Education
Centre for Education
Statistics & Evaluation

LEARNING CURVE ISSUE 14

5 Essentials for Effective Evaluation

All education programs are well-intentioned and many of them are highly effective. However, there are usually more ways than one to achieve good educational outcomes for students. When faced with this scenario, how do educators and education policymakers decide which alternative is likely to provide most 'bang for buck'?

There's also an uncomfortable truth that educators and policymakers need to grapple with: some programs are not effective and some may even be harmful. What is the best way to identify these programs so that they can be remediated or stopped altogether?

Program evaluation is a tool to inform these decisions. More formally, program evaluation is a systematic and objective process to make judgements about the merit or worth of our actions, usually in relation to their effectiveness, efficiency and appropriateness (NSW Government 2016).

Evaluation and self-assessment is at the heart of strong education systems and evaluative thinking is a core competency of effective educational leadership. Teachers, school leaders and people in policy roles should all apply the principles of evaluation to their daily work.

Research shows that:

- Effective teachers use data and other evidence to constantly assess how well students are progressing in response to their lessons (Timperley & Parr, 2009).
- Effective principals constantly plan, coordinate and evaluate teaching and the use of the curriculum with systematic use of assessment data (Robinson, Lloyd & Rowe, 2008).
- Effective education systems engage all school staff and students in school self-evaluations so that program and policy settings can be adjusted to maximise educational outcomes (OECD, 2013).

This Learning Curve sets out five conditions for effective evaluation in education. These are not the only considerations and they are not unique to education. However, if these parameters are missing, evaluation will not be possible or it will be ineffective.

The five prerequisites for effective evaluation in education are:

1. Start with a clear and measurable statement of objectives
2. Develop a theory about how program activities will lead to improved outcomes (i.e. a program logic) and structure the evaluation questions around that logic
3. Let the evaluation questions determine the evaluation method
4. For questions about program impact, either a baseline or a comparison group will be required (preferably both)
5. Be open-minded about the findings and have a clear plan for how to use the results.

1. Start with clear and measurable objectives

It may sound obvious but understanding whether program activities have been effective requires a clear understanding of what the program is trying to achieve. The objectives also need to be measureable.

For some programs or activities this is very easy. For example, reading interventions like Reading Recovery aim to improve students' ability to read. In these instances it is easy to start with a clear statement of objectives (i.e. to improve students' ability to read). It is also quite easy to measure outcomes because reading progression is relatively easy to measure (although the issue of causal attribution is important – more on that later).

However, for some programs, it can be more difficult to develop a clear statement of objectives and it is even more difficult to measure whether they have been achieved. Take the Bring Your Own Device (BYOD) policy as an example. The objective of BYOD is often described as using technology to 'deepen learning', 'foster creativity' or 'engage students'. These are worthy objectives. The challenge for schools and systems is to work out whether they have been achieved. What does 'deep learning' look like and how can it be measured? How will teachers know if a student is more 'creative' or 'engaged' now than they were before? How much of that gain is due to the program or policy (BYOD) and how much is due to other factors?

Make sure you know how you're going to measure success before you start. This will ensure you don't get to the end of your program/funding cycle thinking "oh, if only I'd measured…"

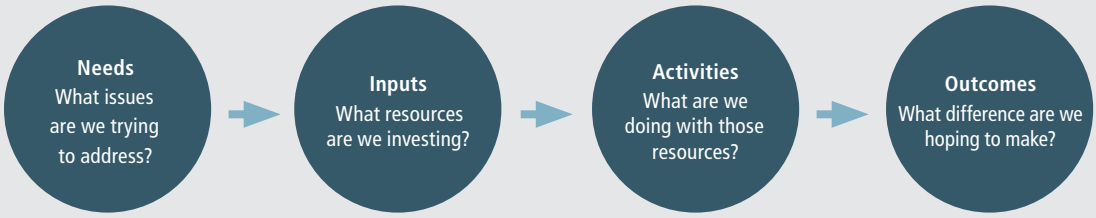
Figure 1 provides some examples of common objectives and possible measures that will inform whether they have been achieved. These are highly idealised examples and the problems that educators are trying to solve are usually more multi-faceted and complex than these. In some cases it may not even be possible to robustly measure outcomes. In other cases, there may be more than one outcome resulting from a set of activities. However, no matter how hard and complex the problem, if there is no clarity about what the problem is, there is also no chance of measuring whether it has been solved.

