

Cynthia Osborne
University of Texas at Austin
LBJ School of Public Affairs

Process to Identify Effective Policies to Strengthen the Prenatal-to-Three System of Care

in

“The Scale-up Effect in Early Childhood and Public Policy: Why interventions lose impact on scale and what can we do about it.”

July 24, 2020

Abstract

The first three years of a child’s life lay the foundation for all subsequent health and development. State policy leaders are increasingly convinced by the science of the developing child that early intervention is prudent, but they need guidance on which state policies are most effective at creating the conditions in which children thrive. This chapter discusses the similarities and differences in identifying effective policies as compared to effective programs, details the process of identifying effective policies, and provides recommendations for policy makers to create an evidence-based policy agenda to strengthen the prenatal-to-3 system of care.

Overview

The brain and body systems develop rapidly during the prenatal-to-age-three period. Science is clear that children's development is strongly influenced by their home and other caregiving environments: Children thrive in environments that are safe, stimulating, warm, and stable, but exposure to repeated stressors with limited support or interaction from a warm, caring adult is associated with life-long negative health and behavioral outcomes (Shonkoff et al., 2012).

State policy leaders are increasingly convinced by the science of the developing child that it is smart to invest in the prenatal-to-three period. Leaders are less clear, however, on what specific policies or programs they should implement to ensure that children get off to a healthy start, that their parents have the necessary skills and resources to care for them, and that early care environments are safe and stimulating for infants and toddlers. States differ widely with regard to their economic, geographic, and demographic characteristics, and their needs and priorities vary as well. Differences across states in the type and generosity of policies they implement lead to substantial differences in the level and type of resources families receive to help their children thrive. Though implementation varies significantly, many if not most states, have increasingly sought to enact *evidence-based policy* in recent decades.

Evidence-based policy is a practical idea to use the best available evidence in combination with practical expertise and input from those affected by a problem (Sackett et al., 1996). Ideally, evidence from rigorous studies can guide state leaders to adopt the most effective policies that strengthen their prenatal-to-three system of care. If this evidence were available, policy makers would know what works and have confidence that implementing a certain policy would lead to desired outcomes and be cost effective. Within the larger bucket of evidence-based policy there

are discussions about different levels of intervention: practices (i.e., discrete actions), programs (i.e., codified sets of practices), and policies (i.e., a set of principles that guide decisions to achieve specific outcomes). Much of the evidence available in the field is focused on practices and programs. Clearinghouses abound, but they tend to focus on providing evidence on which *programs* meet a given standard of effectiveness. There is no authoritative source to highlight which state-level *policies* are effective at fostering the conditions in which infants and toddlers thrive. Policies differ from programs in that policies are typically adopted through legislative or regulatory channels and are implemented at-scale to a population based on stated eligibility criteria (e.g., state minimum wage or a tax credit). In contrast, a program is a well-defined service that is implemented for a targeted group (e.g., home visiting programs or Early Head Start).

The challenges of identifying and scaling effective, evidence-based programs and practices to support early childhood are well established. But, to create a system of care for infants, toddlers, and their parents and move outcomes on a large scale, states need to go beyond programs and also implement effective, evidence-based policies. In this chapter, I explore the challenges in identifying effective state-level policies and provide recommendations to researchers and state policy leaders on how we can develop an evidence-driven policy agenda over time.

Differences Between Evaluating Programs Versus Policies

Many of the challenges identified in the economic model for scaling effective programs in Chapter 6 are similar to the challenges states grapple with in ensuring that policies have their intended impact. For both policies and programs, the aim is to identify for whom the intervention works best and under what conditions, and to identify the core elements of the program or policy that lead to improved outcomes. As noted in the model, this can be a challenge when the

properties of the population reached by a program or policy, or the properties of the situation in which it was implemented, differ across implementation settings. Additionally, the impacts of both programs and policies may be relatively small at scale, given that the impacts are spread across a diverse population who may not benefit similarly from the intervention (Chapter 6). The effects of programs and policies may also spill over to individuals not directly affected by them, causing the overall impact to look smaller than it would were spillover effects measured. For example, if a neighboring school district seeks out additional resources to support students because they saw the success of an initiative implemented in a different district, outcomes for students in the district not reached directly by the initiative may also improve. Despite these shared challenges, programs and policies differ, and how we evaluate them and determine whether they are effective differs as well.

