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

CrossMark

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

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

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).

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).

Table 1
Healthy Families America Critical Elements.

Area	Critical element
Service initiation	1. Initiate services prenatally or at birth.
	2. Use a standardized assessment tool to systematically identify families who are most in need of services. This tool should assess the presence of various factors associated with increased risk for child maltreatment or other poor childhood outcomes.
	3. Offer services voluntarily and use positive outreach efforts to build family trust.
Service content	4. Offer services intensively (for example, at least once a week) with well-defined criteria for increasing or decreasing frequency of service and service over the long term (for example, three to five years).
	5. Make services culturally competent such that the staff understands, acknowledges, and respects cultural differences among participants; staff and materials used should reflect the cultural, linguistic, geographic, racial, and ethnic diversity of the populations served.
	6. Focus services on supporting the parent as well as supporting parent–child interaction and child development.
	7. At a minimum, link all families to a medical provider to ensure optimal health and development. Link families to additional services, as needed.
	8. Limit staff caseloads to ensure that home visitors have an adequate amount of time to spend with each family to meet their unique and varying needs and to plan for future activities.
Program administration	9. Select service providers because of their personal characteristics, their willingness to work in or their experience working with culturally diverse communities, and their skills to do the job.
	10. Train service providers about their role so they understand the essential components of family assessment and home visitation.
	11. Give service providers a framework, based on education or experience, for handling the variety of situations they may encounter when working with at-risk families. All service providers should receive basic training in areas such as cultural competency, substance abuse, reporting child abuse, domestic violence, drug-exposed infants, and services in their community.
	12. Give service providers ongoing, effective supervision so that they are able to develop realistic and effective plans to empower families to meet their objectives; to understand why a family may not be making progress and how to work with the family more effectively; and to express their concerns and frustrations so that they can see that they are making a difference and avoid stress-related burnout.

Source: www.healthyfamiliesamerica.org.

literature at this point that families in home visiting programs tend to receive fewer visits, over a shorter period of time, than prescribed by the program model (Boller et al., 2014; Daro, 2006; Duggan et al., 2013; Duggan, Rodriguez, Burell, Shea, & Rohde, 2005; Howard & Brooks-Gunn, 2009; Jacobs, 2003; Matone et al., 2013).

While there has been quite a bit of research examining model fidelity at the program level, and research examining how individuals utilize services in relation to the home visiting model, to our knowledge there has been little attention paid to how these two arguably disparate phenomena differ and/or overlap. In other words, while the overall service delivery *environment* of a particular program can be understood using aggregated program-level fidelity data, those variables do not necessarily explain how the individuals nested within these programs personally experience services (Loney & Nagelkerke, 2014; Schwartz, 1994). Indeed, the fidelity score that a given program demonstrates across their total clientele may well present a different picture than does the fidelity score that any individual client exhibits on items relative to her or his program participation. And those individuals who are considered low fidelity users, within the context of high fidelity programs, may be particularly important to identify and understand. The study described here—an investigation of a similar set of fidelity indicators from the perspectives of the program, and the individuals within that program—represents an initial attempt at parsing, and making sense of, these two arguably disparate aspects of model fidelity.

1.4. Current study

This study builds upon and extends the existing home visiting implementation literature on program fidelity. Focusing largely on data concerning engagement and service dosage, it uses home visiting program data to create two composite fidelity variables comprising several performance indicators: (1) a *program*-level fidelity score, reflecting the degree to which programs operated as intended by the model, and (2) an *individual*-level fidelity score, reflecting the degree to which individual evaluation participants utilized services as the model intended. There are several advantages to this methodology: it allows for comparisons of program and participant behavior on a similar set of indicators; leverages program data that are reported by the home visitor in a real-time web-based system; synthesizes several individual

indicators into a continuous composite measure of fidelity for use in analyses; and can be systematically applied across programs, models, and participant samples.

2. Methods

This section presents the methods used to construct the program- and individual-level fidelity constructs, beginning with an overview of HFM, including an overview of the performance standards identified by HFM, followed by a description of the sample, data sources, and steps taken to construct the final variables.

2.1. Healthy Families Massachusetts

HFM, an affiliate of the paraprofessional[3] home visiting model, Healthy Families America (HFA), provides home visiting services, including: goal-setting; curriculum-based activities and family support tailored to the individual family; routine developmental and health screenings; and linkages to medical and other services as needed. Its five stated goals are: (1) Prevent child abuse and neglect by supporting positive, effective parenting; (2) Achieve optimal health, growth, and development in infancy and early childhood; (3) Encourage educational attainment, job, and life skills among parents; (4) Prevent repeat pregnancies during the teen years; and (5) Promote parental health and well-being. Implemented by 25 agencies across the state, HFM is available to every first-time parent in Massachusetts under the age of 21 years. Agencies are selected through a request for response process, and represent multiple social service sectors, including community health, early childhood, and child abuse prevention organizations. Parents can enroll during pregnancy until the child's first birthday, with services available through the child's third birthday.

---

[3] The term "paraprofessional" is used by HFA to denote that the model does not require home visitors to have formal education credentials, instead prioritizing other hiring criteria, such as knowledge of the local community, personal characteristics, and experience building trusting relationships. That said, the majority of HFA home visitors, including those employed by HFM, have had at least some college/training in their fields (Holton & Harding, 2007).

2.1.1. Healthy Families America performance standards

HFA is designed to flexibly respond to families', communities', and programs' needs (Kessler, Nixon, & Nelson, 2008); however, there are 12 aspects of implementation that are considered "critical" by the model developers (http://www.healthyfamilies-america.org/). These critical elements (see Table 1) center on three broad areas, including: service initiation, service content, and program administration. To be an accredited HFA provider, all affiliates must demonstrate adherence to these 12 critical elements. As can be seen in the table, some of critical elements lend themselves far more to concrete and objective operationalization and measurement (e.g., *Initiate services prenatally or at birth*) than others (e.g., *Make services culturally competent . . .*).

2.1.2. HFM performance standards

The agency that administers HFM, the Children's Trust of Massachusetts, has developed a statewide quality assurance and improvement infrastructure that includes frequent site visits, shadowing and supervision, and intensive training, aimed at ensuring adherence to the HFA critical elements. There is a strong emphasis on using program data for monitoring and accountability, and the Children's Trust sends each site an annual data report assessing its practices with regard to 21 *performance indicators of excellence* (see Table 2 for a complete list of indicators, including those used for this study). The HFM indicators, which pertain to outreach and referrals, service initiation, service delivery and content, and program administration, map loosely onto those the more tangible HFA critical elements under the categories of service initiation, service delivery and administration. Those six indicators noted as "sentinel" in Table 2 are considered most critical to program success by HFM administrators.

