

CHAPTER 2

D2: Design



Systematically search and agree on high-probability interventions to START and to STOP

2.1 Explore Options in the Design Space



2.2 Build Program Logic Model(s)



2.3 Stress Test Logic Model(s)



2.4 Agree on What to STOP



2.5 Establish a Monitoring and Evaluation Plan



INTRODUCTION

When COVID-19 emerged in late 2019, the first question for governments was whether the virus was a *genuine* risk. The second question was about the *severity* of the impact. If you cast your mind back to those early days before COVID-19 was declared a pandemic, many of us dismissed the news of a new virus from a distant part of the world. It ended up taking a few months before policymakers across most parts of the globe accepted that it was, indeed, a *genuine* risk. And it still is a risk for many citizens and is likely to be with us well into the future.

A common feature in those early closed-door government sessions was that someone (almost) always asked whether COVID-19 was something that required a response or whether it would simply be better to allow the virus to wash over us and quickly generate herd immunity. In most countries, after the numbers were crunched, the consensus was that the severity of risk was extremely high, that (on average) 1–2% of the population could die, and that many more would be left with what we now call long COVID (i.e., lingering health difficulties for the long haul). The consequence of this assessment was that most countries decided to act and protect their populations from these harrowing consequences.

The diagnostic processes that governments went through to assess whether COVID was *genuine* and to calculate the severity of impact are not dissimilar to those we outlined in Stage D1 (Discover). Governments explored data to answer the question, “What’s the worst that could happen if we do absolutely nothing?” And their scientists then began to define and map the key features of the virus (i.e., the breakdown structure). They also undertook their own version of path analysis, exploring

- **transmission pathways** (i.e., how quickly and under what conditions the virus passes from one person to another) and
- **biological interactions** (i.e., how it enters the body, what it does, and how the immune system responds to this).

Armed with this information, the next step was to investigate and agree on interventions to slow, block, or reverse the different nodes or bubbles on that path analysis map. Ultimately, this bit was a design activity that culminated in the identification of a range of high-probability options. With virus transmission, for example, the identified interventions included face mask wearing, handwashing, social distancing, and lockdowns. These interventions were not randomly selected. Scientists looked carefully at successful strategies that had been used to curb the transmission of other similar viruses in the past (for an early account of public policy responses to COVID-19, see Murphy, 2020).

But it didn’t stop there. The next level down was to agree on the **design features and setting levels** for each selected intervention. For example, did face masks need to be worn outside and indoors? Was it okay to reuse masks? Were standard surgical masks sufficient, or was double-masking or even the use of more robust N95 masks required? How would people be encouraged to wear them? Would there be any sanctions if they refused? Would the sanctions be enforced and by whom?

The same questions were asked about the optimal design features and setting levels for social distancing, for lockdown protocols (including whether schools needed to close), and for the design and distribution of both vaccines and treatments for individuals

infected with the virus. And each of these “work packages” was built out into a carefully designed plan, which was then implemented and iteratively evaluated, to decide *where to next*.

The processes and tools that we will outline in **Stage D2: Design** of the *Building to Impact 5D* framework are remarkably similar to those used by the scientists and policymakers to investigate and agree on how they would respond to COVID-19. This is not by chance. Remember that in developing this framework, we explored the successful tools used in a range of settings, including health care and even the business sector.

Rather than selecting random approaches and hoping for the best, the idea is that through the use of these systematic design approaches, you will significantly increase the probability that you push the impact needle on your priority education challenge.

Through the use of these systematic design approaches, you will significantly increase the probability that you push the impact needle on your priority education challenge.

2.1 EXPLORE THE OPTIONS IN THE DESIGN SPACE

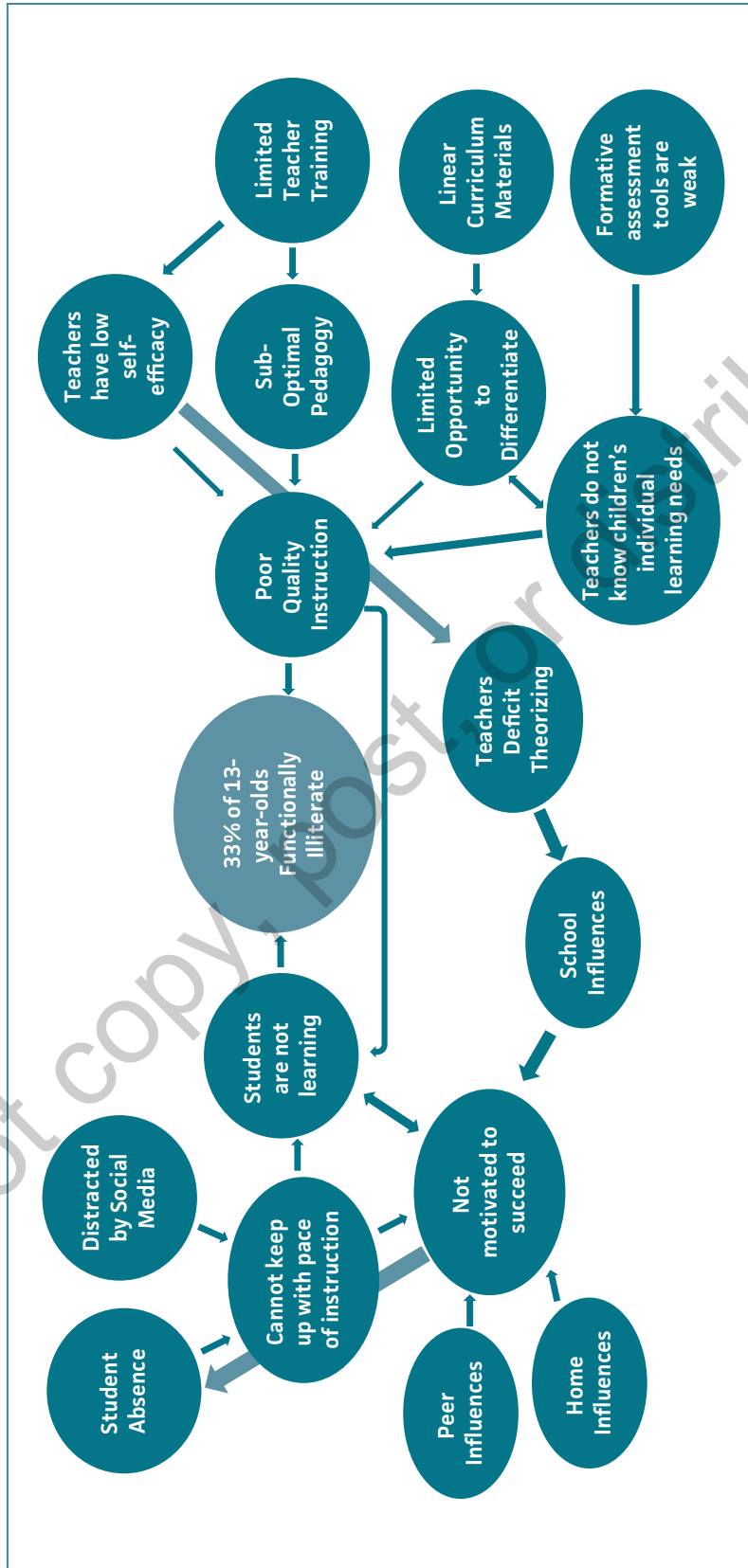
During Stage D1 (Discover), you established your backbone organization (1.1), decided your ONE education challenge (1.2), undertook a path analysis to explain the education challenge (1.3), and then set provisional improvement goals to agree on what better looks like (1.4).

The activity that you undertook in Step 1.3 is especially crucial and directly linked to what you are going to do now. During that specific activity, you mapped the key causal dimensions of your education challenge and then doubled-back to validate these. The outcome of this process is a checked and cross-checked path analysis with arrows and **influence bubbles**, like the one we presented in Figure 1.9, which we recap again in Figure 2.1.

What you are now going to do in Step 2.1 is systematically explore the options in the design space that could potentially be leveraged to block, reverse, or weaken each of the identified influence bubbles on your path analysis. In the example map recapped in Figure 2.1, there are 17 influence bubbles, all contributing to the education challenge at the center. This means that you (ideally) need to search for a range of options or **opportunity sketches** for each of those 17. And for clarity, by opportunity sketches, we mean initiatives, programs, actions, interventions, and so on—things that you can implement that bring you ever closer to your success criteria.

As you undertake this search, you might identify opportunity sketches that could impact multiple influence bubbles. For example, you might identify a specific type of teacher coaching program that potentially addresses the “limited teacher training” + “sub-optimal pedagogy” + “teachers have low self-efficacy” influence

FIGURE 2.1 • Recapping the Path Analysis



bubbles, at the same time. This is good and actively encouraged because the less complex your designs, the less likely that the wheels will fall off during delivery (Stage D3).

However, the key question is *how* should you go about that search through design space to identify potentially viable opportunity sketches? In our work with schools, sadly it is all too often done informally and unsystematically. Someone went to a conference and heard a “guru” talking about the X-Program, saw a blog post, or got a testimonial from a friend who works at another school. And these signals get treated as “evidence” that the X-Program works and that it works for *your* specific education challenge. Daniel Willingham’s (2012) excellent book *When Can You Trust the Experts?* highlights the marketing puffery and questionable claims that are all too often made by commercial education products and program developers. What we need, of course, is a “Crap Detector” or “Crap Avoidance protocol”, so that we vector in on the highest-probability opportunity sketches.

The first step to avoiding junk is to stick closely to your *actual* needs. You have your path analysis with influence bubbles, and you are proactively searching the design space for opportunities that are directly connected to these. We repeat with additional emphasis, you are searching for opportunities that are *directly* connected to these. By contrast, you are *not* engaging in a cognitive bias that’s commonly called *The Law of the Instrument* (and sometimes *Maslow’s Hammer*). This is where you have this pet thing called the X-Program that, for example, teaches children how to ride bicycles in 10 easy steps, when your identified education challenge is cyberbullying and suicide prevention. Yes, you can make a causal leap that exercise increases endorphins, which makes people happy, and that cycling is a form of exercise. But it’s a bit of a stretch.

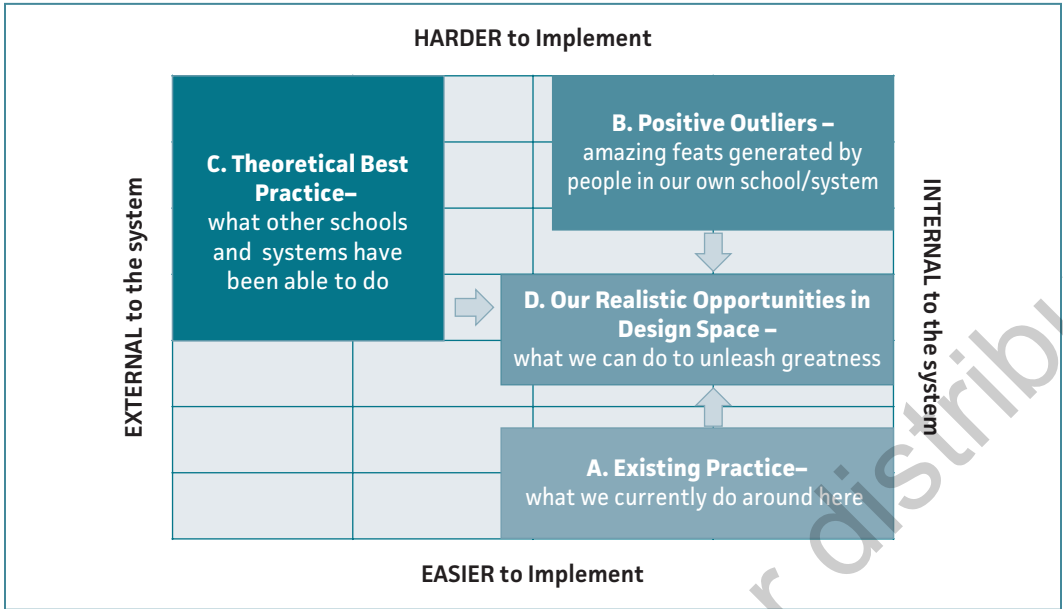
Therefore, the first step is to systematically search the options in the design space to address *your* actual education challenge. In Figure 2.2, we illustrate three key sources of data that you can leverage, and we then go on to explore each in more detail.

These are the three key informational levers:

- A. **Existing practice.** This is what you currently do and what you have learned from it.
- B. **Positive outliers.** These are the behaviors and actions of local stakeholders that significantly buck the current trend, in a good way.
- C. **Theoretical best practice.** This is what you can glean from research about how other schools and systems have generated impact on the same or similar goals.

The idea is that by mining and triangulating these three sources of information, you are then able to identify:

FIGURE 2.2 ● Identification of Options in the Design Space



Source: Adapted from Andrews et al. (2017).

D. Realistic opportunities in the design space. These are the locally feasible activities, based on your current capabilities and resourcing and that, importantly, also have a strong probability of progressing your education challenge. And to have that strong probability, they need both strong evidence of impact and a strong connection to one or more of your influence bubbles (i.e., they address a need that you actually have).

Let’s now explore each of these in turn.

EXISTING PRACTICE

The good news is that you will likely have already made good progress toward mapping what you currently do. During Stage D1 (Discover), you undertook a *challenge breakdown structure* activity to better define your area of inquiry and you also developed a *path analysis* to better understand your challenge context. However, if you feel that you still need to collect more information on your challenge area, you can supplement this with the following:

- Lesson observations, including a collection of video or even audio transcriptions
- Interviews with teachers and students
- The development of process maps, where you can use sticky notes to plot out end to end how existing activities are undertaken

POSITIVE OUTLIERS

No matter what your area of inquiry, there will always be some stakeholders in your school or system who do significantly better than average. If, for example, you are trying to significantly enhance literacy outcomes, you might find that certain teachers consistently generate above-average outcomes or that certain cohorts of students do phenomenally well irrespective of the teachers.

