Maternal and Infant Home Visiting Program Evaluation (MIHOPE)

0970 - 0402

Supporting Statement

Part B: Statistical Methods

Updated July 2012

Submitted By:
Office of Planning, Research and Evaluation
Administration for Children and Families
U.S. Department of Health and Human Services

7th Floor, West Aerospace Building
370 L'Enfant Promenade, SW
Washington, D.C. 20447

Project Officer:
Lauren Supplee

B1. Sampling

Exhibit B.1 summarizes the sample sizes for baseline data collection. Approximately twelve states including approximately 85 sites will be included in the evaluation. The average site will include 60 women (30 assigned to the home visiting program and 30 assigned to the control group) and approximately 6 home visitors, for a total of approximately 5,100 women and 510 home visitors. Each site is expected to have one or two home visiting supervisors and one program manager.

States and their local program sites will be selected for MIHOPE in 2012 based on a variety of characteristics including the type of home visiting model, geography, urbanicity, target population, and research feasibility. As described in Part A, the study team will be collecting information from states early in 2012.

From that information, a list of potential local programs will be compiled. Eligible local programs will meet several criteria: (1) having two or more years experience with one of the four evidence-based home visiting service models that were selected by at least 10 states receiving MIECHV funds, (2) excess demand for their services so that they can provide enough families for a control group, (3) the ability to enroll 30 families in their program over a period of about a year, and (4) locations where there are few other home visiting services in order to ensure a strong service differential between the program and control groups.

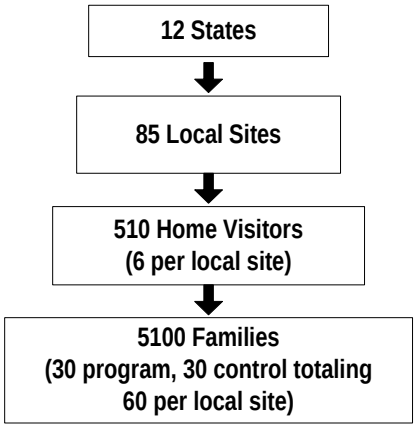
States will be classified in terms of which of four clusters of ACF/HRSA regions the state is in, the number of local sites that appear to be eligible for the evaluation, the urbanicity of the potential program sites, and the national service model of the potential program sites. Once this information is compiled, the study team will choose states so they meet the following criteria: each of four clusters of regions will be represented, the four evidence-based models are represented roughly evenly across the sites, and sites are as representative as possible of the urbanicity of all potential sites. Once 12 states are chosen, 85 sites will be chosen from within those states to meet the same criteria (for example, having the four evidence-based models represented roughly evenly across sites).

Within a site, the evaluation will enroll women who are pregnant or have a child under six months old. Home visiting programs will identify families who appear to be eligible for the study and a field staff person from the research team will go to the family's home to explain the study and obtain informed consent. Families will continue to be recruited until 60 families have been recruited in a site.

Within each state, the evaluation will conduct semi-structured group interviews with program managers and supervisors at one year following the state's recruitment into the study. The group interviews will include the program manager and all supervisors from each site participating in the evaluation. Each site will have only one program manager and most sites will have only one supervisor. Thus, the evaluation needs to include all program managers and supervisors in order to have representation of all sites in eliciting information to explain each site's quantitative results.

Within each state, the evaluation will conduct semi-structured group interviews with one third of the home visitors in each participating site and semi-structured individual interviews with another third of the home visitors in each participating site. The interviews will be carried out at one year following the state's recruitment into the study. Thus, about 2 home visitors from each site will participate in the group interview and another 2 will participate in individual interviews. This sampling plan will allow for two group interviews with home visitors in each state, with about 7 home visitors in each group interview. This sampling plan will also allow for individual interviews within two different home visitors in each site, to permit examination of personal psychosocial attributes as factors for service delivery while holding site characteristics constant. The responses for the individual and group semi-structured interviews may be combined for some qualitative descriptions of how staff describe their roles and the program, with documentation that both individual and group interviews contributed to the conclusions drawn .

Exhibit B.1
Structure of National Home Visiting Evaluation



Statistical power. Exhibit B.2 shows the "minimum detectable effect" of this sampling plan for the full sample and for differences in impacts across subgroups. A minimum detectable effect is the smallest true effect that is likely to generate statistically significant estimated effects. For purposes of the design, calculations were performed to find the smallest effects that would generate statistically significant findings in 80 percent of studies with a similar design, using two-tailed t-tests with a 10 percent significance level. All results are presented as effect sizes, that is, in terms of the number of standard deviations of the outcome being examined. Results are presented both for administrative data, which would be available for all families, and for data such as surveys, which are assumed to be available for 80 percent of families.

Pooled sample. The minimum detectable effect for the pooled sample would be 0.058 standard deviations for administrative records and 0.065 for survey-based or observational outcomes. For example, if a site had a rate of child abuse and neglect of 20 percent in the control group, this design would have an 80 percent chance of finding a statistically significant impact if the true impact is a reduction of 2.3 percentage points (from 20.0 percent of the control group to 17.7 percent of the program group).

These pooled minimum detectable effects provide reasonable statistical power for the evaluation. For example, HomVEE found average effects of this magnitude or larger for four of the domains

of interest: child health, child development and school readiness, maternal health, and referrals and coordination.

Although families within a site may be more similar to one another than to families in other sites, this would not affect the statistical power of the pooled estimates or the subgroup estimates presented below. That is because individuals will be randomly assigned to the program or control group within a site.

Exhibit B.2
Minimum Detectable Effects of Proposed Sampling Plan

	Administrative data	Survey or observational data
Full sample	0.058	0.065
Subgroups (% in larger subgroup)		
50	0.117	0.130
60	0.119	0.133
70	0.127	0.142
80	0.146	0.163

Results are the smallest true impact that would generate statistically significant impact estimates in 80 percent of studies with a similar design using two-tailed t-tests with a 10 percent significance level. No adjustment for multiple comparisons is assumed. Results are based on fixed effects estimates. Administrative data are assumed available for all families, while survey or observational data would be available for 80 percent of families. Baseline data are assumed to explain 30 percent of variation in outcomes across families.

Subgroup differences. In addition to looking at the average effect across sites, the evaluation would assess whether home visiting had larger effects for some subgroups. For purposes of investigating the statistical power of subgroup estimates, it is assumed that the evaluation would be interested in detecting significant differences across subgroups. Since statistical power depends on the number of families in each subgroup, minimum detectable differences are presented for cases where 50, 60, 70, and 80 percent of the sample is in the larger of two subgroups. For a subgroup that divides the sample in half, for example, the minimum detectable differences are 0.117 standard deviations using administrative data and 0.130 using survey data. If 20 percent of control group families had a substantiated case of child abuse and neglect, the study would have an 80 percent chance of finding significantly larger effects for one subgroup than for another if the difference in true effects was 4.7 percentage points (for example, reducing child abuse and neglect by 4.7 percentage points for one subgroup but having no effect for the other subgroup). These minimum detectable differences increase gradually as the proportion of families in one subgroup increases. They are quite similar if 60 percent of families are in one subgroup, but they increase by 25 percent if 80 percent of families are in one subgroup.

Investigating the effect of program features. The evaluation will include 85 sites to allow it to explore the relationship between program features and program impacts. Program features could include any aspects of the community context, implementation system, service models, organizational influences, or home visitor characteristics. For example, this analysis could explore how program impacts vary with the duration of home visits, the background and training of home visitors, the support provided by supervisors for home visitors, the clarity of the goals of the local program, or the intended targets of the national model being used.

A framework for exploring the links between program features and program impacts is described in Greenberg, Meyer, Michalopoulos, and Wiseman (2003). Within this framework, the precision of the estimated relationships between program features and program impacts depends on a number of factors, including (1) the number of sites in the evaluation, (2) the precision of impact estimates within each site (which will increase with the number of families in the site), (3) the variation in characteristics across sites, (4) the number of program features to be investigated, and (5) how related the various program features are to each other. It is easier to detect differences by program feature if there are more sites, if there are more families in each site, if different sites vary more across the program feature being examined, if fewer program features are being examined at any one time, and if the program features are not closely related to one another. As an example of the last point, it may be very difficult to distinguish the effect of planned duration of home visits from the effect of actual duration, since the two are likely to be closely related in a particular site.

Exhibit B.3 shows the minimum detectable effects of program features for several scenarios. The upper half of the table shows results for a program feature that is binary and takes on one value in half of the sites and a different value in half of the sites. For example, half of the sites might plan to visit families weekly while half would visit only every other week. The lower half of the table shows results for a continuous program feature, such as how many weeks home visits would take place. In each panel, results are presented depending on whether 10, 20, or 30 program features would be examined at one time. As noted above, the ability to detect the effects of program features will worsen as more features are examined. Finally, results for each scenario are presented for three assumptions about how highly correlated various program features are with one another. As noted above, the ability to detect the effects of program features worsens as features become more highly correlated with one another.

Consider the first row of Exhibit B.3, which shows the case where 10 program features are being examined simultaneously and there is a low correlation across them. For outcomes measured using administrative data, the model would be able to detect differences of 0.203 standard deviations between sites of one type and sites of another type. If the overall effect on an outcome were 0.15 standard deviations, for example, the study would have an 80 percent chance of finding a statistically significant relationship between the program feature and impacts if the true impact were 0.252 standard deviations in one set of sites and 0.048 standard deviations in the other set of sites.

