WELFARE & ENTITLEMENTS

THE FEDERAL EXPERIMENT WITH EVIDENCE-BASED FUNDING

Standards for the Home Visiting Program are a good first step toward helping children, but even higher standards are needed.

BY PHILIP G. PETERS JR.

Federal and state governments spend hundreds of billions of dollars every year on social service and education programs that are well-intentioned, but typically unproven and often ineffective. As a result, Congress and President Obama have called for greater reliance on evidence-based programs.

One of the few federal programs to condition funding on such evidence is the Maternal, Infant, and Early Childhood Home Visiting Program (better known as the Home Visiting Program), a part of the 2010 Patient Protection and Affordable Care Act (ACA). The law's insistence that program effectiveness be proven with rigorous research provides a promising template for future state and federal funding legislation.

Unfortunately, Congress did not complement its stiff research design requirements for the program with equally meaningful requirements for the minimum outcomes needed for a home visiting model to qualify for federal funds. At present, the law has no minimum thresholds for effect size, duration, consistency, replication, or salience. To insure that it funds the efforts most likely to improve children's lives, Congress should add minimum requirements for each of these outcomes when it reauthorizes this funding program.

PHILIP G. PETERS JR. is the Ruth L. Hulston Professor of Law at the University of Missouri School of Law. He formerly was 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. Portions of this article are from a larger treatment of this subject in "Funding for Programs That Work: Lessons from the Federal Home Visiting Program," *Journal of Legislation*, Vol. 41 (2015).

