FUNDING FOR PROGRAMS THAT WORK:

LESSONS FROM THE FEDERAL HOME VISITING PROGRAM

Philip G. Peters, Jr.[1]

[1] Ruth L. Hulston Professor of Law, University of Missouri School of Law. Former Executive Director of First Chance for Children, a nonprofit working to close the kindergarten readiness gap by, among other things, providing home visits to low income mothers with newborns.

Table of Contents

Introduction .....
Part I. Home Visiting Programs .....
A. The Appeal of Home Visiting Programs .....
B. Research on the Impact of Home Visiting Programs .....
Part II. Legislative History .....
Part III. the Act's Requirements for Design Rigor .....
A. The Preference for Randomized Trials .....
B. The Ethics of Randomization .....
C. Overlooking Synergies .....
D. Efficacy at Scale .....
E. Inability to Serve Families in Need .....
F. Conclusions about Design Rigor .....
Part IV. the Act's Outcome Requirements .....
A. Minimum Effect Size .....
B. Durability of Benefits .....
C. Replication .....
D. Salience of the Benefits Conferred .....
E. Consistency of Outcomes .....
F. Conclusions about Outcome Thresholds .....
V. Conclusion .....

ABSTRACT

Congress spends hundreds of billions of dollars each year on social programs. Many don't work. Congress and the President have called for greater reliance on evidence-based programs. Thus far, however, only one major federal program conditions state access to formula-based federal funding on the use of evidence-based practices: the Maternal, Infant, and Early Childhood Home Visiting Program.[2] *In this Article, I examine the extent to which this initial effort has succeeded and conclude that Congress has taken a promising first step, but attainment of its objective will require more demanding proof standards than those contained in the current Home Visiting Program. These weaknesses can be fixed. In this Article, I offer a roadmap for improving the program and for drafting a new generation of evidence-based federal programs that are much more likely to improve the lives of America's children and families.*

[2] Congress has enacted several smaller, competitive funding programs, but no other program conditions the receipt of formula-based funding for states on the use of evidence-based intervention models.

INTRODUCTION

Federal and state governments spend hundreds of billions of dollars every year on well-intentioned social service and education programs that are typically unproven and often ineffective.[3] There is a better way. By targeting funding to evidence-based programs, lawmakers can continue their efforts to break the cycle of poverty while spending taxpayer funds more productively. In the process, they will change the lives of children and families for the better.

Drafting these statutes will be difficult because lawmakers have no precedent or templates to build upon. In addition, the legislative process will be complicated by the certainty of strong political pressure from stakeholders to dilute the rigor of the evidentiary requirements.[4]

At present, Congress has created only one major federal program designed to restrict state access to formula-based federal funding on the use of a rigorously proven social service model--the Maternal, Infant, and Early Childhood Home Visiting Program (the Home Visiting Program).[5] In this Article, I examine the extent to which this initial effort has succeeded, concluding that Congress can indeed create social service programs that restrict funding to rigorously proven programs, but that doing so will require more demanding proof standards than those contained in the current Home Visiting Program.

The weaknesses can be repaired. In this Article, I offer a roadmap for improving the Home Visiting Program and for drafting a new generation of evidence-based federal programs that materially help the families they serve.

The stakes of this experiment with federal formula funding are enormous. Across the federal budget, over $300 billion is distributed to states annually using formula-based allocations.[6] Services range from refugee resettlement[7] to Community Development Block Grants,[8] Section 8 housing vouchers,[9] and vocational reha-

---

[3] See Our Mission, Coalition for Evidence-Based Pol'y, http://coalition4evidence.org/mission-activities/ (last visited Feb. 26, 2014); see infra note 6 and accompanying text.

[4] See infra notes 87-104 and accompanying text.

[5] Patient Protection and Affordable Care Act, Pub. L. No. 111-148 §§ 2951-2955 (signed into law March 23, 2010). The Home Visiting Program is set out in a new Section 511 of Title V of the Social Security Act. Most major steps toward evidence-based funding have involved competitive grants in which funding is not guaranteed to each willing state (e.g. Investing in Innovation (i3) grants) or grants aimed at nonprofits (e.g. the Social Innovations Fund).

[6] See U.S. Gov't Accountability Office, GAO-09-832T, Formula Grants: Census Data Are among Several Factors That Can Affect Funding Allocations 1 (Statement of Robert Goldenkoff, Director of Strategic Issues) (2009) [hereinafter GAO].

[7] See Release of FY2013 Social Services and Targeted Assistance Formula Allocations, U.S. Dep't Health & Human Servs., http://www.acf.hhs.gov/programs/orr/news/release-of-fy2013-social-services-and-targeted-assistance-formula.

[8] See Community Planning and Development Program Formula Allocations for FY 2013, U.S. Dep't Hous. & Urban Dev., http://portal.hud.gov/hudportal/HUD?src=/program_offices/comm_planning/about/budget/budget13.

[9] GAO, supra note 6, at 4.

bilitation services for people with disabilities.[10] In education alone, over $15 billion dollars were allocated in Fiscal Year (FY) 2014 on a formula basis for Title I assistance for low income schools and another $2.3 billion for Title II teacher quality programs.[11] If even a portion of that funding could be reserved for proven practices, the lives of millions of Americans could be dramatically improved.[12]

Home visiting programs constitute an ideal subject for this experiment because parent education programs are both widely supported and poorly proven. On the one hand, programs that offer home visits to at-risk mothers with young children make sense. On the other hand, a substantial body of research reveals that most home visiting programs fail to provide children with any detectable benefit. Given this evidence, explored in Part I below, a sensible funding program must distinguish between programs that are effective and those that are not.

Congress accepted this challenge when it created the Home Visiting Program as part of the Affordable Care Act (the "Act"). The Act authorized $1.5 billion in funding over five years for states to create evidence-based home visiting programs.[13]

The initial challenge came from an unexpected direction. As explained in Part II, many existing home visiting programs and respected children's advocates strongly opposed tough standards for proving a program's effectiveness.[14] Ultimately, Congress compromised on the issue of design rigor and only a remarkable rescue from the Department of Health and Human Services (HHS) preserved the law's focus on reliably proven programs.[15]

Part III evaluates the research reliability requirements currently imposed under the Act. Although they were criticized as unfairly demanding, they are actually the key strength of the Home Visiting Program. Unfortunately, neither Congress nor HHS imposed similarly meaningful minimum requirements upon program *outcomes*, such as magnitude of impact, consistency of findings, durability of impact, importance of benefits conferred, or replication of positive outcomes. As a result, many of the approved programs have threadbare or troublingly inconsistent evidence of positive impact. As explained in Part IV, these omissions can and should be cured through carefully crafted thresholds that will greatly increase the odds that federally funded programs meaningfully improve children's lives.

---

[10] GAO, supra note 6, at 3.

[11] See U.S. Dep't Educ., Fiscal Year 2014 Congressional Action Table, http://www2.ed.gov/about/overview/budget/tables.html?src=ct (showing Title I at row 73, col. T; Title II at row 118, col. T).

[12] The i3 program is a major step in the right direction because it funds replication at scale of proven models and requires rigorous assessment of the outcomes. However, local school districts still have discretion to use other models for their Title I programs. See Robert E. Slavin, Baby Steps Toward Better Formula Grants in Education, Huffington Post (June 27, 2013), available at http://www.huffingtonpost.com/robert-e-slavin/baby-steps-toward-better_b_3509165.html.

[13] Patient Protection and Affordable Care Act, Pub. L. No. 111-148 §§ 2951-2955 (signed into law March 23, 2010). The Home Visiting Program is set out in a new Section 511 of Title V of the Social Security Act.

[14] See infra notes 95-104 and accompanying text.

[15] Congress also failed to impose any minimum thresholds for the size, durability or consistency of outcomes and HHS could not fix that mistake. See infra Part IV.

With the Home Visiting Program, Congress took an important first step. The law's insistence on highly reliable evidence of effectiveness provides an outstanding template for future state and federal funding legislation. But the work is not yet done. To improve the odds that funded programs will improve children's lives, Congress should impose meaningful minimum outcome requirements as well when it reauthorizes funding.[16] With this change, the Home Visiting Program will offer a promising template for future efforts to fund programs that really work.

PART I. HOME VISITING PROGRAMS

Low income and minority children commonly arrive at kindergarten far behind their classmates.[17] Because infants and toddlers spend most of their time in the care of their parents, common sense suggests that helping parents during the early childhood years would be a way to help their children. To accomplish this, states across the country fund a wide variety of programs that send social workers, educators and paraprofessionals to the homes of mothers with newborns or toddlers.[18]

Yet, studies show that many existing delivery models do not change either parenting practices or child development.[19] As a result, the field is well-suited for a government program that restricts funds to programs with proven effectiveness.

The Appeal of Home Visiting Programs

The achievement gap surfaces among infants and toddlers with shocking speed. By the age of eighteen months, researchers regularly detect diverging trajectories between poor and upper income children[20] and between white and black children. Similar disparities surface in the development of early social skills.[21]

---

[16] In 2014, Congress continued funding for an additional year. See Protecting Access to Medicare Act of 2014, Pub. L. No. 113-93 (Apr. 1, 2014) (reauthorizing the Home Visiting Program only until March 31, 2015).

[17] See infra notes 20-24 and accompanying text.

[18] See infra notes 45-47 and accompanying text.

[19] See infra notes 49-65 and accompanying text.

[20] See Tamara Halle et al., Disparities in Early Learning and Development: Lessons from the Early Childhood Longitudinal Study—Birth Cohort (ECLS-B), 4 (2009) (Executive Summary), available at http://www.childtrends.org/wp-content/uploads/2013/05/2009-52DisparitiesELExecSumm.pdf. At nine months, children from families with an income at or below 200 percent of the poverty were 0.16 standard deviations below the mean of their higher-income peers on cognitive skills and by twenty-four months, the gap had grown to more than half a standard deviation (about half the adult gap). Id. See also Betty Hart & Todd R. Risley, The Early Catastrophe: The 30 Million Word Gap, 27 Am. Educator 4 (2003) (finding that, by age three, children of professional parents have as large a vocabulary as parents in low income families and much larger than the children in those families).

[21] See Jeanne Brooks-Gunn et al., Racial and Ethnic Gaps in School Readiness, in School Readiness and the Transition to Kindergarten in the Era of Accountability 283, 286 (Robert C. Pianta et al. eds. 2007) (white kindergarteners scored higher than blacks on approaches to learning and self-control by 0.36 and 0.38 of a

When children reach kindergarten, the average poor or minority child is half a standard deviation or more behind the mean on academic and social skills and farther behind on vocabulary.[22] That puts the average poor kindergartener at about the 32nd percentile of her more affluent classmates. Sadly, these skills are a good predictor of future school success. Early vocabulary, for example, is strongly associated with later school performance.[23] Yet, black kindergarteners on average have only half the vocabulary of white kindergarteners.[24]

Children who lag on achievement tests during their preschool years are more likely than their higher-performing classmates to be retained in grade, be placed in special education classes, and drop out of school. Even more tragically, they are more likely to become teen parents, engage in criminal activities, and suffer clinically significant depression.[25]

Because the gaps appear early in life, early intervention seems essential. Our experience as parents tells us that "babies are born learning and that parents are their first and most influential teachers."[26] Child development experts widely agree.[27] Psychologist Edward Zigler, a giant in the field and a force behind the creation of Head Start, argues that waiting for preschool at age four is a "huge mistake" because the "first nine months are the most critical."[28] Kindergarten readiness, ac-

---

standard deviation, respectively, and lower on externalizing behaviors by 0.31); see also Halle et al., supra note 20, at 3, Figure 3 (at nine months, the income-rated difference in positive behavior shown by infants and toddlers was 0.16 standard deviations, and by 24 months, the disparity had grown to 0.30).

[22] Estimates of the academic gap between blacks and whites at kindergarten range from around 0.5 standard deviations to over 1.0 standard deviations. A one standard deviation gap would put the black mean at the 16th percentile of the white scores. For a low estimate, see Roland G. Fryer, Jr. & Steven D. Levitt, The Black-White Test Score Gap Through Third Grade, 8 Am. L. & Econ. Rev. 249, 256, 262-63 (2006) (using the EKLS-K data base and finding a kindergarten racial achievement gap of 0.66 standard deviations in math and 0.40 in reading). At the higher end, see, e.g., Christopher Jencks & Meredith Phillips, The Black-White Test Score Gap: An Introduction, in The Black-White Test Score Gap 1-2 (Christopher Jencks & Meredith Phillips eds., 1998) (Figure 1-1) (showing 12 point gap with standard deviation of 10 for three and four-year-olds). See also Meredith Phillips et al., Family Background, Parenting Practices, and the Black-White Test Score Gap, in The Black-White Test Score Gap, Id. at 103, 106 ; Brooks-Gunn, et al., supra note 21, at 285 (reporting that on verbal ability and general cognition, black preschoolers score more than one standard deviation (SD) below white preschoolers, when differences are not adjusted for social, economic, and family background differences); Donald A. Rock & A. Jackson Stenner, Assessment Issues in the Testing of Children at School Entry, 15 Future of Children, Spring 2005, at 15, 15. ("On average, the tests find a gap of about 1 standard deviation."); Richard J. Murnane et al., Understanding the Trends in the Black-White Achievement Gaps During the First Years of School, in Brookings-Wharton Papers on Urban Affairs 97, 109 (Brookings Inst. Press 2006) (reporting a kindergarten race gap of 1.1 standard deviations in mathematics and 1.0 in reading).