Figure 1. Some examples of common objectives and measures that might inform whether they've been achieved

Objective	Possible measures
Increased participation in quality early childhood education	Number of children enrolled in high quality preschools, measured over time
Improved preschool quality	Independent ratings of quality aspects using measurement tools such as the ECERS-E or ECERS-R
Improvements in early literacy among lowest performers	Proportion of students above National Minimum Standards on Y3 NAPLAN Reading
Whole-school improvements in early secondary maths achievement	Year 7-9 value-added score for NAPLAN Numeracy
Improvements in student intellectual engagement	Percentage of students reporting 'high skill – high challenge' on the *Tell Them From Me* student survey
Reductions in bullying behaviour	Percentage of students reporting being bullied on the *Tell Them From Me* student survey
Improved post-school pathways	Proportion of students engaged in further education or training after leaving school

1 Note: The *Tell Them From Me* student surveys coordinated by CESE measure social, institutional and intellectual engagement and could be a useful source of data to measure engagement.

Figure 2. A simple logic model



2. Linking activities and outcomes

Effective programs have a clear line of sight between the needs they are responding to, the resources available, the activities undertaken with those resources, and how activities will deliver outcomes. Logic modelling is one way to put these components on a piece of paper. Wherever possible, this should be done by those who are developing and implementing a program or policy, in conjunction with an experienced evaluator. At its most simple, a logic model looks like that shown in Figure 2.

The needs are about the problem at hand and why it is important to solve it. Inputs are the things put in to address the need (usually a combination of money, time and resources). Activities describe the things that happen with the inputs. Outcomes are usually expressed in terms as measures of success. A logic model is not dissimilar to the processes used in school planning. Needs are usually the strategic priorities identified in the plan. Inputs are the resources allocated to address those needs. Activities are often referred to as processes or projects. Outcomes and impacts are used interchangeably. Figure 3 gives some common examples of needs, inputs, activities and outcomes.

Some of these examples are 'add-on' activities to business-as-usual (e.g. speech pathology) and some simply reflect the way good teachers organise their classroom (e.g. differentiated instruction). Figure 3 merely serves to illustrate that the evaluative process involves thinking about the resources going into education, how those inputs are organised and how they might plausibly lead to change.

Good evaluation will make an assessment of how well the activities have been implemented (process evaluation) and whether these activities made a difference (outcome evaluation). If programs are effective, it might also be prudent to ask whether they provide value for money (economic evaluation).

A simple logic modelling worksheet can be found in the Appendix.

Figure 3. Some examples of program needs, inputs, activities and outcomes

Needs	Inputs	Activities	Outcomes
Low quality instruction among teachers new to the profession	Funding for teacher release time to be mentored by a senior colleague	Teachers have time off class to participate in mentoring and other professional learning	Improved teacher quality among early career teachers
Low attendance among Aboriginal students	Employment of Aboriginal Community Liaison Officer	ACLO organises cultural events at the school	Improved Aboriginal attendance rates in school
Low levels of achievement in numeracy	Implementation of a small-group mathematics program	Teachers differentiate instruction based on ability groupings	Improved mathematics test scores across all ability groupings
Low levels of communication skills among students with a cognitive impairment.	Flexible funding through the resources Allocation Model to employ a speech pathologist.	Speech pathologist works with students with a cognitive impairment and their teachers.	Improved literacy skills for students with a cognitive impairment.

Types of evaluation

Process evaluation is particularly helpful where programs fail to achieve their goals. It helps to explain whether that occurred because of a failure of implementation, a design flaw in the program, or because of some external barrier in the operating environment. Process evaluation also helps to build an understanding of the mechanisms at play in successful programs so that they can be replicated and built upon.

Outcome evaluation usually identifies average effects: were the recipients better off under this program than they would have been in its absence. However, when viewed in combination with process evaluation, it can provide a more nuanced overview of the program. It can explore who the program had an impact on, to what extent, in what ways, and under what circumstances. This is important because very few programs work for everyone. Identifying people who are not responding to the program helps to target alternative courses of action.