2.2. Data sources

This study relied primarily on program data extracted from the Participant Data System (PDS), a web-based management information system administered and maintained by the Children's Trust. The PDS is used by home visitors and supervisors to provide background information about participants (e.g., pregnancy and birth information), details about service planning and utilization (e.g., referral, enrollment, and service levels, the frequency and content of home visits and other services), Individual Family Service Plan (IFSP) goal setting and attainment, child and mother assessments and status reports, and discharge records. These data were used to assess program-level fidelity (i.e., how programs were implemented in accordance with HFM's performance indicators of excellence), as well as individual-level fidelity (i.e., how mothers utilized services in accordance with HFM's performance indicators of excellence). They were also used to create individual service utilization measures (e.g., number of home visits received, length of enrollment).

This study also incorporated several measures of maternal characteristics. These data were collected by the evaluation team when mothers enrolled into the HFM evaluation. Mothers who agreed to participate in the evaluation were asked to complete a half-hour interview on the phone, and sign a release allowing our evaluation team to access her administrative data from several state agencies. Individuals who participated in these initial data collection efforts were then given the option of completing an additional in-person interview. These data collection protocols included a wide variety of assessments and measures, covering topics such as demographic characteristics (e.g., maternal age, race/ethnicity, parenting status at enrollment), maternal wellbeing (e.g., maternal depression, social support, mother's own history of abuse and neglect), maternal employment and education, financial resources, living arrangements, and child care arrangements.

Table 2
[HFM Program Performance Indicators, with Those Used to Calculate Individual- and Program-Level Fidelity Scores Noted.

Performance Indicator (Benchmark)[*]	Program-level Fidelity	Individual- level Fidelity
Performance Area: Outreach and Referrals		
1. Program receives referrals for eligible parents within the catchment area (100% of referrals)		
2. Program receives referrals for eligible parents from each city and town within the catchment area (100% of cities and towns in catchment area)		
3. Program receives referrals for parents during their prenatal period (60% of referrals)	✔	✔
Performance Area: Service Initiation		
4. Program makes first contact with new participants either prenatally or within two weeks of birth (80% of clients)	✔	✔
5. Program makes first contact with new participants, on average, within 10 days from the referral (100% of clients)	✔	✔
6. Program completes a first home visit with new participants, on average, within 20 days from referral (100% of clients)	✔	✔
7. Eligible parents referred to the HF program accept services (90% of clients)	✔	–
Performance Area: Service Delivery and Content		
8. Participants receive weekly home visits for at least six months following the birth of their baby (100% of clients)[**]	–	✔
9. Program provides intensive home visiting services to participants with a minimum average of 18 visits per participant (12 visits in FY08 & FY09; 100% of clients)[**]	✔	✔
10. Participants receive 75% of their visits according to their service level (75% of clients)[**]	✔	✔
11. Program provides home visits to participants with each participant receiving at least one home visit (100% of clients)[**]	✔	✔
12. Participants receive long term services with a minimum average of 18 months of service at discharge (100% of clients)[**]	✔	✔
13. Participants have an active IFSP that has been initiated within the last six months (100% of clients)		
14. Program administers the ASQ developmental screen to all children according to ASQ guidelines (100% of children)		
15. Participants receive at least one visit annually with both parents (100% of clients)		
16. Program provides group services including a parent support group series, a parent child interaction group, and a social activity group		
Performance Area: Program Administration		
17. Program meets capacity expectation for the number of families serviced (based on program capacity)		
18. Program supervisors provide home visitors with weekly supervision lasting 1.5 h (85% of home visitors)[**]	✔	–
19. Program abides by all policies as defined in the HFM Policies & Procedures manual		
20. Program submits all billing by the 10th of each month		
21. Program submits MSDRs by the 10th of each month		

[*] Benchmarks are noted only for those indicators included in the fidelity scales for the current paper.
[**] Denotes "sentinel" indicators, per the Children's Trust.

2.3. Analysis sample

Data used for this study were collected as part of a large-scale, longitudinal impact evaluation of HFM. Participants were recruited from 18 HFM sites across the state. To be eligible to participate in the RCT, participants had to be female, at least 16 years old, conversant in English or Spanish, new to HFM, and cognitively able to provide informed consent. Random assignment occurred at the HFM site level through an algorithm in its web-based management information system. (To minimize denial of services, 40% were assigned to the control group and 60% were assigned to receive services.) Program group participants could receive HFM home visiting; control group participants could not receive HFM, but were provided information about child development and referrals to other services.

In total, 837 participants were randomly assigned (517 program; 320 control). From this group, 704 mothers (433 program; 271 control) participated, at minimum, in an initial (Time 1, T1) phone interview or signed a release allowing access to data from the Massachusetts agencies of child protection, public health, transitional (financial) assistance, and education. Of these, 475 mothers (274 program; 199 control) completed the additional in-person interview. The analysis sample used here includes only program mothers who completed a T1 phone interview (n = 433).

2.3.1. Maternal characteristics at enrollment

Mothers in the evaluation were 18.6 years old, on average, and about two-thirds (65%) of mothers were pregnant at time of enrollment; the remainder was already parenting. Over two-thirds of the sample self-identified as non-Hispanic White or Hispanic (37% and 36%, respectively); others self-identified as non-Hispanic Black (19%) or non-Hispanic "other" (8%). Mothers exhibited a wide range of psychosocial risk and vulnerabilities; over one-third of mothers reported experiencing at least two partner-perpetrated acts of intimate partner violence (36% at T2 and 38% at T3), and more than half (55%) had substantiated reports of maltreatment as a child, according to administrative data. On average, mothers experienced three traumatic events in their lives, and the majority met the criteria for either full or partial Post Traumatic Stress Disorder (PTSD; 39% met full criteria, 29% met partial criteria, and 31% did not meet criteria).

2.3.2. Program utilization

Mothers enrolled in HFM for an average of 448 days (almost 15 months; range = less than one month to 3.5 years) and received 24 home visits over the course of their enrollment in the program, with a range of 0 to 118. Fourteen percent of the sample did not receive any home visits at all. The median values, which perhaps provide a more accurate representation of participants' service use, were 14 visits and 10 months of enrollment. Secondary activities (i.e., non-visit activities conducted by the home visitor or HFM staff with, or on behalf of, the participant such as phone calls to the mother, unannounced drop by visits, phone calls to other social service programs or government agencies) were reported 62 times per mother, on average (median = 43). On average, mothers attended two group-based activities (e.g., social support groups, parenting education), with a median value of 0 and range of 0 to 28, indicating that the average was likely inflated by a small number of mothers who attended many groups.

