**On Research and Development of Social Programs**

Flávio Cunha, Rice University[1]

August 2026

Communities face pressure to act. The problems are real, families need help now, and waiting until every uncertainty has been resolved is neither realistic nor desirable. We do not think research should become an argument for postponing action. We think the opposite.

But we also think that, if organizations are going to invest time and money in a social program, that investment should leave us knowing more than we knew before. We should learn something about the problem, the intervention, the people it serves, the organizations delivering it, or the conditions under which it can succeed.

That is what we mean by research and development for social programs. The central idea is simple: doing and learning are not competing objectives, and the goal is to design what we do so that action also produces knowledge. This requires changing when research enters the process. Evaluation cannot be something added after a program has been designed and launched. The questions we hope to answer must influence diagnosis, program design, implementation, measurement, and ultimately decisions about adaptation and scale.

Conventional evaluation begins with an intervention: an organization has a program and wants to know whether it works. Research and development begins earlier and ends later. It is the process through which we learn what intervention should exist, how it can be delivered, whether it works, and what should happen next. The distinction matters because the argument is easily mistaken for a smaller one.

What links those questions is uncertainty. Each stage begins with something we do not know, and the work of research and development is to design the action so that carrying it out reduces that uncertainty enough to improve the next decision.

Our work with Alief Independent School District over more than a decade illustrates what this process looks like in practice, including how much of what we originally believed turned out to be wrong. The stages that follow are not a recipe. Not every project will need the same sequence or the same methods, and we do not claim otherwise. What they identify is the kind of uncertainty that serious research and development has to confront at each point.

## 1. We Begin With the Problem, Not the Intervention

Research and development begins with diagnosis. A population is not a diagnosis. A neighborhood is not a diagnosis. A disparity is not a diagnosis. None of them tells us what intervention to implement.

[1]**Acknowledgments**. The work with Alief ISD described in this essay was made possible by support from several foundations over many years. The Smith Richardson Foundation supported the randomized controlled trial. The Powell Foundation, the Brown Foundation, the Samuels Family Foundation, and the Episcopal Health Foundation supported implementation of the program and the work required to build Alief's capacity to deliver it. I am grateful for their support and for the patience that made this work possible. The views and interpretations expressed in this essay are my own.

Suppose we observe that a particular group of young adults is much more likely to be disconnected from education and employment. That is an important fact, but it does not tell us what to do. We first need to know how the disconnection happened. Did an event push people who had been connected out of school or work? Were they already disconnected before it? Were they weakly attached and then pushed into persistent disconnection by one additional constraint? Are several different pathways being combined into a single statistic? Each possibility implies a different intervention.

Our work in Alief began this way. Since 2015, our team had been working with the district's Department of Family and Community Engagement, and our original project evaluated JumpStart, a parenting program serving children between 36 and 47 months. That work produced a finding we had not been looking for. Using the Bracken School Readiness Assessment, we found that 48 percent of children entering JumpStart, at a median age of roughly 42 months, were classified as delayed or very delayed. That is about three times the rate the instrument's norming studies would predict. The children were arriving at the district's parenting program already behind.

We did not immediately search for another program. We measured a possible mechanism. During the JumpStart application periods in Fall 2016 and Fall 2017, we recorded children's home language environments using the LENA system, which measures adult words, conversational turns, and screen exposure throughout the day and reports results against three bands constructed so that a third of children should fall into each. Nearly 70 percent of children fell into the lowest band for conversational turns, while only 7 percent were in the highest. That diagnosis changed the problem.

The issue was not simply low school readiness at age three. A specific deficit in the home language environment was already present before children reached the district's existing program, which implied that any response had to begin earlier and had to change parent and child verbal interaction.

Our first principle of research and development is therefore: do not begin by asking which program we should implement. We begin by asking what is producing the outcome we want to change.

Diagnosis rarely settles that question on its own. What it can do, and what it did here, is narrow the plausible mechanisms enough to generate testable hypotheses about what an intervention would have to accomplish. The measurements above establish that the deficit exists and where it sits. They do not establish why the language environment is impoverished in these homes. The next section begins from that ignorance rather than pretending it away.