The economic model (Chapter 6) helps us determine how to take an effective intervention to scale (practice, program or policy). It calls on us to consider whether the available evidence is sufficient to be actionable, whether the studied population is representative of those we aim to support, whether the context or situation differs across implementation sites such that it may affect the causal impact of the program or policy, and whether there are spillover or general equilibrium effects at scale.

For programs, we derive actionable evidence from rigorous randomized control trials of pilot programs that clearly identify a causal link between specific program elements and well-measured participant outcomes. The challenge is often to replicate the evaluation of the program in different contexts and populations to determine if the program is robust and likely to generate similar results when taken to scale.

By contrast, social policies are typically implemented at scale, rather than in a pilot setting. For example, policies to expand Medicaid or implement paid family leave or a state minimum wage will apply to all eligible members of the population when the policy goes into effect. Because the policy is implemented at scale, evaluators cannot use randomized controlled trials to measure its impact in a smaller or controlled setting; rather they have to take advantage of natural experiments or employ quasi-experimental designs or rigorous longitudinal designs to measure the impact of the policy on changes in outcomes. We often use rigorous statistical techniques, such as difference-in-difference, regression discontinuity, or interrupted time series designs, that allow us to compare changes over time in populations that experienced the implementation of a policy with populations that did not experience the policy. These techniques have gotten increasingly sophisticated and often allow for causal inference. However, the studies are not as rigorous or straightforward as randomized controlled trials of a well-defined program. Some policies are evaluated multiple times by multiple scholars, but scholars often employ a variety of research designs, use different samples or study periods, examine different outcomes, and come to varying conclusions about a policy's effectiveness. This robust investigation of the policy is important and provides the field with necessary information about the policy, however, it is often more difficult to compare results across studies of policies compared to replication studies of well-specified programs.

The population of participants and situations discussed in the economic model are also easier to define in well-controlled program evaluations than in evaluations of policies. Policy evaluators must often rely on administrative data, rather than data collected directly from participants, and we often estimate policy participation based on eligibility, rather than measuring actual take-up of the policy. We also seldom track the same individuals over time in

policy evaluation, and therefore rely on changes in outcomes at the population level. Rather than carefully measuring the specific outcomes linked to the theory of change via surveys or direct assessments, we are often limited to measuring what is available in the aggregate or population-level data, and these measures may provide an incomplete picture of the impact of the policy.

Another contrast between evaluating programs and policies is that the properties of the policy context or situation may vary substantially more across two sites as compared to the context in which a program is administered. Effective programs are typically well-defined—that is, you know what you are evaluating, and can take the remaining context or policy environment into account when replicating program evaluations across sites. In contrast, policies across two states may share the same name, but their elements may differ considerably, making it a challenge to assess overall effectiveness across states. Properties of the population can be a driver for some of these variations in policies. For example, decisions about core components included in a policy or funding rates for certain policy features can be shaped by different political environments in states, different needs of the population, or different geography and infrastructure. This makes it difficult to disentangle the efficacy of the policy across states and draw conclusions about best practices.

For example, 31 states have adopted a state Earned Income Tax Credit (EITC) and set their credit as a percentage of the federal EITC. The federal EITC is a refundable tax credit that increases in value as workers earn more, up to a given level, then plateaus, and eventually fades out at income levels about twice the federal poverty level. However, the percentage of the federal EITC that state EITCs provide ranges from a low of 3% of a recipient's federal EITC award in Montana to a high of approximately 50% of the federal EITC award in California, and in six states, the state tax credit is not refundable. Although the research is clear that nonrefundable

credits are ineffective at improving outcomes (Lim, 2009), no study provides guidance on the optimal percentage of the federal EITC that states should offer to achieve the best outcomes. The variation by states in the generosity of their EITC is useful for measuring the marginal impact that additional dollars in a household have on specific outcomes, but the studies do not provide clear evidence on the optimal policy elements.

Finally, determining the impacts of policies at scale may be challenged by the presence of spillover effects. Spillover effects occur when the implementation of a policy in one community or for one group of people affects outcomes for others outside of the community or target group. For example, implementing a paid leave policy for government employees may influence other organization and companies in a community to implement leave policies. Spillovers can be positive or negative. Policies may spill over in different ways across communities and states depending on their political landscape, the resources available to others in the surrounding area, and the needs of adjacent communities. When analyzing the effectiveness of a policy, it is important to consider the extent of spillover effects and whether those spillover effects might occur in a new implementation setting.