[23] Valerie E. Lee & David T. Burkam, Inequality at the Starting Gate 8 (2002).

[24] See Rock & Stenner, supra note 22, at 19 (citing George A. Miller & Patricia M. Gildea, How Children Learn Words, 257 Scientific Am., no. 4, 1987, at 1, 94-97).

[25] See Cecilia Rouse et al., Introducing the Issue, 15 Future of Children, Spring 2005, at 6, 6 n.5.

[26] Mary M. Wagner & Serena L. Clayton, The Parents as Teachers Program: Results from Two Demonstrations, 9 Future of Children, Spring 1999, at 91, 92.

[27] See id.; Edward Zigler et al., The Parents as Teachers Program and School Success: A Replication and Extension, 29 J. Primary Prevention 103, 104 (2008) (stating that home visiting programs are "theoretically" sound).

[28] (Press Release, Pa. Info. & Res. Ctr.), Parents as Teachers Program Once Again Shown to Improve School Readiness, Pa. PIRC (Pa. Info. & Res. Ctr.), (Fall 2008) (on file with author).

cording one team of researchers, is a more important predictor of third-grade achievement than other variables such as poverty and minority status.[29] As a result, both theoreticians and policy advocates "believe strongly that home visiting can be a beneficial and cost-effective strategy."[30]

Researchers have regularly found significant correlations between child readiness for school and parental behaviors such as reading to their children, engaging them in conversations, giving positive reinforcement, and sharing a strong mother-child bond.[31] In one review of the literature, Brooks-Gunn and Markman identified seven parenting practices that improved child well-being: (1) expressing love, affection, warmth, and care, rather than detachment, intrusiveness and negative regard (nurturance); (2) language use; (3) avoidance of harsh disciplinary practices like spanking, slapping or yelling; (4) having materials in the home; (5) monitoring; (6) management of the home; and (7) direct teaching of skills like tying a shoe or sorting blocks by color, and asking questions that encourage the children to find the answer, rather than providing the answers themselves, such as asking "what would happen if you turned that puzzle piece around" or "can you find all the pieces that go on the edge of the puzzle?"[32]

The researchers found that different areas of home life are associated with different aspects of school readiness.[33] Discipline and nurturance tie most closely to behavior and attention, while language and learning materials tie most closely to vocabulary and early school achievement.[34] The authors also found that "[w]hen researchers measuring school readiness gaps control for parenting differences, the racial and ethnic gaps narrow by 25-50 percent."[35] The poverty gap narrows significantly as well.[36]

Backed by this body of research, home visiting programs have received very strong support from many quarters. Conservatives like former Republican Senator Kit Bond liked the idea of helping parents be their child's first teacher.[37] Liberal nonprofits like the Center for Law and Social Policy[38] and the Pew Foundation[39]

---

[29] See Pa. PIRC, supra note 28, at 1; Zigler et al., supra note 27, at 113.

[30] Kimberly S. Howard & Jeanne Brooks-Gunn, The Role of Home-Visiting Programs in Preventing Child Abuse and Neglect, 19 Future of Children, Fall 2009, at 119, 119, 138 (stating that they do so despite questions about efficacy).

[31] See, e.g., Jeanne Brooks-Gunn & Lisa B. Markman, The Contribution of Parenting to Ethnic and Racial Gaps in School Readiness, 15 The Future of Children, Spring 2005, at 139, 140-43; Ross A. Thompson, The Roots of School Readiness in Social and Emotional Development, 1 The Kauffman Early Educ. Exchange 8, 9-10 (2002) (securely attached children do better in school).

[32] See Brooks-Gunn & Markman, supra note 31, at 141-143.

[33] See A Conversation with Jeanne Brooks-Gunn, 10 The Evaluation Exchange, Winter 2004-05, at 12, 12-13.

[34] See id.; Brooks-Gunn & Markman, supra note 31, at 144 (warmth would not be expected to increase vocabulary in the absence of more talking).

[35] Brooks-Gunn & Markman, supra note 31, at 157 (citing evidence that parenting explained 1/5 to 1/4 of the gap, after controlling for parent education, income and mother's test scores).

[36] Id. (citing research that parenting practices account for 1/3 to 1/2 of the poverty achievement gap).

[37] Senator Bond created the Parents as Teachers program as governor of Missouri and sponsored federal legislation to fund home visiting in the U.S. Senate for many years. See infra notes 71-80 and accompanying text.

[38] See generally Child Care & Early Education, Clasp, http://www.clasp.org/issues/child-care-and-early-education (last accessed Mar. 3, 2014). Although it focuses more heavily on early childcare than on parent

hoped to break the cycle of poverty by helping at-risk children arrive at school more ready to succeed.

In an influential book collecting data on the racial achievement gap, Brooks-Gunn and Markman said "[c]hanging the way parents deal with their children may be the single most important thing we can do to improve children's cognitive skills."[40] The U.S. Advisory Board on Child Abuse and Neglect concluded that "no other single intervention has the promise that home visitation does."[41] This was followed by endorsements from the American Academy of Pediatrics, the Task Force on Community Preventive Services, the National Academy of Sciences, the National Governors Association and the World Health Organization.[42]

Support is not universal; some critics fear the risk of cultural imperialism.[43] Nonetheless, support for parent education and support programs remains extraordinarily widespread.[44]

As a result, home visiting programs have grown steadily in number over the past few decades. By the fall of 2009, just before the new federal home visiting program was enacted, home visitation programs operated in all 50 states and the District of Columbia.[45] Total funding from private and public sources was estimated to fall between $750 million and $1 billion annually, supporting home visits for an estimated 400,000-500,000 families.[46] Although home visiting programs differ widely in their operations and target an array of child and family outcomes ranging from child development to family self-sufficiency, they are bound together by the belief that home visits will improve parenting practices and, in this way, enhance child development.[47]

Research on the Impact of Home Visiting Programs

The case for investing public funds in home visiting programs was bolstered by

---

education, CLASP lobbied heavily to get a federal home visiting bill that fit its priorities. See infra note 97 and accompanying text.

[39] See generally Home Visiting Campaign, Pew St. & Consumer Initiatives, Pew, http://www.pewstates.org/projects/home-visiting-campaign-328065 (last accessed Mar. 3, 2014).

[40] Jencks & Phillips, supra note 22, at 46.

[41] Deborah Daro, Home Visitation: Assessing Progress, Managing Expectations 4 (2006), available at http://www.chapinhall.org/sites/default/files/old_reports/323.pdf.

[42] J. H. Filene et al., Components Associated With Home Visiting Program Outcomes: A Meta-Analysis, 132 Pediatrics S100, S101.

[43] See Lisa Delpit, Other People's Children: Cultural Conflict in the Classroom, 30 (1995) ("I do not advocate that it is the school's job to attempt to change the homes of poor and nonwhite children to match the home of those in the culture of power. That may indeed by a form of cultural genocide.").

[44] See Emilie Stoltzfus & Karen E. Lynch, Cong. Research Serv., R40705, Home Visitation for Families with Young Children (2009) (noting a current "phase of broad popularity").

[45] Id. at Summary.

[46] Id. (about 3% of all families (17.4 million) with children under six years of age).

[47] See Howard & Brooks-Gunn, supra note 30, at 120.

early evaluations of programs like Healthy Families America and Parents as Teachers, which showed significant improvements in parenting and child development.[48] In 1995, however, serious questions were raised about the actual benefits.[49] A review of the research by Steven Barnes and his colleagues concluded "there is little research evidence to support the assumption that parent services affect child outcomes."[50] Four years later, Deanna Gomby and her colleagues concluded that most of the studied programs provided no significant benefits for a majority of the developmental domains measured and many showed no positive benefits at all.[51] The authors called the rarity of proven gains "sobering"[52] and concluded that "children's development is better promoted through more child-focused interventions [like preschools]."[53]

Jean Layzer and her colleagues were equally pessimistic in a 1999 report to the Department of Health and Human Services.[54] While family support services in general had a small positive impact,[55] home visiting programs in particular did not. The authors concluded:

> The assumption that parents lack the necessary skills to be effective teachers of their children has led to the widespread use of parenting education in family support programs. There is no evidence of its effectiveness in promoting children's cognitive development. Nor is it clear that adding parent education to direct services to children confers an additional benefit.[56]

Then Dr. David Olds created the Nurse-Family Partnership (NFP) program in Elmira, New York and replicated it in Memphis and Denver. NFP was evaluated in multiple randomized trials and consistently produced sizeable gains in child development among high risk families.[57] Follow-up studies found that gains lasted into

---

[48] See infra notes 149-50 and accompanying text (describing the early research on PAT).

[49] See Howard & Brooks-Gunn, supra note 30, at 139.

[50] Wagner & Clayton, supra note 26, at 112 (quoting H.V. Barnes et al., 1 National Evaluation of Family Support Programs: Review of Research on Supportive Interventions for Children and Families 3-17 (1988)).

[51] Deanna S. Gomby et al., Home Visiting: Recent Program Evaluations – Analysis and Recommendations, 9 Future of Children, Spring 1999, at 4, 12.

[52] Id. at 6.

[53] Id. at 22 (looking at programs without child care components). ("[A]ny new expansion of home visiting programs should be reassessed in light of these findings.") Id. at 24. ("Intensive universal home visiting probably will not lead to broad benefits."). Id. at 21 She reached a similar conclusion in a 2005 article. Deanna S. Gomby, Home Visitation In 2005: Outcomes For Children And Parents 24 (2005) ("Most meta-analyses and literature reviews offer one clear conclusion: large benefits in children's cognitive development are most likely when services focus directly on the child, and not when they rely upon parents to intervene with the child, as most home visiting programs do."). See also Brooks-Gunn & Markman, supra note 31, at 153-54 (contrasting home visiting with center-bases early childhood programs).

[54] Jean I. Layzer et al., National Evaluation of Family Support Programs, Final Report, Vol. A: The Meta-Analysis (2001).

[55] See id at A5-42.

[56] Id. at A5-43 (also noting that adding parent education to preschool programs had not been proven to be effective).

[57] See, e.g., David L. Olds et al., Effects of Nurse Home-Visiting on Maternal Life Course and Child Devel-

adolescence.[58] NFP's success generated great enthusiasm and almost single-handedly widened support for home visiting programs.[59]

Yet, recent literature reviews have sounded only slightly more optimistic about the value of other home visiting programs. In 2009, Kimberly Howard and Jeanne Brooks-Gunn found some evidence that home visiting programs could improve maternal parenting practices and "to a lesser extent" children's cognitive development.[60] The pattern of effectiveness was mixed. Some studies showed impact, but others did not; programs typically affected one outcome but not others and often only for a particular subset of families.[61] They found that the literature is "somewhat conflicting regarding essentially every outcome under study."[62] Overall, demonstrated cognitive gains for children were the exception rather than the rule.[63] Separately, Gomby concluded that the popularity of home visiting "has been driven by the results of a few studies of programs such as the Nurse-Family Partnership."[64]

The response to these disappointing findings has been quite revealing. On the one hand, the authors of the reviews commonly call for reduced expectations[65] and point out the weak research methodologies typically used to evaluate programs in this field, such as the lack of randomized trials.[66] On the other hand, most early childhood experts continue to believe that high quality parent education can change children's lives. Despite concluding that serious questions exist about both short and long term benefits, Howard and Brooks-Gunn insisted that "the evidence base suggests much more strongly [than ever] the important benefits of home-visiting programs for parents and children."[67] They noted that leaders in the field generally agreed, stating that "despite questions about the short- and long-term benefits of home visiting, theorists and policy makers alike believe strongly that it can be a beneficial and cost-effective strategy for providing services to families and chil-

---

opment: Age 6 Follow-Up Results of a Randomized Trial, 114 PEDIATRICS 1550 (2004) [hereinafter Age 6 Follow-Up]

[58] David L. Olds et al., Prenatal and Infancy Home Visitation by Nurses: Recent Findings, 9 Future of Children 44, 56 (1999) (reviewing outcomes of the adolescent children of poor unmarried mothers who received NFP).

[59] See Gomby, supra note 54, at 2 ("The popularity of home visiting has been driven by the results of a few studies of programs such as the Nurse-Family Partnership.").

[60] Howard & Brooks-Gunn, supra note 30, at 138 ("After nearly another decade of research [since the 1999 Gomby review], many concerns remain, but the evidence base suggests much more strongly the important benefits of home-visiting programs for parents and children.").

[61] See id. at 133-34, Table 2.

[62] Id. at 128.

[63] See id. at 133-34, Table 2 (showing most of the reviewed models had no impact unless combined with child care).

[64] Gomby, supra note 54, at 2.

[65] See, e.g., Jennifer Astuto & LaRue Allen, Home Visitation and Young Children: An Approach Worth Investing in?, 25 Soc'y for Res. in Child Dev. 3, 5 (2009) (calling for caution in expectations); Daro, supra note 41, at 12 ("manage expectations"); Gomby, supra note 54, at 2 ("funders should maintain modest expectations for what home visiting alone can accomplish"); Howard & Brooks-Gunn, supra note 30, at 138 ("important to recognize the limits).

[66] See Astuto & Allen, supra note 65, at 5.

[67] Howard & Brooks-Gunn, supra note 30, at 138.

dren."[68]

These divergent views illustrate the need for evidence-based funding.[69] While some delivery models have produced positive outcomes, many have not. Rather than make an all-or-nothing decision about funding home visiting programs, Congress wisely chose to restrict funding to the models with proven impact.

