Evaluation and Program Planning 55 (2016) 163–173

Contents lists available at ScienceDirect

Evaluation and Program Planning

journal homepage: www.elsevier.com/locate/evalprogplan

Measuring program- and individual-level fidelity in a home visiting program for adolescent parents

Jessica Goldberg*, Erin Bumgarner[1], Francine Jacobs

Tufts University, 574 Boston Ave., Medford, MA 02155, United States

ARTICLE INFO

Article history:
Received 18 September 2015
Received in revised form 13 December 2015
Accepted 22 December 2015
Available online 2 January 2016

1. Introduction

Newborn and early childhood home visitation – a two-generation approach to providing support to families with young children – has garnered increasing public attention over the past 30 years. Long supported by state and local governments, as well as private dollars, home visiting first received major federal funding in 2010, through a provision within the Affordable Care Act (Public Law 111–148) entitled the Maternal, Infant, and Early Childhood Home Visiting (MIECHV) program. This amendment to Title V of the Social Security Act funds states, tribes, and territories to implement home visiting services in at-risk communities. At least 75% of the $1.5 billion dollars of funding was reserved for home visiting models that met federal standards for effectiveness[1] (Adirim & Supplee, 2013). As of the end of 2014, based on a review of what the U.S. Department of Health and Human Services considered to be rigorous evaluations, 17 models had been deemed *evidence-based.*

Most of the MIECHV-funded models, which currently operate in 721 counties in all 50 states, the District of Columbia, and five territories, originated as single demonstration projects. The Nurse-Family Partnership, for example, which now operates in 43 states, the U.S. Virgin Islands and six Tribal communities (Nurse-Family Partnership, October 2014), had its beginnings almost three decades ago in one small program in Elmira, NY (Olds et al., 1997). Healthy Families America (HFA), which currently has 600 affiliates in 40 states, the District of Columbia, and all five U.S. territories, started in 1991 with a small handful of sites (Holton & Harding, 2007). This type of expansion or "scale-up" of evidence-based models can be seen as a franchise form of replication, in which there is a central organization or model developer that is responsible for defining standards and, to some extent, monitoring performance in locally-implemented programs (Bradach, 2010; Yoshikawa, Rosman, & Hsueh, 2002).

Two central assumptions undergird this "scaling what works" approach (Bradach & Grindle, 2014): first, that those outcomes observed in the demonstration programs will accrue in the scaled-up replicated programs as well (Dees, Anderson, & Wei-Skillern, 2004; Yoshikawa et al., 2002), and second, that local programs will be able to routinely implement those particular core elements that contribute to the success of the model (Bradach, 2010; Bradach & Grindle, 2014; Carroll et al., 2007; Dees et al., 2004; Yoshikawa et al., 2002). Evidence from the past several years of home visiting expansion, however, calls both assumptions into question. Impact evaluation results have been consistently lukewarm, with only a scattershot of benefits observed across outcome domains, populations, program sites, and evidence-based models (Avellar et al., 2014; Avellar & Supplee, 2013; Azzi-Lessing, 2011; Bobbitt, Osborne, & Bradbury, 2015; Duggan et al., 2013; Rubin, Curtis, & Matone, 2014; Sweet & Appelbaum, 2004). And, even the most tightly-designed and monitored home visiting programs are beset

* Corresponding author.
E-mail addresses: jessica.goldberg@tufts.edu (J. Goldberg), erin_bumgarner@abtassoc.com (E. Bumgarner), francine.jacobs@tufts.edu (F. Jacobs).
[1] Present address: Abt Associates, Inc., 55 Wheeler Street, Cambridge, MA, United States 02140.

[1] HomVee conducted an exhaustive search of the home visiting evaluation literature, assessing program performance in eight outcome domains (child health, child development and school readiness, family economic self-sufficiency, linkages and referrals, maternal health, positive parenting, reductions in child maltreatment, and reductions in family violence and crime). Home visiting models that met one of the following criteria were considered evidence-based: "at least one high- or moderate-quality impact study of the model finds favorable, statistically significant impacts in two or more of the eight outcome domains; at least two high- or moderate-quality impact studies of the model using non-overlapping analytic study samples with one or more favorable, statistically significant impacts in the same domain" (Avellar et al., 2014).

http://dx.doi.org/10.1016/j.evalprogplan.2015.12.007
0149-7189/

by what Jacobs (2003) wryly describes as "intrusions of context": shifting policy environments, variations in agency and community settings, differences in organizational and staff dynamics, and marked variability in how the services themselves are delivered (Damschroder et al., 2009; Matone et al., 2013). In fact, seen in aggregate, few home visiting programs have been implemented precisely as intended by the model (Boller et al., 2014; Daro, Boller, & Hart, 2014).

The current paper builds on this newly emerging literature by presenting results from a descriptive, cross-site analysis of model fidelity in Healthy Families Massachusetts (HFM), a statewide home visiting program for adolescent parents. Here we focus on key aspects of "structural" fidelity (Boller et al., 2014), related to the delivery of direct services to participants. Collected as part of a randomized controlled trial (RCT)[2] of HFM, program data are used to describe fidelity from two perspectives: that of the local program sites (i.e., the extent to which each program, as a whole, performs in accordance with the model) and of the individual study participants (the extent to which each participant receives services in accordance with the model). The relations among these two fidelity constructs, single indicators of utilization, and maternal characteristics, are also discussed.