The ability to detect an effect of a program feature is only slightly worse if the features are more highly correlated or if 20 program features are being examined. The statistical power gets considerably worse, however, if more features are being examined and the correlation across features is high. For example, the minimum detectable difference is 0.317 standard deviation (for an effect of 0.309 standard deviation in one set of sites compared with –0.009 standard deviation in the second set of sites) if 20 program features are being examined and the correlation across them is high, and the minimum detectable difference is 0.348 standard deviation if 30 features are being examined and the correlation across them is medium.

Mother and Infant Home Visiting Program Evaluation

Exhibit B.3
Minimum Detectable Effects of Program Features

Type of variable	No. of variables representing program features	Correlation across program features	Administrative data	Survey or observational data
Binary, half of sites have the feature				
	10	Low	0.203	0.231
		Medium	0.213	0.243
		High	0.226	0.258
	20	Low	0.231	0.263
		Medium	0.264	0.300
		High	0.317	0.361
	30	Low	0.268	0.305
		Medium	0.348	0.397
		High	0.626	0.713
Continuous				
	10	Low	0.101	0.115
		Medium	0.107	0.122
		High	0.113	0.129
	20	Low	0.115	0.131
		Medium	0.132	0.150
		High	0.158	0.180
	30	Low	0.134	0.153
		Medium	0.174	0.198
		High	0.313	0.356

Results are the smallest true impact that would generate statistically significant impact estimates in 80 percent of studies with a similar design using two-tailed t-tests with a 10 percent significance level. No adjustment for multiple comparisons is assumed. Results are based on fixed effects estimates. Administrative data are assumed available for all families, while survey or observational data would be available for 80 percent of families. The correlation across program features is based on the R^2 statistic when one program feature is regressed on all other program features. For purposes of the calculations, a low correlation means the R^2 increases by .01 with every added feature, by .02 with every added program feature for a medium correlation, and by .03 for a high correlation.

The lower half of Exhibit B.3 shows minimum detectable effects if the program feature is continuous and normalized to have a variance of 1.0 standard deviation across sites. Because there can be greater variability in continuous variables than in binary ones, the design would have a greater ability to detect differences for such measures. For example, for a study examining 10 program features that are not highly correlated, the minimum detectable effect size of the program feature would be 0.101 standard deviation using administrative data and 0.115 standard deviation using survey data. Even for the most extreme case shown in the table — 30 highly correlated program features — the design could detect differences in impacts of 0.313 standard deviations using administrative data and 0.356 standard deviations using survey data.

These minimum detectable differences are well within the range found across previous studies of home visiting. For example, the HomVEE review found that prior studies of home visiting have produced impacts on positive parenting practices with a range of 0.82 standard deviations across studies (Michalopoulos et al. 2011). The range in impacts across prior studies is similar for other domains, including child maltreatment (range of 0.75 standard deviations), child health (0.93), child health and school readiness (.48), maternal health (1.14), and referrals or coordination (1.29). Although some of these differences are due to sampling error, a substantial portion of the differences are likely due to differences in program implementation. For example, a review of over 500 studies of prevention and health promotion programs for children and adolescents found that mean effect sizes were at least two to three times higher when programs were carefully implemented and were free of serious implementation problems (Durlak and Dupre 2008).

The Greenberg et al. framework underlying the calculations shown in Exhibit B.3 assumes that impacts are not correlated across sites. This may not be the case in MIHOPE because sites within a state will be funded through the state MIECHV grantee, which may exercise some control over the operation of local programs. The MIHOPE analysis will take this into account by adjusting the standard errors of estimated effects for such clustering.

It is difficult to say how such clustering will affect the statistical power of the analysis that links program features to program impacts. This is true for two reasons. First, there is little information about how similar sites within a state are likely to be in terms of their program implementation or, just as important, their effects on parent and child outcomes. Second, there is no well-established alternative to the Greenberg et al. framework that would provide an analytical derivation of the standard errors of the analysis linking program features to program impacts. For example, a similar analysis of mandatory welfare-to-work programs assumed program impacts were independent across welfare offices even when welfare offices were in the same city or county and run by the same agency (Bloom, Hill, and Riccio 2003).

One means of assessing the possible effect on statistical power of clustering of sites within a state is to assume the analysis of program features would include state “fixed effects” that would, in essence, base the analysis on variation of programs within a state but not use variation in programs across states. With an intraclass (intrastate) correlation in impacts of 0.01, this would increase the minimum detectable differences by about 10 percent, for example, from 0.20 in the first row of Exhibit B.3 to 0.22. With an intraclass correlation of 0.10, the minimum detectable effects would increase by 15-20 percent. Such differences in effects are still well within the range found in HomVEE.

Although typical levels of intraclass correlation would not affect the statistical power much, the study team will try to minimize the similarity of sites within a state by aiming to include states where local programs vary in features such as the evidence-based model that is being used, the urbanicity of the local site, and the type of local implementing agency.

Implementation data. As shown in Table A.2 above, implementation data will be used to answer several different types of research questions, requiring different types of analyses. They will be used to describe the local program models, their implementation systems, local staff, and service

delivered. Such analyses will rely on primarily on descriptive statistics such as means, medians, and ranges across the 85 sites, for which power calculations are not required. These descriptive analyses will be conducted for all sites combined as well as for the four program models separately. Implementation data will also be used in linear regression analyses or multi-level analyses to understand how program inputs (staff characteristics, organizational characteristics, implementation systems, community characteristics, and family characteristics) are associated with program outputs (service delivery). These will generally be conducted for all sites combined, so that we are analyzing approximately 510 home visitor-level estimates for analyses conducted at the level of individual staff and their service delivery behavior, or 85 site-level estimates for site-level analyses, and so on. Finally, the implementation data for each of the 85 sites will be combined with impact estimates to conduct analyses aimed at "getting inside the black box." Power estimates for these analyses are described above.

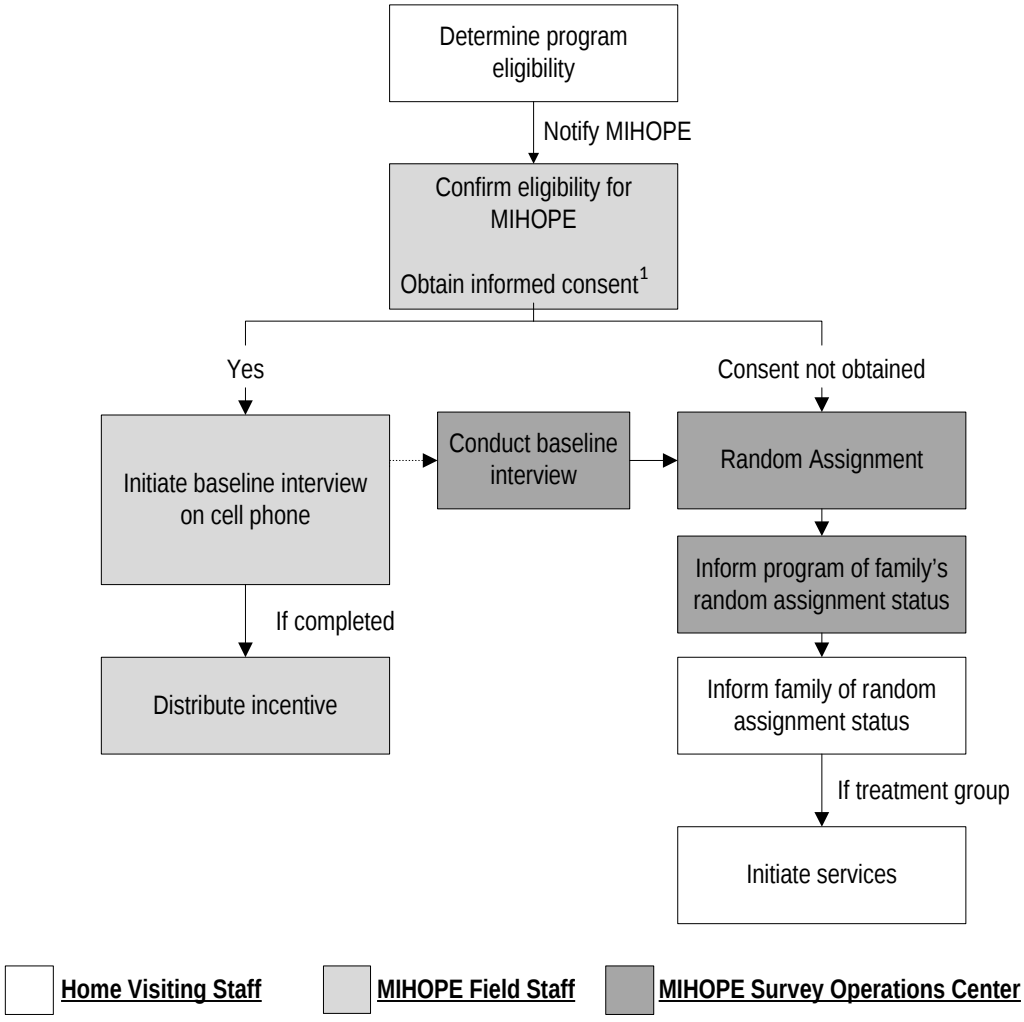
B2. Procedures for collection of information

This section focuses on procedures for quantitative data collection activities: the family baseline survey, the surveys of staff at participating home visiting program sites, and the surveys of administrators of community resources.