You need to know why this is and what it is they are doing differently, so that you can evaluate whether it is something that could easily be scaled and replicated by others. If you discover that the students who do well irrespective of the teacher tend to come from higher socioeconomic backgrounds and that their parents tend to pay extra for tuition or additional tutoring support, you might conclude that this would be extremely difficult to replicate. Whereas if you uncovered that the successful students had established a study group and a range of codified study-skills practices for how this operated, there is significantly more replication potential.

The same goes for teachers. In some of Arran's work, he has explored how to increase student attendance at schools in low- and middle-income countries. Some remote schools have done extremely well at this, even with indigenous students whose parents are sometimes initially reluctant to enroll their children. But some mechanisms are easier to replicate than others. In one school, the success seemed to be down to an inspiring and passionate teacher who perpetually sang in the indigenous language while skillfully strumming his guitar. This was beautiful and brilliant but difficult to replicate. Where do we find 500 guitar-wielding teachers? Whereas in other settings, success had been achieved (1) through structured parental outreach sessions to inform them of the benefits of educating their children and (2) by supplementing this with conditional cash transfers to reduce the economic burdens to these parents of sending their children to school. These approaches are much easier to map, codify, and replicate than the guitar-playing teacher, albeit they are less fun.

Figure 2.3 draws on the rich practice-based research into **positive outliers** (LeMahieu et al., 2017; Pascale et al., 2010). It provides you with a framework to map and record the positive outliers in your context.

FIGURE 2.3 • Identification of Positive Outliers

OUTLIER STAKEHOLDER	OUTLIER OUTCOME	OUTLIER BEHAVIORS	REPLICATION POTENTIAL
Who are they?	How do their outcomes buck the general trend?	What do they seem to be doing differently?	How easy would it be for other stakeholders to replicate the outlier behaviors?

This is all about the identification of *positive* variance, the explanation of that variance, and the ease with which others could do the same. Areas that have high replication potential represent high-potential opportunity sketches.

THEORETICAL BEST PRACTICES

The third place you should look for high-probability bets is **theoretical best practices** identified in the global *what works best* literature. As we explained in the Introduction to this book, there are now more than 1.5 million research articles on the whole gamut of education interventions. Admittedly, it would take you several lifetimes to explore, map, and catalog these—but the good news is that this has (largely) been done already. There are several places you can go to find high-quality **systematic reviews** that synthesize the findings of multiple studies to come to an overall conclusion and that make recommendations, as shown in Figure 2.4.

The reason that we *strongly* advocate explicitly mining high-quality systematic reviews is that they get out of the swamp of *what works* and into the Goldilocks zone of *what works best*. Indeed, one of the unfortunate features of the more than 1.5 million research articles on effective practices is that if you look hard enough, you will be able to find “proof” of anything. You can find “evidence” that homework is ineffective (Kohn, 2006), effective (Roschelle et al., 2016), or sometimes effective (Heffernan, 2019). You will also see many variations in the quality of the research design. The danger, therefore, is that you begin with an idea firmly lodged in your mind about the opportunity sketches that have the most potential for impact (i.e., your pet ideas). And you then only search for data that conform with that view. After enough searching, you will surely find “evidence” that homework works/does not work/sometimes works [*delete as appropriate] or even that the moon landings were faked and that the Earth is flat.

The beauty, however, of going to the systematic reviews is that professional researchers have already done the heavy lifting of mining and aggregating the more than 1.5 million studies to give you an overall probability of impact. This means that you can make a decision based on *all* the relevant studies rather than just your own initial search.

From these and other sources, you will be able to identify high-impact strategies that have worked in other contexts to progress *similar* education challenges. They could *potentially* work in your context, too.

You are especially interested in productized and codified programs that have impact. The beauty of such interventions is that someone else has already done the heavy lifting and had tested, iterated, and refined the protocols in a range of contexts. For better programs, this also includes activation, implementation, and maintenance

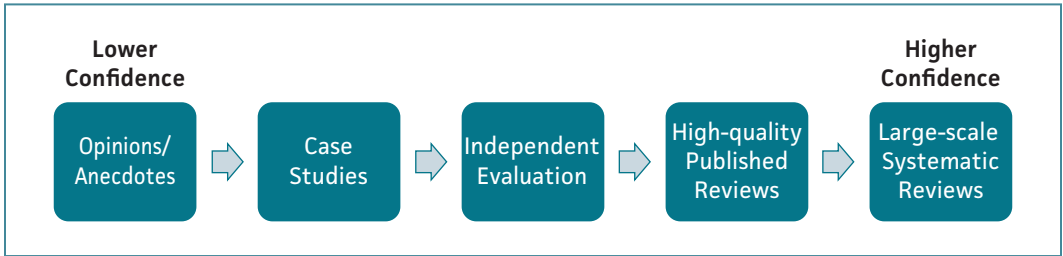
FIGURE 2.4 • Global Education Research Repositories

REPOSITORY	CONTENTS
Visible Learning Meta^x Global	Catalogs 1,800+ meta-analyses of 100,000+ studies, involving 300 million+ students. Findings are segmented into 300+ influences on student achievement across nine domains, including school, classroom, teacher, teaching strategies, and curricula https://www.visiblelearningmetax.com/
Education Resources Information Center (ERIC) Global	Catalogs (Google-style) 1 million+ research articles, some of which are behind publisher paywalls (although most of these are individual studies rather than systematic reviews) https://eric.ed.gov/
What Works Clearinghouse (WWC) United States	Catalogs a range of evidence-based interventions and/or programs across a range of areas, including literacy, mathematics, science, behavior, and teacher excellence https://ies.ed.gov/ncee/wwc/
Best Evidence Encyclopedia (BEE) United States	Synthesizes findings on effective programs for mathematics, reading, science, and early childhood education. https://bestevidence.org/
Education Endowment Foundation (EEF) United Kingdom	Catalogs 30+ common educational interventions, scoring them based on the cost of implementation vs. impact of implementation https://educationendowmentfoundation.org.uk/
Campbell Collaboration Global Campbell-UNICEF MegaMap on Child Well-being Interventions Global	Provides systematic reviews in a range of areas, including education, health, crime, and social justice https://www.campbellcollaboration.org/ https://www.unicef-irc.org/megamap/
Iterative Best Evidence Synthesis (BES) New Zealand	Offers narrative-style systematic reviews on 8+ common education improvement categories, including teacher professional development and high-impact instructional approaches https://www.educationcounts.govt.nz/topics/bes
Australia Education Research Organisation (AERO) Australia	Provides evidence guides on a range of “tried and tested” approaches, including formative assessment, mastery learning, and explicit instruction https://www.edresearch.edu.au/
Ontario Education Research Exchange (OERE) Canada	Catalogs evidence, exemplar resources, and frameworks for effective implementation https://oere.oise.utoronto.ca/

tasks. Why reinvent the wheel? It's better to find the type of wheel that best fits your terrain.

However, as we illustrate in Figure 2.5, the quality of the evidence is key. You can have much higher confidence if you start with the large-scale systematic reviews that bring together the research findings from hundreds, thousands, or tens of thousands of different deployments.

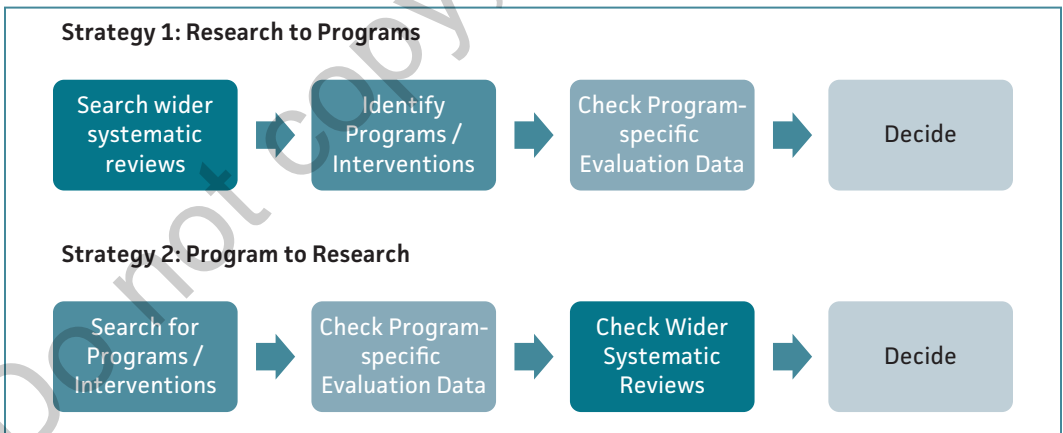
FIGURE 2.5 • Not All Evidence Is the Same



You can also undertake the search the other way around and collect opinions on suitable interventions or programs from colleagues within your wider system. You then look for the research data on each specific recommendation, keeping in view the programs and initiatives that have stronger supporting data from similar contexts to your own and discarding the rest.

In Figure 2.6, we illustrate these two different approaches. Strategy 1 starts with the wider systematic reviews, leveraging these to identify warm leads for programs and then cross-checking program-specific evaluation data in order to decide. Strategy 2 starts with the programs themselves, which may have been brought to your attention as warm leads from the practice-based insights of colleagues and collaborators in the wider system. You then check the program-specific evaluation data for each of these warm leads and, finally, cross-check them against the findings of large-scale systematic reviews to confirm alignment. Then you decide.

FIGURE 2.6 • Evidence to Programs vs. Programs to Evidence



Both of these strategies are perfectly acceptable, as long as you implement them properly—that is, you search for disconfirming as well as confirming data.

However, here is one final look-for as you explore program-specific data. Many education program developers use language

like “research-based” or “based on proven research” in the marketing of their wares. What they are basically saying is this:

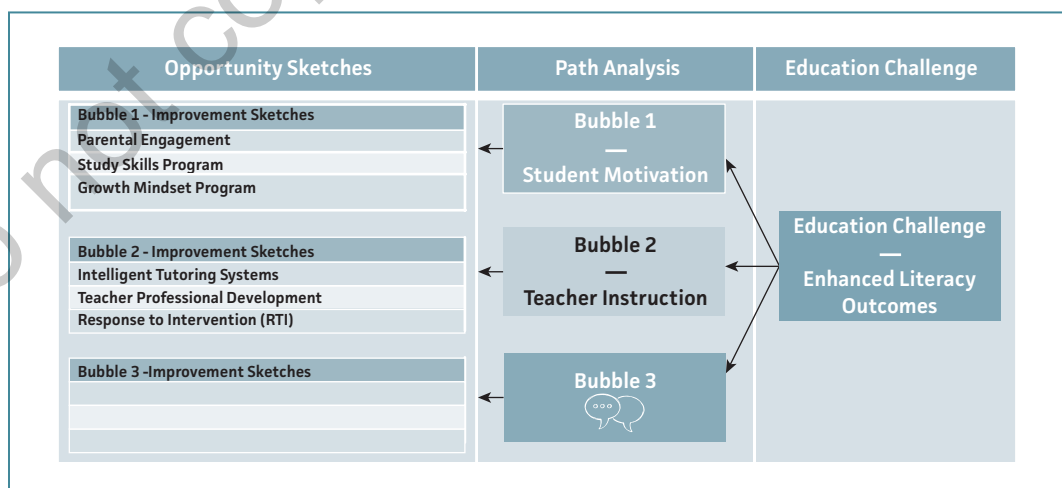
Someone, somewhere, [not us] developed something a bit like what we have built, and they gathered evaluation data that demonstrated impact. Therefore, you can be assured that our thing works just the same—even though it’s not actually the same.

While it is understandable that product developers should engage in this kind of puffery before they have robust impact data about their specific program, these kinds of statements are still at the opinion/anecdote end of the claim spectrum. If their design and implementation protocols are extremely similar to the programs they are emulating and if those other programs have high-quality published reviews or large-scale systematic reviews supporting their efficacy, then yes, you can have higher confidence. But why not just go to the original program?

YOUR REALISTIC OPPORTUNITIES IN THE DESIGN SPACE

As you explore your existing practice, positive outliers, and the theoretical best practices, the idea is that you “longlist” the ones that have the potential to significantly improve your local context. In Figure 2.7, we illustrate one way that you can do this. At the far right, your education challenge is listed. In the middle column, you transcribe each of the influence bubbles from the path analysis that you undertook in Step 1.3. Then in the far-left column, you list your opportunity sketches—that is, the interventions or actions that could

FIGURE 2.7 • Opportunity Sketch Mapping



Source: Copyright © Cognition Education. (2022). All rights reserved.

interact with specific bubbles to improve outcomes. You may find that some of these improvement sketches are relevant and can contribute to more than one bubble or your path analysis. This is even better: remember the adage about one stone and multiple birds?

In the opportunity sketch map in Figure 2.7, we have only listed three potential sketches per **influence bubble**. Depending on how long and carefully you search, you could identify hundreds of potential opportunities. Of course, it all comes back to optimal stopping—that decision about how long you should search before moving on.

Each of your opportunity sketches will likely contribute to enhancing outcomes in a different way, with a different theory of improvement and a different causal pathway. For example, sticking with our student literacy example, let's say two of your influence bubbles were (1) students lacking motivation and (2) misaligned instructional approaches. You might identify a range of potential opportunities for each:

- **Student motivation:** parental engagement, growth mindset programs, study-skills programs, behavior management programs, and so on
- **Misaligned instructional approaches:** intelligent tutoring systems, teacher professional development, Response to Intervention (RTI), scripted direct instruction, and so on

Each of these interacts with its respective influence bubble in a different way. Parental engagement initiatives focus on co-opting parents as partners in the learning, whereas growth mindset programs focus on directly enhancing students' self-efficacy and thereby their motivation and approach. Intelligent tutoring systems bypass teacher instruction, providing an overlay of remediation. In contrast, teacher professional development combined with either RTI or scripted direct instruction is designed to enhance what teachers do and thereby accelerate student learning outcomes.


You will almost certainly identify more potential options in the design space than you could possibly hope to implement. Indeed, the more you attempt to implement at the same time, the more likely that you will drop all your balls. So, you need to select carefully. In Figure 2.8, we provide you with a rubric and scoring sheet that you can use to evaluate all your options. You can also adapt this for a better fit with your local context.