**2. We Turn the Diagnosis Into Hypotheses**

Once the mechanism is clearer, program development becomes a process of hypothesis formation. The Alief diagnosis implied several requirements. Parents needed to understand why early language interaction matters. They needed a practical way to change their behavior without reorganizing the lives of working families. They needed objective feedback about what was happening in their own homes rather than generic advice. We believed group delivery could help parents learn from one another and form relationships. We wanted the program delivered in schools, both to control cost and to create a positive connection between families and the district.

We wanted it delivered by people already employed by Alief so that implementation could continue after the researchers left. And it had to work in both English and Spanish.

LENA Start fit those requirements, but each requirement contained a hypothesis. Would information change parental beliefs? Would personalized feedback change behavior? Would parents respond differently when they could see objective information about their own language environment? Could existing district staff deliver the intervention effectively? Would group delivery matter? Program development becomes much more useful when these assumptions are made explicit. Some can be tested immediately. Others will be tested through implementation. Still others require formal experiments.

The point is not that every design choice requires its own randomized trial. The point is that a program is a collection of hypotheses about why behavior will change, and research and development makes those hypotheses visible.

Those hypotheses come in two kinds, and separating them is worth doing explicitly. Some are about families: that information changes beliefs, that personalized feedback changes behavior, that a group setting contributes something a one-to-one session would not. Others are about the organization: that Family Liaisons can acquire the necessary skills, that supervisors can recognize fidelity when they see it, that the program can be delivered inside existing district structures without new staff. We had a model of family behavior and a model of organizational behavior, and both were exposed to evidence. The organizational model failed first.

**3. We Treat Implementation as Research and Development**

This was the stage we understood least when we began. Together with the district, we drafted a plan in 2019 that assumed that the LENA Foundation team could train Alief's Family Liaisons in one session, provide a handbook, and then support them remotely while district leadership monitored fidelity. That plan was wrong.

The Family Liaisons were experienced and trusted by families, but delivering this particular program required skills they had never been asked to develop. They had to lead sessions aligned with a curriculum. They had to model sensitive caregiving rather than merely explain it. They had to coach parents through practice without taking over. They had to interpret each family's weekly language environment report so that the numbers became meaningful and actionable.

A handbook could describe these behaviors. It could not produce them. And without a measurement system, neither the district nor the researchers could know whether the program was being delivered as intended.

Correcting that problem took three years. When COVID closed the schools and cohorts could not run, we used the interruption to build what the implementation system lacked. We developed an implementation fidelity scale. We created training materials for the program as it would actually be delivered in Alief, in both languages. We trained the Family Liaisons and their supervisors. And once implementation began, we observed every session, scored it, and provided feedback so that the team continued learning while delivering the program.

The first cohort did not enroll until Fall 2022, and no families were served during the three years between the original implementation plan and that cohort. Those years were not a delay between

research stages. They were research and development, and what we learned in them was what human capital the intervention required, how to measure it, how to build it, and how to know whether it was present.

Our original plan treated fidelity as something that could be instructed and then monitored: write it down, explain it once, and check on it periodically. What the three years showed is that fidelity is produced. It has inputs, including training, observation, feedback, supervision, and accumulated practice, and it has a technology that can be described, measured, and improved like any other. Treating it as a matter of compliance rather than production is what made the original plan fail.

This leads to a second principle: implementation is not the mechanical step between program design and evaluation, but an object of learning in its own right. A program that cannot be reliably delivered by ordinary staff under ordinary conditions is not yet ready for evaluation or scale.

**4. We Build Learning Into the Action**

Communities often want quick wins, and there is nothing wrong with that. The mistake is treating a quick win and an opportunity to learn as alternatives.

Suppose an organization is already going to launch a program. The relevant question for a research partnership is not necessarily whether implementation should wait. It is: what would we need to do now so that this implementation teaches us something useful later? The choice is not between a randomized trial and no research. It is a portfolio question, because different designs purchase different kinds of information, in different amounts, at different costs, and the right one depends on what you most need to know and on what the organization can bear.

**Random assignment** buys the most credible causal counterfactual available. It costs coordination, and it requires an organization willing to let chance determine the order in which people are served.

**Phased implementation** preserves much of that value while reducing operational resistance, because it converts the question from who receives the program into who receives it first. It is available whenever a program will run more than once.