Family baseline survey. Exhibit B.4 depicts the process of collecting family baseline data. This process includes determining eligibility for the evaluation, contacting eligible women, and obtaining consent before conducting the family baseline survey. Steps taken to monitor the data quality are also briefly described.

Before recruiting women into the study, following the site's normal procedures, the home visiting program will collect information to determine whether the woman is eligible for the program's services. The program will also make an initial determination that the family is eligible for MIHOPE, for example because they have a child under six months old.

Exhibit B.4
Roles in the MIHOPE Baseline Data Collection Approach



[1]We anticipate that in 15 percent of cases, field staff will be present when eligibility is determined. In the other 85 percent of cases, field staff tasks take place during a separate visit.

For women who appear eligible for MIHOPE, the program will provide survey staff with the woman's name, address, telephone number, and primary language, as well as the child's name and date of birth (if the child has already been born).

During the visit, MIHOPE field staff will conduct the following procedures:

- Staff will distribute attractive introductory materials about MIHOPE (Attachment 5)
- Staff will introduce the study, provide a commitment to confidentiality, explain random assignment, and answer questions.
- Staff will attempt to obtain informed consent for the mother to participate in MIHOPE. Informed consent forms (Attachment 5) will reference the baseline and follow-up data collections and will allow the study team to collect state administrative

data on the family. Based on prior studies such as Baby FACES and studies of home visiting programs in Alaska and Hawaii, 90 percent of families are assumed to provide consent to participate in the study. Thus, the evaluation expects to describe the study and attempt to obtain consent from 5,667 families in order to enroll 5,100 families.

- If an applicant is a minor, it might be necessary to obtain consent from the parent as well, unless the state's emancipated minor laws make this unnecessary. This consent will be obtained by telephone if the parent of the minor is not in the home at the time of the consent process. The study team anticipates that 20 percent of program applicants will be minors. The protocol for obtaining consent from parents of minors is included as Attachment 5.
- If the woman consents to participate in random assignment and MIHOPE, she will be given a copy of the consent form. Field staff will then initiate computer-assisted telephone interviewing (CATI) via cell phone and then give the parent a cell phone to complete the 60-minute interview.
- If the mother has two or more children younger than six months old, survey staff will identify one as the focal child and all child-specific activities will be conducted with or about that focal child.
- While in the home, field staff will complete a selection of observational items about the internal and external physical home environment drawn from the HOME.
- At the completion of the interview, field staff will give the parent a $25 gift card as a token of appreciation and a $15 gift for the child.
- At the end of the CATI program, the woman will be randomly assigned to the program or control group.
- If the woman does not consent to participate in MIHOPE, random assignment will still be completed to determine if program services will be provided. This ensures that participation in the evaluation does not affect the family's ability to receive MIECHV services.
- The result of random assignment will be uploaded to a secure web-based system and an automated generic email will be sent to the home visiting point of contact to check the web-based system for that participant's random assignment status so that appropriate services can be initiated.
- Program staff will inform women whether they were assigned to receive program services. They will initiate home visiting services for women assigned to the program group and provide referrals to other community services for women assigned to the control group.

The proposed approach has a number of advantages.

- Having survey staff visit the family's home will help build rapport and maximize response rates for follow-up data collection.
- Having survey staff in the home allows the study to obtain written consent, which may be needed to obtain administrative records in some states.
- No burden is placed on home visitors to obtain consent or baseline data.

- Using CATI allows for low cost monitoring of data collection to ensure uniformity and data security (for example because data will not have to be transmitted back to the survey operation center).
- CATI can also accommodate complex instruments and different instruments for different families.
- Phone surveys also provide privacy to the parent in answering sensitive questions. For example, mothers will respond to the CATI survey questions with verbal responses or with a numerical value code, whichever they prefer. Neither the field interviewer nor anyone else in the room will know which survey question the mother is answering or what the question is.
- Field staff will collect information for the HOME while waiting for the survey to be completed. The HOME assessment is expected to take roughly 40 minutes to conduct.
- To facilitate answering of longer questions, the field staff will hand the woman a packet of color-coded show cards. The CATI interviewer will prompt the mother when a show card is needed and which show card to use. For example, the CATI interviewer will say, "For this next question, please take out the yellow show card identified with a G6 in the upper left hand corner. It contains the list of response options that will be used for the next several questions."
- While the respondent is on the phone the field staff will be available for any questions the respondent may have, and to troubleshoot any technical difficulties with the completion of the CATI interview, for example, if the interview needs to be broken off, or if the call is dropped.
- Field staff can provide a gift card on the spot rather than making participants wait for a gift card to be mailed. This can increase the willingness to respond to the baseline survey.
- Field staff can monitor the infant for any needs while mother is on the phone doing the baseline interview (if there is an infant and if the infant is awake)
- Completing the baseline survey before random assignment ensures a 100% response rate. As noted above, we assume that 10 percent of women who are eligible for the study will decline to participate, so 90 percent of eligible women will complete the baseline survey.

One critique of CATI is that it is more difficult to establish rapport than when conducting the interview in-person. This is addressed by the presence of field staff in the home to introduce the study and obtain consent. Experienced field staff can build rapport to maximize enrollment and support a high response rate in subsequent follow up data collection activities. When possible, the same field staff will collect data during subsequent data collections.

An alternative to CATI is computer assisted personal interview (CAPI) or computer assisted self interview (CASI). CATI is being used for the family baseline survey because it has several advantages over the other methods:

- It provides greater confidentiality of responses for the woman compared with CAPI since others in the home could overhear a CAPI interview. Many items in the baseline survey are of a sensitive nature (domestic violence, drug and alcohol use, cigarette smoking, depression and other mental health issues, attitudes about parenting and attachment) so providing a confidential method of responding is critical. This is an advantage of CATI over CAPI, but both CATI and CASI interview methods provide

a confidential method for participants to answer sensitive questions without being overheard.

- It provides data security since data are collected and stored in a central secure location. CAPI and CASI data are stored on individual laptops and require field staff to upload the data regularly. If field staff do not upload their data, or their laptops are lost or stolen, data will be lost and security may be breached.
- It allows for real time monitoring of interviewers and higher data quality on the study. All telephone interviewers are recorded and ten percent of each interviewer's work is monitored in real time by supervisors. Feedback is given immediately following a monitoring session.
- The cost of developing and implementing a CATI survey is less than a CAPI or CASI survey. This includes the cost of the equipment (laptops) and the labor to program, upload and maintain the laptops.
- Telephone interviewers can halt or stop an interview if a mother needs a break. If needed, the telephone interviewer can schedule an appointment and call back at a more convenient time to complete the interview. This may be especially important for mothers with infants. This is less expensive than sending a field staff person back to the home to complete a CAPI or CASI survey.
- After completing the baseline interview, the CATI program will randomly assign the mother to the treatment or control group. Information on random assignment will be sent in real time to a secure web-based system that home visiting program staff can access, along with a generic email alert to check it so they can start services for treatment families right away. CAPI or CASI systems require that the data from the laptops be uploaded by the interviewer to a secure server and therefore rely on human transmission of the data. There is often a lag of a day or more with this process which would delay receipt of the random assignment status and sending the information to the home visiting program in a prompt fashion.

The study team has used this method on many large scale studies, such as FACES, Baby FACES, and BSF, sending field staff to participants' homes with cellular telephones to complete a survey via CATI. It is more efficient and cost effective than using CAPI or CASI.

Surveys of Staff at Participating Home Visiting Program Sites. Web-based surveys of the program manager, supervisors, and home visitors in each participating program site will be conducted near the time that the state enters the evaluation (baseline) and 12 months later. Site liaisons will notify each site's point of contact about two weeks prior to the targeted date for each survey to discuss the timeline and review survey procedures. Survey completion will be tracked using the management system. If a survey is not completed within one week of the targeted time-frame, the site liaison will follow up with the point of contact at the site to remind the staff member that the survey response is due. These instruments will be designed to preclude backtracking to change responses or printing of the survey.

Surveys of Administrators of Community Resources. Web-based surveys will be conducted with administrators of two types of community resources: a) services to which participating home visiting programs might make referrals relevant to MIECHV benchmarks and participant outcomes; and 2) home visiting programs not participating in the evaluation but serving the same

community. These surveys will be carried out with administrators of the organizations identified in the Program Manager Survey, Baseline, Part 1. For each community service provider and home visiting program identified, program managers will provide their primary contact's name, email address, telephone number, and street address. Web-based surveys of these administrators will be conducted between Parts 1 and 3 of the Program Manager Survey, Baseline. Administrators will be contacted by email with instructions about how to complete the web-based survey. Survey completion will be tracked using the management system. If a survey is not completed within one week of the targeted time-frame, research staff will send a reminder email with follow up by phone if needed. The survey instruments are designed to be completed in a single session of about 0.10 hours.

Logs Maintained by Supervisors and Home Visitors. Data about service delivery, training and supervision will be collected through weekly web-based logs. For sites in which supervisors and home visitors do not have regular access to the internet, paper versions of the logs will be offered. Home visitors and supervisors can complete the paper forms and a support person in the site can enter these data using the site's computers.