PART II. LEGISLATIVE HISTORY

The intuitive appeal of parent education programs, along with mountains of favorable correlational research, prompted several federal lawmakers to work persistently for a decade to enact a federal funding program. Support was bipartisan and ultimately culminated in adoption of the Home Visiting Program. Its legislative history highlights a central challenge for lawmakers who wish to direct funding toward evidence-based programs. Both lawmakers and lobbyists have widely differing beliefs about what it means for a program to be "evidence-based."

In 2004, Republican Senator Christopher Bond of Missouri, along with Republican Senator Jim Talent, also of Missouri, and Democratic Senator Dick Durbin of Illinois, introduced the first iteration of the Education Begins at Home Act.[70] As governor of Missouri, Bond had helped create a state-funded Parents as Teachers (PAT) home visiting program. His bill would provide federal funding to states for the creation or expansion of a PAT program or "other programs of early childhood home visitation."[71]

In the next session of Congress, the 109th, Bond reintroduced his bill, but this time his bill made its first explicit reference to program quality. Funding would be limited to "quality" programs that were "research-based."[72] In the House, Representative Danny Davis (D-IL) introduced a different version of the Education Begins at Home Act that contained an important additional criterion. [73] The Secretary of Health and Human Services (HHS) was instructed to treat state grant applications more favorably if, as part of their grant evaluation, they "incorporate[d] comparison or control groups in their service delivery model."[74] Under this version of the Act, the grants would serve a dual purpose; they would provide home visiting services

---

[68] Id. (noting impact on parenting and cognitive development); see also Astuto & Allen, supra note 65, at 5 (noting the split in responses to the current evidence).

[69] See Richard V. Reeves & Kimberly Howard, The Parenting Gap 13 (2013) (concluding that "the right question is not whether parenting programs work but rather which parenting programs work").

[70] Education Begins At Home Act, S. 2412, 108th Cong. (introduced May 12, 2004).

[71] Id.

[72] Education Begins at Home Act, S. 503, 109th Cong. § 5(f)(1) (introduced Mar. 3, 2005).

[73] Education Begins at Home Act, H.R. 3628, 109th Cong. § 5(f)(1) (introduced July 29, 2005). The House version mandated that programs be "grounded in empirically-based knowledge related to home visiting and linked to program-determined outcomes." Id. Both the House and Senate versions would have provided $500 million over three years for expansion of state home visiting programs. Id. at § 5(b)(5). See also S. 503 at § 5(b)(3).

[74] H.R. 3628 at § 5(d)(1).

while at the same time producing outcomes research that would inform future funding decisions. The bill implicitly acknowledged the need to find more home visiting models with measurable results. Neither the Senate bill nor the House bill passed.

Bond and Davis tried again in the 110th Congress.[75] Bond strengthened his 2007 Senate bill by borrowing the language in the 2005 House bill that had expressed a preference for the use of comparison groups to evaluate funded programs.[76] Meanwhile, Davis replaced that requirement with an even tougher one in the 2007 House bill, which required *prior* testing of all funded home visiting programs. States could only use home visiting models with at least one study published in a peer reviewed journal showing positive outcomes.[77] Once again, the bills failed to pass.[78]

Both bills were reintroduced in 2009 essentially unchanged,[79] but they would be eclipsed by a proposal from the Obama administration. In April 2009, President Obama announced a proposed budget for FY 2011 that included $8.5 billion over ten years for home visiting.[80] The White House proposed to create a "Nurse Home Visitation program, which would provide funds to States to provide home visits by trained nurses to first-time low-income mothers and mothers-to-be." [81]

The program would have two tiers. Primary funding would be targeted at "home visitation models that have been rigorously evaluated and shown to have positive effects on crucial outcomes for children and families."[82] As a result, only the most clearly proven programs would be eligible for the primary pool of funds. The obvious target here was the Nurse Family Partnership (NFP) home visiting program. NFP had produced dramatic gains for at-risk children in a series of highly rigorous studies using randomized controlled trials.[83] The administration insisted that any other funded programs have equally rigorous proof of effectiveness.

A much smaller pool of funds would be provided for states to implement

---

[75] S. 667, 110th Cong. (introduced Feb. 16, 2007); H.R. 2343, 110th Cong. (introduced May 16, 2007).

[76] H.R. 2343, 110th Cong. § 5(f)(1)(A).

[77] Id. at § 4(b)(5). It is not clear whether the requirement of a "research basis" applies to the content of the program, to the programs outcomes, or both. The bill authorized $150 million for the first of five years. Id.

[78] However, the Bush Administration noted that states did not always follow "proven-effective" models of home visitation and requested $10 million for FY 2008 to help states funnel existing resources into proven models. Congress approved, but stipulated that HHS "ensure that States use the funds to support models that have been shown, in well-designed randomized controlled trials, to produce sizeable, sustained effects on important child outcomes such as abuse and neglect." Emilie Stoltzfus & Karen Lynch, Cong. Research Serv., Home Visitation for Families with Young Children 13 (2009).

[79] *S. 244, 111th Cong. (introduced Jan. 14, 2009)*; *H.R. 2205, 111th Cong. (introduced Apr. 30, 2009). The Senate bill was introduced by Senators Bond, Murray, and Clinton.*

[80] See e.g. Jennifer Astuto and LaRue Allen, Home Visitation and Young Children: An Approach Worth Investing, 23 Social Policy Report, no. 4 (2009) at 5; Ron Haskins et al., Commentary, Home Visiting Programs: An Example of Social Science Influencing Policy, 23 Social Policy Report, no. 4 (2009) at 7 [hereinafter Social Science Influencing Policy].

[81] Office of Mgmt. & Budget, A New Era of Responsibility: Renewing America's Promise 70 (2009) [hereinafter "OMB Statement for FY2010"].

[82] Astuto & Allen, supra note 81, at 5 (quoting from a White House press release).

[83] See infra notes 58-59 and accompanying text.

"promising programs." These were programs with promising preliminary outcomes but insufficient reliable evidence to qualify as evidence-based.[84] When describing this second tier of funding, Peter Orzag, Director of the Office of Management and Budget and a close advisor of the president, said "[l]et's try those too, but rigorously evaluate them and see whether they work. Over time, we hope that some of those programs will move into the top tier — but, if not, we'll redirect their funds to other, more promising efforts."[85]

Early childhood insiders realized immediately that the administration's program was intended to expand the Nurse-Family Partnership program. Only the NFP program had a track record of large, durable, and consistently positive effects in a series of randomized controlled trials.[86] Although the Obama proposal did not mention the Nurse Family Partnership by name, the thresholds it imposed implicitly singled out NFP. For example, the proposal targeted nurses, rather than the wide array of early educators and social workers who commonly provide home visiting services.[87] In explaining the rationale for this funding, the administration also relied on the NFP research, arguing that the program it wanted to expand "has been rigorously evaluated over time and proven to have long-term effects" and that it had conferred a "return-on-investment [of] between $3 and $6 per dollar invested."[88] Both references were clearly to NFP.[89]

The administration's belief that NFP was the best of the existing home visiting programs was widely shared by experts in the field.[90] The Nurse Family Partnership program had produced durable positive impacts in a series of randomized controlled trials in three different cities serving populations with very different demographics.[91] Because its creator, Dr. David Olds, believed that the program's success was tied very closely to faithful implementation of the model, he had only

---

[84] Astuto & Allen, supra note 81, at 5 (quoting a White House web page, "Additional funds will be available for promising programs based on models with experimental or quasi-experimental research evidence of effectiveness that will be rigorously tested to assess their impact."). Peter Orzag endorsed a two-tier approach that would give most of the funding to rigorously proven programs and a smaller portion to promising programs with "some supportive evidence, but not as much" to develop and rigorously assess their model. Peter Orszag, Building Rigorous Evidence to Drive Policy, Off. Of Mgmt. & Budget (June 8, 2009, 8:39 AM), http://www.whitehouse.gov/omb/blog/09/06/08/BuildingRigorousEvidencetoDrivePolicy/ (last visited June 20, 2011).

[85] Orszag, supra note 85.

[86] Social Science Influencing Policy, supra note 81, at 7 (noting three different clinical trials).

[87] OMB Statement for FY2010, supra note 82, at 70.

[88] Id.

[89] Social Science Influencing Policy, supra note 81, at 7 (laying out specific references to the model's features). As of the February 2009 budget blueprint, Obama recommended up to $8 billion over ten years for a nurse home-visiting program targeting low income mothers in order to boost child development. Id. Ron Haskins et al., Policy Brief, Social Science Rising: A Tale of Evidence Shaping Public Policy, Policy Brief, The Future of Children, Fall 2009, at 1, 1-3 [hereinafter Social Science Rising] (quoting the budget as calling for a "Nurse Home Visitation" program). The surrounding language left no doubt that the program had the Nurse-Family Partnership model in mind. Id. at 3 (laying out references to the model's features, such as its limitation to "first-time mothers and mothers-to-be").

[90] Social Science Rising, supra note 90, at 2-3.

[91] Id. at 3.

cautiously allowed replication in other cities and only with close oversight.[92] The combination of dramatic outcomes in multiple randomized trials and careful fidelity to the proven model generated many supporters. The White House was among them.[93]

As a result, many early childhood advocates and supporters of other well-established home visiting models feared that the proposed pool of federal funding would not be available to expand any of the other very popular home visiting models used across the country—programs which early childhood advocacy groups had been supporting for years, like Health Families America (HFA), Home Instruction for Parents of Preschool Youngsters (HIPPY) and Parents as Teachers (PAT).[94]

The early childhood community responded swiftly. Major home visiting programs, like Parents as Teachers, Healthy Families America, and HIPPY USA were shocked to have been left out.[95] The slighted programs were members of influential and effective national coalitions and child advocacy groups. Acting very quickly, they enlisted other coalition members, like the Center for Law and Social Policy and the Children's Defense Fund, to lobby on their behalf.[96] They argued that every home visiting program with evidence of effectiveness should be eligible for the primary tier of funding even if they lacked highly rigorous studies to confirm that impact.[97]

On April 21, 2009, four respected scholars from the field of early childhood wrote a letter to President Obama arguing that it was unwise to single out one program for support.[98] Other commentators noted that narrowly targeting a delivery model that served only first time, low-income mothers who enrolled prenatally or very soon after delivery would leave many at-risk families unserved. Cognizant that most of the other models could not base their claims of effectiveness on the results of randomized trials, the critics also argued that proof of effectiveness should not be limited to randomized trials.[99] Randomized trials "provide too little guidance on how to replicate the model at sufficient scale to serve the national interest."[100]

Worried that Congress would water down its evidentiary standards, the Coalition for Evidence-Based Practice strongly urged the President and Congress to pri-

---

[92] Id. ("Seldom has an intervention program been so carefully tested and expanded with such serious attention to getting new sites to maintain fidelity to the program model."). By 2008, there were NFP programs, often small, in twenty-five states. Id.

[93] Id.

[94] Id. (mentioning Healthy Families America, the Parent-Child Home Program, and HIPPY-USA).

[95] Social Science Influencing Policy, supra note 81, at 7 (listing Parents as Teachers, Healthy Families America, the Parent Child Home Program, and HIPPYUSA as disappointed suitors).

[96] Social Science Rising, supra note 90, at 3.

[97] Social Science Influencing Policy, supra note 81, at 7.

[98] Social Science Rising, supra note 90, at 4.

[99] Letter from Deborah Daro et al., Research Fellow, Univ. of Chi., to President Barack Obama (Apr. 21, 2009) (hereinafter "Letter from Deborah Daro et al.) (on file with the author). See also Astuto & Allen, supra note 66, at 6; Social Science Rising, supra note 81, at 4 (calling other models "highly respected")

[100] Letter from Deborah Daro et al, supra note 100, at 1. Supporters of expanded eligibility would later argue that randomized trials can also be unethical, depriving children assigned to the control group of beneficial services. See infra notes 163-68 infra and accompanying text.

oritize randomized clinical trials (RCTs).[101] It noted that "many existing home visitation models produce weak or no effects on key child outcomes" when tested by RCTs.[102] The Coalition also warned that offering more money to existing programs based on "a diluted evidence standard . . . is unlikely to do much good, and may miss an opportunity to fundamentally improve life outcomes for millions of children born into disadvantaged backgrounds."[103]

Several competing home visiting bills were filed that spring in the House and Senate. From them, Congress enacted a home visiting program using language contained in a bill drafted by Senator Max Baucus. On October 19, 2009, Baucus finished marking up his master health care reform bill.[104] Subtitle I of that bill contained the home visiting program that would eventually become law. It incorporated many of the concepts that had surfaced in prior legislative proposals. Its most important provisions provided that:

Programs would target at-risk pregnant women and children.[105]

Funding would be divided into two tiers based on the quality of the evidence of effectiveness. States would have to spend at least seventy-five percent of their funds on programs with proven effectiveness.[106] Up to twenty-five percent could be used to implement a model which had considerable promise, but had yet to be subjected to rigorous evaluation.[107]

States receiving funding for implementation of a promising, but not yet proven, model would have to assess the program using highly reliable research design.

Each state receiving a grant would be required to make an annual report of its progress, including data on a number of specific "benchmarks" of progress.[108] In addition, a program evaluation would have to be done.

Although Baucus's health reform bill was not enacted, its home visiting provisions were incorporated virtually unchanged on November 19, 2009, as Subtitle L of Title II of another Senate bill, the Patient Protection and Affordable Care Act (also known as the Affordable Care Act).[109] President Obama signed that bill on

---

[101] Coalition For Evidence-Based Policy, Early Childhood Home Visitation: Effectiveness of a National Initiative Depends Critically on Adherence to rigorous Evidence About "What Works," Coalition Policy Proposal (2009), available at http://coalition4evidence.org/wordpress/?page_id=468 [hereinafter Effectiveness Depends].