**A credible comparison group**, drawn from eligible families who could not be accommodated or from a comparable site, buys a weaker answer at lower cost. It is often the only option once a program has already begun.

**Deliberate variation in components** answers a different question altogether, which is which parts of a program are doing the work. It is the least used of these designs and often the most valuable, because most programs are bundles and most evaluations test only the bundle.

**Measuring the right outcomes before the program begins** identifies nothing on its own. It preserves the option of learning later, and it is both the cheapest item on this list and the most frequently skipped.

The specific design depends on the question, but one principle does not change: the learning design has to begin before implementation. Once a program has been launched, many of the most valuable opportunities disappear. Baseline measures may no longer exist. A comparison group may no longer be possible. Important implementation differences may never have been recorded. The program may change during delivery without anyone documenting how.

Five years later, we may know exactly how many people were served and still be unable to answer whether serving them changed anything. That is a lost opportunity, and the problem is not the quick win. The problem is spending scarce resources on the quick win and emerging from it no better able to decide what should be done next.

**The objection that matters**

The objection we hear most often is that a comparison group means denying services to families who need them. It is a serious objection, and it deserves a serious answer rather than a lecture about statistical power.

In many of the settings in which we work, capacity is already the binding constraint. A program that can serve sixty families a semester and has two hundred eligible applicants is denying services to most of them no matter what the researchers do. It is simply denying them by application date, or by who heard about the program first, or by who could attend the sessions at the time they happened to be scheduled. Randomization does not create the shortage. It changes how the shortage is allocated, and it makes the allocation defensible to the families who are not selected.

The design also matters. In Alief the program ran in six cohorts over three years rather than once, and assignment happened within each cohort and language group against a number of places fixed by what the Family and Community Engagement Centers could actually staff. A program that recurs allocates timing as well as access, and that changes what randomization asks an organization to accept.

This is the general point. The ethical difficulty is usually not created by the research design. It is created by the fact that there are more families than places, and a learning design is one of the few ways to make that allocation both fair and informative.

### 5. We Ask a Causal Question, Not a Reporting Question

Between Fall 2022 and Spring 2025, 293 families enrolled in LENA Start across six cohorts. Families were randomly assigned within cohort and language group to the program or to a comparison group, and approximately 70 percent completed the full ten weeks.

Because we had designed the evaluation before implementation, we could answer a different question from the one answered by ordinary program monitoring. We did not simply ask how participating families changed. We asked how they changed relative to comparable families who did not receive the program.

Conversation between parents and children increased by 0.30 standard deviations, and most of the increase came from exchanges initiated by the children rather than from parents simply talking more. Parents' knowledge of early child development increased by approximately 0.50 standard deviations, and their belief that their own engagement mattered increased by approximately 0.70. The home environment changed on the dimensions the program directly targeted, including reading, talking, and interaction in daily activities. It did not change on the dimensions the program did not target, such as the stock of toys and learning materials. Children's language development improved by approximately 0.34 standard deviations over ten weeks.

Standard deviations are not intuitive, so a word about magnitude. The conventional benchmarks in social science treat 0.2 as small, 0.5 as medium, and 0.8 as large. By that standard the effects on

conversational turns and on children's language are small to medium, the effect on parental knowledge is medium, and the effect on the combined belief measure is large. Those conventions were not derived from education research, however, and against the distribution of effects that education field trials actually produce they are demanding. Measured against that distribution, every effect reported here would count as large. We give both yardsticks because the choice between them does real work, and a reader should be able to watch it being made rather than inherit it. All of these describe the average family in the program and not any particular one.

These findings required more than measurement. If we had observed that participating children improved during the ten weeks, we still would not have known how much of that improvement would have happened anyway because the comparison children also improved over the same period, though by considerably less. Measurement tells us what happened. A credible comparison tells us what would likely have happened without the program. Learning requires both.

It also requires knowing what intervention was actually delivered. A randomized evaluation of a program that staff were not prepared to implement would have answered a very different question. This is why implementation and evaluation cannot be separated cleanly: the credibility of the causal evidence depends on the credibility of the implementation.