2.4. Selecting program performance indicators

This study focused on a selected number of HFM performance indicators. There were several principles guiding our decision about which indicators to use. First, we included at least one indicator from each of the four performance areas. Second, we included indicators designated as "sentinel" by the Children's Trust. Third, we chose indicators that have been reported on/correlated with outcomes in other home visiting program evaluations. Fourth, we selected indicators that would lend themselves to fairly straightforward adaptation at the individual level. This last principle meant that indicators such as *17. Program meets capacity expectation* and *19. Program submits all billing by the 10th of each month* were excluded. In theory, it should also have resulted in the exclusion of *Indicator 18. Supervisors provide home visitors with weekly supervision.* However, because we were intent on including indicators that satisfied the first three principles, we included this in our program fidelity measure despite our not being able to include it in our individual-level measure. (We only had access to these data at the aggregate program level—not that of individual home visitors who could be matched to our participants.) And finally, we included only those indicators for which there were reliable data across fiscal years. Because of this last criterion, there were three indicators we would have liked to include—*13. Participants have an active IFSP; 14. Children are screened using ASQ;* and *15. At least one visit annually should include both parents*—but were unable to because of poor data integrity. Table 2 shows the final list of indicators used for the program-level and individual-level fidelity scores.

2.5. Calculating program-level fidelity scores

The program-level fidelity score was derived from data for ten HFM performance indicators of excellence, across 26 program sites, and four fiscal years (FY 2008 through FY 2012). These data were available for all individuals who enrolled in HFM (including parents who were not part of the HFM evaluation). Creating program-level fidelity scores was a multi-step process. First, we calculated a score for each indicator, for each program site, for each fiscal year, by dividing the program's performance indicator score for that year by the benchmark goal. As an example of how this works, consider the fourth indicator: *Indicator 4. Program makes first contact with new participants either prenatally or within two weeks of birth.* The benchmark for this performance indicator is set at 80%, meaning that a program can only meet this performance indicator if it makes contact with at least 80% of referred participants when they are still pregnant, or within two weeks of birth. If a site was only able to contact 70% of its clients prenatally or within two weeks of birth in FY11, this site would receive a score of 0.875 for that indicator for that fiscal year (the quotient of the 70% performance rate divided by the benchmark goal of 80%). If in FY12, the site was able to contact 80% of clients prenatally or within two weeks of birth, its score on that indicator for that fiscal year would be 1.0 (80% performance divided by 80% benchmark goal). Scores were capped at 1.0, even if a program exceeded the benchmark, because there is not sufficient evidence to support the assumption that more is always better. In theory, then, fidelity scores could range from 0.0 to 1.0.

We then averaged indicator scores across fiscal years, to arrive at a four-year average score, for each indicator, for each site. These scores, in turn, were averaged across indicators for each program, resulting in a single site-level composite score reflecting all indicators across all fiscal years. Finally, we assigned a program-level fidelity score to each of the RCT program group participants, according to the program in which she was enrolled the longest. In other words, participants who transferred from one HFM site to another site (i.e., 12% of the sample), were assigned a program-level fidelity score that corresponded to the site in which she was enrolled the longest.

2.6. Calculating individual-level fidelity scores

The indicators we used to develop individual-level fidelity scores were the same as those included in the program-level fidelity composite, with a few exceptions (see Table 2). *Indicator 7* (*acceptance*) was excluded because it was not applicable to individual mothers (who, by definition, had all accepted the program), and *Indicator 18* (*supervision*) was excluded because we did not have access to individual-level data on how much supervision each participant's home visitor(s) received. Further, one additional indicator that had been excluded from the program-level measure due to unreliable program-level data (*Indicator 8. Participants receive weekly home visits for at least six months following the birth of their baby*) was included in the individual-level fidelity scale.

To calculate a total individual-level fidelity score, we created a dichotomous variable to indicate whether the mother met each program indicator (e.g., for *Indicator 1* [*prenatal enrollment*], 1 = yes, mother was referred prenatally; 0 = no, mother was referred postpartum). Then, a total score was created by dividing the number of indicators that were met by the total number of program indicators. Like the program-level fidelity score, possible values on the individual-level fidelity score ranged from 0 (indicating the mother did not meet any program indicators) to 1 (indicating the mother met all program indicators). The rate of missingness on program indicator data was low; of the 433 mothers in the analysis sample, 85% had data on each of the 10 program indicators, 12% were missing data on just one program indicator, and 3% were missing two to three program indicators. Given the low rates of missingness, this study used listwise deletion, such that any participants with missing data were dropped from analyses.

We created two individual-level fidelity subscales; one subscale included program indicators related to *initial* exposure to the program (e.g., HFM made first contact with the participant within ten days from the referral), and the other included those indicators related to *overall* exposure to the program (e.g., participant received 75% of her visits according to her service level). These scores could range from 0 to 1 as well.

3. Results

In this section we review findings pertaining to program-level and individual-level fidelity scores, and results of descriptive analyses exploring the relations between fidelity measures and single indicators of utilization, and participant characteristics.

3.1. Program-level fidelity

Table 3 shows average program performance on each indicator, by fiscal year. For some indicator goals, program performance was consistently close to the benchmark across fiscal years (e.g., *Indicator 3: 60% of referrals are prenatal*) for some, consistently low (e.g., *Indicator 12: 18 month duration*), and for some, there was improvement across fiscal years (e.g., *Indicator 5: first contact within 10 days*). Only for *Indicator 9* (*completed visits per year*) did program performance significantly worsen from FY08 to FY11, likely due to the raising of the goal in FY10 from 12 visits per year to 18.

As seen in Fig. 1, once averaged across fiscal years and indicators, and assigned to participants according to the program in which they spent the most time, this variability in program performance is obscured. When seen in aggregate, participants appeared to be enrolled in consistently high-fidelity programs; the average program fidelity score for participants was 0.74, with a narrow range in scores from 0.71 to 0.80.