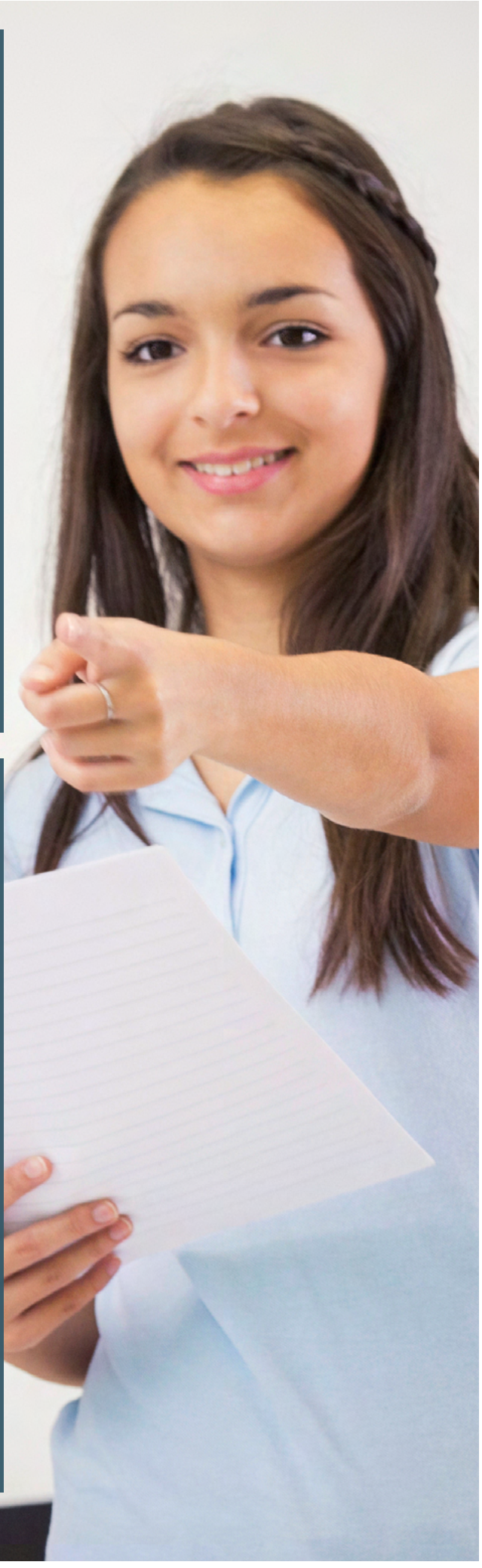
Economic evaluations help us choose between alternatives when we have many known ways of achieving the same outcomes. In these circumstances, the choice often comes down to what is the most effective use of limited resources. If programs are demonstrably ineffective, there is little sense in conducting economic evaluations. Ineffective programs do not provide value for money.

When program logic breaks down – repeating a school year

While repeating a school year is relatively uncommon in NSW, it is quite common in some countries such as the United States. It is a practice that has considerable intuitive appeal – if a student is falling behind (need) the theory is that an additional year of education (input) will afford them the additional instruction (activity) required to achieve positive educational outcomes (outcome). Evidence suggests that this is true only for a small proportion of students who are held back. In fact, after one year, students who are held back are on average four months further behind similar-aged peers than they would have been had they not been held back.

According to research conducted by the UK Education Endowment Foundation, the reason that repeating a year is not effective is that it "just provides 'more of the same', in contrast to other strategies which provide additional targeted support or involve a new pedagogical approach. In addition, it appears that repeating a year is likely to have a negative impact on the student's self-confidence and belief that they can be an effective learner". In other words, for most recipients of the program the activities are poorly suited to the students' needs. In situations like this, well-intentioned activities can actually have a negative impact on a majority of students.

Source: *https://educationendowmentfoundation.org.uk/toolkit/toolkit-a-z/repeating-a-year/*



3. Let the evaluation questions determine the method

Once a clear problem statement has been developed, the inputs and activities are identified, and intended outcomes have been established, coherent evaluation questions can be developed. Good evaluation will ask questions such as:

- Did the program deliver what was intended? If not, why not?
- Did the program reach the right recipients? If not, why not?
- Did the program achieve the intended outcome and were there any unintended (positive or negative) outcomes?
- For whom did it work and under what circumstances?
- Is this the most efficient way to use limited resources?

All too often educational researchers get hung up on using 'qualitative' versus 'quantitative' methods when answering these questions. This is a false dichotomy. The method employed to answer the research question depends critically on the question itself.

Qualitative research usually refers to semi-structured techniques such as in-depth interviews, focus groups or case studies. Quantitative research usually refers to more structured approaches to data collection and analysis where the intention is to make statements about a population derived from a sample.

If you're using quantitative data (e.g. assessment data) to inform whether your program has worked, make sure you use a valid measure (i.e. one that accurately measures the thing you're trying to influence). If your measures are flawed, your evaluation will be too.

