When is Evidence Actionable? Assessing Whether a Program is Ready to Scale*

John PA Ioannidis[1], Zacharias Maniadis[2] and Fabio Tufano[3]

[1] Department of Medicine, Stanford University, Stanford Prevention Research Center; Meta-Research Innovation Center at Stanford (METRICS)

[2] School of Economics, Social and Political Sciences, University of Southampton

[3] School of Economics, University of Nottingham, University Park

* This is a preprint version of "Chapter 7 – When is Evidence Actionable? Assessing Whether a Program is Ready to Scale" first published in 2021 by Routledge in "*The Scale-Up Effect in Early Childhood and Public Policy Why Interventions Lose Impact at Scale and What We Can Do About It*" edited by John A. List, Dana Suskind and Lauren H. Supplee. ISBN: 978-0-367-36044-3 (hbk) • ISBN: 978-0-367-42247-9 (pbk) • ISBN: 978-0-367-82297-2 (ebk)

Abstract

The effects of small-scale interventions often prove much lower than expected when they are implemented at a large scale. We illustrate the problem and its potential causes using a number of examples from the early childhood intervention literature. We delve deeper by introducing a basic logical framework allowing us to discuss the key factors in assessing whether a program is ready to scale, particularly with regards to uncertainty in the potential outcomes of small-scale interventions. We conclude putting forward a set of concrete recommendations on how to bridge *the science of using science* and real-life policy.

Introduction

How can philanthropists and policy makers decide which studies from the vast scientific literature of promising interventions are appropriate to be scaled up? Not all promising interventions are truly effective and, even if they are effective, very often the effects of interventions prove much lower than expected when they are implemented at a large scale. The latter phenomenon is called *the scale-up effect*. Furthermore, some harms of the intervention may not have been detected in the early evidence but may emerge later. Fortunately, there is a field of science that can help us answer the question, "when is evidence scalable?" This chapter will employ tools from this field—called *meta-research*—and its basic insights, provide examples, and discuss how philanthropists and policy makers (who we shall refer to as decision makers here) can utilize this knowledge to help guide their decisions.

The structure of this chapter will be as follows: we will first provide a number of examples from the early childhood intervention literature to illustrate the scale-up problem and its potential causes. We will then discuss a basic logical framework that can help guide our thinking in assessing these causes. One particularly important point will be the treatment of uncertainty in potential outcomes and how to appropriately address it. We will then conclude with a set of concrete recommendations for decision makers on how to connect the *science of using science* with real-life policy.

On Empirical Inference

When evaluating educational interventions, well-designed and conducted randomized controlled trials (RCTs) are considered to provide rigorous empirical evidence to support specific interventions. However, RCTs may fall short of providing actionable evidence—that is, evidence that supports the scaling of the intervention—which could happen for several reasons. For instance, (a) it could be that the RCT's results appear sufficiently robust given specific features of a study

population or context, but the beneficial effects of the intervention decrease or disappear in other populations or new contexts. In a different case, (b) the RCT may involve a number of interventions that are bundled together, with only some of them generating robust effects. In the latter case, the challenge is to distinguish between cases where only a part of the whole intervention bundle is genuinely scalable, from cases where, by the pure mechanics of statistical inference, some effects are found statistically significant by chance. Alternatively, (c) the findings in the original RCT may not be robust, fail to be (fully) reproducible in follow-up studies and, therefore, constitute simply a case of inaccurate statistical inference due to biases and/or chance that may reduce or eliminate any ambition of scalability.

To study these three different cases, we will first discuss three early childhood interventions exemplifying the scale-up problem and some of its potential causes. These interventions are Head Start, a group of home visiting programs evaluated in the Mother and Infant Home Visiting Program Evaluation (MIHOPE), and Family Connects. For the sake of brevity and simplicity, when delving below into these three interventions, we will focus only on dimensions or levels of aggregation that illustrate the aforementioned cases.

Head Start: The Scale-Up Problem, Outcome Heterogeneity, and Challenges to Inference

Our first discussion will be very thorough, since it will illustrate the scale-up problem in the domain of educational interventions and introduce many key aspects of this policy domain that will help us assess some drivers of the problem—particularly issue (a) above, pertaining to study population or context. Head Start is an early childhood program offering preschool for low-income children in the US. Established in 1964 as a summer-only program, it is now a national, primarily federally funded, year-long program. Head Start has spurred a heated debate concerning its effects, a large non-experimental literature, and RCTs for impact evaluations (for recent ones, see Kline & Walters, 2016; Johnson & Jackson, 2019). In what follows, we focus on an RCT called Head Start

CARES and also briefly discuss the Head Start Impact Study. We do so in order to discuss how intervention effects may decrease or disappear, and how study contexts or populations can be considerably heterogeneous.

The Head Start CARES trial (Morris et al., 2014) evaluated three different classroom-based approaches aimed at enhancing children's socio-emotional development. These three approaches were selected based on small-scale evidence of their efficacy and were then scaled up at the national level in 104 Head Start centers (with a total of 307 classrooms) across the US. The three approaches were the Incredible Years Teacher Training Program, addressing teachers' classroom management and children's behavior; the Preschool PATHS program, delivering structured lessons to promote children's appropriate peer interactions and emotional knowledge; and Tools of the Mind—Play, using make-believe play to enhance children's learning. The Head Start CARES project provided a considerable range of support to teachers, aimed at helping them improve their practices and implement the targeted approaches.