In Figure 2.9, we provide an example of how you might collect and record preliminary information on each opportunity sketch in order to undertake the scoring and ranking just described. As

FIGURE 2.8 • Ranking Your Options in the Design Space

FACTOR	CRITERIA
Evidence of impact	<ul style="list-style-type: none"> • Outcomes achieved in other contexts (e.g., effect size data) • Number of studies and population of studies (e.g., in Visible Learning Meta^x, we include a confidence ranking for each influence) • Quality of the research (i.e., opinions/anecdotes → systematic review) • Similarities between the context of the studies and your local environment
Ease of Replicability	<ul style="list-style-type: none"> • Is the intervention “productized” or do you need to build it yourself? • Are the steps easy to follow or open to wildly different interpretations? • Was it developed for your cultural/linguistic context and/or has it already been localized?
Local Capacity to Implement	<ul style="list-style-type: none"> • Do you have access to high-quality internal or third-party technical assistance to support implementation? • Is there buy-in from stakeholders? Does the intervention model conform with local stakeholder beliefs/theory of action? • Do stakeholders have sufficient time to engage/participate at the levels required for success? • Do local stakeholders have the skills to implement the new approach? How easy will it be to upskill them?
Cost of Implementation	<ul style="list-style-type: none"> • Total cost ÷ Total number of <i>Direct Beneficiaries</i> <p>Note: You also need to factor in reoccurring costs, not just the initial setup.</p>



 OPPORTUNITY SKETCHES	EVIDENCE OF IMPACT 1-5 (5=STRONG EVIDENCE)	EASE OF REPLICABILITY 1-5 (5=HIGH EASE)	LOCAL CAPACITY TO IMPLEMENT 1-5 (5=HIGH CAPACITY)	COST OF IMPLEMENTATION 1-5 (5=LOW COST)	TOTAL
Intelligent tutoring systems	5	3.5	2	3	13.5/25
Scripted direct instruction	5	2.5	1	4	12.5/25

Source: Hamilton and Hattie (2022).

you undertake this analysis, a subset of the opportunity sketches you identified will probably stand out as being much better bets for impact. These are the ones you will carry forward to the next stage of Design.

OTHER APPROACHES TO OPPORTUNITY SKETCHING

In addition to the systematic search processes that we outlined earlier, here are other approaches you could consider to triangulate and test your thinking.

- 1. Worst possible idea.** This is where you literally and deliberately come up with as many bad ideas as you can think of for “improvement” in your content. Then you identify all the similarities in those bad ideas, and then search for activities or programs that do the opposite of these bad idea features. You can also attempt to combine different features of the bad ideas together to see if it results in a good idea. One of the potential benefits of the deliberate search for bad ideas is that it’s less stressful and inclusive than asking stakeholders to generate good ideas. Everyone can think of bad ideas!
- 2. Analogy.** Make comparisons to other situations to test the logic of your thinking. You may have noticed that we have used analogy a lot throughout this book (getting to the moon, the Pyramids, high diving, COVID-19, etc.). We consistently find that making our thinking generic and applying this to new contexts helps us to quickly unpack the flaws in our logic. There is also a great deal of research on the benefits of analogy in transferable skills and critical thinking (e.g., Aubusson et al., 2006; Holyoak, 2012).
- 3. Bodystorming.** Here you use roleplay to literally act out the steps of implementing your identified opportunity sketches in order to explore what the practical barriers to delivery might be from the perspective of different stakeholder groups (e.g., teachers, leaders, students, parents) and even personas of different subcategories of each, such as newly qualified teachers vs. experienced teachers. This is very useful for stress testing, which we explore in Step 2.3.
- 4. Creative pause.** If the ideas are not flowing, just stop. Take a break, even for a few days. And start again.
- 5. Get a second opinion.** Speak to colleagues in other schools, districts, and/or systems that have progressed similar education challenges. Draw particularly on their lessons learned and wrong turns. However, be careful not to take their claims at face value. Always be driven by the evidence of impact and form a third opinion on their second opinion.
- 6. Subtract.** Systematically explore whether your education challenge might exist because you are doing too much, rather than too little. Could you make more progress by subtracting activities, programs, and initiatives? Sometimes less is more. However, by some accounts, we are cognitively primed to add rather than subtract (Adams et al., 2021).

All six of these approaches just provide you with *opinions*. You still need to cross-check these warm leads against high-quality evidence of impact—for example, by using the “Ranking Your Options in the Design Space” criteria outlined in Figure 2.8—even if this is locally adapted.

FIGURE 2.9 • Example Opportunity Sketch Analysis Framework

OPPORTUNITY SKETCH	TARGET GROUP	ASSUMED CAUSAL MECHANISMS	RESEARCH EVIDENCE	MODIFICATIONS	EXISTING PROGRAMS	COST	EASE OF IMPLEMENTATION	PROBABILITY OF IMPACT
What is it?	Who will we engage with?	What is the theorized path from A, to B, C, to Impact?	What evidence is available that supports this opportunity sketch?	Based on our review of the research, what tweaks would make it more effective?	Are there any pre-existing programs that we can buy/adapt? Do those programs have strong evidence of impact?	How much will it cost per school, per student, and per target student?	Can this be done quickly and with little effort or will it require protracted effort and buy-in?	How do we rate the chances of it working?
AI-driven early warning system to pre-identify at-risk learners	Students and teachers	Assumes that there are student data patterns that can be used to provide early warning of future absenteeism Assumes that educators can do something with these data and that this will reduce absenteeism in the targeted group	Correlational data from a range of US deployments where back tests show that their systems successfully predicted “challenging” students 18 months prior to emergence of absenteeism and behavioral challenges	The system alone will not bring about change. Requires structured intervention with the identified students	There is a range of off-the-shelf software systems, including: System “1” System “2” System “3”	\$1500 per school per annum/ average of \$2.50 per student per annum	Medium. Key issues: • Data-system integration • Actual use of the system by educators to reduce absenteeism	Moderate. It will provide data to enable educators to target support to at-risk learners but will not on its own solve the absenteeism challenge
Intelligent tutoring system to provide personalized instruction to students and actionable dashboard data to teachers	Students and teachers	Assumes that children’s individual learning needs can be identified through standardized online assessment and that an AI algorithm can select appropriate learning content to enhance each learner’s progress in their area of need	6 meta-analyses of 357 studies, involving 22,000+ students, generating an effect size of $d = 0.51$	These systems are “black boxes” (i.e., we cannot vary how they work). However, we can vary: • Which students have access • Duration of access • Where they access • How we link to classroom instruction	There is a range of off-the-shelf software systems, including: System “A” System “B” System “C”	The average cost is \$2.5 per student per week	Medium. Key issues: • Making time for students to use • Access to devices for all students • Selling the benefits to students, parents, and teachers	High. The systematic reviews are from our country. However, we need to make sure the selected system has high-quality independent evaluation data, rather than just being “based in the evidence”

2.2 BUILD PROGRAM LOGIC MODELS

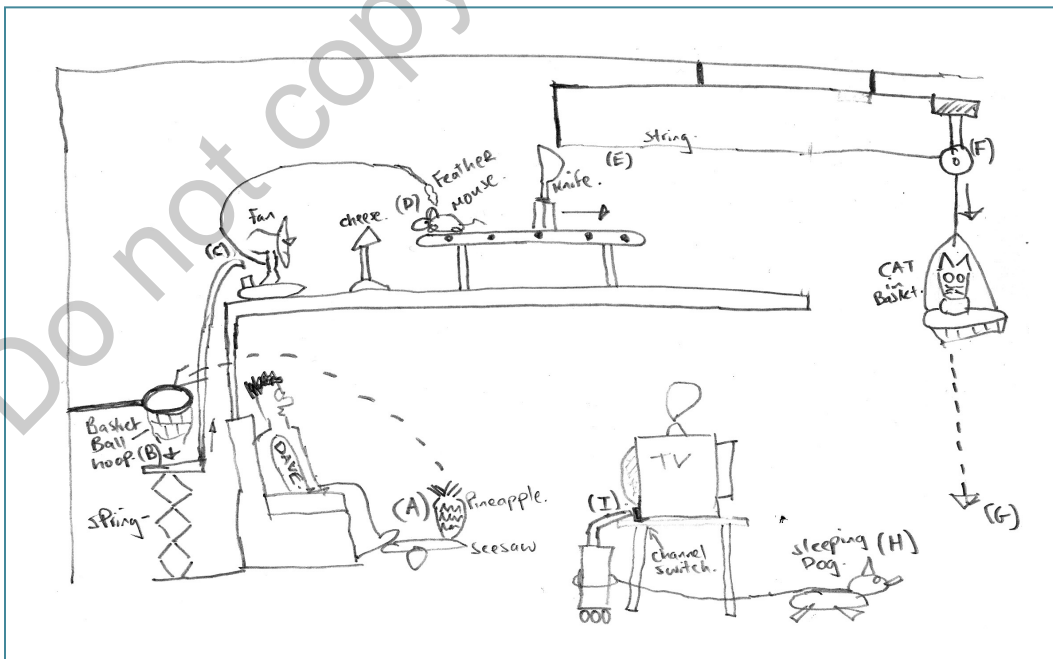
Now that you have identified and agreed on your higher-probability opportunity sketches, the next step is to decide how you will bundle and sequence them together into a coherent and integrated program design that addresses the various influence bubbles on your path analysis.

Back in the 1920s, American cartoonist Rube Goldberg became famous for a genre of cartoons, as illustrated in Figure 2.10.

This is a **Rube Goldberg machine**. It shows a man sitting in a chair, who we will call Dave, and his goal is to switch on the television. There are many ways Dave can go about it. First, he could stand up and walk across the room. Second, he could buy a long stick to poke the buttons on the front of his TV from the comfort of his chair. Third, he could buy a new TV that comes with a shiny remote control. Or fourth, he could build a complex contraption with many moving parts to do his bidding. These are each **theories of improvement**, or high-level ideas about broad types of intervention that could be effective. This is akin to an opportunity sketch. The next level down from this is a **theory of action**. This is significantly more detailed and spells out the end-to-end inputs, activities, outputs, and outcomes that will bring the theory of improvement or opportunity sketch to life.

In Dave's case, he has opted for the "complex contraption" theory of improvement. And the cartoon spells out step by step the specific theory of action. In this case, he flicks a seesaw with his foot (A),

FIGURE 2.10 • A Rube Goldberg Machine



which fires a pineapple through a basketball hoop that then activates a spring-loaded platform (B). This activates a fan that shunts a mouse into action on a conveyor belt (C and D), which in turn pushes a knife through a piece of string (E), thus dropping a cat in a basket (F and G), propelling a dog forward (H), and, abracadabra, changing the TV channel (I)!

Of course, for something as simple as changing a TV channel, this theory of action seems a tad too elaborate. There are too many moving parts—all prone to failure. What if the mouse wanders off or the dog falls asleep? The TV stays stuck on the same channel. Indeed, the sheer beauty of Rube Goldberg machines rests in the fact that we all know they simply won't work; and the joy (and giggling) comes from visualizing all the places where theory and practice are bound to diverge.

However, your selected education challenge is probably a lot more complicated than changing a TV channel. You wouldn't have set up a backbone organization to progress something *that* simple. Instead, it is likely that your implementation requires many complex moving parts—some of which might be prone to failure or at least not do quite what you intended, when you intended. Therefore, it helps considerably if you map out your version of the Rube Goldberg machine end to end to see whether it makes sense and what the potential points of failure could be.

If you want to draw this out like a cartoon, you can. If you want to use sticky notes, you also can. However, one tool that we have found useful in our work is the **program logic model**. This was first formalized by the United States Agency for International Development (USAID) in the late 1960s, based on the thinking tools used at NASA for the moon landings (World Bank, 2000). However, it has taken more than five decades for this approach to catch on to education, and it is still early days for use and adoption.

The program logic model template gives you a structured framework to explore and address the following questions:

1. **What is our education challenge** (i.e., the “problem” we are trying to fix or the moonshot goal we are seeking to progress)?

For Dave: changing the TV channel, without getting out of the chair

For you: whatever it was that you agreed on during the D1 Discover Stage

2. **What activities will be undertaken** with this resource to generate improvement in the education challenge area and with what stakeholders?

For Dave: actions A–I in Figure 2-10. Dave will need to STOP doing some things: to build the machine, to keep it oiled, and

to train and feed the animals. In this case, he's decided to forgo playing within his model railway set for 2 weeks.

For you: again, those you identified in your opportunity sketches (i.e., the specific programs, interventions, actions, etc.)

3. **What resources do we need to deploy** to implement our identified opportunity sketches (i.e., the people, time, budget, etc.)?

For Dave: mouse, cat, dog, and some metal parts and foodstuffs to make the contraptions

For you: those you identified in your opportunity sketches

4. **What assumptions are we making** about how and why this will work?

For Dave: that the mouse will be obedient; that the dog "likes" cats; and that all the springs, levers, and belts will interact perfectly

For you: that the interventions you selected are robust, relevant to your context, and will generate impact

5. **What will the outputs of the activity be** (e.g., the "products" created, the number of people engaged with, etc.)?

For Dave: the device for changing TV channels is successfully built and installed

For you: it could be curriculum materials developed, people coached, training events that have taken place, etc.

6. **What measurable outcomes do we expect to see** from implementing the intervention over the short, medium, and longer term?

For Dave: being able to switch over from *The Simpsons* to *MacGyver* at 7 p.m. daily, without getting out of the chair

For you: an increase in student literacy outcomes or whatever your specific education challenge areas happen to be

7. **How will we collect data and measure** for monitoring and evaluation purposes? And what types of data will we collect?

For Dave: maintain a logbook detailing whether the device successfully switched the channel at 7 p.m. each day. He will also commission an independent evaluator to explore each point of linkage in his machine and identify areas of efficiency (Dave likes to overengineer everything!).