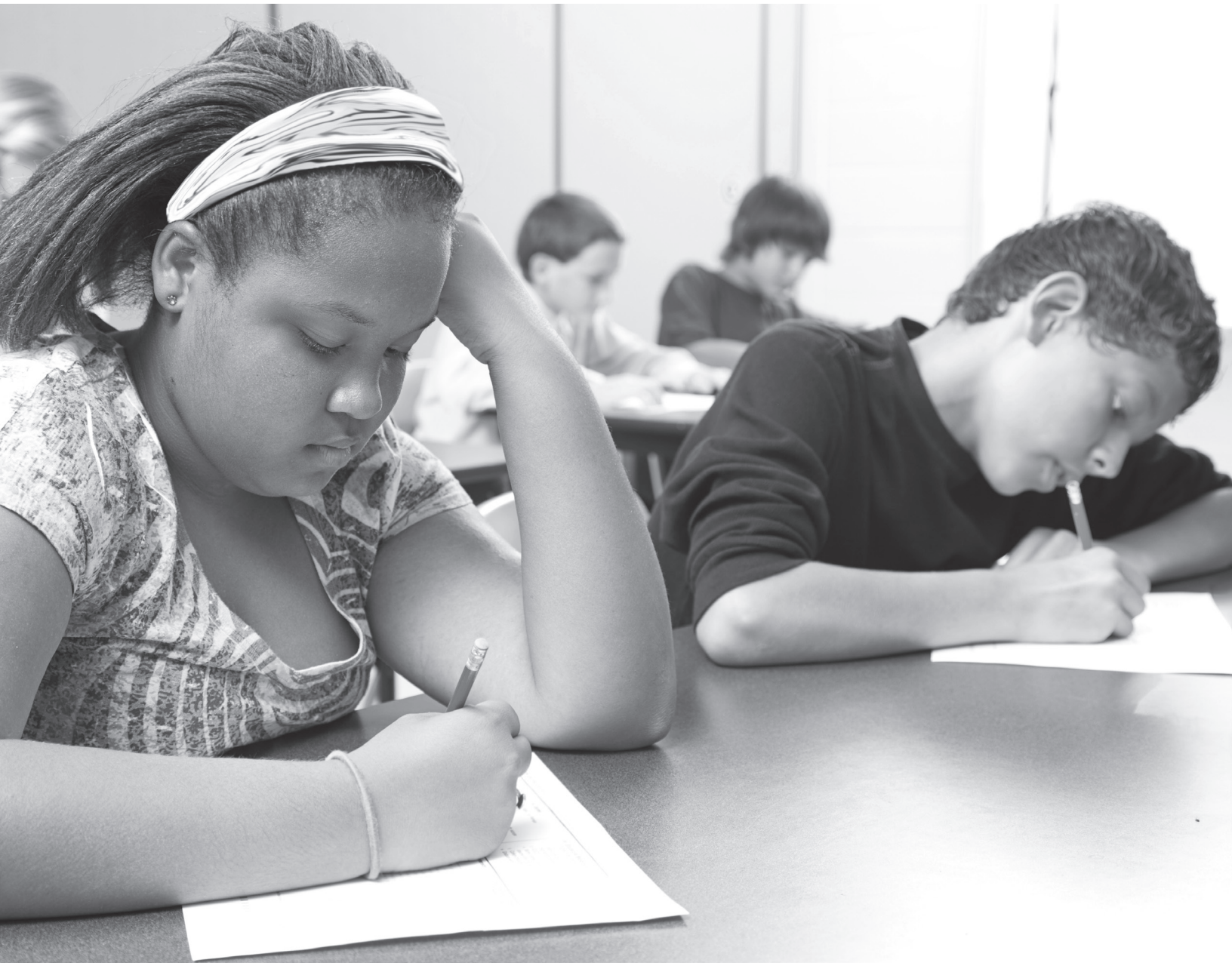
THE ACHIEVEMENT GAP AND HOME VISITS

The achievement gap emerges among infants and toddlers with shocking speed. By age 18 months, researchers regularly detect diverging trajectories between poor and upper-income children and between white and black children. Similar disparities surface in the development of early social skills.

When children reach kindergarten, the average poor or minority child is half a standard deviation or more below the mean on academic and social skills and further behind on vocabulary. 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 success. Early vocabulary, for example, is strongly associated with later school performance. Children who lag on achievement tests during their preschool years are more likely than their higher-performing classmates to be held back a grade, 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.

Because these gaps appear early in life, many child welfare



LISA F. YOUNG

experts advocate early intervention. One type of intervention, home visiting programs, has grown steadily over the past few decades. By the fall of 2009, just before the new Home Visiting Program was enacted, such programs operated in all 50 states and the District of Columbia. Total annual funding for these programs from private and public sources is estimated to be between $750 million and $1 billion, supporting home visits for an estimated 400,000 to 500,000 families.

Effectiveness research/ The initial case for investing public funds in home visiting programs was fueled decades ago by evaluations of programs like Healthy Families America and Parents as Teachers. The evaluations indicated the programs yielded significant improvement in parenting and child development. However, the evaluations rarely used reliable research designs. By 1995, serious questions were being raised about the actual benefits of these programs. A review of the research by Helen Barnes and colleagues concluded that "there is little research evidence to support the assumption that parent services affect child outcomes." Four years later, Deanna Gomby and colleagues were nearly as downbeat. They 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. The authors called the rarity of proven gains "sobering" and concluded that "children's development is better promoted through more child-focused interventions [like preschools]."

Then David Olds created the Nurse-Family Partnership (NFP) program in Elmira, N.Y., and replicated it in Memphis and Denver. The NFP program was evaluated in multiple randomized trials and consistently produced sizeable gains in child development among high-risk families. More encouraging, follow-up studies found the gains lasted into adolescence. The NFP program's success generated great enthusiasm and almost single-handedly widened support for home visiting programs.

The difference in outcomes, as measured by repeated and rigorous studies, between the earlier programs and Olds' NFP program illustrates the need for evidence-based funding. While some delivery models have produced positive outcomes, many have not. Congress quite reasonably decided to restrict funding to the models with proven effect.

CONGRESS ENACTS FEDERAL FUNDING

The intuitive appeal of parent education programs and the remarkable success of the NFP program prompted several lawmakers to work persistently for a decade to enact a federal funding program with an "evidence-based" requirement. Support for this concept was bipartisan and ultimately culminated in adoption of the Home Visiting Program.