The challenges of determining whether a program will be effective at scale are substantial, and they are quite similar to the challenges of determining whether a policy is effective. But given that policies are typically implemented at scale, direct measures of participation and outcomes are difficult to obtain and often estimated using administrative data, and the elements of a similar policy may differ considerably from state to state, determining whether a policy is effective is subject to more professional judgement and may be less precise than program evaluation.

Steps to Determine If a Policy is Effective

As mentioned previously, dozens of clearinghouses exist to identify effective *programs* (Neuhoff et al., 2015). The clearinghouses differ considerably in the criteria they use to determine if a program is effective overall, and differ even more in the nomenclature they use to label a program effective or not. But, nearly all clearinghouses limit their reviews to evaluations that use randomized controlled trials (RCTs) or rigorous quasi-experimental designs (QEDs), and indicate whether the program overall is well supported by the evidence or not. To be labeled the highest level of effectiveness, the program typically has to have demonstrated impacts in two RCTs, and many clearinghouses require that the program effects are sustained for at least 12 months. Most clearinghouses also require the program to demonstrate a positive impact on at least one intended outcome (with no harmful impacts), and the criteria often ignore the preponderance of null effects.

Very few clearinghouses include reviews of *policies*. If a clearinghouse does include a policy (e.g., universal Pre-K), it relies on meta-analyses, which require that a sufficient number of rigorous studies have been conducted on a given policy. Few policies have this level of evidence.

Using a clearinghouse approach to identify effective policies is not as straightforward as it is for programs. Although many of the same steps can be used, such as clearly identifying the policy and its theory of change, and relying only on strong studies with causal inference, it doesn't make sense to use a point system or precise criteria (e.g., two RCT studies with positive effects) to determine policy effectiveness.

For policies, it is important to take all of the research into consideration to conclude whether a policy is effective or not and to highlight the conditions in which the policy is

effective for certain outcomes. Evaluation studies of the same policy may vary based on the time period, sample, policy elements, geographic location, or more, and therefore the results of the studies may vary for the same outcome. Additionally, the policy may be effective at improving one outcome, but not another, and decision makers need to understand these tradeoffs. For example, child care subsidies improve work effort, earnings, and the use of child care, but subsidies do not increase the quality of care that a child experiences (Forry et al., 2013). Mixed findings on outcomes can provide useful guidance for policy leaders in understanding which goals a certain policy might achieve for a certain population; however, they make it challenging to draw clear conclusions about the broad effects of a policy.

Process for Determining Policy Effectiveness

The economic model for scaling recommends researchers scale up programs only when they have a post-study probability of 0.95—in other words, when there are enough replication studies conducted with positive results to indicate that the program has strong evidence of effectiveness. In the policy sphere, however, rigorous replication studies are not common and are often not feasible given the broad scale on which policies are generally implemented. Instead, policy makers must rely on information from the available body of evidence about the impacts of a policy on different outcomes, the strength of that evidence, and the degree to which those outcomes align with their own goals in implementing a policy.

Below I describe the process that my research team at the Prenatal-to-Three Policy Impact Center at the University of Texas at Austin developed to determine whether a state-level policy is effective at strengthening the prenatal-to-three system of care. Our aim is to create a

clearinghouse that provides guidance to states on which policies effectively meet specific prenatal-to-three policy goals.

Identify a State-Level Policy

The first step of the process is to clearly define the state-level policy and determine the leverage that states have to adopt or implement the policy. States often implement a variety of approaches to support the prenatal-to-three period, including programs that target specific populations, strategies to distribute resources to eligible families, and state-level policies that leverage federal policies or are designed to meet a specific state need.

For example, states are often required or allowed to determine their own levers of federal policies, such as eligibility, generosity, and procedures for enrollment and recertification (known as administrative burden). This state leverage leads to considerable variation across states in resources provided to families. States can choose to align their policies with federal policies to take advantage of benefits offered or states can piggyback on federal policies. For example, since 2010, 38 states have chosen to expand Medicaid to provide benefits to adults with incomes up to at least 138% of the federal poverty level, and the federal government pays up to 90% of the expansion costs.

Variation also stems from states' decisions on how to implement federal policies. For example, the proportion of eligible families who receive benefits through the federal Supplemental Nutrition Assistance Program (SNAP) varies from a low of 57% in Montana to over 100% in Hawaii. This variation is largely driven by the eligibility requirements and level of administrative burden states place on residents, such as the length of time between recertification periods.