1.1. The Five-tiered Approach to evaluation

The framework that guided this investigation is Jacobs' Five-tiered Approach (FTA) to evaluation (Jacobs, 1988, 2003). The FTA takes a developmental view of evaluation, moving evaluation activities from a primary focus on descriptive and process-oriented information to an emphasis on program effects. Tier One activities produce needs and demand assessments, and usually are conducted prior to the program's implementation. Evaluation activities at Tiers Two and Three are directed at program processes: they describe program staff, services, clients, and costs; examine program implementation compared to model standards; and provide feedback to programs for improvement. Tiers Four and Five focus on outcome evaluation activities, assessing the extent to which a program is meeting its short-term and long-term goals. The investigation of model fidelity, as described in the present paper, is rooted in analyses conducted at Tiers Two and Three.

1.2. Virtues of fidelity assessment

Meta-analyses across a wide range of child and family programming (e.g., mentoring, anti-bullying, and drug prevention programs) provide evidence that program effects are more robust when interventions are well-implemented – that is, when fidelity is high (Chaffin & Friedrich, 2004; Derzon, Sale, Springer, & Brounstein, 2005; DuBois, Holloway, Valentine, & Cooper, 2002). These studies confirm the proposition that: (1) when programs are developed with an accurate view of how phenomena of interest operate and what an efficacious intervention into that process would be (e.g., the set of corrective measures that reflect that understanding), and (2) the intervention is implemented as intended, then (3) the expected benefits accrue to participants. In broad strokes, the first element of that proposition represents *program theory*; good theory is a necessary, but insufficient, condition for achieving a program's desired results. Here the second element – *fidelity* – is critical as well, in order to achieve the expected participant benefits (Birckmayer & Weiss, 2000; Weiss, 1995, 1997).

Operating with this proposition in mind, researchers use assessments of fidelity to explain negative or null results. In the event that a program "does not work," an understanding of how services were implemented in relation to the model design can help distinguish theory failure from implementation failure (Birckmayer & Weiss, 2000; Durlak & DuPre, 2008; Jacobs, 2003). On establishing implementation failure, it is sometimes argued that the program would have worked if only it had operated correctly. That could, indeed, be the case, but can only be validated empirically, since it is also the case that the program might reflect inadequate theory, alone or in combination with poor implementation.

In addition to contextualizing the internal validity of program evaluation findings, assessments of implementation fidelity can also shed light on the extent to which results can be replicated and generalized. There is great variability in how programs are implemented; assessments of program fidelity can demonstrate this variability and help explain why some studies find positive effects of the program on participants while others find null or negative effects, even when examining identical program models (Carroll et al., 2007; Durlak & DuPre, 2008). In these circumstances, and when programs produce positive outcomes despite inadequate model fidelity, there are also opportunities to refine or reimagine program theory, and ultimately, to improve programs (Durlak & DuPre, 2008).

1.3. Measuring model fidelity in home visiting

Until recently, home visiting evaluations did not routinely include an assessment of model fidelity; instead, researchers would describe a set of program protocols or procedures that programs were expected to follow (e.g., participants receive weekly visits over three years) without documenting the services *actually delivered* (e.g., participants on a weekly service level received an average of 23 visits over 12 months) (Bilukha et al., 2005; Duggan et al., 2013; Duggan et al., 2000; Filene, Kaminski, Valle, & Cachat, 2013; Paulsell, Del Grosso, & Supplee, 2014). In recent years, however, due in large part to funding through the MIECHV initiative, this line of investigation has become more common, with most large evaluations including at least some kind of assessment of program performance in relation to key model standards.

Typically, descriptions of program fidelity in home visiting evaluations are focused on a handful of utilization indicators (Paulsell et al., 2014), along such dimensions as program reach (Mihalic, Fagan, & Argamaso, 2008), adherence to visit schedule (Drotar, Robinson, Jeavons, & Lester Kirchner, 2009; Duggan et al., 2004), curriculum/visit content (Drotar et al., 2009; Saïas et al., 2012), match between program goals and staff training/ability (Duggan et al., 2007), and program differentiation and adaptation (Berkel, Mauricio, Schoenfelder, & Sandler, 2011; Boller et al., 2014; Carroll et al., 2007; Damschroder & Hagedorn, 2011; Durlak & DuPre, 2008). Notable for its breadth and depth, the cross-site evaluation of the Supporting Evidence-Based Home Visiting to Prevent Child Maltreatment initiative (EBHV) represents one of the more comprehensive investigations of fidelity in the home visiting field. This multi-site, multi-model study includes examination of a host of what the authors term *structural* and *dynamic* fidelity indicators; structural components include those elements reflecting basic aspects of adherence (e.g., dosage, caseload numbers, staff retention), and dynamic components refer to aspects of the relationship between providers and participants (Boller et al., 2014). The authors of this study found that, overall, home visiting programs appear to have particular difficulty adhering to structural fidelity elements related to dosage (Boller et al., 2014). This is not surprising, given that it is almost axiomatic in the home visiting

[2] For a full description of the study, see Tufts Interdisciplinary Evaluation Research (2015).