Supervisor Logs. Supervisors will use the web-based system to complete logs each week during the period in which home visitors are also completing logs. Supervisors will be prompted each week to complete a log for each of their home visitors that are participating in the study (about 5-8 home visitors). If the supervisor did not have a supervisory session with a particular home visitor, s/he will record the reason (for example, vacation or sick leave, scheduling conflict). The supervisor should enter all data for a given week no later than the end of the first workday of the following week.

Home Visit Logs. Home visit logs will be the major source of standardized data on actual service delivery to families. Home visitors will complete logs each week for the first 15 months of family enrollment. To ensure that log data are completed every week, home visitors will be asked to record information for every family enrolled in the evaluation and assigned to their caseload. If a family did not receive a visit that week, the home visitor will record the reason (for example, no visit was scheduled or a scheduled visit was cancelled). The home visitor should enter all data for a given week by the end of the first workday of the following week.

Group and Individual Interviews with Staff at Participating Home Visiting Program Sites. Within each state, group interviews of program managers and supervisors from participating program sites will be conducted about 12 months after the state's recruitment into the evaluation. All program managers and supervisors will participate in the group interviews.

Within each state, group and individual interviews of home visitors from participating program sites will also be conducted about 12 months after the state's recruitment into the evaluation. Within each site, one-third of the home visitors (about two home visitors per site) will be randomly selected for participation in the group interviews and another third (about two home visitors per site) will be randomly selected for participation in the individual interviews. The interviews will be carried out as part of the evaluation team's 12-month site visit to the state.

Site liaisons will notify each site's point of contact about one month prior to the site visit to discuss the timeline and to review group and individual interview procedures. Interview content will be audiorecorded and notes will be taken to document content.

B3. Maximizing response rates

This section focuses on strategies to maximize response rates for quantitative data collection activities: the family baseline interview, the surveys of staff at participating home visiting program sites, the surveys of administrators of community resources, and the logs maintained by supervisors and home visitors.

Family baseline survey. Minimizing sample attrition is of paramount concern for any longitudinal study. A number of techniques will be used to achieve high response rates:

- Establishing rapport with women at baseline to ensure consent
- Training field staff in respondent cooperation and refusal-avoidance techniques
- Ensuring privacy of participant information
- Providing adequate information about the study at the time participants are recruited
- Conducting random assignment after the baseline interview
- Designing surveys carefully with pretested questions that are easy to answer
- Providing an incentive for participation and to encourage participation in the follow-up survey
- Using MIHOPE's sample management system to track sample recruitment, response rates, and potential sample attrition

Field staff will be trained in how to establish rapport with and gain the trust of women they visit in order to secure their participation in the study. In-person contact at the beginning of the study will provide a solid basis for obtaining participants' cooperation and for tracking participants and ensuring high response rates for the follow-up data collection. Whenever possible, the same field staff will collect data during subsequent data collections, such as the follow-up survey. Field staff will also be trained in refusal-avoidance techniques. Participants will be assured that the information they provide will be secure, treated confidentially, and used only for research purposes. The family will receive a flier (Attachment 5) with information about the study, its importance, an estimated time line of when subsequent visits to the home for data collection will take place, and who the family can call with questions about the study.

Completing the baseline survey before random assignment ensures a 100 percent response rate. After a woman has agreed to participate, completed the baseline interview, and been randomly assigned, field staff will give her $25 to thank her for completing the interview and a small toy or book for the child of $15 value (if the child has been born at the time of the interview). Other methods of sample retention we propose are to send a birthday card to the parent on her birthday and to send the child a card when he or she reaches six months of age as a way of maintaining rapport with families and keeping their interest in the study. These mailings also provide additional opportunities outside of the tracking mailings to learn of address changes. Examples of these cards are included in Attachment 27.

The women will also be surveyed when her child is 15 months old. Tracking of study participants for follow-up data collection will begin with the initial visit. First, information will be gathered at the baseline interview to allow the study team to stay in touch with families until the follow-up interview is conducted. This information will include names, dates of birth, Social Security numbers (if possible), addresses, and telephone numbers (home and work) for the parents and detailed contact information for at least two relatives or friends who will know how to reach them in case we have difficulty doing so. Families will also be periodically sent cards that ask them to confirm or update their address and telephone information and return it in the self-addressed, postage-paid return envelope, and they will receive $5 for completing and returning the card. (An example of this card is included as Attachment 26.) At the second interim contact, the card will also request the sampled child's date of birth for parents who were pregnant at the time of study enrollment. A tracking database will identify when families are due for their tracking letter and generate these materials for mailing. Letters returned as undeliverable will be sent to the survey unit's tracking department for locating and then remailed to the updated address. The study team will call families that do not return a card within three weeks of the mailing in an attempt to verify their contact information by telephone. The study team will contact the secondary contacts for families that we cannot reach by telephone in an attempt to locate them. This script is included as Attachment 26.

Surveys of staff at participating home visiting program sites. When a site enters the study, the research team will explain to program staff the importance of the web-based surveys for advancing the field of home visiting in general and the MIECHV program in particular. Staff will receive a $30 gift card for each web-based survey they complete. Research staff will closely monitor data completion reports. If a survey is not completed within one week of the targeted time-frame, the site liaison will follow up with the point of contact at the site to remind the staff member that the survey response is due.

Surveys of administrators of community resources. When a site enters the study, the research team will explain the purpose and importance of the community resource survey to the program manager. Our targeting of administrators of community resources nominated by the program manager will maximize response rates as administrators will be more likely to respond to a survey about a program with which they work. Research staff will closely monitor data completion reports. They will send email reminders to administrators who do not complete the survey within a week of the initial contact. They will telephone administrators who do not complete the survey after three weekly email reminders. If research staff reach the administrator by phone, they will offer to complete the survey with the administrator by telephone.

Logs maintained by supervisors and home visitors. Strategies for maximizing response rates are the similar to those described above for the surveys of staff at participating home visiting program sites. When the site enters the study, the research team will explain to program staff the importance of the logs for advancing the field of home visiting in general and the MIECHV program in particular. Research staff will closely monitor weekly log completion reports. They will send program staff two weekly messages (Attachment 28). The first message will remind staff to complete their logs. The second message will document the data that were entered in the previous log by that staff person, thank the staff member for the data provided, and remind those who have not yet completed the previous week's log to do so.

Group and individual interviews with staff at participating home visiting program sites. When a site enters the study, the research team will explain to program staff the importance of the interviews for advancing the field of home visiting in general and the MIECHV program in particular. The MIHOPE team's in-person presence for the group and individual interviews will also motivate strong staff engagement and participation. In past studies by MDRC and the other MIHOPE partners, program staff have been very willing to participate in in-person interviews (both group and individual) and have attended scheduled interviews at high rates.

B4. Pre-testing

This section focuses on pretesting of quantitative data collection activities: the family baseline interview, the surveys of staff at participating home visiting program sites, the surveys of administrators of community resources, and the logs maintained by supervisors and home visitors. Each type of pretest was conducted with 9 or fewer parents or home visiting staff. Therefore, we have not included pretests in the burden estimates.

Pretest of family baseline survey. The study team is conducting an iterative pretest of the baseline interview. The team will conduct two rounds of pretests**.** The first pretest occurred in April 2012. The second will occur no later than four weeks before initiation of sample recruitment and after OMB approval.

The first pretest was completed with six participants via telephone by two Mathematica staff experienced with cognitive interviewing techniques. Participants consisted of three pregnant women and three women with young infants ranging from 5 months old to 15 months old. All participants were Maryland residents currently enrolled in home visiting programs, two each from Early Head Start (EHS), Nurse-Family Partnership (NFP), and Parents as Teachers (PAT). Pretest participants were recruited with assistance from EHS, NFP, and PAT program staff. We were unable to make contact with a Healthy Families American staff member, so we did not recruit families from that program.

The pretest interviews began by introducing the survey and informing women that participating in the survey was voluntary and that the data collected would be kept confidential. Five of the six participants consented to audio recording of the interviews. The interviewers asked each survey question exactly as worded and followed-up with specific probes for prescribed questions or if questions appeared confusing to respondents during the interview. Interviews ended with a short debriefing to solicit feedback on the survey experience from participants. Interviews of pregnant women took an average of 45 minutes to administer and interviews of women with infants took an average of 58 minutes to administer.

As a result of the pretest, a number of changes were made to the instrument, as summarized in Appendix B. Most changes were designed to avoid respondent confusion. Several follow-up questions were added, for example, to learn how long a newborn had stayed in a neonatal intensive care unit. One question was eliminated because respondents could not distinguish this question from another one. Two questions about the use of mental health or substance use services were simplified by reducing a long list of options to six options.

Because the average pretest interview took less than one hour to complete, questions were added to respond to a comment from the Nurse Family Partnership that the project identify a subgroup with low psychological resources. Questions from the Pearlin mastery scale and the Wechsler Adult Intelligence Scales Similarities subtest were added for this purpose. A description of the properties of the Wechsler subtest is provided in Appendix C.

Following OMB approval, the instrument will be programmed for CATI format. The second pretest will focus on testing the CATI program, which enables us to test the flow and skip logic of the instrument and to refine our CATI data collection procedures. As with the first round of pretesting, cognitive interviews will be conducted with parents and interviewers will be debriefed. This iterative approach to pretesting helps to ensure that the programmed instrument is almost final, reducing the need for costly changes to programming specifications.