[102] Id, at 2.

[103] Id.; see also Social Science Rising, supra note 90, at 4.

[104] America's Healthy Future Act of 2009, S. 1796, 111th Cong. (2009).

[105] Id. at § 1801 (d)(4). See also Diane Paulsell et al., Home Visiting Evidence of Effectiveness Review: Executive Summary, Office of Planning, Research & Evaluation, Admin. For Children and Families, United States Dep't of Health & Human Servs., Nov. 2010, at 1.

[106] America's Healthy Future Act §1801 (d)(3)(A)(ii)

[107] Id.

[108] America's Healthy Future Act §1801 (e)(8)(A).

[109] Patient Protection and Affordable Care Act, H.R. 3590, 111th Cong. §§ 2951-2955 (signed into law March 23, 2010), available at http://www.govtrack.us/congress/bill.xpd?bill=h111-3590. When it left the House, the bill was called the Service Members Home Ownership Tax Act of 2009. Service Members Home Ownership Tax Act of 2009, H.R. 3590, 111 Cong. (2009). The Senate substituted the health care bill for the original text. It was amended by the Senate on November 19 and the bulk of the health reform legislation was added at

March 23, 2010.[110] As a result, a new Section 511 of Title V of the Social Security Act created the Maternal, Infant, and Early Childhood Home Visiting program and authorized $1.5 billion in funding over five years for states to create evidence-based home visiting programs.[111]

Later, HHS announced that roughly half of these funds would be allocated on a formula basis to states that agreed to use an evidence-based home visiting model[112] and the rest would be awarded through a competitive process in which the evidence supporting a state's chosen model would be taken into account.[113]

PART III. THE ACT'S REQUIREMENTS FOR DESIGN RIGOR

As enacted, the Home Visiting Program allowed states to select among home visiting models whose effectiveness had been demonstrated in either randomized controlled trials or quasi-experimental studies.[114] While this was a promising start, some RCTs suffer from weaknesses, such as high attrition, that render their findings suspect. Quasi-experimental studies are even more vulnerable to factors, like selection bias, that can skew their results. As a result, HHS announced that it would use the discretion conferred on it by Congress to more specifically identify the kinds of RCTs and quasi-experimental designs (QEDs) that would be taken into account.[115]

Lobbying immediately shifted from Congress to HHS. On July 23, 2010, HHS, acting through the Health Resources and Services Administration (HRSA), proposed criteria to determine whether a home visiting program is evidence-

---

one time, including the home visiting program that had been contained in Baucus's Senate bill. The language was virtually identical to that in the Baucus bill and remained unchanged when enacted.

[110] Patient Protection and Affordable Care Act of 2010, Pub. L. No. 111-148, 124 Stat. 119 (2010); see also Maternal, Infant, and Early Childhood Home Visiting Program Notice, 75 Fed. Reg. 43172 (proposed July 23, 2010) [hereinafter Notice] (reviewing the provisions of the Patient Protection and Affordable Care Act of 2010 and inviting comment on criteria for evidence of effectiveness of home visiting program models).

[111] Patient Protection and Affordable Care Act § 2951.

[112]That part of federal funding will be distributed to the states submit a proposal meeting these requirements using a formula which insures that each state receive a base amount ($1 million in FY 2010) plus additional funding based on the rate of child poverty. See, e.g., Funding Opportunity Announcement: Fiscal Year 2011, United States Dep't of Health & Human Servs., June 2011, at 1, 1 & n.1 (OMB Control No. 0915-0339; HRSA-11-179) (offering $1M for each state, plus amount based on number of children under age five up to 100% of federal poverty level for a total no less than 120% of the FY2010 award, plus the amounts then provided for projects formerly known as the Supporting Evidence based Home Visiting Program).

[113] To insure that this new federal program will have "the greatest impact," the Department announced that it would locate all "funding that exceeds funding available in FY2010" through a competitive process in which funding would go to the states whose home visiting proposals were strongest. Notice, supra note 111, at 43176,.

[114] Patient Protection and Affordable Care Act § 711(d)(3)(A)(i)(I); Notice, supra note 111, at 43173. In addition, the model must be associated with a university or national home visiting program. The state must employ a home visiting model that has been used for at least three years.) Patient Protection and Affordable Care Act § 711(d)(3)(A)(i)(I).

[115] Patient Protection and Affordable Care Act § 711(d)(3)(A)(iii) ("The Secretary shall establish criteria for evidence of effectiveness of the service delivery models and shall ensure that the process for establishing the criteria is transparent and provides the opportunity for public comment.").

based.[116] Under the proposed criteria, studies would be classified as high, moderate, or low in quality, depending on the study's capacity to provide unbiased estimates of program impact.[117] Only high and moderate quality studies would be taken into account in determining whether a model was "evidence-based."[118]

In its initial draft of the guidelines, the agency treated only well-executed randomized controlled trials as high quality studies.[119] As DSS noted, these are the "gold standard" of research design because randomization greatly increases the likelihood that any positive results were produced by the treatment and not by the characteristics of the individuals who were in the group receiving treatment.[120]

Once again, critics argued that randomized trials should not be preferred over QEDs.[121] After a period of public comment, the department made only a very modest change, adding two especially reliable types of quasi-experimental studies to the category of high quality studies:[122] single-case-study designs[123] and rigorous regression discontinuity designs.[124] Both provide stronger evidence of causality than other quasi-experimental designs,[125] but neither is commonly used in educational research at this time.[126]

Few studies of existing home visiting programs had used either randomized trials or the favored forms of QEDs. Nearly all programs relied heavily on other kinds of QEDs to support their claims of effectiveness. Under the proposed HHS rules, these studies would be deemed "moderate," at best. Although moderate qual-

---

[116] Maternal, Infant, and Early Childhood Home Visiting Program Notice, 75 Fed. Reg. 43173 (proposed July 23, 2010).

[117] Id.

[118] Id. at 43174.

[119] Id. High quality RCTs have limited attrition and no reassignment of subjects after randomization; they must control for any statistically significant differences observed in the two groups and lack differences in data collection in the two study arms. Id.

[120] Id.

[121] See e.g. Letter from Rutledge Q. Hutson and Tiffany Conway Perrin on behalf of the Center for Law and Social Policy (Aug. 16, 2010) (on file with author) (requesting consideration of additional research designs).

[122] See Supplemental Information Request for the Submission of the Updated State Plan for a State Home Visiting Program, APPENDIX F: RESPONSE TO PUBLIC COMMENTS ON FEDERAL REGISTER NOTICE ON CRITERIA FOR EVIDENCE OF EFFECTIVENESS OF HOME VISITING MODELS 51, 52, 54 (Feb. 8, 2011), available at http://www.clasp.org/federal-policy/regulations-and-guidance/maternal-infant-and-early-childhood-home-visiting-program-guidance-from-hhs (describing the comments and how the department had responded to them) [hereinafter "SIR Report of Feb 8, 2011"]; U.S. Dept. health & Human Serv., Home Visiting Evidence of Effectiveness, http://homvee.acf.hhs.gov/document.aspx?rid=4&sid=19&mid=5#revieweligible (last visited April 3, 2015) (describing current requirements and noting the use of WWC standards) [hereinafter "HHS Study Ratings"].

[123] The single-case-study model uses a single population and attempts to insure validity by measuring the outcome variable repeatedly within and across different conditions or levels of the independent variable. Kratochwill, T. R., et al., SINGLE-CASE DESIGN TECHNICAL DOCUMENTATION Version 1.0 (Pilot) 2-3 (2010),available at http://ies.ed.gov/ncee/wwc/pdf/wwc_scd.pdf.

[124] The discontinuity designs must meet the What Works Clearinghouse (WWC) design standards. HHS Study Ratings, supra note 123.

[125] Id.

[126] None of the 59 studies initially deemed eligible for consideration by HHS used these designs.

See also Regression-Discontinuity Design, supra note 69 (noting the low use of regression discontinuity).

ity studies could be used to qualify a home visiting model for formula-based funding, they would be given less weight than RCTs in the competitive funding process. As a result, critics were not satisfied with the modest concession that HHS had made in its ranking process.[127]

Furthermore, HHS had imposed an additional requirement, which meant that only a fraction of QEDs would qualify as "moderate." To qualify, a QED would need a matched comparison group whose baseline equivalence had been established at the onset of the study on the attributes of race, ethnicity, socioeconomic status and, where possible, the outcomes being measured.[128] Findings from studies with a well-matched comparison group are much more reliable than those that use a convenience comparison group. Consequently, HHS ruled that they can be used to establish that a home visiting program is evidence-based.[129]

Many stakeholders asked that these QEDs be treated as high quality, like randomized trials.[130] But the department declined,[131] reasoning that QEDs could be no better than moderate in quality because "even if the treatment and comparison groups are well matched based on observed characteristics, they may still differ on unmeasured characteristics" making it "impossible to rule out the possibility that the findings are attributable to unmeasured group differences."[132]

At any rate, most evaluations of home visiting programs that had been done prior to creation of this federal funding program lacked a carefully matched comparison group.[133] As a result, they are classified as low quality and cannot be used to qualify a home visiting program for federal funding. Many studies compared the treatment group to a group of community members and others compared the treatment group's pre-treatment scores and post-treatment scores. Findings from these studies can be skewed by differences between the people who participate and those who do not.[134] As a result, HHS classified them as low quality.

Does this framework strike the right balance? During the legislative and administrative processes, critics of this overall framework had focused primarily on the preference given to randomized trials.[135] They argued that: (1) equally important information comes from observational and quasi-experimental research, (2) RCTs can be unethical to employ because they deprive at-risk families of effective assistance, (3) randomized trials are ordinarily narrow and, thus, overlook the synergies

---

[127] See infra note 137 and accompanying text.

[128] See HHS Study Ratings, supra note 123 (noting that in some cases such as prenatal interventions that may not be possible).

[129] The moderate category also includes random assignment studies, single case designs, and regression discontinuity studies that failed to qualify for a high rating due to a weakness such as significant attrition. Id.

[130] SIR Report of Feb. 8, 2011, supra note 123, at 52, 54; Letter from Hutson and Perrin, supra note 122, at 4.

[131] SIR Report of Feb. 8, 2011, supra note 123, at 51.

[132] HHS Study Ratings, supra note 123.

[133] See infra note 152. For an example of reliance on weak QEDs, see infra notes 149-50 and accompanying text (noting that many of the studies of Parents as Teachers are QEDs)

[134] HSS concluded that studies which lack a matched comparison group "offer no way to assess what the sample's outcomes would have been in the absence of the intervention." HHS Study Ratings, supra note 122.

[135] See infra note 130 and accompanying text.

that occur among multiple social service programs, (4) programs proven at the small scale typical of randomized trials may not scale up effectively, and (5) the Nurse Family Partnership—the only home visiting program with very strong RCT findings--targets only a tiny fraction of children and families who badly need assistance.

The remainder of this Part examines these objections and finds none to be persuasive.[136] To the contrary, HHS's demanding interpretation of the statutory text saved Congress from loose drafting that could have defeated its stated goal of allocating money to programs with a track record of changing children's lives.

The Preference for Randomized Trials

In articulating its concerns about undue emphasis on RCTs, the highly respected Center for Law and Social Policy (CLASP) argued that "QEDs and observational research provide equally important information to develop and implement evidence-based policy."[137] While CLASP is correct that a variety of research methods provide us with useful information, randomized controlled trials are superior for the specific task of determining whether a specific intervention confers benefits on the participants. That is why the National Academies of Sciences calls randomized trials the gold standard for determining effectiveness.[138]

The choice of the best research design turns on the goals of the research. If a scholar wants to expand her understanding of the role that grandparents play within today's extended families, then witnessing or recording the actual interactions of extended families may be the best research design. Ethnographic studies of this kind have shed important light on early childhood development. They have helped us understand the different ways that low income and high income families see the respective responsibilities of parents and teachers (e.g. Lareau[139]) and revealed the huge difference in exposure to language that separates low income toddlers from

---

[136] The critics were correct in one important respect. Models which have not yet been tested at scale in the field should undergo rigorous evaluation as part of their funding under the Act. See infra notes 172-73 and accompanying text.

[137] Hutson & Perrin, supra note 122, at 2; see also Astuto & Allen, supra note 66, at 18 (suggesting that, if we don't have RCT outcomes, we should accept other methodologies and favoring evaluations using "diverse methodological approaches"). CLASP requested that HHS not distinguish between "high" quality studies (high quality RCTs and QEDs) and "moderate" quality studies (weaker RCTs and QEDs). CLASP concluded that "the proposed criteria are inconsistent with the legislation when they make distinctions between these two study designs and give one design a designation of "high" and the other a designation of "moderate." Hutson & Perrin, supra note 122, at 4

[138] Committee on the Prevention of Mental Disorders and Substance Abuse Among Children, Youth and Young Adults, *Preventing Mental, Emotional, and Behavioral Disorders Among Young People: Progress and Possibilities* 371(Mary E. O'Connell et al. eds., 2009).

[139] Annette Lareau, Unequal Childhoods: Class, Race, and Family Life (2003) (finding that middle-class parents take a different approach to interventions in their children's lives than working-class parents).

high income toddlers (*e.g.* Hart & Risley[140]). In this respect, CLASP's claims are correct.