3.2. Individual-level fidelity

Individual-level fidelity scores reflect each HFM evaluation participant's utilization of services, in relation to the HFM indicators. In contrast to program-level fidelity scores, individual-level fidelity scores were widely distributed ($M = 0.54$, SD = 0.24); mothers in the analysis sample met about half of the indicators on average (see Fig. 2). The subscales show that mothers' fidelity scores related to their initial exposure ($M = 0.6$) to the program were slightly higher than their scores related to overall exposure ($M = 0.5$).

3.3. The relation between fidelity and single indicators of service utilization

Table 4 displays correlations between program-level fidelity scores, individual-level fidelity scores, and single indicators of participants' program utilization. First, the individual-level fidelity score and single utilization indicators were consistently correlated with one another, which is not surprising, given that the single utilization variables were elements of the composite individual-level fidelity score. Interestingly, none of these correlations were perfect, and not all were even high, suggesting that individual-fidelity scores offer a more comprehensive depiction of mothers' experiences than any singular utilization variable can on its own. Further, program-level fidelity was positively correlated with

Table 3
Descriptive Information on HFM Statewide-level Fidelity Scores (*n* = 26 programs).

Program Indicator (Benchmark)	FY08	FY09	FY10	FY11
Indicator 3. Program receives referrals for parents during their prenatal period (60% of referrals)	55.4%	57.8%	58.6%	52.9%
Indicator 4. Program makes first contact with new participants either prenatally or within two weeks of birth (80% of clients)	78.8%	78.3%	68.6%	61.4%
Indicator 5. Program makes first contact with new participants, on average, within 10 days from the referral (100% of clients)	55.9%	67.9%	71.1%	75.7%
Indicator 6. Program completes a first home visit with new participants, on average, within 20 days from referral (100% of clients)	25.5%	45.8%	54.6%	58.0%
Indicator 7. Eligible parents referred to the program accept services (90% of clients)	89.4%	85.4%	88.1%	88.0%
Indicator 9. Program provides intensive home visiting services to participants with a minimum average of 18 visits per participant (12 visits in FY08 & FY09; 100% of clients)	53.7%	61.7%	40.6%	33.4%
Indicator 10. Participants receive 75% of their visits according to their service level (75% of clients) [**]	44.7%	68.1%	74.2%	67.0%
Indicator 11. Program provides home visits to participants with each participant receiving at least one home visit (100% of clients)	89.0%	88.0%	87.7%	86.2%
Indicator 12. Participants receive long term services with a minimum average of 18 months of service at discharge (100% of clients) [**]	26.9%	26.6%	31.5%	29.9%
Indicator 18. Program supervisors provide home visitors with weekly supervision lasting 1.5 hours (85% of home visitors) [**]	83.2%	85.1%	87.0%	89.0%

Note: For each fiscal year we report the mean and range of data of the fidelity scores. The sample includes the clientele-at-large at each HFM site during the respective fiscal year. Site report missingness ranged from 0% to 15%, with the majority of indicators missing less than 4%.
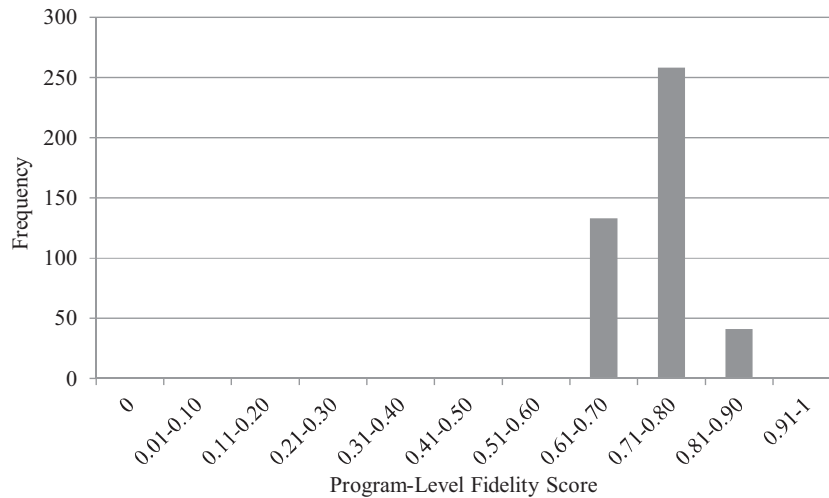[**] Denotes "sentinel" indicators, per the Children's Trust.

Fig. 1. Distribution of Program-level Fidelity Scores.

individual-level fidelity, but not nearly as highly as one might have expected. Relatedly, most single indicators of mothers' utilization were *not* related to program-level fidelity, with a few exceptions: Program-level fidelity was positively associated with the number of home visits that mothers received, and with two indicators pertaining to the IFSPs (number of goal-setting sessions and number of goals met).

3.4. The relation between fidelity and maternal characteristics

Next, we tested whether program- and individual-level fidelity scores were associated with maternal characteristics. The aim of these analyses was to examine whether certain types of participants (e.g., those who had more family support) were more or less likely to use HFM in particular ways. (Note: These analyses were conducted as bivariate correlations, and are not intended to generate causal estimates.)

The majority of maternal characteristics at enrollment did not reveal statistically significant associations with fidelity scores. Associations that were detected (see Table 5) include the following: mothers with higher *individual*-level fidelity scores were more likely to be non-Hispanic Black than non-Hispanic White, to enroll when pregnant (than when parenting), to be above the clinical cut-off for depressive symptoms, to be enrolled in school, and to live with an older relative or guardian. They also were less likely to live with the father of the baby. There were very few associations between maternal characteristics and *program*-level fidelity, and those that did emerge actually conflicted with the individual-level findings. Mothers enrolled in higher fidelity

programs were more likely to be non-Hispanic White than Hispanic and more likely to be living with the father of the baby.

4. Discussion

This paper set out to identify a novel approach to measuring fidelity of implementation by distinguishing between individual- and program-level fidelity, underscoring the unique contribution of each, as well as the intersections and overlaps between the two.

4.1. Discussion of findings

Results show that, when averaged across indicators, programs, and fiscal years, program-level fidelity scores were quite high, with a surprisingly narrow range. Considering that HFM is being implemented by multiple types of agencies across a state with considerable geographic and demographic diversity, the fact that such a high, invariant degree of fidelity has been achieved across programs is noteworthy, and unusual in a statewide initiative (see, for example, Duggan et al., 2005). Researchers have identified several areas that may influence adherence, including well-defined program frameworks, implementation policies, monitoring and accountability systems, built-in feedback loops, and ongoing technical assistance to programs (Boller et al., 2014; Carroll et al., 2007; Durlak & DuPre, 2008). Indeed, the Children's Trust has built a home visiting network that shows strength in every one of the aforementioned areas, and the high program fidelity scores seem to bear out the importance, and effectiveness, of these ongoing quality assurance efforts.