Pretest of Implementation Instruments. Pretesting was carried out in March – May 2012 in preparation for the launch of the study in summer 2012. Appendix D summarizes changes to instruments resulting from pretesting and public comments.

The objectives of pretesting were to:

1. Assess readability and understandability of instructions and questions;
2. Estimate and minimize the time needed for staff completion of each instrument;
3. Confirm that questions and response choices would adequately measure each construct in the study implementation model for each benchmark and participant outcome; and
4. Identify technical problems with web-based administration of the instruments

As a result of pretesting and our own commitment to minimizing respondent burden, nearly all of the instruments have been streamlined, thereby reducing length and eliminating unnecessary repetition across the instruments.

Pretesting focused on the web-based instruments, whose numbers and names are as follows:

- 09 Program Manager Survey Part 1
- 10 Program Manager Survey Part 2
- 11 Program Manager Survey Part 3 / Community Services Inventory
- 13 Supervisor Survey Baseline
- 15 Home Visitor Survey Baseline
- 19 Supervisor Logs
- 20 Home Visitor Logs

Several other implementation study instruments were edited to improve their clarity, minimize the time needed for completion, and assure that questions and response choices would adequately measure each construct. The semi-structured group and individual interviews were edited to delete items that were redundant with the baseline and 12-month web-based surveys and to identify optional items and potential probes. Only a subset of questions will be asked, with the exact subset to be determined by the specifics of the data collected in the other instruments completed by the participating sites. These edits were motivated by public comments and by the need to re-align the instruments with the modified web-based instruments.

1. *Overview of Procedures*

Each instrument was pretested with at least one staff member from each model included in the MIECHV evaluation: Early Head Start (EHS), Healthy Families America (HFA), Nurse Family Partnership (NFP), and Parents as Teachers (PAT). Each item was tested on nine or fewer people in total. As planned, we used an iterative approach to testing. One round of pretesting was carried out in March, and in April and May additional pretests were conducted to address problems identified in the earlier iterations.

2. *Identification and Recruitment of Pretest Sites and Respondents*

For pretesting, home visiting program contacts in Florida, Georgia, Maryland, New Jersey, and Washington state were identified.MIHOPE team members had prior relationships with these contacts, and they were thought to be amenable to participating in pre-testing. These contacts were sent introductory emails that described the pretesting opportunity, and they in turn put the team in touch with individual staff members who might be interested in taking part in pretesting. The first round of pretests occurred in Maryland, New Jersey, and Washington states. The subsequent rounds of pretests occurred in Georgia, Maryland, New Jersey, and Washington states.

3. *Overview of Pretesting and Cognitive Interview Procedures*

A pretest and cognitive interviewing protocol was developed based on best practices from the field (Willis, 2006; Napoles-Springer, Santoyo-Olsson, O'Brien, & Stewart, 2006) and the resources available for pretesting. The cognitive interviews were conducted by MIHOPE implementation study team members over the phone. Team members followed explicit protocols eliciting information to meet the pretesting objectives identified earlier in this section. The content of each cognitive interview was documented in an Excel database for analysis.

Results from Pretest of Implementation Instruments

The results of pretests are summarized in Appendix D.

B5. Consultants on statistical aspects of the design

There are no consultants on the statistical aspects of Phase 1. We have drawn on the expertise of team members including Charles Michalopoulos and Howard Bloom of MDRC.

Appendix A: Justification for Not Including Direct Child Assessments at Baseline

This memo discusses the potential child assessment measures that could be conducted, and presents our recommendations. The recommendations are informed by consultation with Sally Atkins-Burnett, Jerry West, and members of the Secretary's Advisory Committee (SAC).

The recommendations are influenced by the three intended uses of the MIHOPE baseline family survey:

1. Describe the characteristics of families that participate in the study

2. Define the analytic subgroups that will be used in the impact analyses

3. Increase the precision of the impact estimates by including measures of key domains at baseline and follow up

The survey will be administered to pregnant women and women with children from birth to 6 months of age, the key groups targeted by the home visiting programs. At this time we do not know the relative proportion of each group, but estimate that approximately one-third to one-half will not be born at the time of the baseline survey. Of those that are born, it is likely that half will be newborns (0 to less than 3 months) and half will be between 3 and 6 months old at the time of the baseline survey. The baseline participant survey will collect data on baseline family characteristics from two data sources: a baseline interview and the observational items from the Home Observation for Measuring the Environment (HOME; Caldwell and Bradley 2003) assessment. The baseline interview will be conducted by computer assisted telephone interview (CATI) to preserve privacy of study participants and to increase the efficiency and security of data collection. The HOME assessment will be conducted by field staff after they obtain informed consent from the family and while the participant is completing the baseline interview on the telephone.

A. CHILD ASSESSMENT MEASURES

There are a number of child assessment measures that could potentially be used for this study at baseline. We list the measures below and some factors to consider in weighing the challenges and benefits of each one as a baseline measure.

ITSEA/BITSEA: This is a parent report measure of child social-emotional well-being and is being used on Baby FACES. However, it is only normed for children aged 12 to 36 months. This was confirmed in an email exchange with the developer, Margaret Briggs-Gowan, who noted that "purposely did not design the BITSEA for less than 12 months due to concern about implying that psychopathology might "exist" at such a young age." Therefore we cannot use it for the baseline survey. We could potentially use it at follow up.

Bayley Scales of Infant Development (Bayley): This measure can be used as early as 1 month of age. Here we discuss five versions of the assessment, the Bayley-II and the short form based on it developed for the ECLS-B, and the Bayley-II screener, the Bayley-III, and the Bayley-III Screening Test. In addition, we summarize the Social-Emotional Scale included in the

Bayley-III, a parent-completed questionnaire based on the Greenspan Social-Emotional Growth Chart (Greenspan 2004).

Published in 1993, the norming sample from the **Bayley-II** is dated and no longer reflects the population of children in the United States. Short forms of the Bayley-II mental and motor scales for 9-month and 24-month old children (BSF-R, Andreassen and Fletcher 2005) were developed with considerable effort and expense for the ECLS-B to simplify administration and reduce data collection time. The BSF-R was developed in response to a much longer than expected administration time encountered during the 1999 field test of the full BSID-II. Development of the short form was also designed to address difficulties field staff had administering and scoring the items using the standardization rules specified by the test developer. The BSF-R took approximately 36 minutes when the children were 9 months old in the ECLS-B. It would take considerable measurement development and psychometric work to create a short form appropriate for the MIHOPE age range. The Bayley Infant Neurodevelopmental Screener (BINS; Aylward 1995) is based on the Bayley-II and screens infants between the ages of 3 and 24 months for neurological impairments and developmental delays. It takes between 5 and 10 minutes to administer but as a screener, it does not show much variation in typically developing children's development. In addition, it does not extend down to cover the birth through 3 month age range.

The Bayley-III (2006) has not been used in a large-scale national study and was not recommended for the Baby FACES study for that reason and because it has a new, untested approach to separately measuring different outcome domains and computing separate scale scores based on a relatively small number of items appropriate to each age range. The Bayley-III direct child assessment has been organized into three scales and five subtests: (1) the Cognitive Scale is comprised of one subtest, (2) the Language Scale is comprised of the Receptive Communication and Expressive Communication subtests, and (3) the Motor Scale is comprised of the Fine Motor and Gross Motor subtests. In addition, the Social-Emotional Scale and the Adaptive Behavior Scale are two separate parent-report questionnaires. Both questionnaires and any direct assessment items that require the interviewer to speak to the child or parent would have to be translated into Spanish, as neither the Bayley-II nor the Bayley-III are available in Spanish. Other important concerns include (1) the Bayley-III's length, (2) the fact that the test has been normed in English only, (3) the lack of data about how predictive the scales are when used with infants 0-6 months, and (4) the fact that each scale has only a few items in it (which may result in severe floor and ceiling effects). The Bayley-III Screening Test (for 1 to 42 months) maintains the same multi-scale structure of the direct assessments in the full Bayley-III with even fewer items included per subtest (which exacerbates floor and ceiling effects). Given that it is based on the Bayley-III, the same issues described above regarding the norming sample apply.

Information publicly accessible indicates that the National Children's Study (NCS) is piloting a short form of the Bayley-III in four locations across the country using procedures similar to what was done for the ECLS-B that focus on reducing the length of the assessment and increasing the reliability of the administration by field staff. Several consultants suggested that due to these multiple concerns, particularly the lack of predictive validity data and the fact that the NCS version is only expected to extend down to 6 months, which is not far enough for the

MIHOPE baseline (we saw reference to 6-month IRT scores in what was publicly available), the Bayley should not be included at baseline or follow up. The Bayley Infant Neurodevelopmental Screener (BINS; Aylward 1995) is based on the Bayley-II and screens infants between the ages of 3 and 24 months for neurological impairments and developmental delays. It takes between 5 and 10 minutes to administer but as a screener, it does not show much variation in typically developing children's development. In addition, it does not extend down to cover the birth through 3 month age range.

The Greenspan Social-Emotional Growth Chart (Greenspan 2004): This assessment is now part of the Bayley-III and is completed by the child's parent or primary caregiver. It is based on functional emotional milestones that correspond to 8 stages for children from birth to 42 months of age (Bayley 2006). One concern about this measure is the small norming sample and very small sample sizes included in it for ages 0-3, 4-5, and 6-9 months (89, 54, and 51, respectively). In addition, we do not believe it has been used in a large-scale national study of high-risk parents and children.