However, the translation of these observations into effective interventions is an inexact science. Correlational studies of this kind commonly lead experts to propose interventions that they believe will help at-risk children thrive. Home visiting programs are one of these "research-based" interventions. Among other things, they attempt to help at-risk parents adopt the practices observed in homes with children who tend to thrive developmentally and to succeed in school.

However, an observed association between a specific parenting practice or feature of the home environment and desired child outcomes is only the beginning of the search for an effective intervention. Because correlation is no guarantee of causation, the causal role of the observed parenting practice may be specious. And once a causal association is confirmed, an effective intervention must be crafted. In the field of home visiting, most "research-based" interventions fail to achieve their desired objectives. As a result, responsible funders need to insist on proof that a funded intervention is not only "research-based," but also proven in practice.

Nonscientists are often surprised to learn that observed correlations between family circumstances and child outcomes may not be causal. Assume, for example, that families with harsh parenting practices more commonly have children with negative behavior. It would be tempting to assume that harsh parenting causes behavior problems. Yet, a third factor, like maternal depression, may be causing both of these conditions. In that event, maternal mental health, not parenting practices, should be the primary target of intervention.

Correlational research can even get causation backwards. Consider, for example, the observed correlation between corporal punishment by parents and defiant behavior by children.[141] Does this evidence demonstrate the causal connection between corporal punishment and negative child behavior? No. Correlational evidence cannot rule out the possibility that causation runs in the other direction. Extremely defiant children may prompt more extreme responses from their parents.[142] Without further study, we cannot know whether causation runs in only one of those directions or in both.

Similarly, researchers have found that "teenagers who regularly eat dinner with their families are healthier, happier, do better in school and engage in fewer risky behaviors than teenagers who don't regularly eat family dinners."[143] Two researchers decided to dig more deeply and found that the association was cut in half when

---

[140] Betty Hart & Todd R. Risley, Meaningful Differences in the Everyday Experiences of Young American Children (1995) (finding that professional families speak much more often to and with their infants and toddlers than middle class and working class parents and that their children have much larger vocabularies).

[141] Judith Rich Harris, The Nurture Assumption: Why Children Turn Out the Way They Do 23-26, 29, 46 (1998).

[142] Id.

[143] See Ann Meier & Kelly Musick, Is the Family Dinner Overrated?, N.Y Times (June 29, 2012), http://www.nytimes.com/2012/07/01/opinion/sunday/is-the-family-dinner-overrated.html?hp (describing the body of research finding this correlation and noting that the correlation disappears by the time the adolescents become young adults).

family resources and dynamics were taken into account and that the gains faded out altogether as children grew.[144] The authors concluded that the factor most responsible for the short term gains was "the extent to which parents use time to engage with their children and learn about their day-to-day activities," not whether they eat together.[145]

Even when an observed parenting practice is causally related to a positive child outcome, an intervention based on that knowledge may not succeed. The Moving to Opportunity project offers an illuminating example. A strong body of correlational evidence indicated that youth who grow up in poor, high-crime areas have worse outcomes than youth who grow up in wealthier, low-crime neighborhoods. The Moving to Opportunity study tested the sensible hypothesis that enabling families to move out of neighborhoods with concentrated poverty and crime would improve youth outcomes. The results were disappointing.[146] After five years, female youth whose families participated were less likely to have been arrested, but male youth were significantly *more* likely to have been arrested than the control group.

Untested interventions can even be harmful. The Cambridge-Somerville youth project implemented an intervention to help boys at risk for delinquency.[147] The program provided psychotherapy, tutoring, social activities, recreational activities, and family interventions. Many of the boys and their caseworkers praised the program.[148] The value of these services seems so obvious that assessment might seem a waste of resources. Yet, a randomized study found no evidence that the program had helped. To the contrary, in the years after they finished the program, the boys who received the services were more likely to have multiple criminal offenses than the control group.

Because even interventions that are based on a large body of correlational research can be ineffective, funding should require reliable proof that a proposed intervention has actually changed children's lives for the better. Intuitions and even well-established associations are not enough.

Unfortunately, the easiest, cheapest and most common ways to perform a program evaluation do not produce reliable estimates of program impact. One common method is to measure the outcome of interest, such as a parenting practice or a child's behavior, when a family enrolls in a program and again when they complete the program. These assessments are unreliable because children and new parents will mature over time whether or not they are enrolled in a home visiting program. Without a comparison group, it can be impossible to determine how much of their growth was due to their participation in a home visiting program.

---

[144] See Kelly Musick & Ann Meier, Assessing Causality and Persistence in Associations Between Family Dinners and Adolescent Well-Being, 74 J. Marriage & Fam. 476 (2012).

[145] Id.

[146] J.R. Kling, J. Ludwig & L. Klatz, Neighborhood effects on crime for female and male youth: Evidence from a randomized housing voucher experiment, 120 Q. J. Econ. 87 (2005).

[147] Timothy D. Wilson, The Message is the Method: Celebrating and Exporting the Experimental Method, 16 Psychol. Inquiry 185, 189 (2005).

[148] Id. People who undergo an intervention commonly give it a positive review even though randomized trials have found no impact. Id. at 189.

That explains why Congress insisted upon findings from randomized controlled trials and quasi-experimental studies. Both designs compare the experience of the treatment group with that of a comparison group. But having a comparison group only reduces the risk of erroneous attribution if the comparison group is very similar to the treatment group. If not, the differences between the two groups may explain any differences in their outcomes.

In Missouri, for example, multiple quasi-experimental studies have found that children whose families enrolled them in the state's free Parents as Teachers program were more ready for school than children who had not.[149] These studies compared the readiness of children whose families had made the effort to enroll them in the program with the readiness of children whose families who had not made this choice. As a result, it posed the risk that the differences in school readiness observed by the evaluators had been produced, not by the PAT home visits, but by differences between the families who had enrolled their children and the families who had not. Selection bias of this sort can produce gains that are mistakenly attributed to the intervention, as apparently happened in the case of PAT. Later randomized trials could not replicate the very strong findings of the earlier quasi-experimental studies.[150]

To reduce this bias to an acceptable level, quasi-experimental studies must have a comparison group that is closely matched to the treatment group on attributes that are known to affect the outcomes to be measured. For this reason, HHS decided that it would only consider findings from nonrandomized (quasi-experimental) studies when researchers had established the baseline equivalence of the treatment and comparison groups at program commencement with respect to race, ethnicity, socioeconomic status and, whenever possible, the outcomes being measured.[151] This decision was necessary to insure that programs eligible for funding have, in fact, produced material gains in the past.[152]

Randomized controlled studies reduce the risk of selection bias even further. As HHS noted, even if a quasi-experimental design uses a comparison group that is relatively well-matched on several important attributes, the groups "may still differ on unmeasured characteristics" making it "impossible to rule out the possibility that the findings are attributable to unmeasured group differences."[153] Random assignment of participants to one group or the other reduces this risk of omitted variable

---

[149] See, e.g. PAT National Center, The Parents as Teachers program: its impact on school readiness and later school achievement: A Research Summary (2007) (stated to be based on a report by Judy Pfannenstiel, & Edward Zigler, summarizing the results of a QED), available at http://www.parentsasteachers.org/images/stories/documents/Executive20Summary_of_K_Readiness.pdf.

[150] See infra notes 221-222 and accompanying text.

[151] See HHS Study Ratings, supra note 123 (noting that in some cases such as prenatal interventions that may not be possible).

[152] None of the dozens of past QED studies of home visiting programs was relied upon to approve the first nine programs deemed to be evidence-based. Virtually all of the QEDs lacked baseline equivalence of the study and comparison groups. These research restrictions were largely responsible for narrowing the initial field of eligible home visiting programs from about 250 to 9. All of the 59 high or moderate quality studies supporting these programs were RCTs. See id.

[153] HHS Study Ratings, supra note 123.

bias.[154]

As a result, the National Academy of Sciences treats RCTs as the gold standard. In its view, evidence of effectiveness generally "cannot be considered definitive" without ultimate confirmation in well-conducted randomized controlled trials "even if based on the next strongest designs."[155] The Institute for Education Statistics at the U.S. Department of Education has reached the same conclusion,[156] as have many early childhood researchers.[157]

As a consequence, HHS wisely decided to classify only high quality randomized controlled studies and two special kinds of quasi-experimental studies as "high" quality and to classify both well-matched quasi-experimental studies and randomized trials with problems like high attrition as "moderate" quality.

Imposing these tight restrictions on quasi-experimental studies was politically difficult. The history of the Act makes it very clear that major stakeholders in the field wanted a looser threshold.[158] They convinced Congress to specifically allow the consideration of quasi-experimental studies. Congress's failure to expressly restrict this blessing to QEDs with well-matched comparison groups had the potential to defeat its ultimate goal of funding only reliably proven models. HHS and the Administration recognized the danger and took steps to prevent it despite opposition from traditional Democratic allies.

Research methods matter. Insistence on rigorous research can be the difference between interventions that change people's lives for the better and false hope. As the Coalition for Evidence-Based Policy points out, **"**The history of social policy and medicine is replete with interventions that appeared highly-promising in less rigorous evaluations, but were subsequently found ineffective in well-conducted randomized controlled trials."[159] In medicine, examples range from ultra-radical mastectomies in the 1960's[160] to brain stents in the twenty-first century.[161] In social services, they run from job training programs to 21st Century Community Learning

---

[154] Id. (noting assignment by chance). See also Social Science Rising, supra note 90, at 2 (stating that RCTs maximize the chances that the two group will be initially equivalent).

[155] Committee on the Prevention of Mental Disorders and Substance Abuse Among Children, Youth and Young Adults, supra note 139, at 371.

[156] U.S. Dept. Educ., Identifying and Implementing Educational Practices Supported by Rigorous Evidence: A User Friendly Guide, 1 (2003), available at http://www2.ed.gov/rschstat/research/pubs/rigorousevid/rigorousevid.pdf.

[157] See, e.g., Gomby,, supra note 51, at 7 (stating that RCTs are "generally agreed to be the best way to test the causal connection"); Social Science Rising, supra note 90, at 2 (stating that RCTs maximize the chances that the two groups will be initially equivalent;. Howard & Brooks-Gunn, supra note 30, at 134 (stating "that conclusions should be based primarily—if not entirely—on experimental evaluations.").

[158] See supra notes 98-101 and accompanying text; see also supra notes,130 and 137 and accompanying text.

[159] Effectiveness Depends, supra note 102.

[160] Siddhartha Mukherjee, The Emperor of All Maladies: A Biography of Cancer 65, 193-201((2010) (describing physician resistance to tests of less radical, but equally effective surgeries).

[161] *See* Editorial, *Damage for Brain Stents*, N.Y. Times (Sept. 8, 2011) (reporting on study which found that patients receiving wire mesh stents were more than twice as likely than control group receiving more conservative treatment to suffer strokes in the next 30 days, a finding contrary to prior positive finding that promising findings from study which lacked a control group).

Centers.[162] Given that history, HHS's demanding research design requirements are good policy.

The Ethics of Randomization

Critics of the agency's preference for randomized trials also argued that the use randomized trials is often unethical.[163] That concern is vastly overstated. Furthermore, the revised HHS rubric largely moots this issue by permitting the use of rigorous QED designs in lieu of randomized trials when researchers feel an RCT would be unethical. As a result, the Act's current research rigor requirements provide a good template for future evidence-based funding streams.

There are, of course, some circumstances in which a clinical trial would be unethical. A study that proposed to deny *proven* protective services to children who are being abused in order to test a new intervention would be an obvious example. However, ethical researchers can design randomized studies of promising new ideas without denying the control group access to previously proven services. In cancer studies, for example, a promising new treatment is often compared to the existing standard of care—not to the absence of any treatment whatsoever. Researchers can evaluate promising home visiting models the same way.

Furthermore, an ethical dilemma only arises when two conditions are present. First, the study must be denying proven services to children in the control group. Second, the service providers must have sufficient resources to offer the proven services to both sets of children. The first condition will not be met when the control group is provided with current best practices in order to test the value of a new model, as noted above. Nor will it be met when current practices are themselves unproven. The mistake that many critics appear to have made is to assume that their favored home visiting model is, in fact, reliably proven and that denying it to one arm of a clinical trial would be unethical. However, most home visiting models lack this kind of evidence.

When HHS did its initial review of the literature on home visiting in 2011, only 7 of the many existing home visiting programs met the agency's standards for demonstrated effectiveness. Of these, most produced tiny or inconsistent benefits. Rigorous research is necessary to find out *whether a program works.*

The second condition is that sufficient resources be available to provide a promising model to *every* eligible child. That is rare. In the real world, funding is usually insufficient to meet community need and services must be rationed. By us-

---

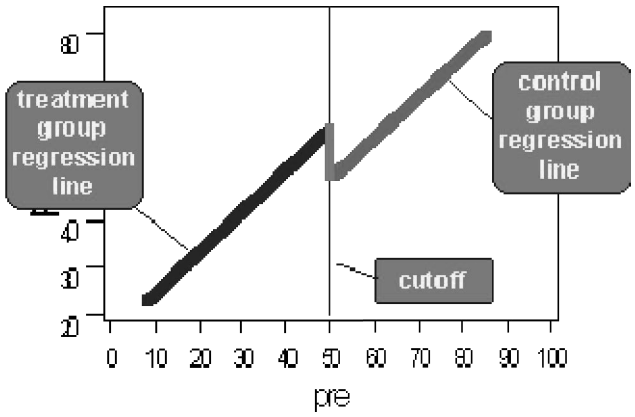
[162] See e.g. James J. Heckman and Jeffrey A. Smith, The Sensitivity of Experimental Impact Estimates: Evidence from the National JTPA Study 33 (July 1997) (For youths, the record of government training programs for the disadvantaged is almost uniformly negative."); James-Burdumy, et al., When Schools Stay Open Late: The National Evaluation of the 21st Century Community Learning Centers Program: Final Report xviii-xvix (2005) (reporting disappointing findings about the 21st Century Community Learning Center program).

