeted outcome behaviour. Whether this outcome changes in terms of either its frequency, duration, magnitude, or quality, will be dependent on the definition of this measurable outcome and the tool employed.

SELECTING EVALUATION TOOLS

A great many measurement tools exist. These can be objective measurements that assess observable data in a highly reliable and verifiable manner, making them resistant to personal biases. Alternatively, subjective measurements have been developed to assess one's perception, feelings or thoughts on a given topic. Whilst they will have been carefully developed to be standardised and reliable, as these tools require introspection, they can be open to biases, such as memory distortions.

Studies which can measure outcomes with objective instruments will provide higher quality data for practitioners to judge the effectiveness of interventions on. However, regardless of the objective or subjective nature of the tool, it is vital that the instrument adequately captures the outcome of interest.

Let's Consider

For example, take our earlier example of an intervention aimed to increase wellbeing as per the earlier operational definition. Focusing solely on the

objective mercury sphygmomanometer to assess blood pressure would be limiting as wellbeing has many psychological components which cannot be assessed by external observation. Well-validated introspective tools measuring wellbeing, such as the DASS-21 in our example, could tell us much about the impact of the intervention on our target population by identifying lower levels of depression and anxiety.

MEDIATORS AND MODERATORS OF CHANGE

Mediators of Change

Studies may also assess a *mediator* of change. These are variables which bridge between the intervention and its' ultimate effect. For example, an intervention which aims to increase academic test scores as its ultimate goal may function by increasing the number of revision hours pupils engage in. In this case, the increase in test scores due to the intervention would have been mediated, or brought about, via the mediator of having had an increased number of revision sessions.

By investigating mediators, studies shed light on the mechanism via which an intervention works. This can be informative when considering our own application of the aforementioned intervention.

Let's Consider

For example, assume that we wish to increase academic test scores in an underperforming high school class. Increasing the number of revision hours as an intervention may seem ideal, though our knowledge of the mediator of effect may make us reconsider if we were short of time before the day of the examination. The limited revision opportunities associated with this would mean that this intervention may have limited utility due to the mechanism of change for our desired effect being stifled.

MODERATORS OF CHANGE

Whilst mediators provide information about mechanisms of action, a *moderator* provides information regarding the conditions under which an effect might be observable.

Let's Consider

For example, our intervention to increase test scores (outcome) via increased revision sessions (mediator) may work best for one sex (moderator) relative to the other. Thus, if the intervention was moderated by sex, we might see a larger effect in females than males.

It is important, however, to keep in mind that this does not imply that the intervention would be useless in males. In our example, one would need to see the extent to which sex moderated the effect of our intervention, if at all.

Research on moderators is beneficial to practitioners as they provide further contextual information when considering the implementation of interventions in their given contexts. Armed with such information, practitioners will be able to make more nuanced decisions as to when and if to intervene, and to consider alternatives, or the tailoring of interventions, to maximise their impact on their target sample.

DETERMINING MEASURES OF SUCCESS

When assessing the quality of evidence for a particular educational intervention, practitioners are advised to think carefully about the measurements mentioned in any given study as these will need to be considered in light of any operational definition provided. In particular, it is suggested that the validity of the outcome measures employed are carefully considered. That is, does the measure used to assess the intervention reflect the outcome or behaviour the intervention was designed to change. Moreover, does the intervention's demonstrated effect rely solely on subjective measurements? Should this be the case, have studies aimed to mitigate or rule out biases? This might be achieved with careful experimental controls, or by means of using varying measurement tools for triangulation[5] purposes. After all, studies which can obtain converging results with varied subjective measurement methods will be more credible than studies reliant on singular subjective measures.

As previously explained, practitioners should also consider if the intervention studies have explored mediators and moderators of the effect as they could provide useful insights into the applicability of the intervention within their professional practice.

PARTICIPANTS

Clarity as to the intended beneficiaries needs to be a key consideration from the offset for any practitioner developing or testing an intervention. Not only

will the selection and composition of participant samples providing the evidence-base of an intervention play a pivotal role in determining the quality of the study, but it will also impact on the *validity* of the findings.

Let's Consider

For example, imagine a mindfulness intervention devised to target behaviours that are a challenge in primary school children. Whilst searching the extant literature, we only find a handful of studies evaluating the effects of said interventions with secondary school students. Practitioners wishing to implement this mindfulness intervention in primary school settings, therefore, should be weary of the evidence detailed in these studies, and of implementing the intervention in question without pause. This is because the evidence-base for said intervention would lack validity (i.e. the evidence for or against the intervention is connected to a group of people who are not the intended recipients of your intervention).