Mullen (1995): This measure can be used from birth through 68 months. However, the norming sample is out-dated and it is only available in English.

Three/Two Boxes/Bags Task and Coding System: This measure examines parenting constructs such as supportiveness, sensitivity, cognitive stimulation, intrusiveness, and negative regard. It also includes scales that examine child engagement of parent, sustained attention, and negativity toward parent. The semi-structured play task and variations of the original coding scheme by Deborah Vandell and colleagues have been used with children 14, 24, and 36 months old in a number of studies, including the Early Head Start Research and Evaluation project, Fragile Families, ECLS-B, and Baby FACES. It was used with children six months and older in the NICHD Study of Early Child Care and Youth Development and in the Early Head Start Newborn Study. Predictive validity data from use of the task and coding system from 0-6 months is scant. This type of task and coding system are being considered for the follow-up assessment with the full sample.

The Nursing Child Assessment Teaching Scale (NCATS) (1995): This observational measure of the quality of the caregiver-child teaching interaction for children from birth to 3 years of age assesses four parent and two child behaviors. The correlations of the total NCATS scores with the total HOME score among children ages 1 to 36 months, in three age groups, ranged from .41 to .44. Given that the HOME is already planned for MIHOPE, NCATS may not add much additional information given the relative cost of training on the assessment. In addition, the adaptations to shorten the observation period made for administering the NCATS in the EHS-REP revealed internal consistency reliability problems inherent in large-scale live or videotaped coding and administration of the measure. There is scant information available about the predictive validity of the NCATS Teaching Task when conducted with children less than 6 months old.

Brazleton (1973): This measure is used with infants and usually in hospital settings. It is a scale for 0-2 months of age, so its use for this study is limited as our sample at baseline will include children 0 to 6 months of age.

Neonatal Intensive Care Unit Network Neurobehavioral Scale (NNNS) (2004). This neurological assessment can be conducted from birth through 48 weeks. The infant should start off in a sleep state that has been maintained for at least 45 minutes. There are 115 items and several position changes are required during which the observer looks for changes in the baby. This assessment requires a highly trained individual, usually a clinician, and does not seem to be suitable for a large-scale study. Although there are a few published articles on the measure, there is little information available on its predictive validity and it has been used primarily for clinical purposes.

Other ECLS-B 9-Month Assessments: The remaining set of measures used at 9 months assesses infant physical development, including weight, length, upper arm circumference, and head circumference. Although a direct assessment may be desirable, we will be getting most of this information from other sources.

B. RECOMMENDATIONS FOR CONDUCTING DIRECT CHILD ASSESSMENT

In developing the MIHOPE baseline survey, we focused on including measures of outcome domains that are most likely to show impacts or that had the potential to mediate or moderate impacts. Direct child assessments of the portion of the sample that includes infants that were born at baseline were considered but rejected because they could only be administered for part of the sample (unborn children would have no data for these measures) and because the developmental experts we consulted with and our SAC recommended against directly assessing children 12 months of age or younger.

For over forty years, the predictive validity of infant assessments, particularly those administered to children less than one year of age, has been an issue for the field of developmental psychology. In the 1970's and 1980's, leading developmentalists debated this issue related to performance of children less than 1 year old on the Bayley and correlations to subsequent cognitive functioning (Lewis and McGurk 1972; Lewis and McGurk 1973; Matheny 1973; McCall 1981; Wilson 1973). Then, as now, researchers have concerns about the predictive validity of assessments conducted with young infants and generally recommend they be used for assessing performance at a given point in time for diagnostic and comparative purposes rather than as predictors of later skills and abilities (for example, Hack et al. 2005). A few measures of information processing for children less than 6 months old have been identified as somewhat more robust predictors to intelligence at 3 years of age (for example, the Fagan Test of Infant Intelligence 2005), but they have not been used in large-scale research projects and are more suitable to laboratory settings than to in-home assessment. The primary arguments against conducting direct child assessments stem from the lack of reliable and valid measures in early infancy; overall, the predictive validity of the measures that are available is either unknown or quite low.

After weighing the information above against the practical issues such as cost, we do not recommend conducting direct child assessments on the MIHOPE study for the following reasons:

1. **Sample Size and Variation.** About one-third to one-half of the sample at baseline will be pregnant women, so we would be able to obtain child assessment data for only part of our sample. The sample of children at baseline will also vary widely, with ages ranging from 1 day old to 6 months. There are few child assessment measures that are suitable for this age group.

2. **Cost.** The cost of conducting child assessments would be high and would require more funds than what have been allocated for the baseline effort. We would need to hire and train a group of staff with experience in complicated direct child assessments. We would need to pay more per hour since they will be doing work that is more difficult. Training will take substantially longer than what we budgeted (4 plus days rather than 2 days). Certification on the measures would be difficult and many staff would not pass, which would require additional hiring, training, and certification.

3. **Logistics.** The logistics of conducting assessments would be more challenging. The baseline visit would be longer, since the field staff would be conducting an assessment. We would need the infant to be awake, which could necessitate going back to the home multiple times to complete the assessment. These logistical considerations would also increase the cost of the baseline data collection.

4. **Low Return on Investment**. There is generally low predictive value of the standard child assessment measures at very young ages (birth to 6 months). We do not believe that the data gathered would provide us with adequate information to make the effort worthwhile. In addition, the measures that could be used at both baseline and follow-up are few and have the limitations described above.

Appendix B: Summary of Changes Made to Family Baseline Survey

Survey Item	Change resulting from pretesting	Rationale
A7 After [CHILD] was born, how long did [he/she] stay in the hospital? A8 After your baby [CHILD] was born, was [he/she] put in an intensive care unit or NICU?	Revised A8 to ask if any of these days were in the NICU, and then if yes, ask for number of days child spent in the NICU.	Will help to clarify how long the baby spent in the NICU.
A13 Do you have a plan to breastfeed?	Revised to "Do you plan to breastfeed?"	Respondents had difficulty with the word "plan." They often responded with "I hope to" or "I'd like to." Revised wording will help match respondent's intent.
A14 How long do you plan to breastfeed?	Revised to "How long would you like to breastfeed?"	
A15 How old was [CHILD] the first time (he/she) ate or drank anything other than (breast milk or) formula?	Replaced with the following item from the ECLS-B 9-month parent interview: "How old was [CHILD] in months when solid food was first introduced? Solid foods include cereal and baby food in jars, but not finger foods."	Revised wording is more specific to ensure respondent understands the question.
B1 The next questions are about your health. In general, would you say your health is…?	For pregnant women, revised to "The next questions are about your health before your current pregnancy. In general, would you say your health is…?	To clarify for pregnant women that they should answer about their health before pregnancy, so they do not consider any pregnancy-related ailments when responding.
B5 During (this pregnancy/your pregnancy with [CHILD]), were you told by a doctor, nurse, or other health care worker that you had gestational diabetes (diabetes that started during this pregnancy)?	Add response option, "haven't been tested yet" for pregnant women.	It is possible that some women may not be far enough along in their pregnancy to have been tested for gestational diabetes.
B8 Is there a place you go for general health care, if you are sick or need advice about your health - that is, any care except prenatal care or family	Added the follow-up item: B8a. What kind of place do you go?	Not all pretest respondents knew that we were asking about a physical location.

Survey Item	Change resulting from pretesting	Rationale
planning?	Clinic Health Center Hospital Doctor's office Some other place	
B9 During the past year, have you ever received family planning or gynecologic services? B9a During the past year, did you ever want or need family planning or gynecologic services? B9b What is the main reason you didn't receive family planning or gynecologic services? B9c Are you currently receiving family planning or gynecologic services?	Replaced with the following items: B9. Is there a place you go, or have gone, for family planning or birth control? B9a. What kind of place do you go/ did you go? The same place I receive general health care Clinic Health Center Hospital Doctor's office Some other place	Some respondents were confused by term "family planning services."
B10 How many more children do you plan to have?	Revised to "How many more children would you like to have?"	Respondents had difficulty with the word "plan." They often responded with "I hope to" or "I'd like to." Revised wording will help match respondent's intent.
C8 What is the highest grade or year of regular school that you have completed?	Removed "regular" from question	Respondents were confused by the term "regular."
Section D items on the woman's spouse or partner	If a woman doesn't have a spouse or partner and doesn't live with the child's biological father, added an item asking if the woman and biological father ever lived together. Added an item asking if the woman is currently in a romantic relationship, and for those who say yes, then ask the intimate partner violence items.	This section didn't flow well during the pretest. This revision will help fill in missing information

Survey Item	Change resulting from pretesting	Rationale
Section E items on household composition and earnings	Revised items about household composition and earnings to accommodate respondents whose household composition is currently different than it was for most of the previous year.	These questions were difficult for respondents to answer if the current household members were not the same as in the prior year (when we ask for total earnings from all household members.) changing these items will make answering them easier for the respondent.
E4 How many months were you employed (did you work for pay) during the past 3 years (including your current job)? RESPONDENT DIDN'T WORK Less than 6 months 7 to 12 MONTHS 13 to 24 MONTHS More than 24 months	Changed format so that interviewer reads the answer choices aloud to respondent, except for "respondent didn't work."	Pretest respondents had trouble calculating number of months; providing answer choices helped them respond.
E19 During the past year, have you received Early Head Start or child care services for [CHILD]?	Revised.	Respondents were confused and wondered if we meant EHS only or child care in general.
E20 During the past year, have you ever received Early Intervention services or [INSERT NAME OF PROGRAM FOR STATE] for (CHILD)? E20a Did you ever want or need Early Intervention services for [CHILD]? E20b Are you currently receiving Early Intervention services for [CHILD]?	Deleted from survey.	Since most babies will be too young to have received early intervention services at baseline, we recommend deleting this question and including it in the follow-up survey.
E21/22 a-c Home visiting items	Moved to end of survey, just before contact information.	Moving the home visiting questions to the end eases the transition from the