Among the different types of outcomes considered, only Preschool PATHS had consistent positive effects in the theory-supported hypotheses. Yet, even these effects tended to be relatively small. The Tools of the Mind—Play approach exhibited zero or very low treatment effects across the range of assessed outcomes. Incredible Years had very low effects for two out of three of its primary targeted outcomes pertaining to children's functions and behaviors. Moreover, none of the three approaches improved children's behavior regulation and executive function skills, despite evidence from previous small-scale trials indicating their potential to improve these outcomes.

In order to delve deeper into this apparent instance of the scale-up problem, let us focus on Incredible Years. A key early study evaluated a small-scale trial of Incredible Years (called Foundations of Learning) that was implemented in Newark, New Jersey and Chicago, Illinois in the school years 2007-2008 and 2008-2009, respectively, with 71 preschool centers and 91 participating

classrooms (Morris et al., 2013). This study found that "Problem behaviors, such as conflict among children, were generally reduced in the intervention classrooms. In addition, [Foundations of Learning] usually improved children's approaches to learning and to executive function skills." (Morris et al., p. ES-8). How can we account for the fact that in the larger study (Head Start CARES) that was based on this earlier study, there is no evidence of the intervention improving children's executive function and behavior problems?

The context of the small-scale study differs substantially from the large-scale one. There were differences in the study target (the first study explicitly targeted children who display challenging behavior), in preschool environments, in the type of intervention, in its dosage, and in its quality. Dosage (amount of service and training) and quality (as rated by teacher) were high in Foundations of Learning. At the same time, fidelity (i.e., how closely the set of implemented procedures matched the program model) was generally satisfying in Head Start CARES. In the absence of additional empirical evidence, we can only speculate that these differences in study contexts and populations may explain (at least partially) the different results between Foundations of Learning and Incredible Years.

The Incredible Years (part of the Head Start CARES demonstration) and its predecessor, Foundations of Learning, thus introduce us to a key issue that informs the scale-up problem in the domain of early childhood interventions: heterogeneity of programs, implementations, and target population/context. To exemplify this further, let us briefly consider also the Head Start Impact Study (Bloom & Weiland, 2015), an RCT implemented with a nationally representative sample assessing the impact of a single Head Start year (earlier year) for 4-year-old (3-year-old) children (see U.S. Department of Health and Human Services, Final Report, 2010). Bloom and Weiland (2015) show that there are important variations in program effects across subgroups, sites, and

individual programs as well as differences in treatment and control groups across National Head Start Impact Study centers.

The discussion of the challenges above teaches us an additional lesson: to be cautious also in interpreting the evidence from multifaceted scaled-up studies like Head Start CARES. In their conclusion of the executive summary, Morris et al. (2014) argue that "The findings suggest, perhaps most important, that scaled-up, evidence-based models can produce impacts in the social-emotional domain during the preschool year of nearly the same magnitude as those from smaller-scale, more controlled studies when the models are supported by strong, comprehensive professional development." However, given the number of programs and outcomes, some statistically significant results will naturally occur due to chance, and one may argue that statements like this do not sufficiently account for this important fact. In our main analysis (see the section, Empirical Inference and Post-Study Probabilities: A Bayesian Framework**,** below), we shall show further how tools from meta-research can illuminate some of these challenges to statistical inference.

The Mother and Infant Home Visiting Program Evaluation—Scalability and Its Many Dimensions

Departing from the educational domain, we shall now discuss a program that clearly exemplifies the scale-up problem in the domain of home visiting, with a special focus on the instances in which only a part of the whole program is genuinely scalable (see item (*b*) above). Over the years, a number of local programs have shown that home visiting improves health, education and psychological outcomes for families. The Mother and Infant Home Visiting Program Evaluation's (MIHOPE) overarching objective was to reaffirm at a national scale that families and children indeed benefit from participating in home visiting programs (Michalopoulos et al., 2019). MIHOPE is an umbrella program including four different home visiting models: Healthy Families America, focused on preventing child maltreatment; Nurse-Family Partnership, focused on improving mother and child

health; and Early Head Start Home-Based Option and Parents as Teachers, both primarily focused on children's school readiness.

MIHOPE examined 12 confirmatory outcomes of interest based on pre-existing evidence. It found that most estimated effects were in the direction signifying improvement for families, but smaller than in previous small-scale studies. The overall pattern of results (nine out of 12 results in the positive direction, including four confidence intervals that did not include zero) indicates that home visiting is generally beneficial but the effects are modest and smaller than the effects found in past studies. For instance, the study found that home visiting improves maternal health and reduces household aggression. No significant differences in the estimated effects were identified across subgroups of families or across program features and services received, indicating that the effects were robust. Once more, we see with this study that when evaluated a large scale, interventions have smaller benefits than indicated by the existing evidence. Is it the case that the initial effects were untrue, or does the context of the new large study drive the disparity? Or, maybe only part of the effects are true?

In light of MIHOPE's empirical evidence, it is possible to assert that, broadly speaking, home visiting appears to benefit families and children. However, if we were to stop there, we would miss the point. MIHOPE has also demonstrated that it is not the case that *any* home visiting intervention may produce the effects hoped for at design stage. In fact, each program features a number of relevant dimensions (e.g., primary outcome; target population; workforce training), which may contribute to the substantial variation of effects across different programs. Therefore, in an attempt to assess whether evidence is actionable, it is important to perform an in-depth evaluation of each program and its dimensions and exert careful judgement in understanding which of those dimensions need to be part of the program model at scale. MIHOPE can be interpreted as an example of an RCT

with a number of programs or interventions in which only some interventions appear to generate robust effects paving the way for possible scalability.