Both approaches will have merit depending on the evaluation question. In-depth interviews and focus groups are often the best ways of understanding whether a program has been implemented as intended and, if not, why not. These methods have limitations when trying to work out impact because, by definition, information is only gleaned from the people who were interviewed. Unless something is known about the people who weren't interviewed, these sorts of methods can be highly misleading. For example, people who didn't respond well to the intervention might also be less likely to participate in interviews or focus groups. This is where quantitative methods are more appropriate because they can generalise to describe overall effects across all individuals. However, combining both qualitative and quantitative methods can be useful for identifying for whom and under what conditions the program will be effective. For example, CESE researchers investigating the practices of high-growth NSW schools used quantitative analysis to identify high-growth schools and analyse survey results, and qualitative interviews to find out more about the practices these schools implemented.

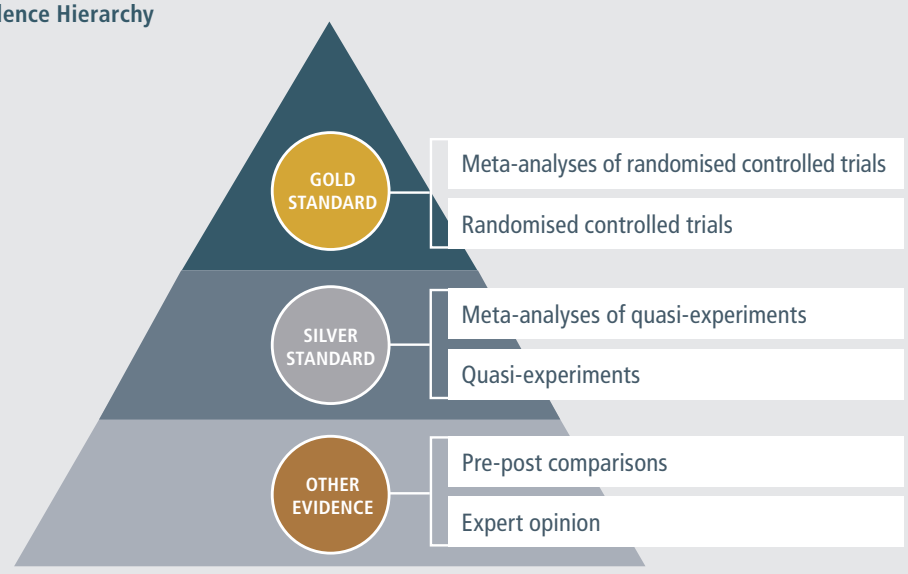
The possible sources of data to inform evaluation questions are endless. The key issue is to think about the evaluation question and adopt the data and methods that will provide the most robust answer to that question.

4. For questions about program impact, either a baseline or a comparison group will be required (preferably both)

The number one question that most evaluations should set out to answer is: did the program achieve what it set out to achieve? This raises the vexing problem of how to attribute activities to any observed outcomes.

No single evaluation approach will give a certain answer to the attribution question. However, some research designs will allow for more certain conclusions that the effects are real and are linked to the program. CESE uses a simple three-level hierarchy to classify the evidence strength, as shown in Figure 4. There are many variations on this hierarchy, most of which can be found in the health and medical literature.

Figure 4. CESE Evidence Hierarchy



Make your findings available. The 'file drawer' problem is the enemy of effective evidence-based policy and practice.

Taking before (pre) and after (post) measures is a good start and is often the only way to measure outcomes. However, simple comparisons like this need to be treated cautiously because some outcomes will change over time without any special intervention by schools. For example, if a student's reading level was measured at two time points, they would usually be at a higher level at the second time point just through the course of normal class and home reading practice.

This is where reference to benchmarks or comparison groups is critical. For example, if the typical growth in reading achievement over a specified period of time is known, it can be used to benchmark students against that expected growth. Statements can then be made about whether growth is higher or lower than expected as a result of program activities.

An even stronger design is when students (or schools, or whatever the target group is comprised of) are matched like-for-like with a comparison group. This design is more likely to ensure that differences are due to the program and not due to some other factor or set of factors. These designs are referred to as 'quasi-experiments' in Figure 4.

Even better are randomised controlled trials (RCTs) where participants are randomly allocated to different conditions. Outcomes are then observed for the different groups and any differences are attributed to the experience they received relative to their peers. RCTs can also be conducted using a wait-list approach where everyone gets the program either immediately or after a waiting period. RCTs allow for strong causal attributions because the random assignment effectively balances the groups on all of the factors that could have influenced those outcomes.

RCTs have a place in educational research but they will probably always be the exception rather than the rule. RCTs are usually reserved for large-scale projects and wouldn't normally be used to measure programs operating at the classroom level. Special skills are required to run these sorts of trials and most of the programs run by education systems would be unsuited to this research design. In the absence of RCTs, it is still important to think about ways to measure what the world looked like before the activity began and what it looked like after some period of activity has been undertaken. This requires taking baseline and follow-up measures and comparing these over time.