Subtitle L of Title II of the ACA, which became Section 511 of Title V of the Social Security Act, created the Home Visiting Program and authorized $1.5 billion in funding over five years for states to create evidence-based home visiting programs. The U.S. Department of Health and Human Services announced that roughly half of those funds would be allocated on a formula basis to states that agreed to use an evidence-based home visiting model 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.

Defining "evidence-based" / As enacted, the Home Visiting Program allows states to select among home visiting models whose effectiveness had been demonstrated in either randomized controlled trials (RCTs) or quasi-experimental studies. While this is 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 bias their results. As a result, the HHS announced that it would use the discretion conferred on it by Congress to identify the kinds of RCTs and quasi-experimental designs (QEDs) that would be taken into account.

On July 23, 2010, the HHS proposed criteria to determine whether a home visiting program is evidence-based. 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 effect. Only high- and moderate-quality studies would be taken into account in determining whether a model is "evidence-based."

After receiving comments, the agency decided that the only studies that would be classified as "high quality" would be well-executed randomized controlled trials and two very specialized and relatively uncommon kinds of quasi-experimental studies (single-case-study designs and rigorous regression discontinuity designs). Randomization was favored because it 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.

Many supporters of existing home visiting programs vigorously objected to the preference for RCTs. Few studies of existing programs had used either randomized trials or the favored forms of QEDs. Nearly all programs relied heavily on other kinds of QEDs, such as the comparison of the treatment group to a convenience sample—that is, a sample that is readily accessible to the researcher—of children or families in the community. Under the proposed HHS rules, the latter type of study would be deemed "moderate-quality" at best. Although moderate-quality studies could be used for formula-based funding, they would be given less weight than RCTs in competitive funding.

The NFP program was evaluated in multiple randomized trials and consistently produced sizeable gains in child development among high-risk families. Follow-up studies found the gains lasted into adolescence.

The HHS imposed an additional requirement that meant that only a fraction of QEDs would qualify as "moderate-quality." 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. Findings from studies with a well-matched comparison group are much more reliable than those that use a convenience comparison group.

Many stakeholders asked that these well-matched QEDs be classified as high quality, like randomized trials. But the HHS decided that even well-matched QEDs were 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."

At any rate, most evaluations of home visiting programs that

had been done prior to creation of this federal program lacked a carefully matched comparison group. As a result, they are classified as low-quality and cannot be used to qualify a home visiting program for federal funding.

CRITICISMS OF RCTS

Critics of these demanding research design requirements made several arguments against the preference given to RCTs. Among those arguments:

- Equally important information comes from observational and quasi-experimental research.
- RCTs can be unethical because they deprive at-risk families of potentially effective assistance.
- RCTs are ordinarily narrow and thus overlook the synergies that occur among multiple social service programs.
- Programs proven at the small scale typical of randomized trials may not scale up effectively.
- The NFP program—the only home visiting program with very strong RCT findings—targets only a tiny fraction of children and families who badly need assistance.

None of those arguments are persuasive, for reasons discussed below. The HHS's demanding interpretation of the statutory text saved Congress from loose language that could have defeated its stated goal of allocating money to programs with a track record of changing children's lives.

Preference for RCTs / Nonscientists are often surprised to learn that observed correlations between family circumstances and child outcomes may not be causal. 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. However, the studies posed the risk of selection bias because the services were only provided to families who had made the effort to enroll in the program and reserve a time each month for hosting the parent educators in their own home. As a result, the differences in school readiness observed by the evaluators could have been produced by unobserved differences between the mothers who had made the effort to enroll their children and mothers who had not, rather than by the home visits. Selection bias of this sort can produce gains that are mistakenly attributed to the intervention. That apparently happened in the case of Parents as Teachers; later randomized trials could not replicate the findings of the earlier quasi-experimental studies.

Untested interventions can even be harmful. The Cambridge-Somerville youth project implemented an intervention to help boys at risk for delinquency. The program provided psychotherapy, tutoring, social activities, recreational activities, and family interventions. Many of the boys and their caseworkers praised the program. The value of these services seemed so obvious that assessment appeared to be 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 interventions whose proof of effectiveness is limited to QEDs often turn out to be ineffective, funders should require reliable proof that a proposed intervention has actually changed children's lives for the better. Intuitions and even well-established correlations are not enough.