Family Connects: When Primary Effects Do Not (Fully) Replicate

Our third example will illustrate a case where even an original intervention with positive effects (according to its evaluation) may fail to be (fully) reproducible and, therefore, may turn out to be potentially ineffective (see item (c) above). Family Connects offers support to families with a newborn child through a home visiting program consisting of one to three nurse home visits during the child's early years. The program screens the family and child's needs and redirects them to the available community services/resources. Family Connects originated in North Carolina, where it was known as Durham Connects (see Dodge et al., 2013), and has since been implemented in several locations in the US. The program is a low-cost intervention that aims to enhance both the parent's and the child's wellbeing. Family Connects has been the focus of two RCT evaluations: the original study and a replication.

The original study reports beneficial effects (compared to *standard of care*, viewed as the control) for families and children exposed to the intervention. An important finding of the original evaluation (which had a sample size of 531) is that the intervention substantially reduces child emergency care utilization in the first 12 months of life (Dodge et al., 2013). By contrast, the replication study (Goodman & Dodge, 2016), with a sample size of 967, does not corroborate this finding. It should be emphasized that the replication considered the same program and the same area in a different time period. Hence, based on these two pieces of evidence, there is no solid ground to believe that Family Connects has definitively beneficial effects on child emergency care utilization. Unless stronger evidence comes to bear, this appears to be a case of invalid statistical inference, maybe due to sheer chance. Therefore, for child emergency care utilization, the lack of reproducibility poses a clear challenge to the scalability of this intervention.

Empirical Inference and Post-Study Probabilities: A Bayesian Framework

The previous section provided a series of examples illustrating the scale-up problem. In summary, there are serious challenges to the rigorous assessment of a candidate intervention to be scaled up. These challenges stem from features of the study population and context, the multidimensionality of an intervention, and the robustness of the reported effects (very often, early evidence is mixed). In the face of the scale-up problem and these challenges, how can we proceed to decide which interventions are worth pursuing on a larger scale? Fortunately, there is a systematic way of making such a decision, which we shall present now. It is based on a rigorous consideration of aspects of the existing literature that affect our confidence in the intervention's reported effects being true. Moreover, it considers the degree to which the implementation of existing studies in the literature departs from the manner in which an actual scaled-up intervention would be implemented in the particular domain under consideration.

As discussed in Chapter 6 of this volume, Al-Ubaydli et al. argue that before an intervention is implemented to scale, we need to ensure that we have proper confidence from the existing evidence that the underlying mechanism behind the intervention works and it is not a *false positive* (i.e., a positive result for a test that should have been a negative result). We shall argue among similar lines and illustrate how existing methodologies and frameworks can be used to systematically and carefully assess whether such confidence exists.

Let us first define a *theory* as a hypothesized association between different phenomena. An example of a theory is the following: let us hypothesize that a given training intervention helps teachers to foster a positive change in the classroom climate (let us call this intervention "Positive Change"), which is in turn associated with less problematic behaviors by pupils. Any new piece of evidence allows us to update our beliefs and the confidence we have in the theory. This insight suggests using a very simple formula for updating our beliefs about the world, the famous *Bayes*

formula. We shall now proceed by explaining how to use this formula to decide which interventions should be scaled-up. For this, we shall need to set the stage and define a few terms, drawing from Ioannidis (2005) and Maniadis et al. (2014).

Let us focus first on what many view as the most reliable form of evidence that we can obtain in science, namely experimental evidence. Generally speaking, an experiment sets up a controlled empirical test of a theory/association and declares success or failure in providing support for the association. Since real data are always a bit noisy, the criterion provides wrong declarations some of the time. There are two types of errors: providing support for the association although the theory is false in reality (*Type-1 error*) and failing to provide support when an association is in fact true (*Type-2 error*).

The Post-Study Probability

We start from a level of confidence in a given theory. This determines the prior probability, which we assign to the theory being true. In the case of our illustrative hypothetical association, on the basis of preexisting knowledge about educational interventions of a similar form, say we expect that Positive Change has a 50% chance of working. We then consider the new evidence (from a new experimental study) and we update this probability. The Bayes formula allows us to infer appropriately from the data how to update our level of confidence in the theory. The most important element here is the posterior probability we obtain after the evidence (*post-study probability*). This probability encapsulates the belief we have in the association/theory, all things considered. The higher the post-study probability is, the more confident we are that the hypothesized association is true. To get things a bit more concrete, let us go back to some of the examples discussed in the previous section. After the Foundations of Learning study, what should the decision makers' confidence be in the Incredible Years intervention? Similarly, after the Durham Connects study,