States also develop their own policies to address the unique needs of their residents. These policies may build on federal policies, such as adopting a state minimum wage that is higher than the federal minimum or instituting paid family leave that builds on the federal Family and Medical Leave Act, or states may develop their own policy, such as universal pre-K, that does not have a federal corollary.

Although multiple states may adopt a given policy with a similar name (e.g., paid family leave), the elements of the policy may differ considerably across states (e.g., number of weeks of leave, reimbursement level, eligibility, funding mechanism) such that parents and children have a very different experience of the policy depending on where they live. This state-level variation in the policy can be an asset when measuring the impact of the policy on specific outcomes, but it makes it difficult to determine which elements of the policy are most effective for which populations and to give an overall stamp of effectiveness.

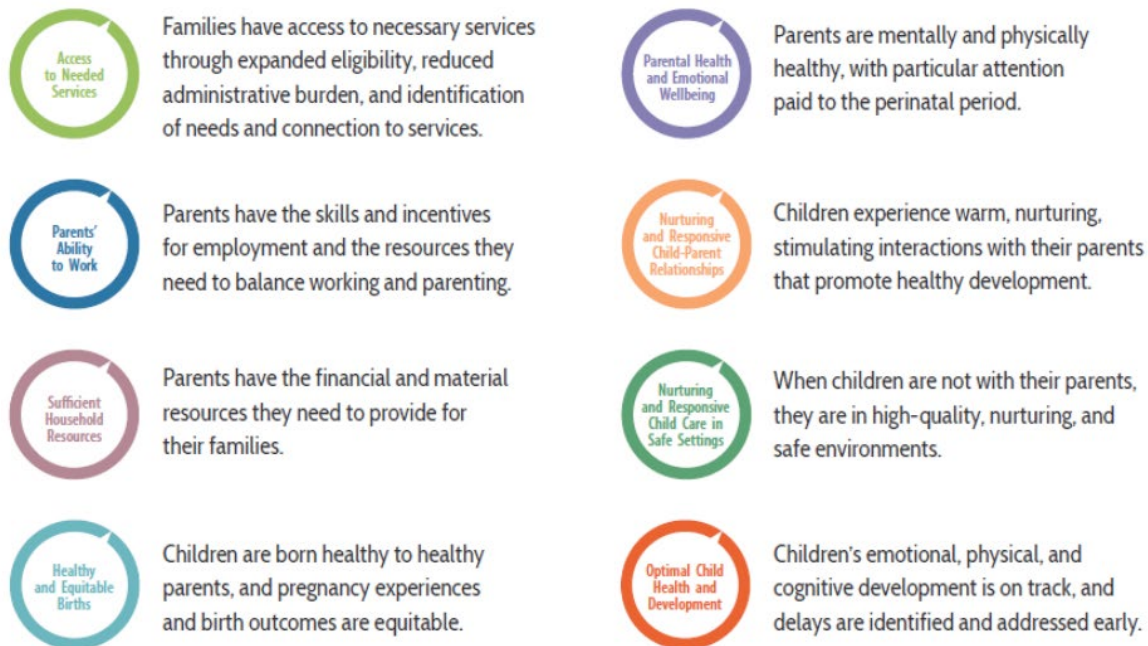
Most research on policies does not provide clear guidance on the ideal thresholds or benchmarks for the features of a policy. For example, child care subsidies are provided by every state to reduce the cost burden of child care on families. The causal impact of a higher subsidy payment on employment and child care quality has been rigorously studied. However, no study on subsidies has linked a specific state's subsidy level to improved outcomes, nor has any study identified the optimal level of child care subsidy that states should provide. Child care subsidies have not been studied as a state-level policy per se, but rather as a mechanism or strategy that states can use to enhance employment and child care use. Therefore, the evidence would allow policy makers to conclude that higher child care subsidies are effective, but it does not provide guidance on the effectiveness of a specific state-level policy.

Articulate a Theory of Change to Improve Prenatal-to-Three Outcomes

After identifying a policy to focus on, the second step in the process is to identify the theory of change or mechanisms through which the policy is expected to improve outcomes for infants, toddlers, and their caregivers. The science of the developing child has clearly identified the conditions under which children thrive and revealed how adverse conditions lead to long-term negative impacts on health and behavior. In the earliest years, children need to be exposed to nurturing relationships, which are fostered by caregivers who have sufficient resources, including mental and physical health, economic security, and the knowledge and skills to care for a child. A state paid family leave policy, for example, may lead to healthier outcomes for parents and infants because it allows new parents time to bond with their child, attend to their own and their child's health needs, and provide greater economic resources and job security.