For You: data on teacher participation in your literacy training program and on enhanced student achievement, or in the specific education challenge you have decided to progress.

THEORY OF THE PRESENT VS. THEORY OF IMPROVEMENT VS. THEORY OF ACTION

THEORY TYPE	DESCRIPTION
1. Theory of the present	<ul style="list-style-type: none"> Your validated explanation about why your education challenge exists (i.e., what drives it, what is the root cause, and what is the path analysis?) Dave cannot change the TV channel easily because the TV does not have a remote control
2. Theory of improvement (i.e., high-level)	<ul style="list-style-type: none"> Your high-level theory of what you will do to improve the situation Dave will build a mechanical contraption to change the channel
3. Theory of action (i.e., detailed)	<p>Your more detailed explanation of the key design features and setting levels</p> <p>Dave's contraption will be made out of steel and contain nine linking elements (pineapple, cheese, mouse, knife, cat, dog, etc.)</p>

All of the previous steps in the *Building to Impact 5D* framework have been explicitly designed to support you to answer these questions and to then build your answers out into an integrated and coherent program logic model.

Let's recap some of the key steps that support your readiness to develop that logic model, after you decided and defined your education challenge. See Figure 2.11.

With this thinking and these outputs, you are now ready to start completing some of the dimensions in the program logic model template, as introduced in Figure 2.12.

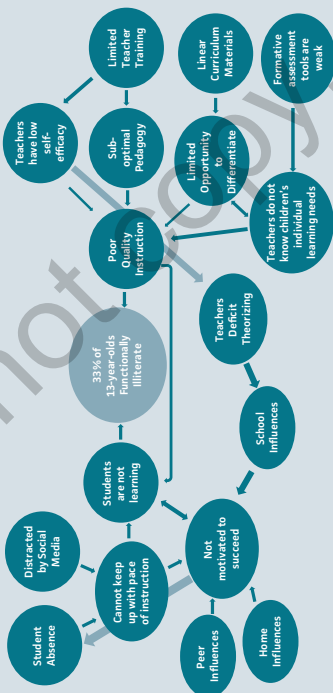
What you insert at this stage might look something like what we present in Figure 2.13.

If you think back to the very start of the book, we introduced the notion of *optimal stopping*. Basically, this is about whether you will luck out and select the "best" option the first time you view a new house, go on a first date, or search for a new car. Most of us do not buy the first house we view, marry the first person we meet, or buy

FIGURE 2.11 • Recapping Key Prior Steps

1. PATH ANALYSIS STEP (1.3)

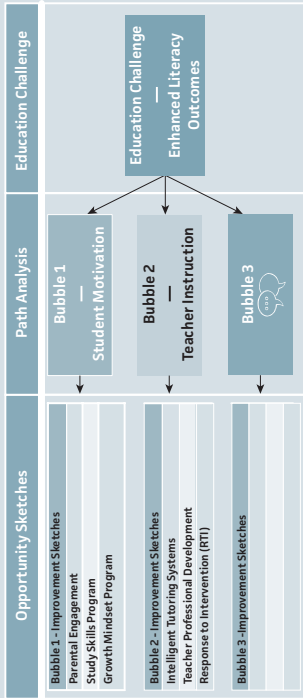
FIGURE 1.9 • Path Analysis



You built a causal model of the present to explain the key dimensions of your education challenge. This contains a number of influence bubbles.

2. OPPORTUNITY SKETCHING STEP (2.1)

FIGURE 2.7 • Opportunity Sketch Mapping



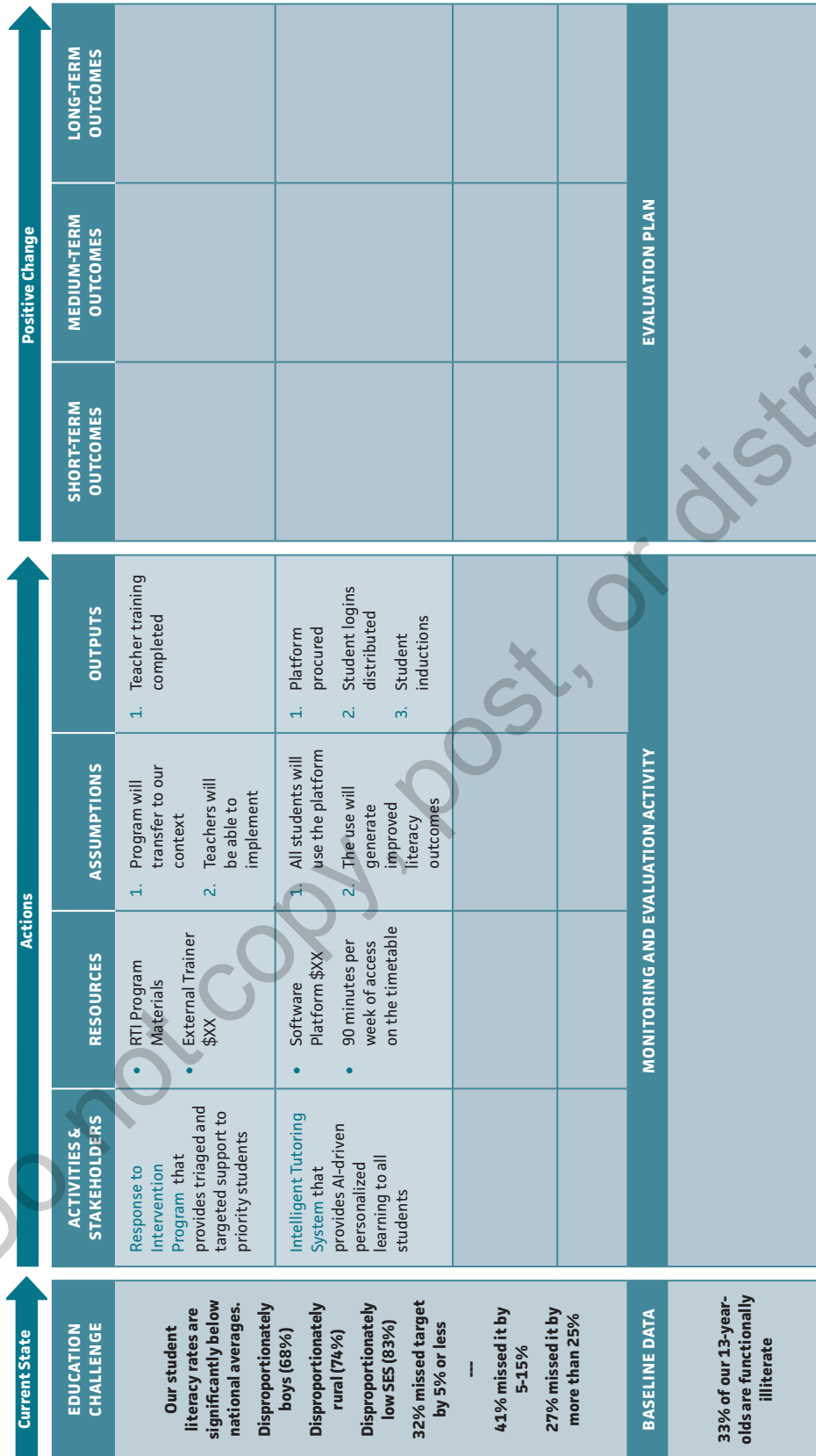
You then developed a longlist of ALL the potential opportunity sketches that could interact with the various Influence Bubbles to negate or reverse their impact.

3. RANKING OPPORTUNITY SKETCHES STEP (2.1)

OPPORTUNITIES SKETCHES	EVIDENCE OF IMPACT 1-5 [5=STRONG EVIDENCE]	EASE OF REPLICABILITY 1-5 [5=HIGH EASE]	LOCAL CAPACITY TO IMPLEMENT 1-5 [5=HIGH CAPACITY]	COST OF IMPLEMENTATION 1-5 [5= LOW COST]	TOTAL
Intelligent Tutoring Systems	5	3.5	2	3	13.5/25
Scripted Direct Instruction	5		1	4	12.5/25

You then scored and ranked the opportunity sketches using criteria to identify those with the most potential for impact across the widest range of influence bubbles. The highest-potential opportunity sketches are now what gets translated into your **Activities and Stakeholders, Resources, Assumptions, and Outputs** in your program logic model.

FIGURE 2.13 • Partially Completed Program Logic Model



Source: Copyright © Cognition Education. (2022). All rights reserved.

the first car we see. And, equally, home builders and car manufacturers do not put their first ideas straight into production. They build multiple prototypes—some just on paper, others in physical form. They do this to double-check and to mitigate the risk that they launch a lemon when they go to market.

The same principle applies to *your* program logic modeling activity. Yes, you *might* hit the jackpot the first time. But it's just as likely that you will land on the lemon. Therefore, we strongly advise that you work up a few different logic models. Going back to our literacy example, some logic models might be simple with only one core work package or opportunity sketch—such as the introduction of an intelligent tutoring system, which children access at home once a week and for one period a week in school. Others might be more complex and involve the introduction of major packages of teacher professional development and perhaps even a whole new instructional approach, like RTI. Obviously, more moving parts = more risk that the dog doesn't play ball. Conversely, fewer moving parts mean that what you are doing is too simple and that it does not properly interact with the various “bubbles” in your path analysis.

You can now either pick your best design(s) and quickly get to implementation or you can spend a little more time **stress testing** before lift-off. If you are working at the district level (or higher), we recommend that you use the tools and processes in Step 2.3 to pre-test your proposed logic models. You will likely be seeking to progress an education challenge across multiple schools, at scale. Therefore, it's profoundly important that you explore your selected activities or interventions from *all* angles before you inadvertently waste the time of stakeholders on ineffective initiatives. If you are working at the school, department, or professional learning community level, we encourage you to understand these tools and perspectives, although they are not mandated. We completely get that sooner (rather) than later you need to get on with implementing *something* and that you can collaboratively adjust it as you go. The longer you procrastinate, the more likely you'll just stop altogether. In which case, draw on Step 2.3 for inspiration and additional considerations and then move on to Step 2.4.

It's profoundly important that you explore your selected activities or interventions from all angles before you inadvertently waste the time of stakeholders on ineffective initiatives.

2.3 STRESS TEST AND IMPROVE ON STEP 2.2

These processes are mandated if you are working at the whole-school or district level. They are highly recommended if you are working as part of a teaching team or professional learning community, but you might undertake them more quickly (i.e., reach optimal stopping sooner).



GOING DOWN THE RABBIT HOLE

While program logic models help you to build a high-level map of how the mouse *should* interact with the feather, cheese, string, and cat, there's a second tier of detail. This might also significantly impact whether your Rube Goldberg machine is effective. This includes the type of mouse, whether it has been trained, how frequently it is fed, whether it actually likes cheese, and therefore whether fruit or seeds would be better bait. And, of course, we can ask the same types of questions about every other link in the machine: the weight and size of the pineapple, the length and sharpness of the knife, the size and color of the cat, and so on.

In our work, we have found that leaving these considerations to chance, just assuming that “any old cheese or mouse” will be just as effective, and also not considering which should come first (the mouse or the pineapple) creates too much risk that you fail to convert your initial energy into a drive that positively impacts all the other links and connectors in your program logic model.

Every potential activity or intervention that you decided to include in your logic model can be varied. Here are some of the generic sources of variation (Hamilton & Hattie, 2022):

- **Dosage** (How much “medicine” do we give?)
- **Duration** (How long do we give it for and at what spacing between “doses”?)
- **Target group** (Who is selected for “treatment”?)
- **Delivery group** (Who implements the initiative?)
- **Fidelity** (How much variation is allowed in how the treatment is delivered locally?)

Other opportunities for variation will be dependent on the specific activity/intervention you plan on implementing (i.e., they are *regimen specific*). For example, if you are opting for an intelligent tutoring system to remediate children's literacy, other considerations will include these:

- Which of the many available systems is selected for use?
- Is use mandatory or optional?
- Is it only for struggling students or for all learners?
- Are parents going to be briefed or even co-opted as activators?

- Is it used at home and/or at school?
- Is it standalone or is the system also used in class for group teaching?
- Will teachers use the formative assessment data to enhance their classroom teaching?
- Will the school leadership use the baseline assessment data as an accountability tool to (secretly or even openly) evaluate teacher performance?
- Will children be allowed to access it on their phones or only on tablets and desktops?
- Will children's use be monitored?
- Will there be any rewards or sanctions for students that under or over-use the platform?

We call these considerations **design features**. For each design feature, there are multiple **setting levels**. Some design features can be switched off entirely or set to zero. And for all the active features, there are several different positions (i.e., setting levels) that the dial can be set to. In our work, we have found it useful to explicitly map *all* the potential design features and setting levels and to use this information to select the optimum ones with great care. Not to do so risks random pineapples, cats, and feathers getting thrown into the mix without much thought for how they can be selected and sequenced with greater care and deeper impact.

In Figure 2.14, we illustrate how you can map the design features and setting levels for each of your opportunity sketches.

The mapping table in Figure 2.14 is only a worked example. For the average activity or intervention, it is possible that there will be 25 or more design features that are worth thinking about. Once you have identified them, initial questions are whether to switch them on or off, whether to leave them to local discretion, or whether to pick and lock a specific setting level.

There is a second level of complexity. For each design feature that you decide to activate and lock, there might be 10 or more setting levels for you to choose from. This means that there are likely 250 or more different settings that you can move the various dials through (i.e., 25+ dials \times 10+ setting positions on each dial). And this is just for one intervention! If you decide to combine intelligent tutoring with a growth mindset program, RTI, *and* teacher professional development, there are equally many design features and setting levels for each of these too!