As a rule, the less rigorous the evaluation methodology, the more likely we are to falsely conclude that a program has been effective. This suggests that stronger research designs are required to truly understand what works, for whom and under what circumstances.

5. Be open-minded and have a clear plan for how to use the results

In all of the above, it is crucial for educators to be open-minded about what the results of the evaluation might show and be prepared to act either way. Evaluation should not be a tool for justifying or 'evidence washing' a predetermined conclusion or course of action. The reason for engaging in evaluation is to understand program impact in the face of uncertainty. It provides the facts (as best they can be estimated) to help make decisions about how to structure programs, whether they should be expanded, whether they need to be adjusted along the way, or whether they need to stop altogether.

Evaluation not only asks 'what is so?' – it also asks 'so what?' In other words, evaluation is most useful if it will lead to meaningful change. Before embarking on any evaluation, it is important to think about what can reasonably be achieved from the research. If continuation of the program is not in question, it may be better focusing on process questions bearing on program efficiency or quality improvement. It is also important to think about stakeholders, how they might react to the evaluation and what needs to happen to keep them informed along the way.

In accordance with the NSW Government Program Evaluation Guidelines (NSW Government 2016), evaluation should be conducted independently of program delivery and it should be publicly available for transparency. Independence might not always be possible where no budget exists or where activity is business-as-usual or small in scale (e.g. classroom-level or school-level programs). Evaluative thinking is still critical in these circumstances as part of ongoing quality improvement.

Where a formal evaluation has been conducted, transparency is a critical part of the process. Stakeholders need to understand the questions the evaluation sought to answer, the methods employed to answer them, any assumptions that were made, what the evaluation found and the consequences of those findings. Transparency also helps people in later times or in other schools or jurisdictions to identify what works.

Conclusion

To embed the sort of evaluative thinking described above into activity across education requires everyone to be evaluative thinkers in one way or another. Everyone designing or implementing a program needs to be clear on what problem they are trying to solve, how they are planning to solve it and how success will be measured.

For smaller, more routine programs and policies, performance should be monitored using the sort of benchmarking described above to determine the effectiveness, efficiency and appropriateness of expenditure. This could be done by an early childhood service Director, by a school teacher, by a principal, school leadership group, Directors Public Schools or Principals School Leadership. If more technical assistance is required, it may be better to bring in that technical expertise.

When in doubt, phone a friend: CESE can be contacted by phone on 1300 972 196 or by email info@cese.nsw.gov.au

References

Centre for Education Statistics and Evaluation 2015, 'Six effective practices in high growth schools', Learning Curve Issue 8, Centre for Education Statistics and Evaluation, Sydney.

NSW Government 2016, 'NSW Government Program Evaluation Guidelines', Department of Premier and Cabinet, NSW Government, Sydney. http://www.dpc.nsw.gov.au/__data/assets/pdf_file/0009/155844/NSW_Government_Program_Evaluation_Guidelines.pdf

OECD 2013, 'Synergies for better learning: An international perspective on evaluation and assessment', OECD Publishing, Paris.

Robinson, V, Lloyd, C & Rowe, K 2008, 'The impact of leadership on student outcomes: An analysis of the differential effects of leadership types', *Educational Administration Quarterly,* vol. 44, no. 5, pp. 635-674.

Timperley, H & Parr, J 2009, 'Chain of Influence from policy to practice in the New Zealand literacy strategy', *Research Papers in Education*, vol.24, no.2, pp.135-154.

Appendix: Logic modelling worksheet

Needs	What issues are we trying to address?	
Inputs	What resources are we investing?	
Activities	What are we doing with the people, time and/or money we are putting in to solve this problem?	
Outcomes	What difference are we hoping to make? How will we measure this?	

Remember:

1. Start with a clear and measurable statement of your objectives
2. Develop a plausible theory about how your activities will lead to improved outcomes (i.e. a program logic) and structure the evaluation questions around that logic
3. Let the evaluation questions determine the method you choose to evaluate your actions
4. If you want to know whether you achieved your objectives, you need a baseline or a comparison group (preferably both)
5. Be open-minded about the findings and have a clear plan for what you're going to do with the results

Centre for Education Statistics and Evaluation
GPO Box 33
Sydney NSW 2001
Australia

02 9561 1211

cese@det.nsw.edu.au

www.cese.nsw.gov.au

NSW GOVERNMENT
Education
Centre for Education
Statistics & Evaluation