Figure 1 provides a simple framework that links any given state-level policy to eight policy goals that are derived from the science of the developing child. A policy may impact one or more of these policy goals, and the impact may be measured by different indicators across studies. For example, paid family leave should increase a parent's ability to work, enhance household resources, improve parental health and emotional well-being, and foster more nurturing parenting practices, which should all lead to better child health and development. Each of these goals may be measured by a variety of indicators: for example, the indicators to measure a parent's ability to work may include labor force participation, weeks worked, employment, and job continuity.

Figure 1: Prenatal-to-3 Policy Goals



Identify Causal Studies

The next stage of the process is to comprehensively review all of the available research on the policy. We thoroughly search databases to identify all studies that have examined the policy and carefully sort through them to identify the studies that have attempted to make a causal link between the policy and a prenatal-to-three policy goal in the framework. A broad scan of the literature will generally yield relatively few studies that attempt to measure a causal link between the implementation of the policy and changes in a policy goal, which makes it difficult to meet the .95 post study probability suggested in the economic model (Chapter 6).

Of those studies designed to detect causality, we then determine if the study has a strong research design with limited threats to validity. To evaluate policies, quasi-experimental designs (QED) or longitudinal studies with adequate controls for selection are common, rather than randomized controlled trials, because the policy was applied to the full population rather than a

sample of the population. The criteria of a strong QED study that adequately measures the causal impact of a policy on a given outcome are well established, but studies vary considerably based on their research design, sample, time period, outcomes measured, and policy elements included.

Determine Direction of Causal Impact

For each strong causal study, we then determine the extent to which the indicators measured in the study align with the policy goals in the framework, and we determine whether the effect of the policy on the indicator is beneficial, null, or detrimental. In the example of paid family leave in Table 1, each letter represents a different study meeting the standards of strong causal inference, and we place the letter in a column to denote the direction of the impact (the reference list for the studies are listed by letter at the end of the chapter). Some studies demonstrate beneficial effects of the policy on multiple indicators (e.g., study B, Baum, & Ruhm, 2016), and some studies find positive and negative effects on various indicators (e.g., study A, Bailey et al., 2019).

Table 1: Evidence of Indicators and Overall Impact on Policy Goal, Paid Family Leave Example

Policy Goal	Indicator	Studies Finding Beneficial Impacts	Studies Finding Null Impacts	Studies Finding Detrimental Impacts	Overall Findings
Access to Needed Services	Leave-Taking	B, N, R			Positive
Ability of Parents to Work	Labor Force Participation	D, F, Q			Mostly Positive
	Weeks Worked	B			
	Average Weekly Work Hours	B, N			
	Employment	B, D	N, O	A, F	
	Attachment to Pre-Birth Employer (Job Continuity)	B, O	A		

Sufficient Household Resources	Household Income	M	N		Mixed
	Risk of Poverty	M			
	Annual Wage Earnings			A	
	Hourly Wages		B		
Parental Health and Emotional Wellbeing	Self-Rated Health	P			Positive
	Overweight	P			
	Obesity		P		
	Alcohol Consumption	P			
	Maternal Mental Health	C			
	Paternal Mental Health		C		
	Coping with Demands of Parenting	C			
	Psychological Distress	P			
Nurturing and Responsive Child-Parent Relationships	Mothers' Time Spent with Children on Reading, Outings, Meals	A, C			Positive
	Abusive Head Trauma	I			
Optimal Child Health and Development	Infant Hospitalizations	L			Positive
	Timely Infant Vaccinations	E			
	Overweight	J			
	ADHD	J			
	Hearing Problems	J			
	Communication Problems	J			
	Breastfeeding	G, H, K			
	Child Health	C, J			
	Asthma	C			
	Allergies (Food and Respiratory)		C		

Note: Each letter represents a study meeting the standards of strong evidence. The studies are listed at the end of the chapter.

Determine the Overall Impact on the Policy Goals

For each policy goal in the framework, we then determine the overall direction of the weight of the causal evidence. For example, if the vast majority of the studies indicate a positive impact of the policy on the outcome, we conclude that the effect on that particular outcome is positive. If most, but not all, of the studies indicate a positive impact, then we conclude that the impact of the policy on the outcome is mostly positive. If a similar number of studies indicate positive, negative, or null findings on an outcome, then we denote the impact of the policy on the outcome is mixed. If most studies indicate null findings, then we denote the impact of the policy is null. If most of the studies indicate that the impact of the policy on the outcome is harmful, then the impact of the policy on the outcome would be considered negative. For outcomes in which there is only evidence from one study, location, author, or data set, we denote the impact of the policy on the outcome is trending in the respective direction.