FIGURE 2.14 • Design Features and Setting Levels

ACTIVITY: INTELLIGENT TUTORING SYSTEM					
	DESIGN FEATURE 1: WHICH SYSTEM DO WE SELECT?	DESIGN FEATURE 2: IS IT MANDATORY?	DESIGN FEATURE 3: HARDWARE	DESIGN FEATURE 4: PARENTAL ENGAGEMENT	DESIGN FEATURE 5: DOSAGE LEVEL
Setting Level 1	Tutoring Platform A	Mandatory for all students	Students' own devices (any)	None (all done in school)	Left to personal choice
Setting Level 2	Tutoring Platform B	Optional for all students	Students' own devices (non-smartphone only)	None (but used at home and school)	Minimum 60 minutes per week
Setting Level 3	Tutoring Platform C	Mandatory for students who have fallen behind. Optional for all others	School tablets	Parent newsletter	Maximum of 90 minutes per week and minimum of 60
Setting Level 4	Tutoring Platform D		Hybrid	Parent briefing session	
Setting Level 5	Tutoring Platform E			Telephone call to parents	
ANALYSIS					
	Platform E has been used in other schools locally. Staff very positive. Also, strong independent evaluation data	If it's optional, no one will take it seriously. If it's mandatory only for students that have fallen behind, this creates stigma	We are not sure it will make much difference. Let's start with hybrid and see what happens	It would probably be helpful to inform parents!! Let's start by putting it in the newsletter and see if any request more information	The global research suggests that 90 minutes a week spread over 3 x 30-minute sessions is about optimal
CONCLUSION					
	Platform E	Mandatory for all	Hybrid	Parent newsletter	90 minutes, split between three sessions

Source: Adapted from Hamilton and Hattie (2022).

One day soon, we will hopefully be able to run all these options through a software engine like IBM Watson or WolframAlpha to help us identify and select between a seemingly infinite number of design options and setting levels. Until this happens, it's important that you give the selection and interaction of design features as much thought as you can. Even seemingly minor details like letting students access intelligent tutoring systems from their personal smartphone devices can have unanticipated implications—with the screens being too small to view the content or to type their responses, plus the feed of K-pop videos on TikTok acting as a constant distractor.

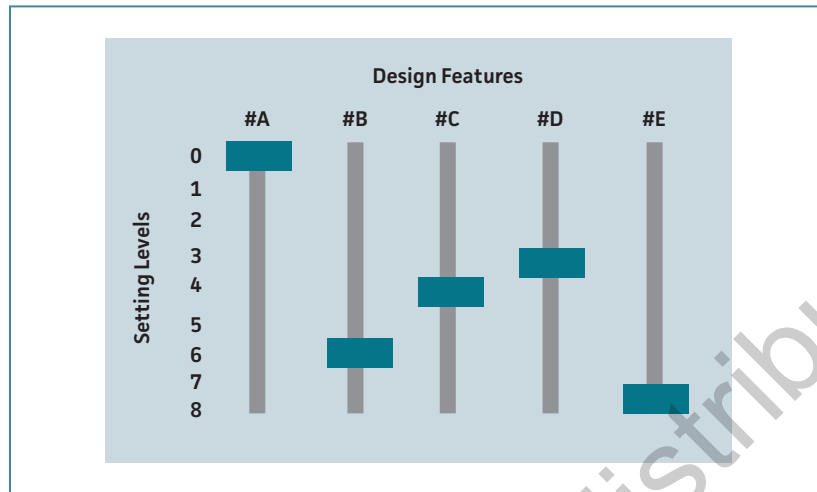
It's important that you give the selection and interaction of design features as much thought as you can.

In our work, we find those initiative designers often only work down as far as the high-level program logic model—that is, agreeing that there will be mice, cats, cheese, and industrial-looking machinery connecting them. They never quite get to the detail of whether all these features are needed, whether they are connected in the right order, and whether it should actually be a kitten rather than cat. Our message to you is that one of the key reasons that implementation often fails is that each of these microfeatures is left to chance, being considered as an unimportant detail. Of course, another source of failure is spending so long on this that you end up in analysis-paralysis! This is why **optimal stopping considerations** are so important. At some point you need to make a judgment call about when it's time to follow Elvis's sage dictum: *A little less conversation, a little more action, please.*

However, even if you spend a relatively small amount of time going down the rabbit hole, one of the benefits of mapping (at least some of) the various design features and setting levels is that if during implementation and evaluation you are dissatisfied with the degree of impact, you can go back to your mapping and identify aspects of your design that could be iterated to enhance the overall efficacy. It may be that you subsequently decide to activate or deactivate specific design features or to incrementally adjust the setting levels on those that are activated.

We also find it helpful to think about design features and setting levels as being a little like the graphic equalizer deck that you used to find on old-fashioned HiFi music centers and that you still see in sound studios. Figure 2.15 illustrates what we mean by this. The idea is that you carefully consider the “best” position for each of the sliders and you explicitly lock those setting levels before you start Stage D3 (Deliver).

FIGURE 2.15 ● The Graphic Equalizer



Source: Copyright © Cognition Education. (2022). All rights reserved.

THINK (A LITTLE BIT) LIKE EEYORE

Many things look good on paper. In fact, the process of typing them up in pretty fonts and colors and inserting icons and infographics can often give hair-brained notions a self of authority or legitimacy that they don't deserve. There's a risk that as you undertake the 5D processes in this book—potentially using our templated tools—you (initially) become seduced by the words on the page and (much later) surprised when your initiative comes apart at the seams.

You probably know or have heard of the *Winnie-the-Pooh* books by A. A. Milne. One of the characters is a gray overstuffed donkey called Eeyore, who is renowned for his pessimism. He always expects bad things to happen and imagines them in advance. Unfortunately, he also stoically accepts what happens and never (usually) tries to prevent what happens.

We now want you to think (a little bit) like Eeyore. Before you get busy implementing your dashing design(s), it helps if you take the time to explore all the ways what you are about to do could go hideously wrong. Unlike Eeyore, you are not going to stoically accept and embrace these impending visions of doom. You are doing this so you can preempt and build mitigations and contingencies into your program logic model—to reduce or even to “design out” the risk.

To get your Eeyore-like mental juices flowing, here are some of the different types of implementation risks you could consider.

THEOREY 1: STAKEHOLDER BELIEFS

From research across a range of sectors, we know that people's existing beliefs are a key determinant of whether their collective future actions will generate impact (Knoster, 1991; Robinson, 2018; Robinson et al., 2009). There are two dimensions to this:

- **Self- and collective efficacy.** Where you believe that you have the individual and collective power to make a difference, generally you do! This positive belief drives positive action. Of course, you are more likely to hold those beliefs where you have confidence that (1) what you are about to embark on is within your existing wheelhouse or capability set and (2) it builds on and stretches your existing “superpowers” rather than requiring you to, say, learn Arabic overnight.

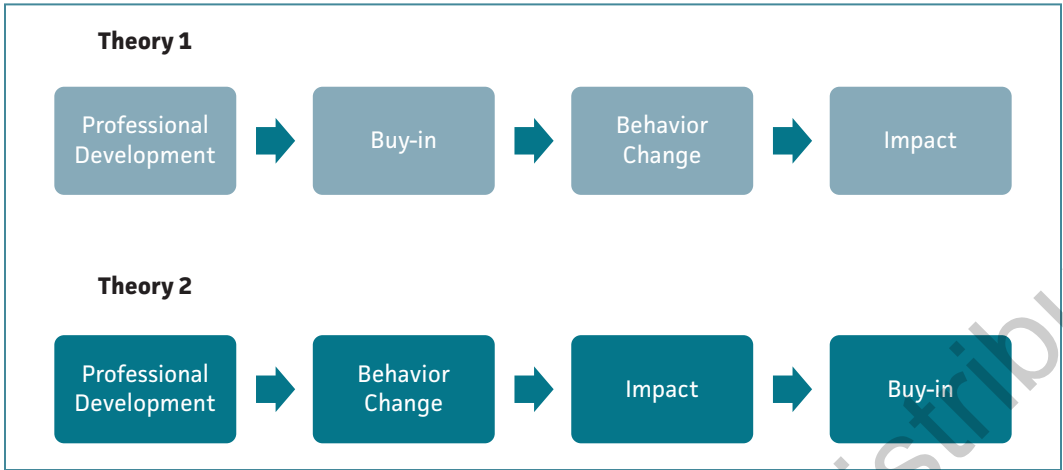
Implication: You need to make sure that your program logic model is predicated on “desirable difficulty” (Bjork, 1994), that it builds on existing capabilities in your team, and that you factor in time, support, and love for people to fall into a few bear traps and learn from this along the way.

- **Worldview.** We all have an implicit theory about human nature, what is important in life, and what being a good educator is all about. Even if we can't consciously articulate those beliefs, they implicitly drive our actions. For example, in some of our work we have engaged with school leaderships that want to implement scripted instructional approaches. While there is a strong evidence base for some of these types of intervention (depending of course on what it is exactly that you are trying to achieve), experienced teachers often do not see themselves as actors reading a script. So, their worldview about teaching and the execution of their professional competency is implicitly misaligned with the philosophy of the intervention. And all of us, when asked to do something that we don't believe in, either go along grudgingly or (perhaps) even attempt to drill holes in the side of the boat while everyone else is rowing away.

Implication: You either need to select actions that are strongly aligned with your team's existing mind frames, to avoid the jarring dissonance, or to build in time to positively engage, build bridges, and shared understanding (i.e., thesis – antithesis – synthesis).

Note that there are two contrasting perspectives on how we confront the misalignment of people's existing beliefs with the actions we are asking them to carry out, which we illustrate in Figure 2.16. Theory 1 suggests that we need to spend a significant period of

FIGURE 2.16 • Theory 1 vs. Theory 2



Source: Copyright © Cognition Education. (2022). All rights reserved.

Behavior change and impact come first, and belief changes a lagging second.

time engaging with those prior beliefs, convincing stakeholders that they are wrong (or coming to some compromise perspective) and getting their buy-in *before* they will go on to adopt and implement. It is a common perspective in generic business improvement texts; and the Japanese call it *Nemawashi*, which roughly translates as ‘laying the groundwork’. Theory 2 is championed by Thomas Guskey (2020) and Doug Reeves (2021a). It suggests that *seeing is believing*—that most of us are only weakly convinced by dialogue and data; and that by shining a light on the difference of perspective, we might inadvertently encourage people to hunker down further into their pre-existing mind frames (i.e., the Backfire Effect). Theory 2 suggests that we are only strongly convinced/converted *after* we have put something new into practice and seen the positive impact with our own eyes. In other words, behavior change and impact come first, and belief changes a lagging second.

The research on these two different theories on the relationship between beliefs and action is still a work in progress. However, recent research involving students suggests that initial task success (i.e., impact) often precedes and primes the motivation to continue to invest more time and energy (Sinha & Kapur, 2021). This offers support for Theory 2.

In the context of *Building to Impact 5D*, we offer the following guidance. If there is a low buy-in for what you are seeking to implement but there is (a) extremely strong evidence that it has been effective in similar contexts to yours and (b) the new approaches can be learned and/or acquired to a satisfactory level with relatively low investments in training, then you could opt for the Theory 2

approach. You ignore the noise, push on with mandated implementation, and you wait (with bated breath) for people to report back that “it works!” and to ask for more.

EEYORE 2: MOTIVATION

From the research on deliberate practice and elite performance (Ericsson & Harwell, 2019; Ericsson et al., 1993), we know that across the full range of sectors and professions, becoming highly proficient in technically complex fields like teaching takes 10 years or around 10,000 hours of *effortful* practice. We also know that educators are often more motivated to do this early in their careers and more prone to plateau later (Papay & Kraft, 2015; Rice, 2013). Therefore, you may find that if you expect stakeholders to do something radically different to their current repertoire, those who are earlier in their careers might be more open to it. In contrast, more established teachers might require additional support and motivation to encourage them to make the leap. Or you might need to design features into your intervention that enable it to be effective even without that leap.

EEYORE 3: FRICTION

This is about the quantity of change and the level of (personal) time and effort required to achieve it. If you are introducing a new step to an existing process (i.e., juggle *one* extra ball), the implementation friction is likely to be lower. Whereas if you expect stakeholders to quickly learn to juggle *five* extra balls or to shift from juggling to “cake decorating,” the level of friction is likely to be significantly higher.

EEYORE 4: MAINTENANCE

Many of us make New Year’s resolutions to lose weight or increase our fitness. However, how many times have you heard someone else making such a resolution and then inwardly thought to yourself “by February all that gym equipment will be in the cupboard collecting dust”? And how many resolutions have you made yourself and subsequently failed to keep or maintain? Maintenance is *really* hard. In the domain of weight loss alone, the research tells us that around 80% of successful dieters rebound to their previous weight (or more) within 5 years (Wing & Phelan, 2005). If something as simple as watching what we eat is so darned hard, you need to consider whether what you are asking or expecting stakeholders to do will be easy or difficult to maintain. And you may also need to consider what ongoing support measures you could build into your program logic model to keep everyone on the up-and-up. We come back to this in Stage D5 (Double-Up).

EEYORE 5: MUTATION

Dylan Wiliam (2018) argues that through professional development it's relatively easy to get teachers to adopt new ideas: the hard part is getting them to stop what they were doing before. To us, this is profound. Often, "new ways" are designed to be implemented with fidelity—much like the insertion of a medical intravenous line, as we discussed in the Introduction. There are highly effective ways to do this that significantly reduce the probability of a bacterial infection. However, many, many line infections still occur; and the reason is that medical practitioners do not always follow the training or the protocols to the letter. They sometimes adapt, cutting corners to enhance the efficiency or blend bits and pieces of their new training with their prior practice.

We often see the same kind of challenge in the implementation of new education initiatives. Larry Cuban (1984, 1990), in his analysis of the transition from "traditional" to "progressive" pedagogies, concluded that very few teachers made a genuine transition from one "way" to the other. It was more common that educators cherry-picked the bits that they liked and blended these with their existing repertoire.

In stress testing your program logic model, you need to consider whether fidelity of implementation is important and/or the level of mutation that you can live with and accept. If fidelity is critical, you will need to build in a support infrastructure to ensure that in implementation, what happens at the chalk face does not end up becoming a grainy imitation that bears a passing resemblance to the original (i.e., a photocopy of a photocopy of a photocopy) (Elmore, 1996).