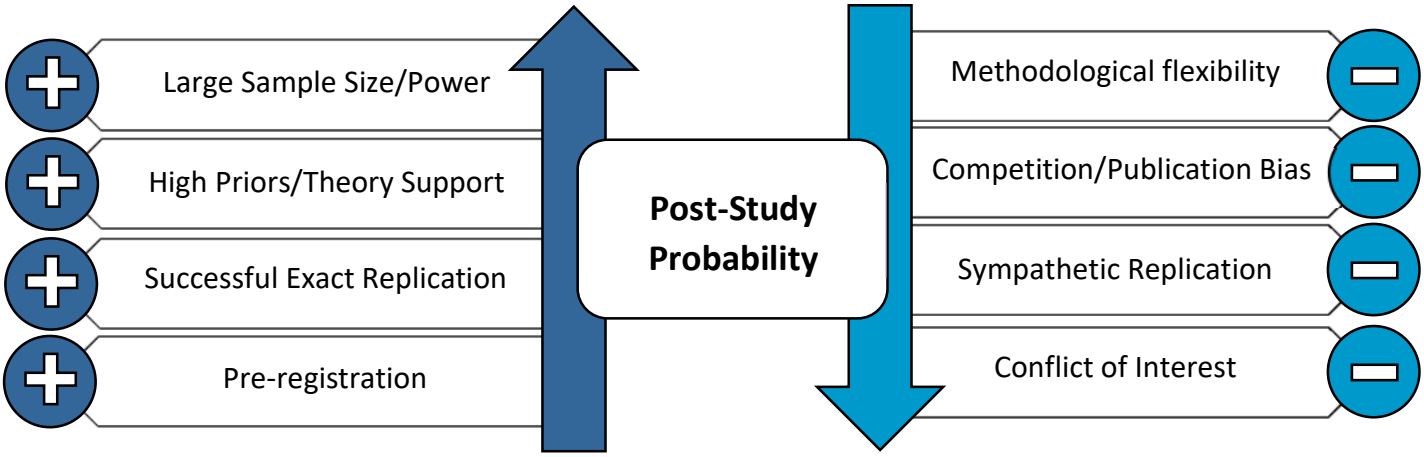
what should the decision makers' confidence be in the Family Connects intervention? In both cases, the answer lies in the post-study probability.

It is clear that decision makers should try to assess the post-study probability that an intervention works as cautiously as possible before deciding to scale up this intervention. To understand and assess it, we need to consider the different elements of the post-study probability.

The Components of Post-Study Probability

Figure 1 provides a general illustration of the drivers of post-study probability, some of which contribute positively, by increasing the post-study probability, and some negatively, by decreasing it.

Figure 1. The Determinants of Post-Study Probability



Prior Beliefs

A first, critical factor that affects the post-study probability is the prior beliefs in the association, (i.e., before the new experimental evidence). What did we know about this association before? What did prior studies indicate? Is this association predicted by any of our trusted theories? A few general examples from science can illustrate the importance of such priors. Consider the evidence connecting genes with particular types of diseases. Given the immense number of genes, what is the probability that a particular one is associated with cancer in the absence of a theoretical

link? It should be minuscule. Similarly, consider the controversy that concerns extrasensory perception. Given the fact that this phenomenon, if true, would violate much of what we know from established theories of the natural sciences, again we should attribute a very low prior. Going back to the Incredible Years intervention, in the Head Start CARES demonstration, several secondary outcomes of an exploratory nature were examined. Even where effects were strong, these effects cannot be accompanied with a high post-study probability after Head Start CARES because of an absence of (ex-ante) theoretical justification. Accordingly, our approach tells us to be skeptical of statements such as "While the findings show that The Incredible Years did not reduce children's behavior problems or improve their executive function skills, the enhancement did improve children's emotional and social skills and their learning behaviors" (Morris et al., 2014, p. ES-8). Such improvements were not hypothesized in advance, and hence our Bayesian framework tells us that there is a higher chance of them being false positives.

That all means that for surprising associations, the evidence in their support should be very strong and robust if we are to assign them a high post-study probability. On the other hand, we should require less demanding evidence from hypotheses that are supported by the predictions of accepted theories that have empirically worked well in similar situations in the past. For instance, in Head Starts CARES, some measured outcomes were primary, meaning that they were predicted by a theory of change and explicitly hypothesized in advance. In particular, Preschool PATHS found statistically significant effects with relatively modest effect sizes for a range of social and emotional behaviors predicted by the theory. The priors for these effects should be relatively high, so our approach teaches us to view them as true effects despite their small size.

It should be emphasized, however, that the development of techniques for determining scientific priors is still in an early stage. For instance, Camerer et al. (2016) illustrates how surveys or prediction markets (a sort of privately informed betting market) may elicit the prior beliefs of the

scientific community. One should avoid in particular *optimism bias*, which may lead to unrealistically favorable prior beliefs. Even excellent scientists may suffer from optimism bias, especially if they work in a field and they have developed a belief that what they work on may be very important. Various types of allegiance biases are also likely to distort this process of determining appropriate priors.

Power

A second key component in the post-study probability is what in the statistical jargon is called *power*. This corresponds to the probability of avoiding Type-2 error when the association is in fact true. This means: is your experimental design strong enough to detect and declare support for the association if the association were true? Clearly, a principal aspect of this strength is the number of observations that an experiment has. If power in some experiment is particularly low, then we can assume that findings indicating a positive association between the intervention and the outcome are likely a false positive. Accordingly, a scholar of the scientific evidence should be skeptical for associations declared in studies with low sample sizes. The power of a study can (and should) be formally calculated before running the study. Regardless, pilot studies are tremendously underpowered, but even pivotal trials that lead to licensing in domains such as biomedical sciences may often be underpowered. For instance, when it comes to educational interventions such as Incredible Years, we need to be very attentive to power considerations and their interplay with the level (or unit) of intervention. Even in medium-size assessments such as Foundations of Learning, given that the intervention has to be at the teacher or even school level, one needs to be careful to have a large enough number of independent observations (schools or teachers) in order to achieve sufficient power.