The example of paid family leave illustrates the range of findings within a policy across multiple policy goals. The impact of the policy is positive or mostly positive on most policy goals, including the ability for parents to work, but the impact is mixed with regard to sufficient household resources.

Each of the studies considered in the review is a strong causal study, but some studies are more recent, include more states, or provide greater specificity on the impact of the policy on indicators. These factors are also considered when making a determination of the overall impact of the policy on a particular policy goal. Additionally, we look at the size and direction of the coefficients associated with each indicator, particularly the null results, to determine the strength and general direction of the findings. Merely counting the number of studies would be insufficient to make a sound conclusion (Higgins & Green, 2011).

Determine Overall Effectiveness

The penultimate stage of the process is to determine the overall support for the effectiveness of the policy on prenatal-to-three outcomes. We examine the findings across all of the policy goals studied to determine if the weight of the evidence supports that the policy is effective, needs further study, or is not effective at improving outcomes for infants, toddlers, and their parents. If the findings demonstrate that a policy has mostly negative impacts on prenatal-to-three policy goals, we would label the policy as harmful. We develop general criteria on how to determine the overall effectiveness, but some level of professional opinion is needed. It wouldn't be appropriate to require that a policy has a positive impact on all policy goals or indicators studied to consider it effective, but it is important to be clear as to what goals the policy impacts and which it does not.

Table 2: Nomenclature of Overall Policy Effectiveness

Overall Support for Policy Effectiveness	Definition
Effective	Positive, mostly positive, or mixed* impacts on at least one policy goal
Needs Further Study	Trending impacts only or no causal studies
Ineffective	Mixed* or null impacts only
Harmful	Mostly Negative impacts

*We include mixed impacts in effective policies if the weight of the evidence falls in the beneficial and null categories, and the null results lean positive. We include mixed impacts in ineffective policies if the evidence includes detrimental effects or null results that lean negative.

In the example of paid family leave, the policy has positive or mostly positive impacts on five prenatal-to-three policy goals. The impact of the policy on sufficient household resources is mixed, but the null findings lean toward positive outcomes, and the impact on lower annual earnings is nominal. Thus, we conclude that paid family leave is an effective prenatal-to-three state-level policy.

We also note whether a policy has been insufficiently studied to determine the causal impact of the policy on an outcome. Evaluators need to be cautious to avoid sending a message to advocates and policy leaders that a policy is ineffective if it has not been sufficiently studied. We draw a big distinction between policies that have been well-researched with rigorous studies and found to have no or very small impacts on an outcome, and policies that have not been rigorously evaluated. The bar for using an evidence-based policy approach to state policy making is quite high. Before such an approach can be used, at least one state or large city has to implement the policy without evidence that it is effective, and evaluators have to study the policy using a rigorous research design. Many policies that may strengthen the prenatal-to-three period have not been rigorously evaluated, and therefore there is no sufficient evidence to suggest the policy is effective or not. We designate these as policies that “need further study.”

Conduct an In-Depth Policy Analysis

The final stage of the process is to conduct an in-depth policy analysis. Determining whether a policy is effective overall is only one factor that state policy leaders need to consider when deciding which policies to adopt. If a clear state-policy lever has been sufficiently studied, we provide states with that guidance, and we are clear on the limits of that guidance. For example, most studies on the state minimum wage have examined increases in the minimum wage to \$10.00 or \$11.00 per hour, but have not studied increases beyond that level. For paid family leave, most of the research is based on six weeks of paid leave following the birth, adoption, or fostering of a child. We cannot provide sufficient evidence on the effectiveness of paid leave policies that provide more weeks of leave. The evidence on the state EITC provides

guidance that a refundable EITC is a necessary condition for the policy to be effective, but we cannot provide further guidance on what percentage of the federal EITC states should offer.

State leaders also need information on the size of the expected impact, the number of families that will be affected, and how the impact of this policy compares to the impact of another policy on the same outcomes. State leaders also need to know the funding mechanisms, costs, benefits, and expected return on investment. We also review the evidence to determine if the policy helps to close racial and ethnic gaps in outcomes, or if the policy will maintain the status quo in disparities. Policy makers also need to have as much information as possible on how feasible it is to implement the policy. Some policies have relatively little burden on government infrastructure (e.g., a state minimum wage), whereas others may require considerable investment and ongoing oversight (e.g., childcare quality standards).