Survey Item	Change resulting from pretesting	Rationale
		end of the survey to collecting contact information.
E22a-c What do you think will be the three most important benefits of home visiting for you and your family?	Deleted these items.	Responses here were the same as those captured in E21.
F15-F18 questions on receipt of mental health and substance abuse treatment services during past year	Shortened the list by grouping similar items together and using broader categories	The list of items was long, categories were redundant, and the list was cumbersome to administer
G6 Please tell me whether you or any other members of your household received income from the following sources in the past month. This includes anyone who you support and/or supports you and lives in your household. G7 During the past year, have you ever received help in applying for public benefits, including TANF, SNAP, or WIC?	Added "WIC" to the list of sources in G6.	Four respondents said yes to G7 because they received WIC, but didn't understand that the question was asking if they had received *help in applying* for services like WIC. Add WIC to G6 to capture this benefit.
	Added questions from the Pearlin mastery scale	Added in response to an NFP comment suggesting the measurement of low psychological resources.
	Added questions from the Wechsler Adult Intelligence Scale Similarities subtest	Added in response to an NFP comment suggesting the measurement of low psychological resources.

Appendix C: Measuring Cognitive Ability

To measure cognitive ability, the MIHOPE baseline survey will contain the Similarities subtest of the Wechsler Adult Intelligence Scales – Third Edition (WAIS-III; Wechsler, 1997). The Similarities subtest is designed to capture abstract reasoning and verbal comprehension abilities, which are two principal dimensions of intellectual abilities (Flanagan and Harrison, 2005; Flanagan, Ortiz and Alfonso, 2007). In the Similarities subtest, respondents are asked a series of questions about how two things are alike. For example, "How are a snake and an alligator alike?" Each item is then scored on a 0 to 2 scale according to general scoring principles and examples that are provided in the testing manual.

This measure is proposed to assess parents' cognitive and intellectual abilities for a variety of reasons:

- The Wechsler Adult Intelligence Scales are among the most widely used measure of intellectual abilities in the United States and in other counties. The WAIS-III Similarities subtest is also one of the few measures of abstract reasoning and verbal comprehension that is available in both English and Spanish that can be readily administered over the telephone or in person.

- Compared with most other assessments of intellectual abilities, the Similarities subtest is relatively brief – consisting of only 18 items – which places substantially less burden on study participants than most other measures of cognitive and intellectual abilities. Furthermore, study participants need not receive all of the items because the testing includes a discontinuation rule when respondents get three consecutive items incorrect. Thus, the amount of time required to administer the subtest can be quite brief and varies the study participants' intellectual aptitude thereby reducing the burden of the measure on study participants.

- The English and Spanish versions of the Similarities subtest have been shown to have good psychometric properties. The publishers of the English version of the subtest found that its split-half reliability is 0.87, the test-retest reliability is 0.83, and the inter-rater agreement on scoring the items of the subtest (ICCs) is 0.93 (Tulsky et al., 1997). Elsewhere, Renteria et al. (2008) found the Spanish WAIS-III Similarities subtest had an internal consistency of 0.79 using a sample of primarily Spanish-speaking adults recruited from Chicago neighborhoods.

- The Similarities subtest has been shown to have good validity and demonstrated capabilities for differentiating individuals with qualitatively different levels of intellectual abilities. In numerous studies the Similarities subtest is a strong predictor of the full-scale score of intellectual functioning that can be created when the full battery of subtests from the WAIS-III. Jones et al. (2006), for example, found that the Similarities subtest loads onto the WAIS full-scale score of overall intelligence at 0.81 in a factor analytic model. Moreover, using a sample of adults who are diagnosed with mild intellectual disabilities according to the DSM-IV-TR criteria for intellectual disabilities (e.g., IQs of 40 – 70), the publishers found that this group on the Similarities subtest scored about 2.5 standard

deviations lower than a matched comparison group with average intelligence (Tulsky et al., 1997). Using a sample of adults who meet the DSM-IV-TR criteria for Borderline Intellectual Functioning (e.g., IQs of 71 – 84), the publishers also found that the group scored about 1.4 standard deviations lower on the Similarities subtest than a matched comparison group with average intelligence (Tulsky et al., 1997).

Appendix D: Implementation Study Instruments – Content in Paired Instruments and Revisions per Pretesting and in Response to Public Comments

Instrument (Number)	Comparison of Content in Paired Instruments	Revisions Resulting from Pretesting	Response to Public Comments
State administrator interview			
Baseline (7)	The baseline survey gathers data on MIECHV- and state-level factors for service delivery, from the perspective of the state's lead agency for MIECHV.	Sections K and L were reformatted to improve clarity.	*Comments*: None
12 Month (8)	The content of the 12-month interview parallels that of the baseline interview. Items elicit information on changes in factors since the baseline survey.	The 12 month interview was edited to align with the revised baseline instrument.	*Comments*: None
Program manager survey			
Part 1, Baseline (9)	The content of each of the three parts of the baseline survey is unique. The three parts are complementary. Together, they gather baseline data on the full set of hypothesized program site factors for service delivery, from the perspective of site	As possible, sections on site policies and procedures were edited to make data collection more efficient by using questions about policies in lieu of requests for copies of the policies. Items on current staff were moved to Part 2 because they fit better with its content.	*Comment*: It is unclear which survey instruments will be completed by a program manager who is also a supervisor. *Response*: A program manager who is also a supervisor will complete the program manager survey and sections of the supervisor survey that are not redundant with the program manager survey.
Part 2, Baseline (10)		Items that could be answered more efficiently via other instruments were	*Comments*: None

Instrument (Number)	Comparison of Content in Paired Instruments	Revisions Resulting from Pretesting	Response to Public Comments
	leadership.	eliminated. Items were reworded as needed to improve clarity and to maintain alignment with parallel items in other instruments. A few items were added to fill identified gaps and eliminate ambiguity in responses. Items on referrals to community resources were moved to Part 3 because they fit better there.	
Part 3, Baseline (11)		Items were reworded as needed to improve clarity and to maintain alignment with parallel items in other instruments. A few items were added to fill identified gaps and eliminate ambiguity in responses.	*Comment*: Questions about referral are redundant with questions in the supervisor survey. *Response*: We have eliminated this redundancy by dropping these questions from the supervisor survey. The questions are now a part of only the program manager survey. A site can choose to have a supervisor or other staff member help the program manager answer these questions if the site feels that is more efficient.
12 Month (12)	The content of the 12-month survey parallels that of parts 1 and 2 and a small portion of part 3 of the baseline survey. Thus, comparison of responses from baseline to the 12 month survey allows assessment of change over time.	The 12 month survey was edited to align with the revised baseline instrument.	*Comments*: None
Supervisor survey			
Baseline (13)	The baseline survey gathers data on hypothesized program	Items were reworded as needed to improve clarity and to maintain alignment with parallel items in other instruments.	*Comment*: It is unclear whether a supervisor who is also a home visitor will complete both or only one survey.

Instrument (Number)	Comparison of Content in Paired Instruments	Revisions Resulting from Pretesting	Response to Public Comments
	site factors for service delivery from the perspective of supervisors, and on supervisor-specific factors for service delivery.	Items that could be answered more efficiently via other instruments were eliminated. In Sections L-S, items were reorganized, reworded, and some items were eliminated to improve efficiency. Some items were added to fill identified gaps and to eliminate ambiguity in responses.	*Response*: A supervisor who is also a home visitor will complete the supervisor survey and portions of the home visitor survey that are not redundant with the supervisor survey. *Comment*: It is unclear which survey instruments will be completed by a replacement supervisor. *Response*: A replacement supervisor will complete a baseline survey upon joining the study. S/he will also complete the 12 month survey if s/he joins the study at least 6 months prior to the 12 month survey.
12 Month (14)	The content of the 12-month survey parallels that of the baseline survey. Thus, comparison of responses from baseline to the 12 month survey allows assessment of change over time.	The 12 month survey was edited to align with the revised baseline instrument.	*Comment*: Both the Baseline and the 12 month surveys ask about program expectations, which is unnecessarily repetitious. *Response*: We have deleted a few of the redundant items. Redundancies are by design, to capture expected site-level changes in program models and implementation systems over time. The MIECHV program has already given rise to substantial changes in home visiting at the national, state, local and program site levels. We expect this will continue in the years ahead. Thus, we have designed the 12-month staff surveys to assess changes in both organization- and individual-level factors for service delivery.
Home visitor survey			
Baseline (15)	The baseline survey gathers data on hypothesized program site factors for service delivery from the perspective of home visitors, and on home visitor-specific factors	Items were reworded as needed to improve clarity and to maintain alignment with parallel items in other instruments. Items that could be answered more efficiently via other instruments were eliminated. In Sections L-S, items were reorganized,	*Comment*: It is unclear which survey instruments will be completed by a replacement home visitor. *Response*: A replacement home visitor will complete a baseline survey upon joining the study. S/he will also complete the 12 month survey if s/he joins the study at least 6 months prior to the 12 month survey.