At the same time, this is a voluntary program, meant to be responsive and adaptive to participants' needs (Boller et al., 2014; Kessler et al., 2008). As such, there is a great deal of flexibility built into the model; the expectation is that the home visitor will work with each participant to establish individual goals, settle on a service delivery plan, and adjust home visit content and schedule in both anticipation of, and reaction to, the participant's needs. It is perhaps not surprising, then, that results of the individual-level fidelity analyses using essentially the same set of indicators reveal a markedly different story of engagement and adherence. Individual-level fidelity showed much more variability, with mothers meeting, on average, only around half of the performance indicators. And when it came to the HFM sentinel benchmarks – what we term in the report as their *overall* exposure (i.e., duration, number of home visits, etc.) – fidelity was even lower. Taken together, these findings suggest that even a program operating at
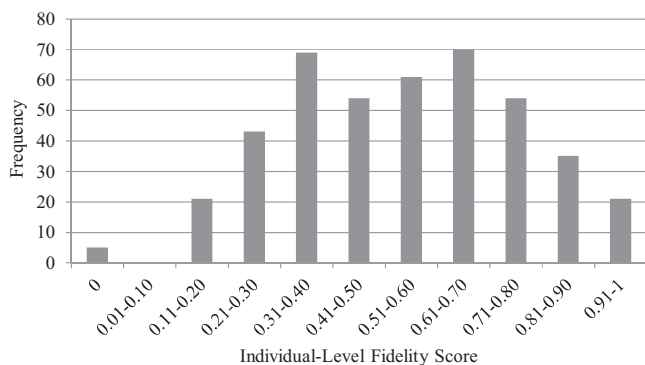

Fig. 2. Distribution of Individual-level Fidelity Scores.

Table 4
Correlations among Program Utilization and Fidelity Indicators.

	Individual fidelity	Program fidelity	Duration (in days)	Home visits	Groups	Secondary activities	IFSP goal sessions	IFSP goals set	IFSP goals met
Individual Fidelity	1.00								
Program Fidelity	.15	1.00							
Duration (in Days)	.68		1.00						
Home Visits	.73	.15	.93	1.00					
Groups	.33		.47	.49	1.00				
Secondary Activities	.49		.69	.65	.39	1.00			
IFSP Goal Sessions	.61	.17	.92	.88	.40	.56	1.00		
IFSP Goals Set	.54		.82	.81	.40	.52	.86	1.00	
IFSP Goals Met	.37	.18	.62	.61	.22	.18	.68	.72	1.00

Note: Only significant correlations are reported ($p < .05$). Correlations equal to 0 indicate no association; correlations closer to 1 indicate stronger, positive associations (i.e., as one measure increases, so does the other); correlations closer to -1 indicate stronger, negative associations (i.e., as one measure increases, the other decreases). IFSP = Individual Family Service Plans.

Table 5
Statistically Significant Associations between Fidelity and Maternal Characteristics.

	Individual-level Fidelity	Program-level Fidelity
Race/Ethnicity[(W, B, H, O)]		
White Non-Hispanic	0.50[B]	0.74[H]
Black Non-Hispanic	0.58[W]	0.74
Hispanic	0.55	0.73[W]
Other Non-Hispanic	0.59	0.75
Timing of Enrollment		
Pregnant	0.63	
Parenting	0.39	
Maternal Depression (continuous score)		
Low	0.52	
Average	0.54	
High	0.57	
Received Food Stamps After Enrollment		
No		0.74
Yes		0.73
Mother Lives with Adult Relative/Guardian		
No	0.49	
Yes	0.57	
Mother Cohabitates with FOB		
No	0.57	
Yes	0.49	

fairly consistently high standards will not always be able to engage its target population as intended by the model. Or stated another way, each participant's utilization and/or receipt of services may or may not be reflective of the overall service delivery environment of her program. As noted by the evaluators of the Evidence-Based Home Visiting to Prevent Child Maltreatment (EBHV) initiative, "the pattern underscores the difficulty in establishing firm expectations for service dosage within the context of a *voluntary program*" (Boller et al., 2014, p.4). That the HFM participants are teenagers probably compromises utilization even further.

4.2. Who are the low-fidelity users, and how did they fare?

The associations between program- and individual-level fidelity scores and maternal characteristics were weak, and told a conflicting tale. Mothers who received the program less in accordance to the model seemed to exhibit both vulnerability (e.g., not in school) and strengths (e.g., less depressed) at enrollment. Indeed, as described elsewhere (Tufts Interdisciplinary Evaluation Research, 2015), a similar pattern (or rather, lack thereof) emerged in our analyses of program outcomes related to fidelity, with mothers who used the program more faithfully less likely to have a

repeat birth (a positive outcome), but also less likely to be employed, and more likely to be in a relationship characterized by domestic violence (negative outcomes). This lack of a clear dose-response relationship has been reported in other studies (e.g., Holland, Xia, Kitzman, Dozier, & Olds, 2014). This is likely because there are mothers at both ends of the utilization spectrum who begin and end their duration in HFM exhibiting both strengths and vulnerabilities. There are low-users who are faring quite poorly at enrollment, and perhaps would have done much better had they been able or inclined to stay on course with the program. But there also are likely low-users who were doing quite well to begin with, achieved their program goals early on, and used as much of the home visiting service as they felt they needed. Conversely, it is probable that the high users comprise both high-functioning mothers who are able to stay on course with the program and subsequently achieve more favorable outcomes, and low-functioning mothers whom home visitors work harder to engage and keep enrolled. In this case, worse outcomes might be observed among some women with more home visits, even if they demonstrate relative improvements over time.

If the program theory is on target, then tight program fidelity should produce expected positive results, or at least it is one of the

core ingredients of success. And indeed, within the RCT design, several critical outcomes were achieved by HFM[4] in the absence of a consistently high level of individual fidelity. In other words, for these goal areas at any rate, being in a high-performing program environment may have been more relevant to participant outcomes than how the participants, as individuals, utilized the services. Equally noteworthy, however, is that other goals attempted within this particular high-program fidelity condition were not met; either the program theory does not hold for these goals, or implementation fidelity alone was insufficient to carry the day.

4.3. Lessons learned

Here we describe some of the conceptual and logistical challenges that characterized our study, followed by implications for home visiting programs to consider.