[163] Daro et al., supra note 100, at 3 (making this objection).

ing a lottery to assign spots among the eligible children, rather than other options such as first in time, an agency can do a randomized comparison without ethical transgressions.

Perhaps the most challenging situation for researchers will arise when they believe that a promising, *though unproven*, intervention should be targeted to the children most in need of its anticipated benefits. In those situations, a lottery is not ideal. Fortunately, one of the quasi-experimental designs approved by HHS is well-suited for use in this context.

A regression discontinuity study can create treatment and control groups by separating the children who score below a cutoff score, such as a score for early language skills, from those who score above.[164] Only the children falling below the cutoff would receive the intervention being studied. By targeting the children most at risk, this design avoids another ethical issue potentially posed by RCTs.[165] After the intervention, researchers assess whether the scores of the intervention group have risen more than those of the comparison group. If the intervention works, the regression line for the treatment group should be higher than that for the comparison group, as shown in this graph.



The disconnection between the two lines is the "discontinuity" that gives this design its name.[166]

To sum up, the vast majority of home visiting programs can be ethically studied

---

[164] See The Regression-Discontinuity Design, http://www.socialresearchmethods.net/kb/quasird.php (last updated Oct. 20, 2006) (noting this advantage of regression-discontinuity designs).

[165] Id.

[166] Id. Although this study design is not yet commonly used in education and social sciences research, its use is likely to grow because this design is now being successfully used to assess the impact of state-funded preschool programs. See, e.g. William T. Gormley, Jr. & Ted Gayer, Promoting School Readiness in Oklahoma: An Evaluation of Tulsa's Pre-K Program, 40 J. Hum. Resources 533, 544 (2005).

using RCTs. When researchers reasonably believe that RCTs would be unethical, HHS rules permit the use of rigorous QEDs. In particular, regression discontinuity designs may provide a good alternative. Like RCTs, regressive discontinuity models are classified as high quality by HHS. No loosening of the current rules is needed.

In medicine, cancer surgeons vigorously resisted randomized trials of ultra-radical mastectomies because surgeons considered it unethical to deny the control group access to a treatment that was theoretically unassailable. They were wrong. Thousands of women were needlessly and horrifically disfigured before randomized trials proved that the technique was not needed.[167] Let's not repeat that mistake. It's not unethical to find out if a popular program works.[168]

Overlooking Synergies

A third criticism voiced against the emphasis placed on RCTs by the Obama Administration is that RCTs force evaluators to focus narrowly on very specific outcomes and that this narrow focus is likely to overlook synergies that a home visiting program may have with other community programs. Together, the programs may generate a whole that is greater than the sum of the parts.[169]

This argument has a fatal weakness. If "the whole" is generating wonderful outcomes, then measure them, too. If they cannot be reliably measured, then why should we believe that they really exist? Taken to its logical end, this argument would support the continuation of virtually every unproven program ever created.[170]

Ironically, the very program used to illustrate the existence of these larger synergies has recently been studied. The hoped-for synergies could not be detected. The Harlem Children's Zone (HCZ) provides an array of services to children who live within its service area (the "Zone"), one of which is a charter school. The laws governing charter schools in New York City require that residents living outside the Zone be allowed to enter the enrollment lottery. As a result, some of the chosen students lived within the Zone and some did not. Only those living within the Zone were eligible for HCZ's comprehensive services. As a result, the lottery provided a natural experiment in which the value of the extra services provided only to children living within the Zone could be tested. The team of researchers found that the students who attended the HCZ charter school were making exceptional gains, but

---

[167] See Mukherjee, supra note 160.

[168] See also Wilson, supra note 147, at 190 ("[T]here is an ethical problem in not conducting experiments to test interventions that could have harmful effects").

[169] Hutson & Perrin, supra note 122, at 3 (referencing the Harlem Children's Zone and stating that RCTs "provide little information about what it takes to combine multiple interventions to achieve stronger outcomes or to scale up such layered, comprehensive approaches to working with children and families")

[170] Research has established that participants routinely think programs are more effective than they are. See Wilson, supra note 147, at 187, 189 (noting research findings and describing one case study).

that the students who lived inside the Zone had no better school outcomes than the students who did not.[171] Thus, the authors found no evidence that the comprehensive social services spilled over into greater school success.

Synergies will sometimes exist, but they can be rigorously studied. Wishful thinking is no substitute.

Efficacy at Scale

Critics also complain that RCTs are typically too small to provide reliable evidence that a program will be equally effective on a citywide or statewide scale. This point is well-taken, but the critics draw the wrong inference from it. Lawmakers may reasonably insist that a program be proven on a small scale before it is funded at a much larger scale. While success in the hothouse of a randomized trial is certainly no guarantee that a program will do as well when it is operated on a much larger scale,[172] programs which have no proven effect in the hothouse are highly unlikely to thrive in the field. When a program is replicated at scale, fidelity to program design is more difficult to insure, staff may lack the motivation associated with testing of a promising new idea, and the creator of the program is no longer at the helm, tirelessly working to insure that the project is well run. As a result, success in a high quality study, like a randomized trial, should be viewed as necessary, but not sufficient, to firmly establish a home visiting program as field tested.[173]

To spend our funds wisely, funders should start with programs that have been reliably tested on a small scale and then insist that scaled up programs be rigorously evaluated as well. A follow-up assessment is essential to determine whether the program's initial promise could be reproduced in new locations and on a larger scale.

---

[171] Vilsa E. Curto, Roland G. Fryer, Jr. & Meghan L. Howard, It May Not Take a Village: Increasing Achievement Among the Poor. Social Inequality and Educational Disadvantage 26-27 (2011), available at http://scholar.harvard.edu/files/fryer/files/it_may_not_take_a_village_increasing_achievement_among_the_poor.pdf.

[172] See SIR Report, supra note 123, at 52 (noting suggestions that it take into account whether a model had been implemented in the "real world"); Jennifer Kahn, Can Emotional Intelligence Be Taught?, N.Y. Times (Sept. 11, 2013), http://www.nytimes.com/2013/09/15/magazine/can-emotional-intelligence-be-taught.html?pagewanted=all (noting the "Hawthorne effect" in which the attention focused on an educational experiment is enough to cause "a temporary uptick in performance").

[173] Large scale randomized trials sometimes contradict the positive findings of pilot projects. See, e.g., Nancy McCall & Jerry Cromwell, Results of the Medicare Health Support Disease-Management Pilot Program, 365 New Engl. J. Med. 1707, 1704 (2011) (finding that having nurses call patients to help manage multiple chronic conditions did not reduce costs).

Inability to Serve Families in Need

The home visiting programs being operated across the country serve a wide variety of families with a wide variety of needs.[174] When the federal Home Visiting Program was first proposed, only the Nurse Family Partnership was sufficiently supported by RCTs to be assured eligibility.[175] Yet, the Nurse Family Partnership model only serves low-income, first-time mothers who enroll within the first weeks of their baby's life. Thus, it will not reach first-time mothers who don't learn about the program in time, mothers with other children, or mothers who exceed the income threshold. In addition, it offers a very specific, mixed package of information and services. That package differs from the services provided by other models, some of which target people in need of mental health services or children at high risk of child abuse.

When legislation was proposed that appeared to target NFP, home visiting programs like Parents as Teachers, Healthy Families America, and HIPPY USA feared they would be left out.[176] They enlisted national child advocacy organizations to make the case for looser eligibility requirements.[177] Their supporters argued that tough research standards would be bad public policy because too many needy families would fall outside the Act's reach, thwarting the goals of Congress.[178] Some communities may need mentoring services for teen mothers; others may feel that mental health assistance should be prioritized.[179] If the law's evidentiary standards are too strict, states seeking funding under the Act would be less likely to find a home visiting program that fits their needs on the list of eligible models.

In reaction, Congress revised the draft legislation to include quasi-experimental studies. If HHS had not used its authority to impose rigorous criteria on qualifying QEDs, a Congressional revision intended to make a wider array of services available would, at the same time, have eviscerated the primary goal of targeting funds to programs with proven effectiveness.

This dispute usefully highlights a significant problem in this field and in many others. Very few existing programs can provide rigorous evidence of their effectiveness. Most home visiting programs were created and evaluated in an era when simple, but unreliable, research designs were sufficient to satisfy funders. As a result, Congress attempted to fund evidence-based programs in a field where the necessary rigorous research had not yet been done. In hindsight, Congress should first

---

[174] See supra notes 45-47 and accompanying text.

[175] See supra notes 58-60, 87-94 and accompanying text (discussing how the text of the proposed statute seemed to target NFP).

[176] Haskins et al., supra note 81, at 7 (listing Parents as Teachers, Healthy Families America, the Parent Child Home Program, and HIPPY USA as disappointed suitors).

[177] See Astuto & Allen, supra note 66, at 6.

[178] See Daro et al., supra note 100, at 2.

[179] Haskins et al., supra note 81, at 8-9; Astuto & Allen, supra note 66, at 6 (noting that various models have different goals, populations, strengths, and weaknesses and that communities should be permitted to pick the one that fits "their particular needs.").

have funded a generation of high quality studies to identify the most powerful home visiting programs. Instead, it created a funding stream to expand access to evidence-based programs without a deep pool of evidence-based programs from which to draw. This understandably produced a dogfight over the definition of "good" research.

Conclusions about Design Rigor

None of the criticisms of the Act's current research design requirements are persuasive. The law now contains well-crafted research design standards—an impressive accomplishment for an innovative program whose legislative authors and agency implementers had to mediate a debate over the rigor of those requirements. This part of the Home Visiting Program provides a useful template for future evidence-based funding efforts.

PART IV. THE ACT'S OUTCOME REQUIREMENTS

Unfortunately, HSS failed to complement its demanding research design requirements with equally tough requirements for the minimum outcomes needed to qualify for federal funding. The current rules contain no requirements with respect to the minimum magnitude of the benefits conferred, the consistency of the findings, the durability or salience of the benefits, or the replication of positive outcomes. As a result, many of the approved programs have threadbare or troublingly inconsistent evidence of positive impact. Only a few would qualify under more defensible standards.

Minimum Effect Size

Under current DHSS rules, any statistically significant positive finding counts toward the agency's proof requirements no matter how trivial the impact. In fact, current rules do not even require the calculation of an effect size at all. As a result, the EIP and Project 12-Ways/SafeCare Augmented programs were approved despite the absence of any estimate of effect size.[180] When the Act is reauthorized, Congress should set a minimum effect size to help ensure that the funded models

[180] See Project 12-Ways/SafeCare: Effects Shown in Research & Outcome Measure Details for Reductions in Child Malttreatment Outcomes, U.S. Dep't of Health & Human Serv. (2011) http://homvee.acf.hhs.gov/effects.aspx?rid=1&sid=39&mid=5&oid=7; Early Intervention Program for Adolescent Mothers: Study Search for Family Economic Self-Sufficiency Outcomes, U.S. Dep't of Health & Human Serv. (2011), http://homvee.acf.hhs.gov/Effects.aspx?rid=1&sid=18&mid=4&oid=4. Statistical significance is not a substitute for impact.While it reduces the odds that a positive finding is simply a matter of chance, statistical significance is not about magnitude.

are capable of having a meaningful, rather than perfunctory, impact.

At present, trivial effects suffice. In a study of Healthy Steps, for example, researchers found that a program component called PrePare increased children's early language skill by 0.03 standard deviations after two-and-a-half years of services.[181] That is equivalent to moving the participating children from the 85th to the 85.66th percentile. Yet, that modest finding qualifies as one of the two positive findings needed to make Healthy Steps eligible for federal funding. Not only is an impact this small unlikely to change a child's life materially in the short run, but it is also virtually certain to fade out soon after program completion.[182] Federal funding based on this finding would be squandered.

Fade-out is normal. As a result, only gains with very large effect sizes are likely to be durable. Studies of ordinary preschool attendance have shown that even larger short-term gains routinely disappear within a couple of years. Not even the famous early childhood programs, like the Abecedarian Project and the Perry Preschool, avoided substantial fade out. Instead, they produced initial gains so large that a substantial residual effect remained several years later despite the loss of roughly half the initial gain. A recent meta-analysis of high quality preschool studies found that initial effect sizes usually shrank by half[183] and a national study of ordinary preschools found that the short-term gains disappeared entirely.[184] In a recent rigorous study of Head Start, initial cognitive gains of about a 0.10 standard deviation disappeared almost completely by the end of kindergarten.[185] Researchers have found similar fade out when studying the advantages of full-day kindergarten,[186] class size reduction,[187] an extra daily class session in math,[188] and even hav-

---

[181] See Healthy Steps: Study Search for Child Development and School Readiness Outcome, U.S. Dep't of Health & Human Serv. (2006) http://homvee.acf.hhs.gov/effects.aspx?rid=1&sid=12&mid=5&oid=3 (describing positive effects for combining two or more words and for vocabulary, but providing an effect size for only the first).

[182] See infra notes 183-89 and accompanying text.

[183] Gregory Camilli et al., Meta-Analysis of the Effects of Early Education Interventions on Cognitive and Social Development, 112 Teachers College Record 579, 600-01,table 7 (2010) (finding that cognitive effect sizes faded by roughly half under each of a variety of assumptions).