Were the findings of these studies to show the intervention had failed to change behaviour in these secondary schoolers, therefore, we could not be certain if this was due to issues with the intervention itself, or its application outside of its initially intended setting. Alternatively, had the studies reported the intervention to have changed behaviour as anticipated, we would remain non-the-wiser about its potential impact in primary-aged children as the evidence would pertain to children in a secondary school setting only.

CONTEXT MATTERS

As I am sure you are already thinking, generalising the findings from secondary school to primary school students could be inappropriate for a variety of reasons all too familiar to educational practitioners. For instance, success with secondary schoolers could be, at least in part, due to the quality of instruction and the language used which may have resonated with these older children. Said language might be ill-understood by younger children in primary settings, leading to disengagement and, ultimately, a lack of effect. Alternatively, the behaviours that challenge in secondary schools (e.g. truancy, drug use, mental health issues) will differ greatly from those observed in primary (e.g. ease of distraction, or undiagnosed special educational need, such as dyslexia). As such, this makes the intervention a 'poor fit' for the behaviours it aims to change.

Similarly, a study reporting the impact of the intervention on a group of primary children, but which included children without behaviours that challenge,

could dilute the intervention's effect and lead to inaccurate conclusions about its usefulness. Therefore, clearly defining the participant characteristics and ensuring that they align with the study's objectives is crucial for obtaining valid results.

SAMPLE VERSUS POPULATION

All research is constrained by resources, opportunity, and ethics. It is, either, impractical or impossible to study an entire population. As such, studies will generally report findings based on a smaller group of participants, *a sample*, that ought to be representative of the population.

Let's Consider

For example, referring to the previous example, our sample could be primary school children in a particular school. Alternatively, and if resources allow, several schools in a given local authority.

The quality of this sample will impact the study's findings and their generalisability. Should our mindfulness intervention designed to target challenging behaviour be shown to improve problem behaviours in schools across a local authority, we might feel confident in implementing it to primary schools across the country. However, should our desire be to implement the intervention in a school from a deprived local authority, our confidence would likely wane upon discovering that the local authority, where testing occurred, was the most affluent in the country. The study findings being derived from an unrepresentative sample means that they may not be applicable to schools in different local authorities.

To mitigate such issues of validity, researchers frequently aim to use *random sampling*. This is where each member of the total population has an equal chance of being selected. Such a technique ensures the sample is representative of the total population, making findings more likely to generalise. However, within educational settings, random sampling could be difficult without clear support from gatekeepers and sufficient resources.

Despite this, not all is lost if random sampling is unfeasible. Instead, it requires the intervention to be shown to work in multiple studies, and from a variety of settings, to ensure generalisability. This will be more time consuming and could, in the long run, be more resource intensive. However, it may be necessary to account for pragmatic constraints in the research process.

SAMPLE SIZE

Complicating matters further, *sample size* needs to be considered. Assuming random sampling has occurred, larger samples are more representative of the total population. Further, and all things being equal, our ability to detect the effect of an intervention on a given outcome increases as a function of the sample size. Smaller samples will be unlikely to detect small changes in an outcome; only large changes in outcomes will be detectable to a level meeting *statistical significance*.

STATISTICAL SIGNIFICANCE

The term 'statistically significant' has appeared quite a bit throughout this chapter, so it is important to explain what it means.

When differences in outcomes are labelled as statistically significant through statistical comparisons, it simply means that, under a specific statistical model (e.g. comparing the mean scores of two groups), and assuming our null hypothesis is correct (i.e. that we should observe no difference between the two groups), the observed effect, or an even more extreme effect, has a low likelihood of happening by chance. Thus, the effect is attributed to the intervention when experimental controls have ruled out alternative explanations. However, this does not provide information on the magnitude of the change in the outcome. The magnitude of the difference, or 'effect size', between groups, or the change from one time-point to the next, should very much be considered by those evaluating interventions.

Increasingly larger samples will be able to detect increasingly smaller effects in outcomes from an intervention. Researchers, therefore, usually aim to recruit the largest sample possible to detect the smallest relevant effect. What makes for a relevant effect, sometimes described as a *clinically significant effect*, is also context-dependent.