The distinction between the two kinds of hypotheses matters here as well, because a program can fail in two quite different ways. If families receive the intervention as intended and outcomes do not move, the behavioral model is in question. If staff cannot reliably produce the intervention, what has been learned concerns implementation technology, and the behavioral model has not really been tested. This is why a null result from a poorly implemented program is so difficult to interpret: it is consistent with a theory that is wrong and with a theory that was never given a chance. In Alief the implementation model failed three years before the trial began, which is the only reason the trial answered a question about families rather than a question about training.

**6. What We Could Not Learn**

A document about learning should say what it failed to learn. We wanted to know who benefited most, so we looked, using methods designed to detect differences in effects across families, and the confidence intervals overlapped across every subgroup we examined.

That is not a finding that the program works equally well for everyone. It is a finding that a study of this size cannot resolve the question. Detecting differences between groups requires far more statistical power than detecting an average effect, and very few single-site evaluations have it. Anyone who reports confident subgroup findings from a few hundred families should be read with care.

We also cannot say which components did the work. The program is a bundle of information, coaching, shared reading, and weekly feedback on a family's own recordings, and our design tested the bundle. We have a theory about which parts matter and why, but this trial does not adjudicate it. Answering that question would require deliberate variation across cohorts, which is exactly the design we listed above and did not use.

We cannot yet say whether the effects persist. Ten weeks is a short horizon, and the outcome that matters is not conversational turns but what a child can do in school years later. And we cannot

say how much of the result depends on Alief specifically: on this district's staff, this population, these two centers, and a research team present throughout.

We cannot say what the program costs in any way that would support a comparison with alternatives. We cannot separate the cost of delivering ten sessions from the cost of building the capacity to deliver them, and the second is the larger and less familiar figure. What the project does establish qualitatively is that the relevant cost of an intervention is broader than the cost of running it. Training, observation, supervision, fidelity measurement, and accumulated practice were all inputs into producing the program that was evaluated. Any costing that omits them describes a different program from the one that produced these results.

Naming these limits is not modesty, and they are not defects in the study. A limitation is usually a question that a different design and another investment could answer. Each of the four above is a research design waiting to be built, and stating them plainly is how the next investment gets scoped.

**7. We Let Learning Change What Happens Next**

Research and development is incomplete if evidence is produced but does not change subsequent decisions. The relevant comparison is therefore not only between what happened to a treatment group and a comparison group. There is another comparison that matters: what did we expect to happen, and what actually happened? The difference between expectation and realization is where learning begins.

What we are describing is learning from prediction errors. A model makes a prediction, reality confirms or contradicts it, and the discrepancy is information. Causal identification tells us whether an intervention changed an outcome. Prediction errors tell us whether our account of why things happen needs revision. Research and development needs both.

This requires a discipline that sounds trivial and is not. Expectations have to be written down before anyone can see the results. Once results are in hand it is nearly impossible to reconstruct honestly what you would have predicted, because explanations arrive too easily and a finding that would have surprised you in advance feels inevitable in retrospect. If expectations were never recorded, the comparison cannot be made, and the most informative part of the exercise is lost.

In our evaluation, the model we had built specified four predictions before the endline data were analyzed: three effects it required to appear, and one it required to be absent. The last one matters most. A theory that predicts only successes cannot be embarrassed by evidence, and committing in advance to a set of outcomes that should not move is what makes the exercise a test rather than a description. All four predictions held.

The same discipline applies to implementation, where it is rarer still. Our 2019 plan did record an expectation, although not in the form of a hypothesis: it stated that a handbook and a single training session by the LENA Foundation team would ensure high levels of adherence to the program. Because that expectation was written down, we can point to exactly where our model of implementation was wrong. Most projects cannot, because nobody wrote down what was supposed to happen in enough detail to be contradicted.

In Alief, the largest discrepancy between expectation and realization appeared before the evaluation ever started. We expected existing district staff to be able to deliver the program after conventional training, and they could not. The appropriate response was not to declare the model unsuccessful. It was to revise the implementation technology.

A prediction error is only information if someone can act on it. Research and development therefore requires institutions capable of adaptation, because evidence creates value only when someone has both the authority and the willingness to change what happens next. In many organizations an assumption can be shown to be wrong without anything changing as a result. Measurement occurs and learning does not. Alief is the counterexample: the implementation prediction failed, the partners accepted that it had failed, and the implementation model changed.