[184] Katherine A. Magnuson, et al., Does Prekindergarten Improve School Preparation and Performance? 26 Econ. Education Rev. 33 (2007) (finding that preschool math and reading gains had "largely dissipated by the spring of first grade).

[185] Michael Puma et al., Head Start Impact Study: Final Report, U.S. Dep't of Health & Human Serv. (Jan. 2010) http://files.eric.ed.gov/fulltext/ED507845.pdf (despite using a loose standard of statistical significant (0.10), the study found that initial gains had disappeared on six of seven metrics by end of first grade).

[186] Elizabeth Votruba-Drzal, et al., A Developmental Perspective on Full- Versus Part-Day Kindergarten and Children's Academic Trajectories Through Fifth Grade, 79 Child Development 957, 974 (2008) (finding initial math and reading gains of one-fifth a standard deviation, that shrank by 25-50 percent after controlling for several factors, had entirely disappeared by spring of third grade).

[187] Alan B. Krueger & Diane M Whitmore, The Effect of Attending a Small Class in the Early Grades on College-Test Taking and Middle School Test Results: Evidence from Project STAR, 111 Econ. J. 1, 10 (2001) (showing that benefits shrank steadily between first and ninth grade)

[188] Eric Taylor, Spending More of the School Day in Math Class: Evidence from a Regression Discontinuity in Middle School, 117 J. Pub. Econ. 162 (2014) (finding that initial gains had shrunk by two-thirds over two years).

ing a highly effective teacher.[189]

Initial effects commonly shrink over time. As a result, only substantial short-term effects offer a reasonable hope of lasting benefit. If we want to change a child's trajectory, only substantial and durable benefits will do. The lack of a minimum effect size means that states may spend substantial sums on models that produce evanescent benefits.

A minimum effect size is advisable for another reason as well. In a carefully designed pilot study, the effects will often be larger than they will be when the program is scaled up. The staff of a high quality experiment is typically carefully selected and is often aware that team success will be gauged by the program's outcomes. Often, the creator of the model supervises the trial and has a powerful incentive to work tirelessly to guarantee that the program is implemented as intended and that barriers and surprises are overcome immediately. Of the last five models approved by HHS, all were approved on the basis of studies that were overseen by the developer of the program. This kind of motivation and fidelity is difficult to duplicate when a program is scaled up. In fact, the founder of NFP was so concerned about loss of fidelity that he carefully limited the program's rate of expansion.[190] As a result, evidence-based programs should insist on proof of a substantial impact in the initial demonstration studies before concluding that a model is likely to confer meaningful long-term benefits when taken to scale.

There is no bright line for what this minimum effect size should be, but an effect size of at least 0.25 standard deviations would be a very reasonable place to experiment.[191] The race and poverty gaps in academic achievement and emotional development are estimated to exceed 0.50 standard deviations at kindergarten entry[192] and a full standard deviation by twelfth grade.[193] Under those circumstances, funders can reasonably insist that funded programs confer an initial gain of at least 0.25 standard deviations in core child developmental skills, anticipating that the long term benefits will be roughly half that—closing 10-15% of the gap. Happily, most of the currently approved programs have effect sizes that meet or exceed that

---

[189] Thomas J. Kane & Douglas O. Staiger, Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation (NBER, Working Paper, 2008) (finding that effects faded out by 50 percent per year in the two years following teacher assignment); Raj Chetty, et al., The Long-Term Impacts of Teachers Value-Added and Student Outcomes in Adulthood 4-5 (NBER, Working Paper, 2011) (noting complete fade out in prior studies but finding that teacher impact it stabilized in their data set at about one-third the initial impact after 3 years); Jesse Rothstein, Teacher Quality in Education Production: Tracking, Decay and Student Achievement, 125 Quarterly J. Econ. 175, 175 (2010). Chetty and colleagues found that, despite the fade-out, earnings were higher for students with more effective teachers. Id.

[190] See supra note 93 and accompanying text.

[191] See e.g., Best Evidence Encyclopedia, http://www.bestevidence.org/aboutbee.htm (requiring an effect size of 0.20 for its highest rating).

[192]See, e.g., Fryer & Levitt, supra note 22, at 256, 262-63 (finding a kindergarten racial achievement gap of 0.66 standard deviations in math and 0.40 in reading). See also supra note 22 and accompanying text for additional studies.

[193] See Larry Hedges & Amy Nowell, Black-White Test Score Convergence Since 1965, 48 J. Hum. Resources 149, 151-53 (Christopher Jencks & Meredith Phillips, eds., 1998) (reviewing the literature and concluding that, among 17-year-olds the gap has been measured between 0.82 and 1.18 standard deviations in composites of vocabulary, reading, and math).

threshold.[194] That is the good news. The bad news is that current rules impose no barrier to funding future home visiting models with much less impact.

Until Congress imposes a minimum effect size, HSS is wise to allocate a substantial portion of the home visiting funds using a competitive model. A competitive grant process can take effect size into account.[195] In the longer term, however, the funding criteria need to be revised.

Durability of Benefits

The current rules do not require any evidence that a program's positive effects persist past program completion. For a few of the desired outcomes, such as reductions in child maltreatment, temporary improvements may constitute a sufficient benefit to warrant funding. But for other goals, like improved child development and better school readiness, the objective is to alter a child's long term trajectory. Meaningful programs like NFP help at-risk children construct strong cognitive and emotional foundations on which to build further during the K-12 years. Only durable gains in cognitive and social development accomplish that end.[196]

Unfortunately, the current HHS approval rubric gives durability a very limited role. The agency's hands were tied by Congress. The Act contains an oddly *de minimis* durability provision which inexplicably applies only to RCTs.[197] Positive findings from RCTS must be observed one year after program *enrollment*.[198] The Act calls these gains "sustained"[199] even though a gain that lasts only as long as the treatment is being delivered can hardly be considered sustained.

---

[194] Only two are close calls. Effect sizes for HFA were only calculated for four of its fifteen positive effects and they ranged from 0.14 to 0.25. Effect sizes for EIP ranged from 0.12 to 0.25.

[195] See e.g., Model Programs Guide, Office of Juvenile Justice and Delinquency Prevention, http://www.ojjdp.gov/mpg/ratings.aspx_The OJP allocates from zero to 3 points in its competitive process on the basis of effect size. Id.

[196] "Sleeper effects" constitute an unusual kind of durable effect that should be included in this calculus. They arise when a program lacks large short-term measurable benefits but the students receiving the intervention demonstrate gains later in life. Head Start, for example, appears to confer some socio-emotional sleeper effects, even though its small cognitive benefits fade by early elementary school. Grover J. "Russ" Whitehurst, Can We Be Hard-Headed About Preschool? A Look at Head Start (Jan. 16, 2013), http://www.brookings.edu/research/papers/2013/01/16-preschool-whitehurst (describing "sleeper effects").

[197]Patient Protection and Affordable Care Act , 42 U.S.C. § 511(d)(3)(A)(i)(I),(2014); Maternal, Infant, and Early Childhood Home Visiting Program Notice, 75 Fed. Reg. 43175 (proposed July 23, 2010) (stating that the model must one that has demonstrated significant positive outcomes, and in the case of the service delivery model described in item aa, those outcomes must be sustained). Only RCTs are included in subsection "aa." Affordable Care Act § 511(d)(3)(A)(1)(II)(aa). The result is sadly ironic. Programs whose effectiveness have been demonstrated by studies which meet the gold standard must also meet two extra requirements that are not imposed on home visiting programs that rely upon less reliable studies. See Home Visting Program Notice, 75 Fed. Reg. at 43174. One hopes that this quirk was an inadvertent error, but language earlier in the Act which specifically exempts quasi-experimental studies from the very modest durability requirement applicable to RCTs suggests that lobbying produced this otherwise inexplicable difference in treatment.

[198] Home Visiting Program, 75 Fed. Reg. at 43174.

[199] Id.

In fact, gains measured at the one year mid-point of a two year program could theoretically qualify that model for funding even if the gains had faded out entirely by the end of the program's second year. In a study of the HFA San Diego program, for example, the gains observed at age one had disappeared by age three.[200] Requiring that gains last a year from program initiation is the thinnest imaginable durability requirement. Furthermore, quasi-experimental studies are exempted from even this minimal requirement, perhaps reflecting strong industry lobbying.[201]

The agency's ability to consider this factor in a competitive grant-making process provides yet another strong reason for allocating a sizable portion of the funds competitively.[202] For the portion of funding that is distributed through a formula, however, the lack of a meaningful durability requirement is a significant weakness. Fade-out is simply too common. To increase the odds of meaningful impact, Congress should require proof of impacts that last at least a year after program completion.

States may object to this change because it will shrink the number of programs that qualify for evidence-based funding. Only six of the fourteen approved programs can currently demonstrate that they confer benefits that last at least one year after program completion.[203] States would have to spend at least seventy-five percent of their grant on these models. The other eight models could only be funded using the twenty-five percent that can be spent on "promising" models. As a result, a meaningful durability requirement could impair the ability of states to find a model that serves the specific needs of the families in that state. This is a legitimate concern.

The best solution to this predicament is to temporarily expand the portion of state formula funding that can be spent on promising programs, not to loosen the standards for qualifying as an evidence-based program. Under the statute, promising models, unlike evidence-based models, must be rigorously evaluated as part of the grant process. As a result, shifting the borderline models into the "promising" category and requiring that they be evaluated for their long-term impact will help fill the current knowledge gap. States will still have the choices that they want, but rigorous evaluation of these promising programs will be mandatory. The resulting studies will reveal whether these promising programs have a durable and material

---

[200] See Healthy Families America: Study Search for Child Development and School Readiness Outcomes, U.S. Dep't. of Health & Human Serv. (2007) http://homvee.acf.hhs.gov/Effects.aspx?rid=1&sid=10&mid=5&oid=3. However, this finding was not relied upon for approval.

[201] But see id. HHS staff recognized the emptiness of this requirement and decided to report whether models has demonstrated effects that last at least one year from program cessation. It calls them "lasting" impacts. However, proof of lasting impact is not required for program approval. Id.

[202] This would resemble the process used by the Department of Justice, which assigns between zero and three points depending on the time between program completion and follow-up. Program Evidence Rating Instrument, U.S. Dep't of Justice, http://www.crimesolutions.gov/pdfs/ratinginstrument_part2.pdf.

[203] But see Sarah Avellar et al., Home Visiting Evidence of Effectiveness Review: Executive Summary 9 (September 2013—Revised June 2014), OPRE Rep. 2013-42, (listing eight programs: EHS, EIP, HFA, HIPPY, NFP, PAT, PALS infant and SafeCare Augmented). The effects from the latter two programs were measured at 9.5 and 10 months after completion, so I do not include them in my total.

impact. Over time, these rigorous studies will expand the pool of meaningfully proven programs, while eliminating those models that do not fulfill their initial promise.

Replication

The statute also lacks a replication requirement. The replication of a positive finding in a second sample increases the odds that a positive finding in the first study was caused by the intervention and not chance. As a result, successful replication of positive results greatly increases the likelihood that sites funded with federal grants will confer the same benefits on their participants as were detected in the qualifying studies.

Yet, the current approval rubric does not require replication.[204] Although replication of a positive finding in one domain is one route to approval, it is not the only one. A single study with positive findings in two domains will also suffice.[205] Only five of the fourteen approved programs have replicated positive findings in the same domain.[206] The other nine programs were approved on the basis of a single study that found at least one positive effect in two different domains, such as child development and positive parenting. The seven most recently approved programs all qualified on this basis. While positive findings in two domains can increase the odds that at least one of them was not a statistical fluke,[207] this criterion is no substitute for repeated and consistently favorable findings in the same domain.

Consider the research on HFA. After a 2002 study found that HFA increased the number of well-baby visits attended by mother and baby,[208] studies in 2005 and 2007 using different samples could not replicate that finding.[209] Had the research on HFA stopped sooner, we would have a dramatically different –-and misleading–- impression of its efficacy. The researchers observed that "[o]ur findings also alert us to the importance of replication studies and caution us about generalizing positive or negative results from a single-sample, single-site evaluation."[208] The initial results did not tell the whole story.

In addition, the definition of a replicated finding should be tightened. At present, HHS requires only that two studies find positive impacts in the same *domain*. However, the domains are so broad that the "replicated" findings can actually involve very different attributes. At present, it is possible for a model to be approved on the basis of a study finding a positive impact on child cognitive development but

---

[204] Only subgroup findings must be replicated in a different sample. DHHS Criteria for Evidence-Based Models, U.S. Dep't of Health & Human Srev., http://homvee.acf.hhs.gov/document.aspx?rid=4&sid=19&mid=6 (last accessed 4/3/15).

[205] Id.

[206] See Avellar et al., supra note 203, at 9 tbl 2 (listing Family Check-UP, HFA, HIPPY, NFP, and PAT).

[207] Two random positive findings are less likely than one unless the number of measurements in doubled.

[208] At 584.

no impact on emotional development combined with a second study finding a gain in emotional development, but not in cognitive skills. This home visiting model would qualify for funding because each study found at least one positive impact in the domain of child development. Yet, neither study replicated the findings of the other. In fact, they reach directly inconsistent conclusions. Nevertheless, the model would qualify as evidence-based because the rules do not require a repeated positive impact on the *same construct* within a domain.