Ethics of randomization / Critics of the agency's preference for RCTs also argue that the tests are often unethical. There are, of course, some circumstances in which that judgment is appropriate; 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. Researchers can evaluate promising home visiting models the same way.

These critics appear 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. Out of 250 home-visit models initially identified by the HHS, only nine had positive findings on high-quality studies. Rigorous research is necessary to find out whether programs, in fact, work.

Perhaps the most challenging situation for researchers arises 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 the HHS is well-suited for 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. Only the children falling below the cutoff would receive the intervention being studied. By targeting the children most at risk, this design avoids the ethical issue associated with randomization. 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.

To sum up, the vast majority of home visiting programs can be studied using RCTs. When researchers reasonably believe such tests would be unethical, HHS rules permit the use of rigorous QEDs. In particular, regression discontinuity designs may provide a good alternative.

Overlooked synergies / A third criticism against RCTs is that they force evaluators to focus narrowly on very specific outcomes, which could result in overlooking 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.

But if "the whole" is generating wonderful outcomes, why can't researchers measure them, too? If they cannot be reliably measured, then why believe that they really exist? Taken to its logical end, this argument would support the continuation of virtually every unproven program ever created.

Ironically, the very program used to illustrate the existence of these larger synergies has recently been studied and the hoped-for synergies could not be detected. The Harlem Children's Zone 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 city residents living outside the Zone be allowed to enter an enrollment lottery for places in the program. 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 the program's other 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 Harlem Children's Zone charter school were making exceptional gains, but the students who lived inside the Zone had no better school outcomes than the students who lived outside. Thus, the authors found no evidence that the comprehensive social services resulted in 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 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. This should not be interpreted as an argument for "scaling up" programs that showed no effect in a small-scale trial; they are not promising candidates for scaling up. As a result, success in a high-quality study, like a randomized trial, should be viewed as necessary, but not sufficient, to firmly qualify a home visiting program for scale-up funding.

Inability to serve families in need / When the Home Visiting Program was first proposed, only the NFP program was sufficiently supported by RCTs to be assured eligibility. Yet, the NFP 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, and mothers who exceed the income threshold. In addition, it offers a very specific 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.

Critics of the HHS research design requirements correctly contended that tough research design requirements would leave some categories of parents and children without a proven program to serve them. Some communities may need mentoring services for teen mothers; others may feel that mental health assistance should be prioritized. If the law's evidentiary standards are too strict, states seeking funding under the legislation will be less likely to find a home visiting program that fits their needs.

Home visiting programs like Parents as Teachers, Healthy Families America, and Home Instruction for Parents of Preschool Youngsters feared they would be left out. They enlisted national child advocacy organizations to make the case for looser eligibility requirements. 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.

In reaction, Congress revised the draft legislation to include quasi-experimental studies. But that potentially opened the door for the funding of programs that had only been "proven" in unreliable forms of QEDs. Fortunately, the HHS used its regulatory authority under the act to impose rigorous criteria on qualifying QEDs, including the requirement of a well-matched comparison group. If it had not, a congressional revision intended to make a wider array of services available would have inadvertently eviscerated the goal of targeting funds to programs with proven effectiveness.

As expected, the initial consequence was to limit the number of programs eligible for funding. Since then, however, evaluators seem to be making greater use of the most reliable research designs. Several additional programs have qualified for funding by participating in the kinds of studies that provide reliable evidence of their effectiveness. Congress should accelerate this process by permitting more of its funds to be used to do these high-quality assessments of promising new service models. Then states and citizens will have proven programs to fit their local needs.

Conclusions about design rigor / None of the criticisms of the legislation's current research design requirements are persuasive. The law now contains well-crafted research design standards that provide a template for future evidence-based funding efforts.

OUTCOME REQUIREMENTS

Unfortunately, the HHS 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 effect. Only a few would qualify under more defensible standards.