Biases and Conflicts of Interest

As Ioannidis (2005) noted, biases and conflicts of interest also play an important role. Often researchers have strong incentives to receive a certain type of result in the experiment. Despite what the public might perceive, experimental studies in the social and biomedical disciplines may be affected by strong influences of the exact context in which a study has been conducted. For instance, in experimental economics and psychology much discussion has been devoted to the problem of the experimenter demand effect: often the experimenter's mere hypotheses (if surmised by the participants) or their physical presence may affect the results. In fact, the exact way of conducting a study may greatly matter for the results. This is why, in the biomedical disciplines, there is an emphasis on double-blind designs, where those who administer the study and those who participate in the study are not aware of the intervention assignments. However, double-blinding is not always easy or even feasible, especially if the intervention is such that it is not possible to create a sham control that would not be distinguishable. Moreover, an effort to blind a trial may diminish its pragmatism (its relevance to real-world circumstances), since it creates artificial circumstances that will not be the same as real life.

Getting back to one of our three examples, in the replication of the Family Connects program by Goodman and Dodge (2016), informed consent was not obtained (see Goodman & Dodge, p. 11). This means that in this study, unlike the original study by Dodge et al. (2013), families receiving treatment did not know that they were participating in a study. Since the replication found substantially lower effects on child emergency care utilization, it is an open question whether the strong effects in the initial evaluation were associated with experimenter demand effects (caused by the unavoidable use of standard protocols dictating informed consent).

With or without blinding, there are many other ways that bias can be introduced. Often, there is a large number of possible statistical specifications and methodological flexibility about which one to choose, i.e., a large number of analytical degrees of freedom. If both the outcomes are carefully and unambiguously defined and the exact analysis plan prespecified in all its important details, the

many options on how to measure outcomes and how to analyze them may be reduced. In all three of our examples of early childhood interventions, the research design carefully specified primary questions with clear research hypotheses. Accordingly, this should increase our confidence in the assessment of the respective programs.

All the aforementioned parameters increase the *bias* parameter in the calculation of the post-study probability. In turn, higher bias is associated with lower post-study probability, everything else equal. Accordingly, important parameters that one should consider for a given intervention are the following: In the relevant studies, how much methodological flexibility was there? Are methods in the particular scientific domain mature and standardized? Are experiments and analyses pre-registered? How strong are incentives to receive a certain type of results? Are there strong norms for sharing data and material from the experiments? These considerations should receive substantial weight.

One may think that with a large number of studies and competition, many of the factors that affect such biases might wash out, and that the overall literature may show an objective picture. Unfortunately, this is not always so, and different domains in science exhibit different types of patterns. At the aggregate level, additional types of biases may emerge. As discussed in Chapter 6 of this volume (Al-Ubaydli et al.), a very famous one is publication bias, namely the fact that studies that find certain types of results are more likely to be published, for instance because society assigns greater value in experiments that report higher treatment effects. If this is the case, the published literature does not paint a representative picture of all studies that have been conducted. Studies may not remain entirely unpublished, but they may be reported with bias—e.g., if some of the most favorable outcomes and analyses get reported, while others are not. In all these cases, the standard Bayes formula for inference cannot be used, and alternative ways to gauge the probability of the association (that account for the bias), need to be used (see Ioannidis, 2005 and Maniadis et al., 2014 for examples).

Meta-Analysis and Pre-Registration

The statistical tool of meta-analysis may offer considerable insights concerning these aggregate patterns of bias. In particular, it may be used to assess the degree to which a literature suffers from publication bias. However, available tests for publication bias are only modestly successful in this task and they can also give wrong impressions (Sterne et al., 2011). Ideally, one would like to make effective study pre-registration, where *all* the trials on a given question are entered in a registry system before they start, with details on their protocols, outcomes, and pre-specified analyses. In the real world where not all trials are registered, one can place more trust in those that have been pre-registered, even more so if the pre-registration record did provide explicit details on these specific features. Trial registration has made substantial progress for clinical trials in medicine and some other related fields (Zarin et al., 2017), but many areas of experimental research still have gaps in their registration. Moreover, even when a trial is registered, often its outcomes are changed between conception and final reporting (Goldacre et al., 2019).

Meta-analyses can help us assess the degree to which studies are supported by subsequent replications. Although the absolute number of replications matters for the reliability of a literature, it is also important to ascertain the *regime* in which they operate (Maniadis et al., 2017). This means that, for example, all or most studies may be done with the same sponsorship and/or the same authors who may have pervasive allegiance biases. What does meta-research evidence say about the prevalence of replications and the regime in which replication takes place? If the domain is considered to have a regime of *sympathetic replicators*—meaning that they are biased in the direction of finding a replication result compatible with the original study—then we should be cautious in how we update our beliefs on the basis of replications. Meta-analyses may offer some indirect hints about the nature of the regime, but ideally large-scale replication initiatives and pre-registered replications can help us assess the overall regime.

Post-Study Probability and the Scale-up Problem. How Do They Relate?

While the post-study probability can inform us (prior to scaling-up) about the degree to which treatment effects are real for the type of studies considered, this is not the end of the story. Post-study probability gives us one piece of information: that the intervention works when applied at small scale. However, this is not the only criterion for choosing what needs to be scaled up. As Al-Ubaydli et al. emphasize in Chapter 6, we also need to understand the degree to which a true effect discovered using small-scale studies is likely to translate at large scale (the pure *voltage effect*).