As a result, the sponsors of a home visiting model now have a strong incentive to measure as many aspects of each domain as possible in both the initial study and in any replication study. That is because any combination of positive findings in a given domain will satisfy the replication requirement. At the same time, measuring more constructs means that the odds of a random positive finding will increase. The standard "p" value for statistical significance is 0.05. The chance of a false positive is, therefore, one chance in 20. At this level of statistical significance, 100 measurements would be expected to produce as many as 5 positive findings based on chance alone, even if the program were totally ineffective.

To minimize this gamesmanship, HHS should tighten its replication requirements to require repeated positive findings within a given domain. If the current rules did so, only five of the fourteen approved models would qualify.[209] That number would shrink still further if programs had to show a replicated positive impact on the same construct.[210]

Salience of the Benefits Conferred

At present, each positive finding counts as much as any other. In the domain of positive parenting, for example, programs that increase the use of safety latches get the same credit as programs that greatly increase the number of parents who read daily to their children.[211] Surely, this is not how Congress intended to spend its money.

---

[209] See Home Visitng Program Model Effects, U.S. Dep't of Health & Human Serv., http://homvee.acf.hhs.gov/EvidenceOverview.aspx (listing Family Check-UP, HFA, HIPPY, NFP, and PAT).

[210] The current absence of a replication requirement further justifies the agency's decision to allocate some of the funds in a competitive process. Replication can be taken into account in the ranking system. See e.g., Early Childhood Education: Review Methods, Best Evidence Encyclopedia, http://www.bestevidence.org/early/early_child_ed/methods.htm (requiring, for the highest ranking, "at least two studies, one of which is a large randomized or randomized quasi-experimental study, or multiple smaller studies"); See also Program Review and Rating from Start to Finish, Nat'l Inst. of Justice, http://www.crimesolutions.gov/about_starttofinish.aspx (last visited Aug. 14, 2012) (accepting only RCTs and high quality QEDs, with preference for RCTs and requiring replication in a second sample for top rating, scoring based on effect size and durability of demonstrated effect, and threats to internal validity).

[211] See Study Search for Positive Parenting Practices Outcome, U.S. Dep't of Health & Human Serv., http://homvee.acf.hhs.gov/Effects.aspx?rid=1&sid=12&mid=5&oid=6 (listing these among the constructs assessed for the HFA program).

The importance of a model's positive outcomes can currently be considered in the competitive grant process, but not in the formula allocation process. Fortunately, the list of core benchmarks being created by DHSS to evaluate its grants can be used in the future as a list of outcomes that must be tracked in the studies submitted for approval of the model.

Consistency of Outcomes

The current approval process ignores studies finding that a home visiting model failed to provide any measurable benefit and even studies concluding that the model had a negative impact. Only positive findings are considered. If twelve studies have evaluated a home visiting program and only one of the twelve found any positive impact of any kind, the model would nevertheless qualify for approval if the study found positive impacts in two domains. In fact, the model would be eligible for funding even if the ten other studies found that the program impaired child development. Only the positive findings count.

The body of research on Health Families America (HFA) illustrates this problem. HFA has been studied many times. It can now boast at least one positive finding in each of the eight domains used by HHS. Yet, its batting average is much less glorious than this statistic would imply. Rigorous studies have found a positive impact on less than 10% of the constructs measured (43 of 494).[212] In two domains, the rate was so low that the few positive findings could easily have been due to chance (1 of 30 in family violence and 3 of 72 for maternal health.) [213]

Consider another example from the HFA research. This study found a favorable effect on well-baby visits.[214] Yet, the same study found no effect on 12 other measures of child health. Was the single positive finding a statistical artifact? Two later studies found no impact on well-baby visits.[215] Should the single positive finding count toward approval despite two null findings on the same construct? It does now.

The PAT research also raises the issue of inconsistent findings. PAT qualified for approval on the basis of replicated findings in a single domain. One study found that PAT had a positive impact on self-help and another found a gain in mastery motivation. Yet, two other studies found that PAT did *not* improve children's "self-help" skills. Despite the even split in the studies, PAT was approved.

---

[212]See Healthy Families America, U.S. Dep't of Health & Human Serv., http://homvee.acf.hhs.gov/document.aspx?rid=1&sid=10&mid=1 (last visited July 21, 2012). The impact on four constructs was unfavorable or ambiguous. Id.

[213] HFA's repeated positive outcomes in the domain of child maltreatment justify its approval. But the studies of HFA illustrate the dilemma that is posed under the current approval rubric when many measurements have been taken.

[214]See Healthy Families America, supra note 215.

[215] See id.

Similarly, HHS approved the Family Check-Up program because one study found statistically significant impacts in two domains. In the domain of child development, the study found positive effects on emotional development.[216] Unfortunately, neither of the two other high quality studies found any impact on emotional well-being.[217] Yet, the two studies with null findings are not taken into consideration even though they outnumbered the single study with positive findings.

Negative findings are also ignored. PAT again offers an illustration.[218] Rigorous studies of PAT have taken 208 measurements. In 196 of those measurements, PAT had conferred no benefit.[219] The remaining twelve assessments found seven negative or ambiguous effects and five positive effects. Fortunately for PAT, the seven unfavorable findings were ignored and two of the five favorable findings were in a single domain. No weight could be given to the multiple negative findings and nearly two hundred findings of ineffectiveness.

The failure to take negative and inconsistent findings into account is ill-advised. The odds that a federally-funded program using these models will confer substantial and durable benefits on participating children go down with each study failing to find an impact. Careful stewardship would take this into account.

The dilemma for HHS, of course, is what to do when the outcomes are inconsistent. How many null findings should it take to offset a positive finding? How many negative findings? What about negative findings in the same domain, but not for the same trait? What if the few positive findings are for very important constructs, like severe physical abuse or early literacy? It may be impossible to develop a rubric that provides bright line guidance when multiple studies have each taken multiple measurements and the results are not consistently positive. Doing so is especially difficult in the absence of standard metrics for each domain.

HHS has taken two important steps to minimize this weakness in its rubric.[220] First, the agency has reserved a portion of the statutory funding for a competitive award process in which overall efficacy of the state's chosen model can be taken into account. Second, it is using the grant evaluation process to identify a set of core

---

[216] See Family Check-Up: In Brief, U,S, Dep't of Health & Human Serv., http://homvee.acf.hhs.gov/document.aspx?rid=1&sid=9. Positive impact was detected for three measures of child emotional development: internalizing, 0.21; externalizing, 0.23, and problem behavior, 0.23. Id. See also, Family Check-Up: Study Search for Maternal Health Outcomes, U.S. Dep't of Health & Human Serv., http://homvee.acf.hhs.gov/effects.aspx?rid=1&sid=9&mid=5&oid=1.

[218] Other examples of offsetting negative findings appear in the studies of the Healthy Steps supplement PrePare and of HFA. The single study of PrePare which found any positive effects found more negative impacts (7), than positive ones (5). See Healthy Steps: In Brief, U.S. Dep't of Health & Human Serv., http://homvee.acf.hhs.gov/document.aspx?rid=1&sid=12&mid=1 (last accessed July 21, 2012) (summarizing outcomes). HFA had similar inconsistency in the domain of family economic self sufficiency, where it had three positive outcomes, two negative outcomes, and thirty-seven absences of impact. Healthy Families America: In Brief, U.S. Dep't of Health & Human Serv,, http://homvee.acf.hhs.gov/document.aspx?rid=1&sid=10&mid=1#title#title (last accessed July 21, 2012).

[219] See Parents as Teachers, U.S. Dep't of Health & Human Serv., http://homvee.acf.hhs.gov/Model/1/Parents-as-Teachers--PAT--sup---sup-/16/1.

[220] As an additional measure, the agency posts negative and null findings on its webpage. This may help the states make informed decisions when they select the models to use.

outcomes that each state must measure.[221] The constructs that emerge from this process can potentially become future benchmarks in the *ex ante* approval process. With this list of common benchmarks, HHS could require that agencies seeking approval demonstrate positive, durable, replicated and consistent impact on one or more of the key constructs.[222]

Conclusions about Outcome Thresholds

At present, the requirements for classification as an evidence-based service model contain no minimum thresholds for the magnitude, durability, replication, salience, or consistency of favorable findings. Null findings and even negative findings are ignored. As a result, home visiting programs can qualify as "evidence-based" despite sparse or troublingly inconsistent findings. The interests of children and the goals of Congress would both be better served if minimum requirements were imposed for all of these outcomes. Doing so would funnel public funding to the programs most likely to change children's lives for the better.

Adoption of these minimum outcomes requirements would shrink the current list of approved programs. Only five of the fourteen programs approved as of June 2014 had replicated positive findings in the same domain.[223] Of this five, four--HFA, HIPPY, NFP and PAT--would remain eligible if the law required proof that positive impacts could still be observed one year after program completion. Only HIPPY, NFP and PAT would also meet a minimum effect size requirement of 0.25 standard deviations. HFA would qualify under a threshold of 0.20 standard deviations. Because the outcomes for PAT and HIPPY are troublingly inconsistent, these models would not qualify if a minimum level of consistency were required. HFA is

---

[221] See Affordable Care Act, 42 U.S.C. § 511 (d)(1)(B)(ii) (requiring improvement in at least four areas); Funding Opportunity Announcement: Fiscal Year 2011, U,S, Dep't of Health & Human Serv.,., June 2011, at 1, 20 (OMB Bull. No. 0915-0339; HRSA-11-179). For purposes of program evaluation, Congress condensed the eight domains of child, parent and family well-being targeted by the grant program into six "benchmark" areas, such as school readiness. In addition, states must show improvement in at least half of the "constructs" measured in each benchmark area. Id. ("[S]tates must collect data for all constructs under each benchmark area"). Id. App C, 43-52 (listing mandatory constructs for each area). For example, in the area of child readiness for school, each state must measure key constructs such as "language and emergent literacy" and "social behavior, emotional regulation, and emotional well-being."

[222] See e.g., Program Evidence Rating Instrument, Nat'l Inst. of Justice, http://www.crimesolutions.gov/pdfs/ratinginstrument_part2.pdf. To earn the top rating ("effective") or the middle tier ("promising"), the program can have no studies showing a negative impact on the targeted problem behavior and only 1 showing a null effect"). Even then, the reviewer must determine whether the null findings should disqualify the program. Id. ("In some cases, the evidence for a program may be inconsistent, for example, if there is one study indicating a statistically significant positive effect (i.e., Class 1 or Class 2); one study indicating a statistically significant null effect (Class 4); and no third study is available for consideration. In such cases, the Lead Researcher will also review both studies and make a final determination on whether a final evidence rating can be assigned.") Id.

[223] They are Family Check-Up, HFA, HIPPY, NFP and PAT. Sarah Avellar, et al., Home Visiting Evidence of Effectiveness Review: Executive Summary 10, tbl. 2(November 2014) (listing these five and a sixth program that was approved in September 2014).

a closer call. Although its outcomes are hopelessly inconsistent in most domains, its repeated positive findings for child maltreatment warrant approval despite the marginal effect sizes. Thus, NFP and HFA would qualify as evidence-based under the tougher standards proposed here.[224]

Using the stiffened approval requirements recommended here, most of the currently approved programs would be treated as promising, rather than evidence-based. Congress, it turns out, raced ahead of the research. To remedy this mismatch, Congress should temporarily expand the fraction of funding which can be used to try out and study promising models. Expanding the pool of money available to rigorously evaluate promising programs in real-world settings is precisely the right solution. It will provide states with more flexibility to find models whose services match the state's needs, while insisting that solid evidence be generated before confirming the model's eligibility for ongoing funding. Over time the list of proven programs will grow and federal funds will be allocated to programs that materially change children's lives.

V. CONCLUSION

The Home Visiting Program is Congress's most ambitious effort to create a funding stream in which every state can receive funding if it spends the funds on evidence-based programs. So far, however, this experiment with formula-based funding is a mixed success. Approval is simply too easy. At the same time, a landmark initial step has been taken and the weaknesses can be fixed.

On the positive side, Congress's halting effort to define high quality research was rescued by the staff at HHS. The agency produced research design standards that are remarkably strong.

On the negative side, the law's outcome requirements are much less robust. The Act's lack of tough requirements for effect size, duration, salience, consistency, and replication greatly weaken the Act's promise. Too many of the approved programs have minimal evidence of positive impact. Few would qualify under more defensible standards.

In the short run, HSS has reduced the harm done by these weaknesses by allocating nearly half of the funding through a competitive grant process, which can consider factors like the evidence of lasting impact and whether positive findings have been replicated. But the decision to allocate those funds competitively is only a temporary solution. In the long run, every state has infants and families who need effective services. As a result, Congress must find a way to insure that its formula-based funding is funneled to social service programs that make a meaningful difference in the lives of the people they serve. Its first step should be toughening the

---

[224] If some of the prerequisites suggested here were not adopted, then the number of additionally approved programs would turn on the requirement that was weakened or omitted. If replication were omitted, then ChildFIRST would be a strong candidate for approval because its only study found multiple and large positive effects in highly salient domains like child development and maternal mental health.

outcomes requirements in the Home Visiting Program.

Because that change will greatly shrink the pool of eligible programs and limit the ability of states to find programs that fit the needs of their residents, Congress should also temporarily allow a larger fraction of the funding to be used for promising programs that will be rigorously evaluated.

Only a combination of strong design requirements, meaningful outcomes thresholds, and rigorous evaluation of promising approaches has the potential to reconcile our desire to fund programs that work with the reality that rigorous evaluation of social services and educational interventions is only now becoming the norm. It offers a promising template for future evidence-based funding. Over time this combination of ingredients will identify innovative models that are even better than those we have today and direct federal funds to programs that change the lives of children, youth, and families.