4.3.1. Conceptual challenges related to fidelity measurement

The variables presented here only comprise a slice of fidelity indicators, primarily focused on adherence to service level and family engagement. While we included every performance measure considered "sentinel" by the program administrators, there were several performance indicators (e.g., child assessments, goal setting, groups) that could not be included in the measures, either because the data were not available, or the data format or quality did not lend itself to the type of analyses we were conducting. In addition, there are numerous, less tangible critical program elements (e.g., cultural competence, quality of supervision, staff ability to build trusting relationships) that are not assessed through the HFM performance indicators, and therefore also were excluded from our fidelity scales. Future scale development would benefit from including a broader representation of performance elements.

Another related limitation pertains to our decision to collapse program performance across fiscal years, and to assign fidelity scores according to the program in which the participant had spent the most amount of time, rather than calculating scores based on the proportion of time the participant had spent in each program. While this approach was necessary for creating a single measure to assess participants overall experience, and for using the scales in multilevel models, it is also likely that some critical nuances in program performance and change over time were obscured by so doing.

Finally there are myriad potential threats to impelementation fidelity, particularly when it comes to service dosage. One can imagine elements of model fidelity along a continuum of sorts, ranging from those elements that, in theory, are under the control of the program administrators (e.g., hiring, training, and supervision), to those elements for which adherence to model standards in many ways depends on the joint behaviors of home visitors and participants (e.g., retention and scheduled visit completion). Fidelity at this dosage-oriented end of the spectrum, then, is highly susceptible to threats from both constituents. Paulsell et al. (2014) have described these as *service delievery threats* (e.g., home visitor turnover, caseload too large), and *service receipt threats* (e.g., participants' schedules, transience, loss of interest in program, crises). Indeed the literature posits that family engagement can be affected by home visitor and supervisor characteristics, staff turnover, and the quality of relationships (Barak, Spielberger, &

Gitlow, 2014; Boller et al., 2014; Daro et al., 2014; McCurdy & Daro, 2001; McFarlane et al., 2010). Examinations of both types of fidelity threats were beyond the scope of this study, but certainly would be a necessary next step in attempting to understand the distinctions between program- and individual-level fidelity.

4.3.2. Implications for programs

With the necessary caveats regarding which critical elements were included in the fidelity measures, HFM has demonstrated that it is possible to achieve high program fidelity across a significant number of sites, located in urban, suburban, and rural settings. This, in itself, is a contribution to the field, and as mentioned above, highlights the robustness of the ongoing quality assurance efforts that are a hallmark of the Children's Trust's administration of this statewide system. It also highlights, in many ways, the importance of the collaboration between the Children's Trust and our evaluation team. Along the lines suggested by the FTA, over our 18-year relationship, the evaluation team consistently has fed back observations of program operations to the program, so that modifications in program processes, and revisions of program goals and objectives, could be implemented. The Children's Trust is a learning organization, committed to honest review of its practices and consistently responsive to evaluation findings, a posture that is arguably critical for the program's continued growth and success (Powell et al., 2015; Supplee & Metz, 2015). This partnership also has yielded what Chambers and Azrin (2013) term a "multipurpose data infrastructure"; over the past decade our evaluation team has worked with the Children's Trust to develop a management information that is equally relevant to researchers and practitioners (Jacobs & Goldberg, 2009). Program providers have full access to the same data points the evaluation team and the administrators use to assess program performance, and in fact are encouraged by the Children's Trust to run their own reports, facilitating a real-time feedback loop. This systemic approach has allowed programs to hone their practices with regard to implementation fidelity.

On the other hand, results from this study underscore the complexities of adopting measures of model fidelity as a primary expression of program quality. Indeed, it would be important to learn how the home visitors, as the frontline workers delivering the services (Lipsky, 1980), make their own decisions around how much to follow the participant's lead, how to balance flexibility and fidelity (Seay et al., 2015). Researchers have conducted a series of studies suggesting that allowing participants more decision-making power over when and how services will be delivered to them can result in more intensive, longer-term participation (Ingoldsby, 2010; Ingoldsby et al., 2013; Olds et al., 2015). And in their study of the intersections between home visitor/client relationships and fidelity to the model, Barak et al. (2014) note that home visitors who allowed clients to prioritize texting/phone calling may have been able to foster more trusting, satisfying, relationships with the clients, albeit ones that were characterized by lower home visit completion rates. These studies, as well as ours, suggest that reliance on performance standards should be balanced by a continued revisiting and interrogation of their usefulness, relevance, and feasibility within local contexts,

In conclusion, it is useful to interpret findings from this study outside the constraining assumption that a faithfully implemented program model would, or should, work the same way for all participants. Indeed, the countless hours home visitors spend, ultimately unsuccessfully, with these reluctant clients may represent hours of service that more willing or eager families do not receive. More research is needed to establish whether this persistence—this imbalance between time spent by home visitor and program uptake by participant—ultimately pays off. More generally, until we better understand which components of the home visiting approach are most essential for which populations

---

[4] HFM was successful in achieving results in several goal areas of critical importance for young, first-time mothers: learning to control stress, curbing externalizing and risky behaviors, and increasing educational attainment (Jacobs et al., 2015).

under which circumstances (Morrison et al., 2009), it is hard to draw conclusions about how to appropriately weigh the many, seemingly critical aspects of program fidelity.

References

Adirim, T., & Supplee, L. (2013). Overview of the federal home visiting program. *Pediatrics*, 132(Suppl. 2), S59–S64.

Avellar, S. A., Paulsell, D., Sama-Miller, E., Del Grosso, P., Akers, L., & Kleinman, R. (2014). Home visiting evidence of effectiveness review: Executive summary. *Washington, DC : Office of Planning, Research and Evaluation. Administration for Children and Families*. US Department of Health and Human Services.

Avellar, S. A., & Supplee, L. H. (2013). Effectiveness of home visiting in improving child health and reducing child maltreatment. *Pediatrics*, 132(Suppl. 2), S90–S99.

Azzi-Lessing, L. (2011). Home visitation programs: critical issues and future directions. *Early Childhood Research Quarterly*, 26(4), 387–398.

Barak, A., Spielberger, J., & Gitlow, E. (2014). The challenge of relationships and fidelity: home visitors' perspectives. *Children and Youth Services Review*, 42, 50–58.

Berkel, C., Mauricio, A. M., Schoenfelder, E., & Sandler, I. N. (2011). Putting the pieces together: an integrated model of program implementation. *Prevention Science*, 12(1), 23–33.