That is a demanding requirement, and it is worth being explicit about what it asked of our partner. Alief ISD allowed families to be randomly assigned. It allowed outside observers into every session and accepted scores on how well its own staff delivered the program. It continued through three years in which the project served no one. And when the implementation model proved wrong, it agreed to change how its staff were trained rather than defend the original plan. Not every organization can do those things, and the ones that can are not always the ones with the most resources. In our experience a research partnership is limited less by the sophistication of its methods than by whether the delivering organization can tolerate being measured.

There is a parallel condition on the funding side, and it is stated less often. Implementing organizations have to tolerate measurement and the possibility of being wrong. Funders have to tolerate investments whose immediate output is knowledge rather than families served, and whose eventual finding may be that an attractive idea does not work. Neither condition is technical. Both are dispositions, and a partnership missing either one will produce activity without learning.

One implication follows that we did not appreciate at the start: partner selection is part of research design. A scientifically interesting intervention is not automatically a good research and development opportunity. If an organization cannot accommodate measurement, variation in delivery, randomization where it is appropriate, observation of its own staff, or a change of course when evidence contradicts expectations, then certain questions cannot be answered there, however good the design looks on paper. The quality of the partner is not an operational detail settled after the research question has been chosen. It determines which research questions can credibly be asked.

**8. We Treat Scaling as Another Research Question**

Scale is often treated as the final reward for producing evidence of effectiveness. We think that is too simple. The Alief program has not yet been scaled; what exists today is a district that is ready to scale it.

Alief continues to run LENA Start without us and without changing the program model. The district retains the fidelity scale, the training materials in both languages, and staff and supervisors who have been trained, observed, and coached over multiple cohorts.

That is important because scaling changes the environment in which an intervention operates. New staff may have different skills. New families may face different constraints. New organizations

may have different cultures. Costs may change. Implementation quality may fall. The assumptions that were true in one location may not be true in another.

Scale therefore creates a new set of hypotheses. The question is not simply whether the program can be made larger. It is which features of the original environment were necessary for the program to work, and whether those features survive when the program moves. That question should be answered with the same discipline used in the original development process.

**9. What This Means for Social Investment**

The purpose of research and development is not simply to produce programs that work. It is to produce knowledge that makes the next decision better.

Every social investment creates an opportunity to reduce uncertainty. We can learn about the problem. We can learn about the mechanism. We can learn about implementation. We can learn about impact. We can learn about heterogeneity, who benefits and who does not. We can learn about cost, though only if the research is designed to measure it. We can learn about whether the result survives at scale. But those opportunities exist only when learning is built into the work.

This has implications that differ by actor. Researchers have to formulate their questions before they can see the answers, keep measurement distinct from causal inference, and be willing to discover that their own theories are wrong. Service providers have to permit measurement of their own implementation and retain the capacity to change it. Funders have to finance infrastructure and learning that may produce little immediately visible activity. Policymakers have to resist treating positive evidence from one setting as sufficient warrant for unrestricted scale.

Put negatively: a grant should not be judged only by how many people it serves. A data system should not be judged only by how many records it connects. An evaluation should not be commissioned only after the implementation decisions have already been made. And a successful program should not automatically be scaled before we understand what made it successful.

One further distinction follows from all of this. A research and development system has to assign value to informative failure. That does not make failure desirable, and an intervention that fails while teaching nothing is simply a failure. But an intervention that fails while resolving an important uncertainty improves the next decision, and a program can generate impressive activity while leaving decision makers knowing no more than they knew before about what should happen next. Judged by output, the second looks better. Judged by what it contributes to the next choice, the first often is.

Research and development requires patience because some of its most consequential work produces very little that is immediately visible. In Alief, the stage that consumed the most time, served no families, and was hardest to fund was the stage that determined whether everything that followed would mean anything. That experience changed how we think about innovation.

A program is not an innovation when it is designed. It becomes one only after it has been delivered repeatedly by ordinary staff, to families for whom participation carries a real cost, and has held up. And even then, the work is not finished.

The final objective is not the program itself. It is the knowledge accumulated through building it, testing it, changing it, and learning what should come next. We do not believe communities must

choose between acting and learning. The central idea of research and development is to design the action so that we learn from it.