EEYORE 6: VOLTAGE DROP

When the great industrialists of the 19th century laid cables across the lands to transport electricity, they quickly discovered that the voltage dropped or dissipated over greater distances. Hence, distribution is required at much higher voltages than are needed in domestic settings and for substations to dilute the juice, so that it doesn't blow up your TV.

A common challenge when implementing new initiatives at scale (e.g., across multiple schools) is that, metaphorically, the initial "voltage" is too low. That is, it's strong enough to "power" one or two schools but when you try and hook up 50 to the same "grid," the power that is transmitted is too low (e.g., see Kilbourne et al., 2007).

So, if you are planning on implementing at scale, you will need to consider how you boost the juice or whether you can accept lower levels of current as you support more and more settings with adoption. The same thinking also applies *within* a single school. Perhaps

your program logic model envisages training a small group of pathfinder teachers and you then *assume* they will “pass it on” without voltage drop to their colleagues.

EEYORE 7: SIDE EFFECTS

When you buy medicine from the pharmacy, there is usually a leaflet inside the packet that contains the small print. A key element of that text is usually a list of all the known side effects that *could* materialize if you take the pills and what you should subsequently do if they occur. It is well known and accepted in medicine that occasionally the cure can be worse than the disease.

Yong Zhao (2017, 2018) has applied the same notion of side effects to education interventions, noting that every opportunity comes with a potential cost. For example, problem-based learning *might* increase student creativity, engagement, and attendance but with a side effect that the overall speed of learning is slower, that misconceptions may be inadvertently reinforced, and that students may not be exposed to key bridging concepts that they required to develop advanced subject matter knowledge. Ditto for direct instruction. This *might* increase the efficiency of learning and ensure content is appropriately sequenced and staged. But it *might* also come with the side effect of inducing boredom and stifling creativity.

You need to consider what the potential side effects of your interventions could be. Whether those are acceptable or whether they require countermeasures. In medicine, doctors often respond by either (a) identifying a different treatment whose side effects are more acceptable or (b) by prescribing an additional intervention that is *for* the side effects (e.g., *medicine A makes me nauseous, so I also take antinausea meds*).

HOW DOES THIS EYORE THINKING HELP?

The idea is that you use this Eeyore thinking (1) to identify all the things that could go wrong in the implementation of your program logic model and (2) to map out mitigations. You can also use the bodystorming technique we introduced in Step 2.2 to roleplay the implementation steps of your logic model, particularly from the perspective of stakeholder reaction. With this technique, you can literally act out the revulsion, misunderstandings, and horde of villagers descending with their burning pitchforks. Of course, your reason for doing this is to develop mitigations and countermeasures and then to consider whether these will be strong enough to hold the Eeyores at bay.

You can use Figure 2.17 to map all these Eeyores. In the first column, you describe the risk. In the second and third columns, you rank the probability of the risk occurring and the severity

FIGURE 2.17 • Program Logic Model Risk Mitigation Planning

NO.	RISK DESCRIPTION (OR "EYORE")	STAKEHOLDERS	LIKELIHOOD (L) 1-5	SEVERITY (S) 1-5	IMPACT L x S	MITIGATION
1	Project design includes a requirement for teachers recording and sharing videos of their lessons. They may be extremely uncomfortable doing this and/or interpret it as accountability rather than an improvement initiative	Teachers	4	3	12	<ul style="list-style-type: none"> Use a specialist video capture platform, so that teachers can control when and to whom they share their videos Lead by example. The SLT will film an exemplar lesson and share video for review at a film club event
2	We will be providing teacher professional development in RTI via a third-party training program. Our educators may not engage with it because they did not build it and it was developed overseas	Teachers	5	5	20	<ul style="list-style-type: none"> Have the identified teachers [opinion leaders] conduct a study visit to other schools nearby that are using the program. They will then report back to the whole teaching body on their findings Consider the option of localizing the materials (i.e., running workshops with our teachers to make modifications to materials, particularly key terms and linkage to organizational vision)
3	Project design involves the use of an intelligent tutoring system by all students. They may not want to participate	Students	4	3	12	<ul style="list-style-type: none"> Make it optional at first to fully test the system and generate buzz and buy-in via trail-blazers Make it recommended and then publish league tables of top users and give them school house-points Include each student's usage ranking in their parental report card

Note: SLT, senior leadership team.

Source: Adapted from Hamilton and Hattie (2021).

of impact. Then in the final column, you outline your mitigation strategy.

Assuming that you stress test your preferred program logic models carefully (and we really think you should), you will likely identify many bear traps you could have stepped into and many improvements that reduce your probability of ankle pain. The idea is that you circle back around, and you adjust your initial logic model to incorporate all this learning.

2.4 AGREE ON WHAT YOU ARE GOING TO STOP

As we are sure you have noticed, there are 24 hours in a day. Not 26 or 37. You will also have noticed that not all those hours are amenable to being leveraged to progress your education challenge. For a start, you need to ringfence 8 hours of shut-eye. You probably also want to have a life outside your work. And within your working day, there are undoubtedly a myriad of business-as-usual activities, pre-existing special projects, and the occasional bit of ad hoc firefighting filling up your time. We have yet to meet an educator who has an hour or two allocated each day for navel gazing or with flex time waiting to be filled. In fact, when we look at the comparative Teaching and Learning International Survey (TALIS) data, it's pretty clear that no matter where you are in the world, you are likely to be working long, grueling hours already (Organization for Economic Cooperation and Development, 2020).

Therefore, before you start progressing your new education challenge, you really, really, really need to take stock of all the other special projects you are already progressing. And you need to do this to find some to STOP, so that you can reallocate the time to this new and more pressing agenda. In Figure 2.18, we present a four-column tool that you can use to support this audit.

You may think this process extreme and, yes, it is. The whole point is to get you to think about all your competing priorities and the level of data you have on hand that demonstrates they are worth continuing rather than deep-sixing. If you take a hard-core approach to this, you will only continue to progress the projects in column 4 of your table and you will drop everything else. However, as an absolute minimum, we propose the *RULE of TWO-for-ONE*. In other words, for every ONE change initiative that you propose to start, find TWO initiatives of similar *time commitment* that you are going to stop.

FIGURE 2.18 ● The STOP Audit 

ALL OUR CURRENT PROJECTS	PROJECTS WITH SYSTEMATIC EVALUATION DATA	PROJECTS WITH REALLY POSITIVE EVALUATION DATA	POSITIVE PROJECTS THAT STILL NEED PUMP-PRIMING
<p>In this column, you list all your active special projects.</p>	<p>Here, you narrow down to those that you have bothered to systematically collect evaluation data for.</p> <p>If you have not set up evaluation protocols, your initiative is more likely to be busywork that's not worth your time; and we recommend that you assume that anything you are NOT evaluating is having no impact.</p>	<p>Now you narrow down even further to the projects you are systematically evaluating and where the data show extremely strong returns.</p> <p>Elsewhere we have suggested the use of effect size statistics. If your pre/post-assessments don't show a gain of at least $d = 0.40$, then consider carefully whether they are worth continuing.</p>	<p>Of the positive projects that are generating profound impact, how many still need centralized support to keep them going?</p> <p>It may be that many of the changes have already become engrained and sustained, or that the original need no longer exists.</p> <p>Put the projects that <i>still</i> need continuing/backbone team oversight here.</p>

FIGURE 2.19 ● The Cognitive Bias Codex

COGNITIVE BIAS	DESCRIPTION	KEY REFERENCE
Optimism bias	The tendency to be overoptimistic about the probability of success and not to develop contingency plans/mitigations (e.g., "I'm sure it's working and that we need to continue with it. Otherwise, why would we have even started it, right?")	Sharot (2011)
Plan continuation bias	Failure to recognize that the original plan/design is no longer relevant and to adapt to the changing situation (e.g., "I know the building is on fire, but we still need to hold the parent-teacher conference")	Heath (1995)
Sunk cost fallacy	Continuing to implement even where data show lack of impact: because so much time, effort, and money have already been invested and it is too emotionally distressing to conclude it has all been in vain: the show must go on!	Arkes and Blumer (1985)
Anecdotal fallacy	Treating anecdotal evidence as being of equivalent value to more rigorous evaluation protocols (e.g., "Everyone likes it, so we should carry on")	Gibson and Zillman (1994)

COGNITIVE BIAS	DESCRIPTION	KEY REFERENCE
Continued influence effect Conservatism bias Confirmation bias	Holding on to prior beliefs about the efficacy of an intervention, even when systematically collected data contradict the misinformed prior belief (e.g., “I don’t care what the data say. I know what I can see and feel. I believe it’s working!”)	Nickerson (1998)
Expectation bias Observer expectancy effect	Tendency for evaluators to believe, collect, and publish data that conform with their prior expectations and to treat contrary data with disbelief/skepticism (e.g., “The data aren’t looking so rosy. They must be wrong. I’ll delete them and focus on the two positive anecdotes”)	Rosenthal (1966)
Ostrich effect	Avoidance of monitoring/collecting data that might cause psychological discomfort. Originally identified in the financial sector, where investors stop monitoring their investment portfolios during market downturns (e.g., “This isn’t looking so good. Let’s just stop collecting data. It’s too painful to look. We’ll keep going with the initiative though. People will be upset if we stop”)	Galai and Sade (2006)

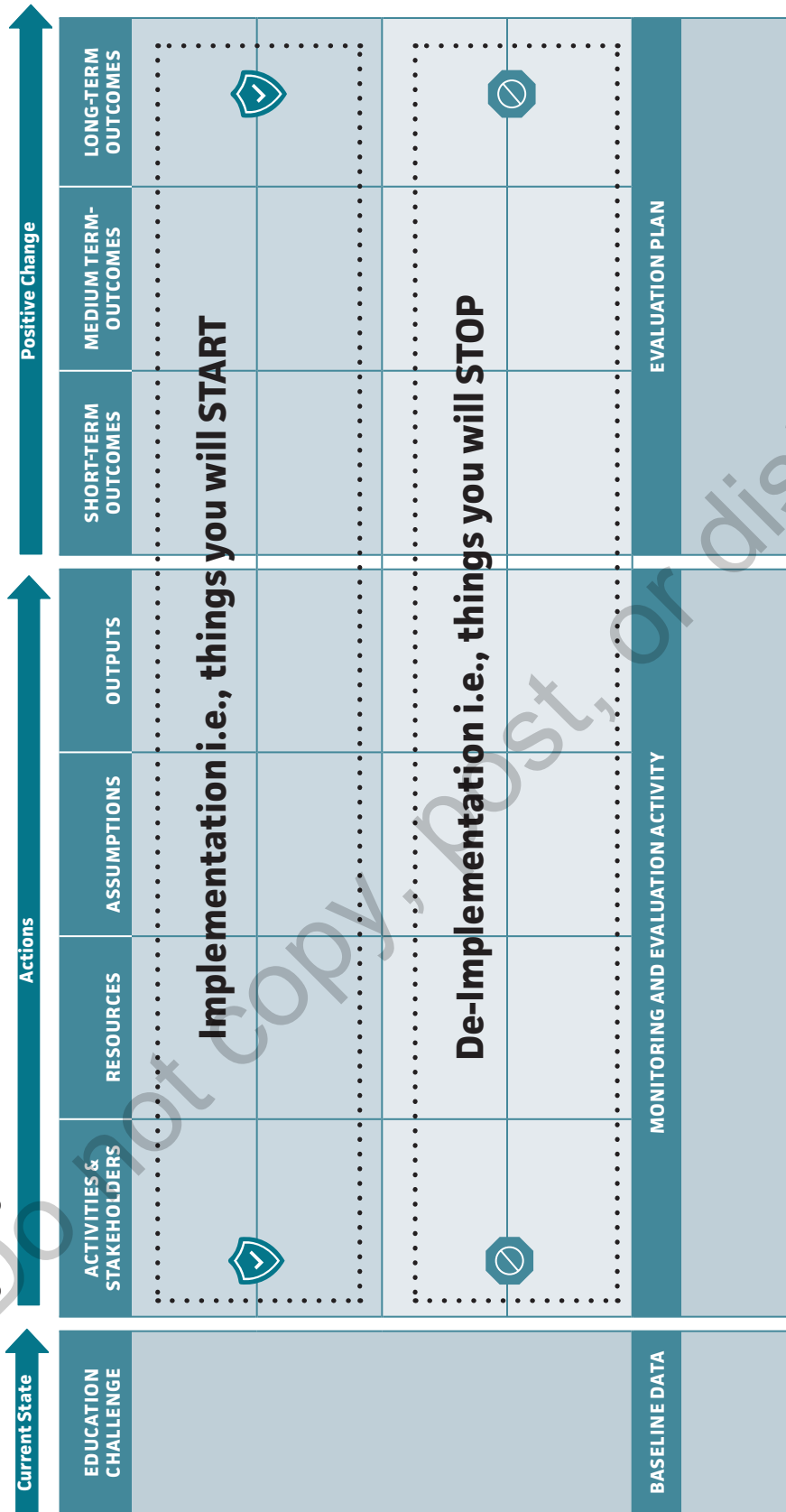
Source: Adapted from Hattie and Hamilton (2020a).

Of course, we recognize that stopping is extremely hard. There is a range of cognitive biases that seem to prime us to continue with things that really should be stopped. These are biases that make it hard for us to just say, *No! Enough is enough!* We list some in Figure 2.19 (for more details, see Hattie & Hamilton, 2020a, 2020b).

Obviously, de-implementing is just as hard as implementing. People become emotionally attached to the work they have engaged in—and to the effort and the long hours. No one wants to admit that it has all been pointless. So, you actually need a strategy for de-implementation and this also needs to confront the seven Eeyores that we unpacked in Step 2.3.

Therefore, we recommend that you divide your program logic model in half and that 50% of the rows or work packages are about implementing your new agenda and 50% are focused on the initiatives you are concurrently dismantling. This is why half the program logic model template that we already introduced was in a different color: half for starting and half for stopping! We illustrate this in Figures 2.20 and 2.21.