Let's Consider

For example, in the field of clinical psychology and psychiatry, guidance from the UK's National Institute for Health and Clinical Excellence (NICE) suggests that an effect size of 0.5 is clinically significant when treating depression via pharmacotherapy (Moncrieff & Kirsch, 2005). An effect of 0.5 is deemed to be a medium effect within the realm of psychology (Cohen, 1988). Contrastingly, in sports science, effect sizes of between 0.5 and 1.25 are deemed small when

looking at changes in previously untrained individuals (Rhea, 2004). Researchers will need to reconcile the pros of larger samples for the detection of intervention effects, against the pragmatic cons of resource intensive recruitment when designing their studies.

An intervention is more likely to work if it has been tailored to account for the characteristics and context of the intended recipients. As such, how well an intervention works must be determined by testing it on the population for which it has been devised. Adapting an intervention for a different setting or population should always be done carefully to ensure the resulting modified intervention is acceptable to the new population. Ideally, it should also be tested to ensure it remains effective longer term.

DETERMINING THE BEST SAMPLE SIZE

It is recommended that educational practitioners carefully consider the sample described in any intervention study they read and reflect on how generalisable the findings may be to their given setting. Is the evidence for the intervention derived from a sample that shares characteristics with the practitioner's intended recipients? And, if not, can the intervention be adapted whilst remaining acceptable to the new intended recipients without losing its potency? Has the study been performed using a suitably large sample to detect the effect size of interest?

New and untested interventions will likely have been run with smaller samples due to ethical concerns due to the lack of track record for the intervention. As such, practitioners will need to consider this and cautiously interpret findings from small sample studies. In particular, practitioners considering the state of evidence must be cautious of sample size as small samples can inflate observed effects (Button et al., 2013). This makes interventions appear as promising in initial pilot studies, but lack-lustre with smaller or non-significant effects in larger representative samples. It will likely be easier to obtain answers to these questions for well-established interventions.

For newer interventions, however, practitioners may need to rely on their judgement, or perform their own study. Regardless, a thoughtful evaluation of the applicability of a study's findings to the practitioner's own context should provide them with a balanced perspective and realistic expectations of the interventions they encounter throughout their literature searches.

Some of the earlier chapters in this book describe some of the challenges of carrying out PPIs in schools and, if you have read any of these, you will note several pragmatic issues related to sample size.

TO SUMMARISE…

This chapter has aimed to provide educational practitioners with pragmatic guidance on good practice in intervention research. In brief, to better equip them with critical insight into whether or not an intervention works, and some of the issues they need to be mindful of. What follows next, therefore, is a simple, non-exhaustive, checklist in Table 6 that can help practitioners

Table 6. A Checklist for Assessing Educational Interventions.

Intervention Study Features	+	–
Study Design		
The intervention considers the baseline level of the outcome of the interest		
The intervention has a control group		
The control group is plausible		
Allocation to control groups is unbiased (randomised)		
The intervention is clearly described		
The study has checks that the intervention was delivered as intended		
The study clearly describes the context in which the intervention is delivered		
Measurements and Analysis		
Measurement tools are valid (adequately capture what the intervention purports to target)		
Measurement tools' reliability is considered and deemed adequate		
Includes objective measures (not entirely reliant on self-report)		
Measurements are varied to mitigate against bias from the same tool		
Mechanism of effect or contexts under which the intervention works are considered		
The study uses appropriate statistical methods (clear rationale provided for statistical technique)		
The study describes how much missing data there was and how this was handled		
Participants		
The sample is representative of the interventions intended target population		
The sample is large enough to detect a 'clinically significant' effect		
Participants are not made aware that they are in the intervention or control group (masking / blinding procedure is described)		
Participants in the intervention group and control groups have been checked at baseline and are similar on key characteristics (e.g. sex, age, other study relevant traits, and demographics)		

Notes: Instructions for use:
• Practitioners are encouraged to tick the '+' column if the feature is present in the intervention study they are reading.
• Tick the '–' column if the feature is absent.
• Tally the number of '+' and '–'.

determine if an intervention has the sought-after evidence-base warranting its application. Suggested instructions for use are in the Notes below.

Studies with more '+' will carry more weight than those with less '+', or those with many '–'. Whilst the '–' may appear redundant at first, they are here so that practitioners looking at multiple studies can make repeated use of the checklist (one per study). Also, to assess if features that were ticked as '–' previously appeared as '+' in subsequent studies, demonstrating increasing quality of evidence. Doing this also allows practitioners to compare two different interventions and their relative evidence-base.