The factors that lead to high post-study probability need not be the same as the factors that contribute to a high voltage effect, and sometimes they might conflict. For instance, social and behavioral science studies conducted in sterilized laboratory environments with very tight control and careful blinding of experiments may result in very replicable results and a high post-study probability. However, when they are implemented as a policy at scale, many aspects of this controlled environment may be lost, and the results are likely to be different. In addition, the quality and dosage may be hard to scale-up, negatively affecting the intervention fidelity (see Caron et al.'s Chapter 9 in this volume for a more comprehensive discussion of intervention fidelity).

Let us illustrate all this using Incredible Years and its implementation in the two relevant trials. In addition to the teacher training received in Head Start CARES, Foundations of Learning had classroom-level consultations, stress management workshops, and individual child-centered consultations. Given also the variation in geographical location of the centers participating in Head Start CARES (four in the Northeast, four in the West, three in the South, and six in the Midwest/Plains) relative to Foundations of Learning, it is reasonable to expect non-negligible differences in study populations, too. Accordingly, the two trials clearly differ in context and target population, and the question is whether we can say something systematic and rigorous about these differences.

This general issue, of course, is related to the trade-off between internal and external validity. This is why we should be careful about using just one criterion to measure study validity. One

alternative is to use the post-study probability along with another quantitative criterion that describes the divergence of the environment in which the existing studies have taken place with the scaled-up actual practice environment. In particular, medical researchers have developed a tool called PRECIS-2 (for recent assessments see Lipman et al., 2017; Dal-Re et al., 2018). The objective of this tool is to categorize experimental studies in a continuum from *entirely explanatory* to *entirely pragmatic*. *Explanatory* means adhering to standards required from initial evidence proving the existence of a new causal mechanism. That means, for example, that interventions should be delivered by expert researchers, informed consent must be carefully obtained, placebos must be used, etc. All these factors tend to differentiate such a study from the current standard of how an intervention would typically be delivered in practice. The PRECIS-2 tool describes nine domains of trial characteristics that need to be assessed at a scale from 1-5, with larger scales denoting more pragmatic trials (see also Davis et al.'s Chapter 8 in this volume on PRECIS-2 and scalability). As far as we can tell, such tools have not been widely used in early childhood interventions, but the logic of rigorously assessing the scalability of a new intervention—on top of establishing the truthfulness of the initial findings—could be extremely useful. However, one should acknowledge that such subjective metrics need strong validation of their reliability across time and across different coders, and on this dimension, there is still much to be desired (Forbes et al., 2017).

Al-Ubaydli et al. raise a number of points regarding potential trade-offs of internal and external validity in Chapter 6 of this volume, which will be important to consider. Importantly, they consider costs as well as benefits of a treatment. Our analysis focuses principally on the benefit side, or when are interventions likely to have truly large effects as indicated by small-scale studies. We shall simply comment on the likely interdependencies between the voltage effect, the post-study probability, and the cost side.

First of all, moving from experimental participants to the general population might lower the effect size because the participants in the small-scale studies might have been selected or self-

selected according to the likely efficacy of the intervention. In terms of our terminology, such selection would cause considerable bias, which would tend to lower the post-study probability (for a more in-depth discussion of the threats to scalability due to differences between initial study population and the target population, see Stuart's Chapter 14 in this volume). For instance, Sweet and Appelbaum (2004) note about the sixty early home-visiting studies that they synthesize: "The majority of programs targeted families at some type of environmental risk (75%)." If these data were interpreted as generalizable for the whole population, bias would be introduced. This may partly explain the lower effects found in the large MIHOPE evaluation. The relatively low effects found in the Head Start CARES evaluation can be viewed in a similar light. In particular, previous small-scale studies were mainly with low-income children. Such children may be especially susceptible to treatment, and again generalizing these effects to the whole US population may introduce bias and lead to non-replicable results.

Another problem is that in the general population, high-risk participants may be less willing to receive or continue treatment, and this may lead to lower effects at scale. For instance, in MIHOPE programs, high-risk families had a higher rate of exiting home visiting programs before completing them. Accordingly, non-representativeness in terms of the direct treatment effect would be a problem in studies that already have a low post-study probability, because of high bias.

Secondly, as an intervention is implemented at a large or national scale, there are potential economies of scale in participation and implementation costs. These would tend to make the intervention more or less cost-efficient at scale, depending on the type of the scale economy. Economies of scale in participation cost means that the scaled-up intervention will have naturally fewer selection problems relative to smaller trials. If the probability of exhibiting scale economies in participation is positively correlated with the participation costs in the small study (as seems very reasonable) then there is an interesting tradeoff. Specifically, higher participation costs entail higher probability of selection problems (e.g., only strongly motivated participants taking part) in the small-

scale study—so, higher bias and lower post-study probability. Accordingly, the expected benefits of scaling up as determined by results from the small-scale intervention go down. However, a higher probability of economies of scale imply that expected participation costs would decrease as an intervention scales up. Such potential interactions deserve careful consideration.

Finally, there may be scale effects in administration quality. In terms of the PRECIS-2 tool, this would correspond to two items labelled "Flexibility in Delivery" and "Flexibility in Adherence." The first refers to "How different is the flexibility in how the intervention is delivered from the flexibility anticipated in usual care?" When interventions are applied in large scale, they may be streamlined, and the loss of important details may downgrade their performance. The second means "How different is the flexibility in how participants are monitored and encouraged to adhere to the intervention from the flexibility anticipated in usual care?" Some interventions require very strict, faithful adherence to be effective, and this may require a commitment of resources to maximize adherence that is no longer seen upon scaling up.