FIGURE 2.20 • Program Logic Model—STOPPING AND STARTING



Source: Copyright © Cognition Education. (2022). All rights reserved.

FIGURE 2.21 • Program Logic Model—Example of Stopping and Starting

Current State	Actions				Positive Change			
EDUCATION CHALLENGE	ACTIVITIES & STAKEHOLDERS	RESOURCES	ASSUMPTIONS	OUTPUTS	SHORT-TERM OUTCOMES	MEDIUM TERM-OUTCOMES	LONG-TERM OUTCOMES	
<p>Our student literacy rates are significantly below national averages.</p> <p>Disproportionately boys (68%)</p> <p>Disproportionately rural (74%)</p> <p>Disproportionately low SES (83%)</p> <p>—</p> <p>32% missed target by 5% or less</p> <p>4.1% missed it by 5-15%</p> <p>27% missed it by more than 25%</p>	<p>Response to Intervention Program that provides triaged and targeted support to priority students</p> <p>Intelligent Tutoring System that provides AI-driven personalized learning to all students</p> <p>STOP: Learning Styles Teacher Professional Development Program</p> <p>STOP: Student Co-curricular Project</p>	<ul style="list-style-type: none"> RTI Program Materials External Trainer \$XX Software Platform \$XX 90 minutes per week of access on the timetable Briefing sessions on why we are stopping and what we are starting Vice Principal time for communications sessions with students and parents 	<ol style="list-style-type: none"> Program will transfer to our context Teachers will be able to implement All students will use the platform The use will generate improved literacy outcomes Program is ineffective Time better spent on RTI A 'happy program' without evidence of impact Stakeholders will accept cancellation. 	<ol style="list-style-type: none"> Teacher training completed Platform procured Student logins distributed Student inductions Teachers have additional 2 hours per week Time transferred to RTI Students have additional 3 hours a week 1 hour saving for teachers 				
BASELINE DATA	MONITORING AND EVALUATION ACTIVITY							EVALUATION PLAN
<p>33% of our 13-year-olds are functionally illiterate</p>								

Source: Copyright © Cognition Education. (2022). All rights reserved.

2.5 ESTABLISH A MONITORING AND EVALUATION PLAN

Warning: This section is long and requires concentration. Consider taking a break before pushing on!

Do not skip this section. If you do, you are not implementing the 5D methodology properly.

When long-distance runners train for marathons, they have a *distance* goal in mind: to successfully run 26 miles and 385 yards. They usually also have a *time* goal—the current speed record, for example, is just over 2 hours. Generally, they don't just turn up on the day and hope to wing it. Professional runners work with a coach to prepare for the race. The coach uses a range of tools to gauge (i.e., *evaluate*) the runner's current performance, including a stopwatch, heart rate monitor, weighing scales, and even AI-driven video analytics to assess posture, technique, and gait. The coach and runner then use these data during training to decide whether the training strategy is working and what to do next. The decision could be to carry on as is or to change footwear, adjust stride length, eat more protein, or a host of other adjustments. Then, once a change is made, the measuring tools are used (yet) again for a bit more evaluation and a bit more iterative variation until the runner is (hopefully) able to complete the course in the desired time.

The same principle applies to the evaluative rules of deciding who has won the race. As a thought experiment, imagine that the starting gun has gone off while at the same time World Athletics (the global governing body for running sports) is still debating the rules and is still deciding what constitutes *success*. Imagine further that some committee members are arguing for the critical measure to be speed (i.e., who passes the finish line first), while others argue that performance should be measured against an agreed standard of technique (i.e., who has the best running gait). Yet others wade in and suggest that score points should be added or subtracted depending on the footwear of the athletes, their social background, or the length of their respective legs. While this is a good debate to have, it happens (and has happened) well *before* the starting gun has been fired. No competition organizer would contemplate having the debate *after* the race was in play. This is (literally) moving the goal posts.

However, in our work with schools, we see this thought experiment playing out for real. Some of the horrors include the following:

1. **Not evaluating at all.** Yes, this happens and all too frequently. In the mad rush to get an initiative out of the starting blocks, everyone forgets to define what success means, to agree on how they will measure it, or what they will do if the collected data aren't rosy.

Takeaway: Unless you systematically evaluate, you will have no idea whether you have generated *meaningful* impact or how you can grow this further.

2. **Using the wrong evaluative tools.** Running coaches tend not to include water pressure gauges in their evaluative toolkit: knowing the pressure in the stadium pipes does not help to make athletes run faster. Equally, medics no longer use mercury thermometers—if they can help it. While, yes, body temperature is a useful indicator of health, the current preference is for digital devices that are more accurate and less inclined to toxic spillage.

Takeaway: You need the right tool for the right job. Select your (evaluative) divining rods with great care and be aware of any potential side effects (like seeping mercury) or perverse incentives, particularly if they are linked to accountability systems or performance appraisals.

3. **Not implementing the agreed evaluation plan.** Here, the plan gets created and (sometimes) to a very high standard. But it gets locked in the draw. No one has the urge to measure—the fear is they won't like what they see. This is linked to a cognitive bias called the *Ostrich Effect*, which we discussed earlier: this is literally the act of burying your head in the sand to avoid looking at disconcerting data.

Takeaway: You have to implement the plan and look at the data. Get your head out of the sand!

4. **Not measuring before, during, and after.** Weight loss 101: Get on the scales and take a baseline reading. Implement your slimming strategy. Get back on the scales. Do more implementation, with variation. Get back on the scales. Repeat, repeat...

Takeaway: Unless you measure regularly and take an initial baseline value, you cannot gauge your success.

5. **Cherry-picking the data, if you don't like what you see.** This is possibly the worst sin of all: You've established a robust evaluation plan and you are regularly collecting data, of the *right* sort. But rather than using the data to enhance your program logic model and your impact, you keep doing the same old thing. Instead, you put your energy to work on mining the data looking for some random thing that has

gotten better—even if it isn’t connected to your original education challenge (e.g., “Our girls’ literacy program has had a brilliant impact in enhancing boys’ numeracy”).

Takeaway: You need to use evaluative data for evaluative purposes. The whole point is to get better. Your initial program logic model won’t be perfect; it might even be riddled with faulty assumptions. It’s far, far better for you to confront these (quickly) and to improve than to waste effort on actions that generate scant impact.

You may be wondering why we have written so much about evaluation here. You might also have flicked ahead and noticed that there are many more pages of this to come in the remainder of this chapter. Potentially, you might be confused by this, given that the Double-Back Stage (D4) is entirely focused on evaluation. But if you have fully processed the five evaluative horrors that we just unpacked, we hope you will see that the key is to confront them *now* (!!!) before you get anywhere near the Stage D3 (Deliver). To select the appropriate tools, establish a baseline, and implement your evaluation plan, you need to build that plan in the first place. And you need to do this *before* you get anywhere near delivery. If you tack evaluation on as an afterthought once your initiative is already underway, then you have a serious problem. We repeat: a serious, serious, serious problem.

Now that you understand why you need to think about evaluation at this juncture—and not delay it until later—here are some mind tools or lenses to help you with that process.

LENS 1: THE PURPOSE

Figure 2.22 compares **monitoring** and **evaluation**.

FIGURE 2.22 • Purpose: Monitoring vs. Evaluating

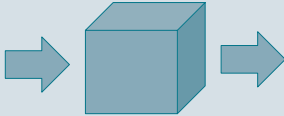
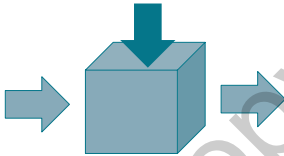
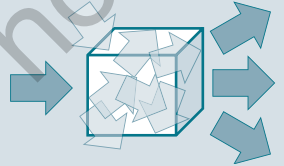
MONITORING	EVALUATION
<p>Monitoring is about checking that you have done the things you set out to do. When you think from a monitoring perspective, you are asking questions like these:</p> <ul style="list-style-type: none"> • Did we do what we set out to do? • Did we do it on time? • Did we do it within budget? <p>Monitoring is a project management activity, focused on keeping your initiative moving</p>	<p>Evaluation is about whether the things you did actually improved outcomes. When you are thinking from an evaluative perspective, you are asking questions like these:</p> <ul style="list-style-type: none"> • Did our actions improve outcomes in our target area? • Is the improvement more or less than we expected? • What have we learned that we can feed forward to further enhance our impact? <p>Evaluation is an improvement focused activity</p>

You need to plan for and do both of these things. You monitor to check that you are doing the things you said you would do. And you evaluate to check whether those things are worth continuing with. Too many initiatives measure impact solely in terms of the former: “We were successful! We achieved all our milestones and deliverables. All the training sessions were run, and all the teachers attended.” But not in terms of the latter: “Yes, we met our milestones but there has been no noticeable improvement in student literacy outcomes.”

LENS 2: EVALUATIVE APPROACHES

Figure 2.23 describes the **Black-Box**, **Gray-Box**, and **Clear-Box** evaluation approaches.

FIGURE 2.23 • Black-Box, Gray-Box, and Clear-Box Evaluative Approaches

APPROACH	DESCRIPTION
<p>Black-Box Evaluation</p>  <p>Did it work?</p>	<p>You get on the scales at the start and again at the end to measure the degree of “weight loss” or learning gain.</p> <p>This tells you whether your intervention generated impact, but the inner workings of the machine are literally a black box. You have no insight into <i>why</i> what you did was or wasn’t effective, which makes it challenging to identify areas for improvement.</p>
<p>Gray-Box Evaluation</p>  <p>Why do we think it worked?</p>	<p>In addition to collecting before, during, and after outcomes data, you also attempt to prize open the lid of the machine and peek inside.</p> <p>You conduct interviews and focus groups with stakeholders to gather their opinions or <i>perceptions</i> about why the initiative was or wasn’t successful.</p>
<p>Clear-Box Evaluation</p>  <p>What worked for whom, in what context, to what extent, through what mechanisms and how can it be improved?</p>	<p>This is more about the <i>rigor</i> with which you use the collected data. It includes</p> <ul style="list-style-type: none"> • Segmenting outcome data by category of stakeholder (e.g., gender, age, SES, teacher, etc.) • Going back over every link in your Rube Goldberg machine and your map of design features and setting levels to identify and agree to variations that have a high probability of enhancing impact.

Source: Adapted from Hamilton and Hattie (2021).

If you are working at a district level (or higher), our suggestion is that you will want to be undertaking a Clear-Box Evaluation. If you are working at a school or professional learning community level, at the very least, you want to be working at a Gray-Box level. We admit that the Black-Box level is better than nothing. And, too often, there is nothing.

LENS 3: THE LEVELS OF EVALUATION

Donald Kirkpatrick (1993) and Thomas Guskey (2000) have done some excellent work on mapping the types of evaluation questions that are worth asking and the types of tools that are worth using. We have adapted these in Figure 2.24.

FIGURE 2.24 ● Levels of Evaluation

LEVEL	FOCUS	TIME HORIZON	EVALUATIVE TOOLS
1	Monitoring Did we do the things we said we were going to do? Did we do them according to the anticipated timelines and with the anticipated level of resources?	Short term	<ul style="list-style-type: none"> • Project plan monitoring • Budget monitoring • Time tracking • Product acceptance criteria
2	Engagement Did stakeholders engage positively with the improvement initiative? Did they <i>like it</i> , and did they participate at the expected level/frequency?	Short term	<ul style="list-style-type: none"> • Satisfaction surveys • Interviews • Focus groups
3	Learning Did stakeholders (usually teachers) successfully learn new skills/ techniques/approaches that have the <i>potential</i> to enhance their collective performance?	Short term	<ul style="list-style-type: none"> • Portfolio evidence aligned (e.g., to teaching standards) • Lesson observation • Interviews and focus groups • Questionnaires
4	Change Was there noticeable change in stakeholders' performance behaviors? Did they (usually teachers) put the learnings into practice in their classrooms?	Medium term	<ul style="list-style-type: none"> • Lesson observation (e.g., with video tools) • Questionnaires • Structured interviews • Self-/collective efficacy psychometrics
5	Impact Did the (L2) engagement, (L3) learning, and (L4) change actually result in improvement in the targeted area? Did outcomes from the students improve?	Longer term	<ul style="list-style-type: none"> • Student achievement data • Student attendance data • Student voice • Structured interviews with teachers, parents, students, and leaders
6	Improvement and Sustainability How can we further enhance the impact, and what do we need to do to stop backsliding?	Continuously	All the above—for the purpose of reviewing and enhancing your program logic model

LENS 4: THE LEVEL OF ADOPTION

Early in your implementation, you are unlikely to be able to capture outcomes and impact-type evaluation data, simply because of the time lag between implementation and impact. However, you will be able to gather a great deal of engagement data. A basic way of doing this is simply to ask people whether they *like* what they are being exposed to. Many training providers use “happy sheets” to evaluate the level of satisfaction from those that they are supporting. However, liking something does not mean it’s good for you. The four of us like cake but that doesn’t make eating truckloads of it healthy. And there are many things we don’t like that are profoundly good for us and that with repeated exposure we might also eventually come to like.

Therefore, we need to get beyond measuring *like* to measuring engagement in terms of level of adoption. We present a rubric for this in Figure 2.25.

FIGURE 2.25 • Level of Adoption

LEVEL	DESCRIPTOR
Unaware	“I don’t know what it is. Never even heard of it.”
Aware	“I vaguely know what it is. But I don’t have time to engage and am not sure it’s relevant to me. I’m probably doing it already.”
Considering	“I’m reading some materials on it and thinking about applying it at some future stage.”
Priming	“I’ve done the workshops and have set aside dedicated time each week to practice implementing.”
Deliberate practice	“I’m attempting to implement but it’s requiring major cognitive effort to juggle all the balls. My head hurts.”
Effortless execution	“It used to be hard to implement but I don’t really have to think about it anymore.”
Adaptation	“I’ve started making tweaks to the protocols to better fit my context. I couldn’t really do this before now, because it was hard enough just remembering and implementing the steps.” <i>***Note the risk that this adaptation might be mutation that reduces efficacy.***</i>
Spread	“Some other teachers have joined the school who don’t know how to use the protocols. I’ve been coaching them so that they understand why it’s important and so they can do it.” <i>***Note the risk that spread might also be mutation/dilution that reduces efficacy.***</i>
On to the next thing	“I’ve been implementing the program for a few years now and have made several improvements, so it better fits our local context. Although I am still interested in it, I’ve started looking at other approaches for other more important education challenges.” <i>***Note the risk of backsliding and see our discussion in Chapter 5 (Stage D5: Double-Up).***</i>

Source: Adapted from Hall and Hord (2011) and Hall and Loucks (1977).