Bilukha, O., Hahn, R. A., Crosby, A., Fullilove, M. T., Liberman, A., Moscicki, E., & Schofield, A. (2005). The effectiveness of early childhood home visitation in preventing violence: a systematic review. *American Journal of Preventive Medicine*, 28(2), 11–39.

Birckmayer, J. D., & Weiss, C. H. (2000). Theory-based evaluation in practice: what do we learn? *Evaluation Review*, 24(4), 407–431.

Bobbitt, K., Osborne, C., & Bradbury, K. (2015). Reality check: can the effects of home visiting programs found in RCTs be replicated in the real world? *Child and Family Research Partnership working paper W15006*. Retrieved from http://childandfamilyresearch.org/publications/w15006/.

Boller, K., Daro, D., Del Grosso, P., Cole, R., Paulsell, D., Hart, B., & Hargreaves, M. (2014). Making replication work: building infrastructure to implement, scale-up, and sustain evidence-based early childhood home visiting programs with fidelity. *Children's Bureau, Administration for Children and Families*. Washington, DC: US Department of Health and Human Services.

Bradach, J. L. (2010). *Foreword: from scaling organizations to scaling impact scaling social impact: new thinking*. New York: Palgrave Macmillan (pp. pp. iv–xiv).

Bradach, J. L., & Grindle, A. (2014). Transformative scale: the future of growing what works. *Stanford Social Innovation Review*.

Carroll, C., Patterson, M., Wood, S., Booth, A., Rick, J., & Balain, S. (2007). A conceptual framework for implementation fidelity. *Implementation Science*, 2(1), 40.

Chaffin, M., & Friedrich, B. (2004). Evidence-based treatments in child abuse and neglect. *Children and Youth Services Review*, 26(11), 1097–1113.

Chambers, D. A., & Azrin, S. T. (2013). Research and services partnerships: partnership: a fundamental component of dissemination and implementation research. *Psychiatric Services*, 64(6), 509–511.

Damschroder, L. J., Aron, D. C., Keith, R. E., Kirsh, S. R., Alexander, J. A., & Lowery, J. C. (2009). Fostering implementation of health services research findings into practice: a consolidated framework for advancing implementation science. *Implementation Science*, 4(1), 50.

Damschroder, L. J., & Hagedorn, H. J. (2011). A guiding framework and approach for implementation research in substance use disorders treatment. *Psychology of Addictive Behaviors*, 25(2), 194.

Daro, D. (2006). *Home visitation: assessing progress, managing expectations*. Chicago, IL: Chapin Hall Center for Children.

Daro, D., Boller, K., & Hart, B. (2014). Implementation in early childhood home visiting: Successes meeting staffing standards, challenges hitting dosage and duration targets. *Mathematica Policy Research*.

Dees, J. G., Anderson, B. B., & Wei-Skillern, J. (2004). Scaling social impact. *Stanford Social Innovation Review*, 1(4), 24–33.

Derzon, J. H., Sale, E., Springer, J. F., & Brounstein, P. (2005). Estimating intervention effectiveness: synthetic projection of field evaluation results. *Journal of Primary Prevention*, 26(4), 321–343.

Drotar, D., Robinson, J., Jeavons, L., & Lester Kirchner, H. (2009). A randomized, controlled evaluation of early intervention: the Born to Learn curriculum. *Child Care Health Development*, 35(5), 643–649.

DuBois, D. L., Holloway, B. E., Valentine, J. C., & Cooper, H. (2002). Effectiveness of mentoring programs for youth: a meta-analytic review. *American Journal of Community Psychology*, 30(2), 157–197.

Duggan, A., Windham, A., McFarlane, E., Fuddy, L., Rohde, C., Buchbinder, S., & Sia, C. (2000). Hawaii's Healthy Start program of home visiting for at-risk families: evaluation of family identification, family engagement, and service delivery. *Pediatrics*, 105(Suppl. 2), 250–259.

Duggan, A., McFarlane, E., Fuddy, L., Burrell, L., Higman, S. M., Windham, A., & Sia, C. (2004). Randomized trial of a statewide home visiting program: impact in preventing child abuse and neglect. *Child Abuse & Neglect*, 28(6), 597–622.

Duggan, A., Caldera, D., Rodriguez, K., Burrell, L., Rohde, C., & Crowne, S. S. (2007). Impact of a statewide home visiting program to prevent child abuse. *Child Abuse & Neglect*, 31(8), 801–827.

Duggan, A., Rodriguez, K., Burell, L., Shea, S., & Rohde, C. ([103_TD$DIFF]2005). *Evaluation of Healthy Families Alaska program: Final report: January 21, 2005*. Anchorage: Alaska Department of Health and Social Services.

Duggan, A., Minkovitz, C. S., Chaffin, M., Korfmacher, J., Brooks-Gunn, J., Crowne, S., & Harwood, R. (2013). Creating a national home visiting research network. *Pediatrics*, 132(Suppl. 2), S82–S89.

Durlak, J. A., & DuPre, E. P. (2008). Implementation matters: a review of research on the influence of implementation on program outcomes and the factors affecting implementation. *American Journal of Community Psychology*, 4, 327–350.

Filene, J. H., Kaminski, J. W., Valle, L. A., & Cachat, P. (2013). Components associated with home visiting program outcomes: a meta-analysis. *Pediatrics*, 132(Suppl. 2), S100–S109.

Holland, M. L., Xia, Y., Kitzman, H. J., Dozier, A. M., & Olds, D. L. (2014). Patterns of visit attendance in the Nurse–Family Partnership program. *American Journal of Public Health*, 104(10), e58–e65.

Holton, J. K., & Harding, K. (2007). Healthy Families America®: ruminations on implementing a home visitation program to prevent child maltreatment. *Journal of Prevention & Intervention in the Community*, 34(1-2), 13–38.

Howard, K. S., & Brooks-Gunn, J. (2009). The role of home-visiting programs in preventing child abuse and neglect. *The Future of Children*, 19(2), 119–146.

Ingoldsby, E. M. (2010). Review of interventions to improve family engagement and retention in parent and child mental health programs. *Journal of Child and Family Studies*, 19(5), 629–645.

Ingoldsby, E. M., Baca, P., McClatchey, M. W., Luckey, D. W., Ramsey, M. O., Loch, J. M., & Smith, B. (2013). Quasi-experimental pilot study of intervention to increase participant retention and completed home visits in the nurse–family partnership. *Prevention Science*, 14(6), 525–534.