Toward Scalable Evidence

The aforementioned issues are by their nature somewhat technical. Accordingly, decision makers should work closely with independent advisors to decide on such matters. By their nature, the arguments put forward by researchers involved in the previous literature and who have a stake in proving that their discovery was important should not be considered sufficient because of the possible conflict of interest. While field experts can be useful advisors, independent validation of claims of benefits and overall appraisal of the evidence is essential.

Decision makers should interact with meta-research experts in order to decide about each particular treatment they are interested in scaling. They should jointly consider the factors specified in the literature (e.g., Ioannidis, 2005; Maniadis et al., 2017). Was there theory behind the treatment effect, and what were the priors? Has preregistration been employed? Has the result been replicated

by independent teams? They should also compare the features of the intervention, setting, and other aspects of running the original studies and examine whether these features allow scalability to start with. They could also commission a set of independent replications prior to scaling.

At the same time, bodies that decide the funding of scientific research should interact with decision makers and make funds available for projects that examine the replicability, robustness, and scalability of interesting interventions. Especially in social sciences, this is an area where research efforts should be directed. This includes funding directed toward helping and incentivizing preregistration and registered reports.

Another critical factor that decision makers should consider is whether the descriptions of interventions, especially complex ones, are adequate. Very often, different trials have tested interventions that are not exactly the same and they vary in components and features. For example, MIHOPE home visiting programs had several difficulties caused by inadequate descriptions of the implementation details of previous home visiting programs (Michalopoulos et al., 2019). These hampered the ability of researchers to assess the validity and robustness of the effects. Accordingly, MIHOPE implementation research emphasized how important it is to provide information on several critical dimensions of the intervention, which facilitates the accumulation of knowledge about what works and what does not.

It can be difficult to decide which features and components (among many) to scale up. For these reasons, decision makers, when considering which studies are likely to be appropriate for scaling up, should consider the quality of the accompanying descriptions. Once more, medicine provides good benchmarks for best practices: TIDieR (Hoffmann et al., 2014; Hoffmann et al., 2017) is a set of reporting guidelines concerning the adequate description of medical interventions. Decision makers should consider such checklists whenever possible and prefer interventions that adhere to their standards. Such adherence signals the high quality of reporting and allows us to

understand what exactly was done as part of the intervention and which are its essential components. As Hoffmann et al. (2014) argue: “Properly endorsed and implemented reporting guidelines offer a way for publishers, editors, peer reviewers, and authors to do a better job of completely and transparently describing what was done and found. Doing so will help reduce wasteful research and increase the potential impact of research…” (p. 9). Without knowing what really was done, it is unlikely to be able to scale the intervention successfully.

Evidence based policy is extremely important, and safeguarding its credibility is especially crucial for modern governments. Policy developments such as early childhood initiatives and the establishment of behavioral intervention (nudge) units need to be accommodated by appropriate theoretical developments of the science of using science. Otherwise, there is a risk of undermined credibility. Indeed, a recent meta-study by DellaVigna and Linos (2020) showed that behavioral interventions may have much lower effects at scale, which should serve as a cautionary tale. In this chapter, we illustrated how a simple framework can help us identify the roots and possible remedies of this problem.

References

Al-Ubaydli, O., Lee, M. S., List, J. A., & Suskind, D. (in press). The Science of using Science: A New Framework for Understanding the Threats to Scaling Evidence-Based Policies. In J. List, D. Suskind, & L. Supplee (Eds), *The scale-up effect in early childhood & public policy: Why interventions lose impact at scale and what we can do about it.* Routledge.

Bloom, H. S., & Weiland, C. (2015). Quantifying variation in Head Start effects on young children's cognitive and socio-emotional skills using data from the National Head Start Impact Study. *Randomized Social Experiments eJournal*.

Camerer, C. F., Dreber, A., Forsell, E., Ho, T. H., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., Chan, T., Holzmeister, F., Imai, T., Isaksson, S., Nave, G., Pfeiffer, T., Razen, M., Wu, H. & Heikensten, E. (2016). Evaluating replicability of laboratory experiments in economics. *Science* (New York, N.Y.), *351*(6280), 1433-1436.

Caron, EB, Bernard, K., Metz, A. (in press). Fidelity and Properties of the Situation: Challenges and Recommendations. In J. List, D. Suskind, & L. Supplee (Eds), *The scale-up effect in early childhood & public policy: Why interventions lose impact at scale and what we can do about it.* Routledge.

Dal-Ré, R., Janiaud, P., & Ioannidis, J. P. (2018). Real-world evidence: How pragmatic are randomized controlled trials labeled as pragmatic?. *BMC medicine*, *16*(1), 49.

Davis, J., Guryan, J., Hallberg, K., & Ludwig, J. (in press). Studying Properties of the Population: Designing Studies that Mirror Real World Scenarios. In J. List, D. Suskind, & L. Supplee (Eds), *The scale-up effect in early childhood & public policy: Why interventions lose impact at scale and what we can do about it.* Routledge.

DellaVigna, S., & Linos, E. (2020). *RCTs to Scale: Comprehensive Evidence from Two Nudge Units*. Working Paper, UC Berkeley.

Dodge, K. A., Goodman, W. B., Murphy, R. A., O'Donnell, K., & Sato, J. (2013). Randomized controlled trial of universal postnatal nurse home visiting: impact on emergency care. *Pediatrics*, *132*(Supplement 2), S140-S146.