CONCLUSION

There is a lot to think about when selecting and deciding on an appropriate intervention, and other chapters in this book provide more information to help with this too. For example, deciding on which model of wellbeing is covered in Chapter 7, and what works best in primary schools, secondary schools, or higher education is covered in Chapters 2–4. The case studies have useful tips and recommendations to follow-up on too.

ADDITIONAL RECOMMENDED READINGS AND RESOURCES TO FOLLOW-UP ON

For those wishing to read and learn more about evaluation, you might find the following useful.

- *Evaluating What Works*. https://bookdown.org/dorothy_bishop/Evaluating_What_Works/ by Bishop and Thompson (2023)
- *Improving Your Statistical Inferences* by Lakens (2022). Retrieved from https://lakens.github.io/statistical_inferences/. https://doi.org/10.5281/zenodo.6409077
- *Designing Clinical Research* by Hulley et al. (2013)
- Identifying good measurements, in *Research Methods in Psychology: Third International Student Edition* by Morling (2017)
- *Adolescents and Health-related Behaviour: Using a Framework to Develop Interventions to Support Positive Behaviours* by Pringle et al. (2018).

NOTES

1. A wait list control group, also called a wait list comparison, is a group of participants included in an outcome study that is assigned to a waiting list and receives intervention after the active treatment group.

2. For example, when participants might experience the same intervention procedure, but not the actual intervention itself.

3. An *extraneous variable* is any variable not under investigation as part of the study that can potentially affect the outcome (e.g. time of day).

4. Internal validity describes *the extent to which a cause-and-effect relationship established in a study cannot be* explained by other factors.

5. Triangulation in research is the *process of using multiple research methods and perspectives to study a particular* topic.

REFERENCES

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365–376. https://doi.org/10.1038/nrn3475

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates, Inc. https://doi.org/10.1017/CBO9781107415324.004

Gartlehner, G., Hansen, R. A., Nissman, D., Lohr, K. N., & Carey, T. S. (2006). A simple and valid tool distinguished efficacy from effectiveness studies. *Journal of Clinical Epidemiology*, *59*(10), 1040–1048. https://doi.org/10.1016/j.jclinepi.2006.01.011

Henry, J. D., & Crawford, J. R. (2005). The short-form version of the Depression Anxiety Stress Scales (DASS-21): Construct validity and normative data in a large non-clinical sample. *British Journal of Clinical Psychology*, *44*(2), 227–239. https://doi.org/10.1348/014466505X29657

Hochard, K. D., Hulbert-Williams, L., Ashcroft, S., & McLoughlin, S. (2021). Acceptance and values clarification versus cognitive restructuring and relaxation: A randomized controlled trial of ultra-brief non-expert-delivered coaching interventions for social resilience. *Journal of Contextual Behavioral Science*, *21*, 12–21. https://doi.org/10.1016/j.jcbs.2021.05.001

Moncrieff, J., & Kirsch, I. (2005). Efficacy of antidepressants in adults. *British Medical Journal*, *331*(7509), 155–157. https://doi.org/10.1136/bmj.331.7509.155

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716–aac4716. https://doi.org/10.1126/science.aac4716

Patsopoulos, N. A. (2011). A pragmatic view on pragmatic trials. *Dialogues in Clinical Neuroscience*, *13*(2), 217–224. https://doi.org/10.31887/dcns.2011.13.2/npatsopoulos

Pringle, J., Doi, L., Jindal-Snape, D., Jepson, R., & McAteer, J. (2018). Adolescents and health-related behaviour: Using a framework to develop interventions to support positive behaviours. *Pilot and Feasibility Studies*, *4*, 1–10.

Rhea, M. R. (2004). Determining the magnitude of treatment effects in strength training research through the use of the effect size. *Journal of Strength and Conditioning Research*, *18*(4), 918–920.

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology*, *54*(6), 1063–1070. http://www.ncbi.nlm.nih.gov/pubmed/3397865

White, C. (1990). Research on smoking and lung cancer: A landmark in the history of chronic disease epidemiology. *Yale Journal of Biology and Medicine*, *63*(1), 29–46. https://doi.org/10.1016/0169-5002(91)90089-o

Worrall, J. (2010). Evidence: Philosophy of science meets medicine. *Journal of Evaluation in Clinical Practice*, *16*(2), 356–362. https://doi.org/10.1111/j.1365-2753.2010.01400.x