Recommendations

Although policy leaders must always make decisions with imperfect information, most decision makers strive to implement policies that have the strongest evidence of effectiveness. The science is clear on the conditions in which children thrive from their earliest years; however, the evidence base on the most effective policies to bring the science of the developing child to life is somewhat slim and it is constantly evolving. Many policies have not been rigorously studied, such as strategies to reduce racial and ethnic disparities in maternal mortality and morbidity, and new studies are constantly being conducted, such as evaluations of policies to provide more stability in work schedules.

The economic model discussed in Chapter 6 of this volume provides guidance to states and evaluators on when a program is ready to scale. In this chapter, I built on the model to

provide additional guidance to states and evaluators on how to identify effective policies that states should prioritize to improve outcomes for infants, toddlers, and their parents.

I encourage state policy leaders to draw on the evidence base (provided in the Prenatal-to-Three Policy Clearinghouse) and prioritize implementing policies that have a robust and clear impact on prenatal-to-three policy goals. Additional evaluations of the policies that states implement are needed, however, so that we can better determine which policy elements are most effective, in which context, and for which populations. We also need to learn how effective policies and programs interact with each other (e.g., is paid family leave more effective when coupled with home visiting programs that support new parents?).

I also encourage state policy leaders to be innovative and try new policies that have a promising evidence base and strong theory of change. If we only implement policies or programs with an existing strong evidence base, then we may miss opportunities to learn how to more effectively support children in their earliest years. Researchers should strive to build the evidence base and fill the numerous gaps in what we know about effective policies to support the prenatal-to-three period. Voices and experiences from the field should be infused in research, and implementation realities should be taken into account when conducting experiments. This collaboration will make studies more relevant and useful for state leaders and foster a culture of continuous innovation, evaluation, and learning that will enhance prenatal-to-three outcomes.

References

- Al-Ubaydli, O., List, J.A., & Suskind, D. L. (2017). What Can We Learn from Experiments? Understanding the Threats to the Scalability of Experimental Results. *American Economic Review*, 107(5): 282 – 286.
- Al-Ubaydli, O., List, J.A., & Suskind, D. (2019). The Science of Using Science: Towards an Understanding of the Threats to Scaling Experiments. *NBER Working Paper* 25848.
- Bollinger, C., Gonzalez, L., & Ziliak, J. P. (2009). Welfare reform and the level and composition of income. In J. P. Ziliak (Ed.), *Welfare Reform and Its Long-Term Consequences for America's Poor* (p. 59–103).
- Forry, N. D., Daneri, P., Howarth, G. (2013). Child care subsidy literature review. OPRE Brief 2013-60. Washington, DC: Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.
- Higgins, J. P. T., & Green, S. (2011). *Cochrane Handbook of Systematic Reviews of Interventions*, Version 5.10. Cochrane Collaboration.
- Neuhoff, A., Axworthy, S., Glazer, S., & Berfond, D. (2015). The What Works Marketplace: Helping Leaders Use Evidence to Make Smarter Choices. *The Bridgespan Group*.
- Sackett, D. L., Rosenberg, W. M., Gray, J. A., Haynes, R. B., & Richardson, W. S. (1996). Evidence based medicine: what it is and what it isn't. *BMJ*. 312 (7023): 71–72. doi:10.1136/bmj.312.7023.71.
- Shonkoff, J.P., Garner, A.S., Siegel, B.S., Dobbins, M.I., Earls, M.F., McGuinn, L., ... & Wood, D.L. (2012). The Lifelong Effects of Early Childhood Adversity and Toxic Stress. *Pediatrics*, 129 (1), 232-246.
- Strully, K. W., Rehkopf, D. H., & Xuan, Z. (2010). Effects of prenatal poverty on infant health: State Earned Income Tax Credits and birth weight. *American Sociological Review*, 75(4), 534–562. <https://doi.org/10.1177%2F0003122410374086>
- Weiss, M. J., Bloom, H. S., & Brock, T. (2014). A Conceptual Framework for Studying the Sources of Variation in Program Effects. *Journal of Policy Analysis and Management*, 33(3).