Minimum effect size / Under current rules, any statistically significant positive finding counts toward the agency's requirements for funding eligibility, no matter how trivial the effect. In fact, current rules do not even require the calculation of an effect size at all. As a result, some programs have been approved despite the absence of any estimate of effect size. When the act is next reauthorized, Congress should set a minimum effect size so that the funded models are limited to those likely to have a meaningful effect.

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 2.5 years of services. 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 effect 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.

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.

Initial effects from an intervention commonly shrink over time. As a result, only gains with very large effect sizes are likely to be durable. For example, 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 and a national study of ordinary preschools found that the short-term gains disappeared entirely. Cognitive gains of about 0.10 standard deviation detected in a recent rigorous study of Head Start also virtually disappeared in elementary school. Researchers have found similar fade-out when studying the advantages of full-day kindergarten, class size reduction, an extra daily class session in math, and even having a highly effective teacher.

Fade-out is normal. 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 the 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, NFP founder Olds was so concerned about loss of fidelity that he carefully limited the program's rate of expansion. As a result, funders should insist on proof of a large effect 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. The race and poverty gaps in academic achievement and emotional development are estimated to exceed 0.50 standard deviations at kindergarten entry and a full standard deviation by 12th grade. 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 percent of the gap.

Happily, most of the currently approved programs had effect sizes that meet or exceed that threshold. That is the good news. The bad news is that the current rules impose no barrier to funding future home visiting models with much smaller effects.

Until Congress imposes a minimum effect size, the HHS 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. 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 the NFP help at-risk children construct strong cognitive and emotional foundations on which to build during the K–12 years. Only durable gains in cognitive and social development accomplish that end.

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 that inexplicably applies only to RCTs. Positive findings from RCTs must be observed one year after program enrollment. The act calls these gains "sustained" even though a gain that lasts only as long as the treatment is being delivered can hardly be considered sustained.

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 fade out entirely by the end of the program's second year. In a study of San Diego's Healthy Families America program, for example, the gains observed at age 1 disappeared by age 3. Requiring that gains last a year from program initiation is the thinnest imaginable durability requirement.

The HHS should require repeated positive findings within a given domain. If the current rules did so, only five of the 14 approved models would qualify. That would shrink further if results had to be in the same constructs.

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 by 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 standard does not require replication. Only five of the 14 approved programs have replicated positive findings in the same domain. None of the seven most recently approved programs have them.

Consider the research on Healthy Families America. After a 2002 study found the program increased the number of well-baby visits attended by the mother and baby, studies in 2005 and 2007 using different samples could not replicate that finding. The researchers observed, "Our 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."

In addition, the definition of a replicated finding should be tightened. At present, the HHS requires only that two studies find positive effects 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 effect on child cognitive development but no effect on emotional development, combined with a second study finding a gain in emotional development but not in cognitive skills. Yet, neither study would actually have 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 effect on the same construct within a domain.

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 any replication study because any combination of positive findings in a given domain would satisfy the replication requirement. But 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 five positive findings based on chance alone, even if the program were totally ineffective.

To minimize this gamesmanship, the HHS should require repeated positive findings within a given domain. If the current rules did so, only five of the 14 approved models would qualify. That number would shrink still further if programs had to show a replicated positive effect on the same construct.

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 to their children daily. Surely, this is not how Congress intended to spend its money. The HHS has drafted a set of core outcomes for its post-grant evaluation of funded programs. Congress should authorize the use of this list at the front end of the funding process as well, so that programs and researchers know the outcomes that matter most.

Consistency of outcomes / The current approval process ignores findings of zero or negative benefits. Only positive findings are considered. If 12 studies evaluate a home visiting program and only one finds any positive effect, the model would nevertheless qualify for approval if the study found positive effects in two

domains. In fact, the model would be eligible for funding even if the 10 other studies found that the program *impaired* child development! Only the positive findings count.