You can capture this progression via surveys, interviews, focus groups, and lesson observation. While this will not tell you whether the program you are implementing is generating impact, adoption is an important precursor to impact.

BUILDING A MONITORING AND EVALUATION PLAN

Now that we have introduced the four key lenses, the next key question is what you do to bring them alive within your improvement initiative. You need to approach this from two key dimensions:

- **Indicators** (what evaluative tools will you use for measuring?)
- **Targets** (what readings from these tools would we consider to be “good” progress?)

INDICATORS

You will, no doubt, have already noticed that everything within the *Building to Impact* 5D framework is extremely systematic. It’s all about searching for options in the design space, mapping those options, and then considering which are likely to be better bets for progress and improvement. The exact same logic applies to the selection of your evaluative indicators. So rather than (randomly) selecting a couple of tools that you happen to have on hand, we want you to think *deeply* about what types of tools will help you to evaluate the *specific* program logic model that you have crafted—subject, of course, to your local constraints related to optimal stopping.

In Figure 2.26, we illustrate how you could record and analyze each potential indicator or tool, within the context where the goal is to improve children’s literacy outcomes. You will see that we list the following in the figure:

- **Potential indicators.** This is your shopping list of *all* the potential measuring tools that *could* be leveraged (i.e., the educational equivalent of weighing scales, stopwatch, blood pressure monitor, etc.).
- **Linkage to education challenge.** Indicators are only useful if they indicate something that is relevant to what you are trying to improve. This column is about you spelling this out to double-check that the identified tool measures a useful thing.
- **Ease of data collection.** This is about assessing whether you need to create the tool (which requires more time and energy) or whether it is something you have on hand or even perhaps already use and already have data for.

FIGURE 2.26 • Evaluative Indicator Selection—Worked Example 

NO.	POTENTIAL INDICATOR	LINKAGE TO EDUCATION CHALLENGE	EASE OF DATA COLLECTION	VALIDITY AND RELIABILITY	PERVERSE INCENTIVES?	CONCLUSION
1	National literacy assessment	Direct and strong. Our agreed education challenge is about student achievement levels in national literacy assessment.	Easy. We already have a national assessment testing infrastructure.	Requires further investigation. External consultants have correlated our internal/national assessment grades to assessments and raised concerns about validity (i.e., we might not be measuring the right things). Our challenge is also with younger learners who have not yet taken the national assessment. We need to be able to screen earlier.	If the measure becomes a performance target for schools, there is risk that they may be incentivized to game outcomes.	
2	Student attendance	Indirect and strong. Students need to be regularly attending school in order to receive literacy instruction.	Easy. We already collect attendance data twice per day.	High validity. Direct measure of the education challenge. High reliability. Binary measure that is not open to subjective interpretation.	As above. Need to consider whether there is any incentive for any stakeholder to falsify attendance data.	
3	Lesson observation data	Indirect and causal. Our theory of change is that students are not achieving L3 literacy because they do not enjoy their lessons and are therefore neither learning nor attending school regularly. Our assumption is that as lesson observation scores increase, we should also see corresponding increases in student attendance and test scores.	Medium. We already undertake two lesson observations each year. However, we are not currently using a structured rubric or training observers to increase inter-rater reliability.	Medium validity. We are not sure whether students are not attending/not achieving because they do not enjoy their lessons. Low reliability. The current rubrics are open to different interpretations when used in classrooms.	Possible that educators will specially prep and deliver their “best lesson” for the observation (i.e., what is observed is not representative of daily classroom practice).	

- **Validity and reliability.** This asks whether the tool measures the right thing in a way that gives you *consistent and accurate* measurements.
- **Perverse incentives.** Is there any danger that there could be unanticipated consequences from using the tool (i.e., that stakeholders dance to its tune and this makes it look like things have been improved but that nothing has changed)?

The idea is that you weigh each of these considerations and then select appropriate tools that will give you short-term, medium-term, and longer-term insights into the effectiveness of your program logic model. You want a mixture that gives you **leading indicators** (i.e., quick data on engagement, learning, and change) and **lagging indicators** (i.e., slower data on change, impact, improvement, and sustainability).

Once you have agreed on your indicators, you can then set out the *what, why, when, where, and how* of your evaluation approach in an Evaluation Plan Methods Grid, as outlined in Figure 2.27.

TARGETS

Once you have selected your indicators, the next step is to baseline your take-off values and set your short-, medium-, and longer-term targets. In Figure 2.28 we outline six different methods you could adopt to set your targets, each a significant improvement on guesswork.

If you are working at the district level (or higher), you might even seek to benchmark against all six of these methods. Of course, there is still some “art” to the process of selecting your target. You also need to consider the Goldilocks principle of desirable difficulty. Your target needs to be challenging enough that it’s genuinely worth doing but not so challenging that achieving it seems nearly impossible.

FIGURE 2.27 Evaluation Plan Methods Grid

EVALUATION DOMAIN	INDICATOR	INSTRUMENT	DATA SOURCE	FREQUENCY	RESPONSIBILITY
L1: Monitoring L2: Engagement L3: Learning L4: Change L5: Impact L6: Sustainability and Scale Example: L4: Change	The category of instrument you have selected to help answer each evaluation question Example: student attendance	The specific instrument that will be utilized Example: student attendance register	What type of data you will generate Example: frequency of student absence	When you are going to do it Example: daily	Who is going to do it Example: all teachers, with monitoring and oversight from XX

FIGURE 2.28 • Six Approaches to Target Setting

NO.	METHOD	DESCRIPTION
1	Improvement on previous year (%)	Using the local benchmarked value to set incremental percentage increases over time
2	Peer average	Using the mean average performance of comparator schools that share similar features (e.g., similar size, cohort, geography, education challenge, etc.)
3	Regional average	Using the mean average performance of comparator schools in the same region (or whole-region average) as the long-term target
4	National average	As per approach 3 but based on the mean average of all institutions within a country/state
5	International average	As per approach 4 but based on the global average of data (e.g., World Bank EduStats Data; UNESCO Institute of Statistics; OECD PISA, etc.)
6	Theoretical best	Using logical reasoning to postulate what the maximum possible improvement that <i>could</i> be achieved

Note: OECD, Organization for Economic Cooperation and Development; PISA, Programme for International Student Assessment; UNESCO, United Nations Educational, Scientific and Cultural Organization.

Source: Adapted from Bryk et al. (2017).

Once you have deliberated and agreed on realistic (but stretching) targets, you can then use a column table like the one in Figure 2.29. This delineates the indicator, the instrument, the baseline value (i.e., the current status), and then successive targets over time.

LOCKING EVALUATION INTO YOUR PROGRAM LOGIC MODEL

The final step is to record your agreed evaluative actions within your program logic model. In the illustration in Figure 2.30, you can see that there are a number of “zones” within the tool that relate to this:

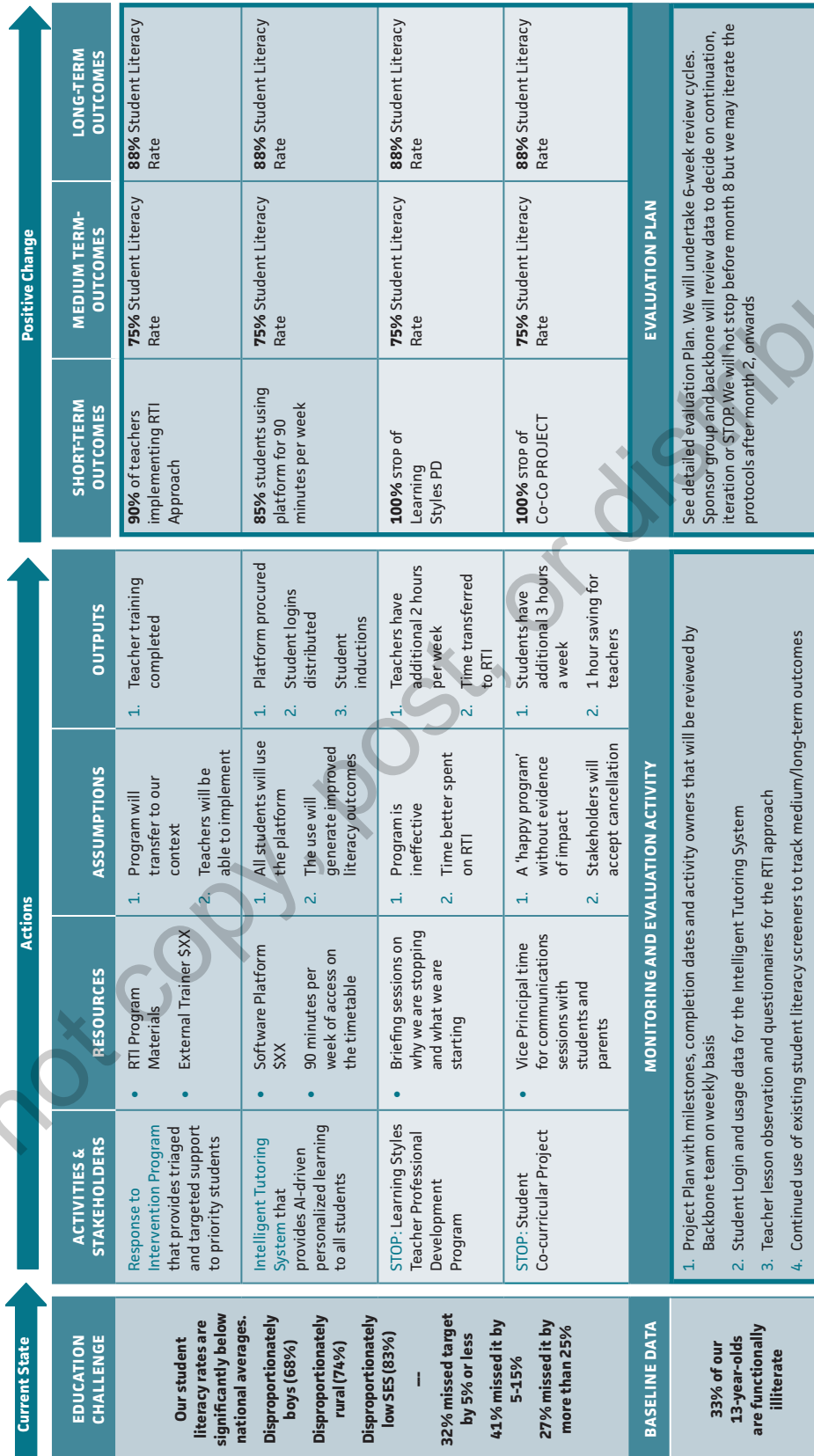
- **Baseline Data.** This is where you record or map to your current status (i.e., your starting point on the weighing scales).
- **Monitoring and Evaluation Activity.** Here you record or map to the tools you will use and frequency of use.
- **Evaluation Plan.** This is about the frequency (1) with which you will look at the evaluative data collected from the monitoring and evaluative activity, to make decisions about whether any of your actions need to be iterated, and (2) with which you will iterate (e.g., will you take an agile approach where you make micro-adjustments all the time, or will you let things play out for several months and collect lots or robust data, then explore the pros and cons of change carefully, before deciding what to do?).

FIGURE 2.29 • Target Setting—Worked Example

NO.	INDICATOR	INSTRUMENT	BASELINE VALUE	TARGET VALUE T1	TARGET VALUE T2	TARGET VALUE T3	TARGET VALUE T4
1	Overall student absenteeism	Student attendance register	12% absence	9%	7%	5%	5%
2	Target group boys (13–17) absenteeism	Student attendance register	17% absence	12%	10%	8%	8%
3	Teachers' implementation of the RTI protocols	RTI lesson observation Rubric compliance (%)	TBC	50%	60%	75%	90%
4	Student usage of intelligent tutoring system	System time log	TBC	45% students 90 minutes per week	55% students 90 minutes per week	75% students 90 minutes per week	85% students 90 minutes per week
5	Student literacy rate	State screening tool	66%	68%	73%	78%	88%

Note: TBC, to be calculated.

FIGURE 2.30 • Monitoring and Evaluation in the Program Logic Model



Source: Copyright © Cognition Education. (2022). All rights reserved.

- **Outcomes.** These are your short-, medium-, and longer-term targets. These are likely to be linked to the Lens 3 levels of the evaluation framework:
 - **Short-term targets** will more often be focused on whether you did what you said you would (i.e., monitoring whether you delivered the outputs) and whether stakeholders engaged and learned anything.
 - **Medium-term targets** are more likely to be focused on levels 2 and 3 (i.e., learning and change).
 - **Longer-term targets** take us to levels 4 and 5 (i.e., outcomes and iterative improvement).

Remember that you are setting an evaluation plan for implementation and de-implementation. Half of your program logic model will be focused on stopping activities to free up time that you can better devote to your agreed education challenge, so it is just as important that you monitor and evaluate whether you are successful with this de-implementation.

D2: Design Summary

You have now reached the end of the D2 Design processes. During this stage of your inquiry:



You will have systematically searched and agreed on high-probability interventions to START and to STOP
You will have done this by

2.1 Exploring Options in Design Space



2.2 Building Program Logic Model(s)



2.3 Stress Testing Logic Model(s)



2.4 Agreeing on What to STOP



2.5 Establishing your Monitoring and Evaluation Plan



In the next chapter, we shift our focus to D3: Deliver. The designs come to life!