Studies That Meet Standards of Strong Causal Evidence: Paid Family Leave Example

- A. Bailey, M., Byker, T., Patel, E., Ramnath, S. (2019). *The long-term effects of California's 2004 paid family leave act on women's careers: Evidence from U.S. tax data* (No. w26416). National Bureau of Economic Research. <https://www.nber.org/papers/w26416.pdf>
- B. Baum, C. & Ruhm, C. (2016). The effects of paid leave in California on labor market outcomes. *Journal of Policy Analysis and Management*, 35(2), 333-356. <https://onlinelibrary.wiley.com/doi/epdf/10.1002/pam.21894>
- C. Bullinger, L.R. (2019). The effect of paid family leave on infant and parental health in the United States. *Journal of Health Economics*, 66, 101-116. <https://doi.org/10.1016/j.jhealeco.2019.05.006>
- D. Byker, T. S. (2016). Paid parental leave laws in the United States: Does short-duration leave affect women's labor-force attachment? *American Economic Review*, 106(2), 242-246.

- E. Choudhury, A.R. & Polachek, S.W. (2019). *The impact of paid family leave on the timing of infant vaccinations* (No. 12483). IZA Institute of Labor Economics. <https://www.iza.org/publications/dp/12483/the-impact-of-paid-family-leave-on-the-timing-of-infant-vaccinations>
- F. Das, T. & Polachek, S. W. (2015). Unanticipated effects of California's paid family leave program. *Contemporary Economic Policy*, 33(4), 619-635. <https://doi.org/10.1111/coep.12102>
- G. Hamad, R., Mordrek, S. & White, J.S. (2019). Paid family leave effects on breastfeeding: A quasi-experimental study of US policies. *American Journal of Public Health*, 109(1), 164-166. <https://doi.org/10.2105/AJPH.2018.304693>
- H. Huang, R. & Yang, M. (2015). Paid maternity leave and breastfeeding practice before and after California's implementation of the nation's first paid family leave program. *Economics and Human Biology*, 16, 45-59. <https://doi.org/10.1016/j.ehb.2013.12.009>
- I. Klevens, J., Luo, F., Xu, L., Peterson, C., & Latzman, N. E. (2016). Paid family leave's effect on hospital admissions for pediatric abusive head trauma. *Journal of the International Society for Child and Adolescent Injury Prevention*, 22(6), 442-445. <http://dx.doi.org/10.1136/injuryprev-2015-041702>
- J. Lichtman-Sadot, S. and Bell, N. P. (2017). Child health in elementary school following California's paid family leave program. *Journal of Policy Analysis and Management*, 36, 790-827. <https://doi.org/10.1002/pam.22012>
- K. Pac, J., Bartel, A., Ruhm, C. & Waldfogel, J. (2019). *Paid family leave and breastfeeding: Evidence from California* (No. w25784). National Bureau of Economic Research. <https://www.nber.org/papers/w25784>.
- L. Pihl, A. & Basso, G. (2018). Did California paid family leave impact infant health? *Journal of Policy Analysis and Management*, 38(1), 155-180. <https://doi.org/10.1002/pam.22101>
- M. Stanczyk, A.B. (2019). Does paid family leave improve household economic security following a birth? Evidence from California. *Social Service Review*, 93(2), 262-304. <https://doi.org/10.1086/703138>
- N. Rossin-Slater, M., Ruhm, C., Waldfogel, J. (2013). The effects of California's paid family leave program on mothers' leave-taking and subsequent labor market outcomes. *Journal of Policy Analysis and Management*, 32(2), 224-245. <https://doi.org/10.1002/pam.21676>
- O. Bana, S., Bedard, K., Rossin-Slater, M. (2018). *The impacts of paid family leave benefits: Regression kink evidence from California administrative data* (No. w24438). National Bureau of Economic Research. <https://www.nber.org/papers/w24438.pdf>
- P. Lee, B., Modrek, S., White, J., Batra, A., Collin, D., Hamad, R. (2020). The effect of California's paid family leave policy on parent health: A quasi-experimental study. *Social Science & Medicine*, 251, 1-8. <https://doi.org/10.1016/j.socscimed.2020.112915>
- Q. Jones, K., and Wilcher, B. (2019). *Reducing maternal labor market detachment: A role for paid family leave*. American University Working Paper. <https://ideas.repec.org/p/amu/wpaper/2019-07.html>
- R. Bartel, A. P., Rossin-Slater, M., Ruhm, C. J., Stearns, J., & Waldfogel, J. (2018). Paid family leave, fathers' leave-taking and leave-sharing in dual earner households. *Journal of Policy Analysis and Management*, 37(1), 10-37. <https://doi.org/10.1002/pam.22030>