Jacobs, F. (1988). The Five-tiered Approach to evaluation: context and implementation. In H. Weiss, & F. Jacobs (Eds.), *Evaluating family programs* (pp. 37–68).Hawthorne, NY: Aldine de Gruyter.

Jacobs, F. (2003). Child and family program evaluation: learning to enjoy complexity. *Applied Developmental Science*, 7(2), 62–75.

Jacobs, F., Easterbrooks, M. A., Goldberg, J., Mistry, J., Bumgarner, E., Raskin, M., & Fauth, R. (2015). Improving adolescent parenting: results from a randomized controlled trial of a home visiting program for young families. *American Journal of Public Health*. http://dx.doi.org/10.2105/AJPH.2015.302919 [e-View Ahead of Print].

Jacobs, F., & Goldberg, J. (2009). Evaluating contemporary social programs: challenges and opportunities. In M. E. Kenny, A. M. Horne, P. Orpines, & L. R. E. Reese (Eds.), *Realizing social justice: the challenge of preventive interventions* (pp. 97–122).Washington, DC: American Psychological Association.

Kessler, S. R., Nixon, A., & Nelson, C. (2008). Don't throw out the baby with the bath water a novel way of evaluating outcomes in the Healthy Families America programs. *American Journal of Evaluation*, 29(3), 288–300.

Lipsky, M. (1980). *Street-level bureaucrats: The dilemmas of the individual in public services*. New York, NY: Russell Sage Foundation.

Loney, T., & Nagelkerke, N. J. (2014). The individualistic fallacy, ecological studies and instrumental variables: a causal interpretation. *Emerging Themes in Epidemiology*, 11(1), 1–6.

Matone, M., Curtis, C., Chesnokova, A., Yun, K., Kreider, A., Curtis, M., & Rubin, D. (2013). Evaluation of maternal and child home visitation programs: lessons from Pennsylvania. *PolicyLab Center to Bridge Research, Practice, and Policy*. Philadelphia, PA: The Children's Hospital of Philadelphia Research Institute.

McCurdy, K., & Daro, D. (2001). Parent involvement in family support programs: an integrated theory. *Family Relations*, 50(2), 113–121.

McFarlane, E., Burrell, L., Fuddy, L., Tandon, D., Derauf, D. C., Leaf, P., & Duggan, A. (2010). Association of home visitors' and mothers' attachment style with family engagement. *Journal of Community Psychology*, 38(5), 541–556.

Mihalic, S. F., Fagan, A. A., & Argamaso, S. (2008). Implementing the lifeSkills Training drug prevention program: factors related to implementation fidelity. *Implementation Science*, 3(5), 1–16.

Morrison, D. M., Hoppe, M. J., Gillmore, M. R., Kluver, C., Higa, D., & Wells, E. A. (2009). Replicating an intervention: the tension between fidelity and adaptation. *AIDS Education and Prevention*, 21(2), 128.

Nurse-Family Partnership Snapshot. (2015, October). Retrieved November, 2015, from http://www.nursefamilypartnership.org/.

Olds, D. L., Baca, P., McClatchey, M., Ingoldsby, E. M., Luckey, D. W., Knudtson, M. D., & Ramsey, M. (2015). Cluster randomized controlled trial of intervention to increase participant retention and completed home visits in the Nurse-Family Partnership. *Prevention Science*, 16(6), 778–788.

Olds, D. L., Eckenrode, J., Henderson, C. R., Kitzman, H., Powers, J., Cole, R., & Luckey, D. (1997). Long-term effects of home visitation on maternal life course and child abuse and neglect: fifteen-year follow-up of a randomized trial. *JAMA*, 278(8), 637–643.

Paulsell, D., Del Grosso, P., & Supplee, L. (2014). Supporting replication and scale-up of evidence-based home visiting programs: assessing the implementation knowledge base. *American Journal of Public Health*, 104(9), 1624–1632.

Powell, B. J., Bosk, E. A., Wilen, J. S., Danko, C. M., Van Scoyoc, A., & Banman, A. (2015). Evidence-based programs in "real world" settings: Finding the best fit. In D. Daro, A. C. Donnelly, L. A. Huang, & B. J. Powell (Eds.), *Advances in child abuse prevention knowledge: The perspective of new leadership*. Switzerland: Springer (pp. 145–177).

Rubin, D. M., Curtis, M. L., & Matone, M. (2014). Child abuse prevention and child home visitation: making sure we get it right. *JAMA Pediatrics*, 168(1), 5–6.

Saïas, T., Lerner, E., Greacen, T., Simon-Vernier, E., Emer, A., Pintaux, E., & Falissard, B. (2012). Evaluating fidelity in home-visiting programs: a qualitative analysis of 1058 home visit case notes from 105 families. *PloS One, 7*(5), e36915.

Schwartz, S. (1994). The fallacy of the ecological fallacy: the potential misuse of a concept and the consequences. *American Journal of Public Health, 84*(5), 819–824.

Seay, K. D., Byers, K., Feely, M., Lanier, P., Maguire-Jack, K., & McGill, T. (2015). Scaling up: Replicating promising interventions with fidelity. In D. Daro, A. C. Donnelly, L. A. Huang, & B. J. Powell (Eds.), *Advances in child abuse prevention knowledge: The perspective of new leadership.* Switzerland: Springer (pp. 179–201).

Supplee, L. H., & Metz, A. (2015). Opportunities and challenges in evidence-based social policy. *Social Policy Report, 28*(4).

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

Tufts Interdisciplinary Evaluation Research (2015). The Massachusetts Healthy Families Evaluation – Phase 2 (MHFE-2): a randomized controlled trial of a statewide home visiting program for young parents. *Final report to the Children's Trust of Massachusetts.* Medford, MA: Tufts University.

Weiss, C. H. (1995). Nothing as practical as good theory: exploring theory-based evaluation for comprehensive community initiatives for children and families. In J. P. Connell, A. Kubisch, L. Schorr, & C. Weiss (Eds.), *New approaches to evaluating community initiatives: concepts, methods, and contexts. Roundtable on comprehensive community initiatives for children and families* (pp. 65–92). Washington, DC: Aspen.

Weiss, C. H. (1997). How can theory-based evaluation make greater headway? *Evaluation Review, 21*(4), 501–524.

Yoshikawa, H., Rosman, E. A., & Hsueh, J. (2002). Resolving paradoxical criteria for the expansion and replication of early childhood care and education programs. *Early Childhood Research Quarterly, 17*(1), 3–27.