The body of research on Healthy Families America illustrates this problem. The program has been studied many times. It can now boast at least one positive finding in each of the eight domains used by the HHS. Yet, the program's batting average is much less glorious than this statistic would imply. Rigorous studies have found a positive effect on less than 10 percent of the constructs measured (43 of 494). In two domains, the rate was so low that the few positive findings could easily have been the result of chance (one of 30 in family violence and three of 72 for maternal health.)

Negative findings are also ignored. The Parents as Teachers program again offers an illustration. Rigorous studies of the program have taken 208 measurements. In 196 of those, the program was found to have conferred no benefit. The remaining 12 assessments found seven negative or ambiguous effects and five positive effects. Fortunately for the program, 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 200 findings of ineffectiveness.

Consider another example from the Healthy Families America research. This study found a favorable effect on well-baby visits. Two later studies found no effect. Should the single positive finding count toward approval, despite two null findings on the same construct? It does now.

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 effect. Careful stewardship would take this into account.

Summary of recommended evidentiary changes / 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.

In the interim, the HHS has taken two important steps to minimize this weakness. 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 outcomes that each state must measure. The constructs that emerge from this process can become future benchmarks in the ex ante approval process. With this list of common benchmarks, the HHS could require that home visiting programs demonstrate positive, durable, replicated, and consistent effect on one or more of the key constructs.

CONCLUSION

The Home Visiting Program is Congress's most ambitious effort to create a funding stream in which every state can receive federal money if it spends the funds on evidence-based programs. So far, however, this experiment with formula-based funding is a mixed success. On the positive side, Congress's halting effort to define high-quality research was rescued by the HHS staff. 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 legislation's promise. As a result, many of the approved programs would fail to qualify under more defensible standards.

In the short run, the HHS 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 effect 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 legislate that its formula-based funding is funneled to social service programs that make a meaningful difference in the lives of the people they serve. *R*

READINGS

- "Disparities in Early Learning and Development: Lessons from the Early Childhood Longitudinal Study— Birth Cohort (ECLS-B)," by Tamara Halle, Nicole Forry, Elizabeth Hair, et al. Washington, D.C.: Child Trends, 2009.
- "Effects of Nurse Home-Visiting on Maternal Life Course and Child Development: Age 6 Follow-Up Results of a Randomized Trial," by David L. Olds, Harriet J. Kitzman, Robert E. Cole, et al. *Pediatrics*, Vol. 114 (2004).
- "Home Visiting: Recent Program Evaluations—Analysis and Recommendations," by Deanna S. Gomby, Patti L. Culross, and Richard E. Behrman. *Future of Children*, Vol. 9, No. 1, 1999.
- *Inequality at the Starting Gate*, by Valerie E. Lee and David Burkham. Economic Policy Institute, 2002.
- "Introducing the Issue," by Cecilia Rouse, Jeanne Brooks-Gunn, and Sara McLanahan. *Future of Children*, Vol. 15, No. 1 (2005).
- *National Evaluation of Family Support Programs: Review of Research on Supportive Interventions for Children and Families*, Vol. 1, by Helen V. Barnes, Barbara D. Goodson, and Jean I. Layzer. Abt Associates, 1995.
- "Prenatal and Infancy Home Visitation by Nurses: Recent Findings," by David L. Olds, Charles R. Henderson Jr., Harriet J. Kitzman, et al. *Future of Children*, Vol. 9, No. 1 (1999).
- "School Readiness: Closing Racial and Ethnic Gaps," by Jeanne Brooks-Gunn, Cecilia Elena Rouse, and Sara McLanahan. In *School Readiness and the Transition to Kindergarten in the Era of Accountability*, edited by Robert C. Pianta, Martha J. Cox, and Kyle L. Snow. Brookes Publishing, 2007.
- "The Black-White Test Score Gap through Third Grade," by Roland G. Fryer Jr. and Steven D. Levitt. *American Law & Economics Review*, Vol. 8, No. 2 (2006).
- "The Black-White Test Score Gap: An Introduction," in *The Black-White Test Score Gap*, by Christopher Jencks and Meredith Phillips. Brookings Institution Press, 1998.