Forbes, G., Loudon, K., Treweek, S., Taylor, S. J., & Eldridge, S. (2017). Understanding the applicability of results from primary care trials: lessons learned from applying PRECIS-2. *Journal of clinical epidemiology*, *90*, 119-126.

Goldacre, B., Drysdale, H., Dale, A., Milosevic, I., Slade, E., Hartley, P., Marston, C., Powell-Smith, A., Heneghan, C., & Mahtani, K. R. (2019). COMPare: a prospective cohort study correcting and monitoring 58 misreported trials in real time. *Trials*, *20*(1), 118.

Goodman, W.B., & Dodge, K.A. (2016). *A Low-Cost RCT of a Universal Postnatal Nurse Home Visiting Program: Durham Connects* (Final Report). https://mfr.osf.io/render?url=https://osf.io/3ys4m/?direct%26mode=render%26action=download%26mode=render

Hoffmann, T. C., Glasziou, P. P., Boutron, I., Milne, R., Perera, R., Moher, D., Altmn, D.G., Barbour, V., Macdonald, H., Johnston, M., Dixon-Woods, M., McCulloch, P., Wyatt, J. C., Chan, A. W., & Michie, S., & Lamb, S. E. (2014). Better reporting of interventions: template for intervention description and replication (TIDieR) checklist and guide. *BMJ (Clinical research ed.)*,, *348*, g1687.

Hoffmann, T. C., Oxman, A. D., Ioannidis, J. P., Moher, D., Lasserson, T. J., Tovey, D. I., Stein, K., Sutcliffe, K., Ravaud, P., Altman, D.G., Glasziou, P.& Perera, R. (2017). Enhancing the

usability of systematic reviews by improving the consideration and description of interventions. *BMJ (Clinical research ed.), 358,* j2998. https://doi.org/10.1136/bmj.j2998

Ioannidis, J. P. (2005). Why most published research findings are false. *PLos med*, *2*(8), e124.

Johnson, R. C., & Jackson, C. K. (2019). Reducing inequality through dynamic complementarity: Evidence from Head Start and public school spending. *American Economic Journal: Economic Policy*, *11*(4), 310-49.

Kline, P., & Walters, C. R. (2016). Evaluating public programs with close substitutes: The case of Head Start. *The Quarterly Journal of Economics*, *131*(4), 1795-1848.

Lipman, P. D., Loudon, K., Dluzak, L., Moloney, R., Messner, D., & Stoney, C. M. (2017). Framing the conversation: use of PRECIS-2 ratings to advance understanding of pragmatic trial design domains. *Trials*, *18*(1), 532.

Maniadis, Z., Tufano, F., & List, J. A. (2014). One swallow doesn't make a summer: New evidence on anchoring effects. *American Economic Review*, *104*(1), 277-90.

Maniadis, Z., Tufano, F., & List, J. A. (2017). To replicate or not to replicate? Exploring reproducibility in economics through the lens of a model and a pilot study. *Economic Journal*, *127*(605), F209-F235.

Michalopoulos, C., Crowne, S. S., Portilla, X. A., Lee, H., Filene, J. H., Duggan, A., & Knox, V. (2019). *A summary of results from the MIHOPE and MIHOPE-Strong Start studies of evidence-based home visiting* (No. 23171787025b46589a5e545fb45db441). Mathematica Policy Research.

Morris, P., Mattera, S., Castells, N., Bangser, M., Bierman, K., & Raver, C. (2014). *Impact Findings from the Head Start CARES Demonstration: National Evaluation of Three Approaches to*

Improving Preschoolers' Social and Emotional Competence (OPRE Report 2014-44). Washington, DC: Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.

Morris, P., Lloyd, C. M., Millenky, M., Leacock, N., Raver, C. C., & Bangser, M. (2013). *Using Classroom Management to Improve Preschoolers' Social and Emotional Skills: Final Impact and Implementation Findings from the Foundations of Learning Demonstration in Newark and Chicago*. MDRC Working Paper. http://dx.doi.org/10.2139/ssrn.2202401 (last accessed: 18 March 2020)

Stuart, E. A. (in press). Accounting for Differences in Population: Predicting Intervention Impact at Scale. In J. List, D. Suskind, & L. Supplee (Eds), *The scale-up effect in early childhood & public policy: Why interventions lose impact at scale and what we can do about it.* Routledge.

Sterne, J. A., Sutton, A. J., Ioannidis, J. P., Terrin, N., Jones, D. R., Lau, J., Carpenter, J. Rucker, G., Harbord, R.M., Schmid, C.H., Deeks, J.J., Peters, J., Macaskill, P., Schwarzer, G., Duval, S., Altman, D.G., Moher, D., Higgins, J.P.T., & Tetzlaff, J. (2011). Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. *BMJ*, *343*, d4002.

Sweet, M. A., & Appelbaum, M. I. (2004). Is home visiting an effective strategy? A meta-analytic review of home visiting programs for families with young children. *Child development*, *75*(5), 1435-1456.

U.S. Department of Health and Human Services, Administration for Children and Families. (January 2010). *Head Start Impact Study*.Final Report.. Washington, DC.

Zarin, D. A., Tse, T., Williams, R. J., & Rajakannan, T. (2017). Update on trial registration 11 years after the ICMJE policy was established. *New England Journal of Medicine, 376*(4), 383-391.