Instrument (Number)	Comparison of Content in Paired Instruments	Revisions Resulting from Pretesting	Response to Public Comments
	for service delivery.	reworded, and some items were eliminated to improve efficiency. Some items were added to fill identified gaps and eliminate ambiguity in responses.	*Comment:* The home visitor baseline survey remains lengthy. *Response:* Editing as part of pretesting has reduced the number of items by about 20%. Pretesting has established that home visitors can complete the survey within the projected time. *Comment*: 105 items are embedded, not fully shown. *Response*: The source instrument, which is proprietary, was identified by name – the Organizational Social Context (OSC) scales. The commenting organization is familiar with this instrument, having reviewed its items and approved its use in December, 2011 for another home visiting study conducted by MIHOPE team members, in which its sites participate. *Comment*: There was concern that the number of items measuring home visitor psychosocial functioning (n=105 + 39) is burdensome and intrusive. *Response*: This section of the survey includes three instruments: the OSC (105 items), the short form of the CES-D (10 items), and the Attachment Style Questionnaire (29 items). We did not change this section, for several reasons. First, the three instruments in this section measure different constructs, all of which are hypothesized to have independent influences on service delivery and impact. There is theoretical and empirical support for the independence influence of each of these constructs on service delivery and impact. Second, the OSC measures not only individual level factors

Instrument (Number)	Comparison of Content in Paired Instruments	Revisions Resulting from Pretesting	Response to Public Comments
			(morale and burnout) but is the study primary measure of two key organization-level factors (culture and climate). Third, depressive symptoms and relationship security have been shown to influence service delivery, *and* to have interactive effects on family engagement. Fourth, leaders of other evidence-based home visiting models expressed their support for assessing staff psychosocial well-being at the ACF/HRSA-sponsored MIHOPE meeting of model developers on October 27, 2011. *Comment*: Questions about home visitors' background as a parent or home visiting recipient seem judgmental. *Response*: These items have been deleted. *Comment*: Questions about referral are redundant with questions in the program manager survey. *Response*: We have kept the referral questions in both instruments. The questions are similar by design, but they serve different purposes. We use answers to referral questions in the program manager survey to assess the site's awareness of and relationship with community resources. We use answers to referral questions in the home visitor survey to measure each home visitor's knowledge of, attitudes toward, and interactions with community resources.
12 Month (16)	The content of the 12-month survey parallels that of the baseline survey. Thus, comparison of responses from baseline to the 12	The 12 month survey was edited to align with the revised baseline instrument.	*Comment*: Both the Baseline and the 12 month surveys ask about program expectations, which is unnecessarily repetitious. *Response*: We have deleted a few of the redundant items. Redundancies are by design, to capture

Instrument (Number)	Comparison of Content in Paired Instruments	Revisions Resulting from Pretesting	Response to Public Comments
	month survey allows assessment of change over time.		expected site-level changes in program models and implementation systems over time. The MIECHV program has already given rise to substantial changes in home visiting at the national, state, local and program site levels. We expect this will continue in the years ahead. Thus, we have designed the 12-month staff surveys to assess changes in both organization- and individual-level factors for service delivery.
Community service provider survey (17)	This survey is conducted at baseline only. Its content parallels that of Part 3 of the program manager baseline survey for each type of service provider listed. It elicits the community service provider's perspective on referral and coordination with a specific home visiting site and on service availability, service accessibility and inter-agency agreements as factors for referral and coordination.	We did not pretest this instrument. Two items were added to the survey to address identified gaps (agency address and cost of services). Items were reworded as needed to improve clarity and to maintain alignment with parallel items in other instruments.	*Comments*: None
Other home visiting program survey (18)	This survey is conducted at baseline only. It documents key characteristics of other home visiting or	We did not pretest this instrument.	*Comments*: None

Instrument (Number)	Comparison of Content in Paired Instruments	Revisions Resulting from Pretesting	Response to Public Comments
	parenting programs for infants in the community in which control group members might enroll.		
Supervisor logs (19)	These logs are completed weekly to measure supervisor training and actual supervision from the perspective of the supervisor as factors that influence actual service delivery.	Items were reworded as needed to improve clarity and to maintain alignment with parallel items in other instruments.	*Comment*: The logs are burdensome because staff are expected to complete them weekly and because they are duplicative of forms that staff complete routinely as part of (NFP) model requirements. *Response*: We reduced the number of items. We expect that each supervisor will complete weekly logs only for home visitors with one or more active families participating in the evaluation. For this reason, repetitiveness is limited. Although the content of the MIHOPE logs overlaps slightly with NFP logs, most items in the MIHOPE logs ask for content different than that in NFP logs. *Comment:* The frequency of log completion should be reconsidered, perhaps to a monthly summative reporting across all home visitors. *Response*: To understand variation in actual services to families and factors that influence service delivery, the study must collect uniform information across all outcome domains for all models and program sites. The logs provide key information about individual-level service delivery and supervision for "black box" analyses as well as for documenting variations in program costs for participant subgroups. No national model requires sites to collect the full set of supervision variables needed for MIHOPE; some sites might not collect any of this information in a systematic way. Our previous research using logs suggests that less

Instrument (Number)	Comparison of Content in Paired Instruments	Revisions Resulting from Pretesting	Response to Public Comments
			frequent completion will negatively impact staff recall of events. Our previous research highlights substantial variability in the intensity and content of both home visits and supervision. We need to measure supervision at the home visitor level and service delivery at the client level. These measures will be key variables in analyses factors explaining variations in service delivery and fidelity. Variation in service delivery and fidelity will, in turn, be tested as a moderator of program impacts.
Home visitor logs (20)	These logs are completed weekly to measure actual service delivery and home visitor perspectives on actual training and supervision as factors for service delivery.	Items on approaches to service delivery within each content area were dropped to reduce respondent burden. Items were reworded as needed to improve clarity and to maintain alignment with parallel items in other instruments.	*Comment*: The logs are burdensome because staff are expected to complete them weekly and because they are duplicative of forms that staff complete routinely as part of (NFP) model requirements. *Response*: We have reduced the number of items in the logs. Home visitors will complete weekly logs only for active families participating in the evaluation. On average, this will be only about five families, a small portion of the home visitor's caseload. For this reason, repetitiveness is limited. Although the content of the MIHOPE logs overlaps slightly with NFP logs, most items in the MIHOPE logs ask for content different than that in NFP logs. *Comment:* The frequency of log completion should be reconsidered, perhaps to a monthly summative reporting across all home visitors. *Response*: To understand variation in actual services to families and factors that influence service delivery, the study must collect uniform information across all outcome domains for all models and program sites. The logs provide key information about individual-level service delivery and supervision for "black box" analyses as well as for

Instrument (Number)	Comparison of Content in Paired Instruments	Revisions Resulting from Pretesting	Response to Public Comments
			documenting variations in program costs for participant subgroups. No national model requires sites to collect the full set of supervision variables needed for MIHOPE; some sites might not collect any of this information in a systematic way. Our previous research using logs suggests that less frequent completion will negatively impact staff recall of events. Our previous research highlights substantial variability in the intensity and content of both home visits and supervision. We need to measure supervision at the home visitor level and service delivery at the client level. These measures will be key variables in analyses factors explaining variations in service delivery and fidelity. Variation in service delivery and fidelity will, in turn, be tested as a moderator of program impacts.
Semi-Structured Interviews			
Group interview – program managers (21) Group interview – supervisors (22) Group interview – home visitors (23)	These group interviews are conducted at 12 months to elicit staff perspectives for interpreting data collected in the surveys and logs, that is to explain the how and why behind quantitative results.	For group interviews with program managers, supervisors and home visitors, we deleted items that were redundant with the staff surveys, added a few questions to fill identified gaps, and edited questions to elicit participants' perspectives on the reasons and mechanisms for results obtained through the surveys.	*Comment*: There is considerable duplication of questions across the 12 month surveys and interviews. *Response*: We have eliminated the Interview participant questionnaire (formally Instrument 24), as it was duplicative of items asked on the baseline surveys. We deleted items from the group and individual home visitor interview instruments that were redundant with the baseline and 12 month surveys (Instruments 13-16). In instruments for both the group and individual interviews, most items are, in fact, either optional or
Interview participant questionnaire (24)	This questionnaire elicits basic information to characterize group interview participants	This instrument has been eliminated .	

Instrument (Number)	Comparison of Content in Paired Instruments	Revisions Resulting from Pretesting	Response to Public Comments
	(Instruments 21-23)		potential probes. We will ask only a subset of questions, with the exact subset to be determined by the specifics of the data collected in the other instruments completed by the participating sites. We've edited the instruments to identify optional items and potential probes.
Individual interview – home visitors (25)	These individual interviews are conducted at 12 months to elicit staff perspectives for interpreting data collected in the surveys and logs, that is to explain the how and why behind quantitative results. The individual interviews seek to elicit views that home visitors are less likely to share candidly in group interviews.	For the individual interviews with home visitors, we deleted items that could be answered adequately in the group interviews, added a few questions to fill identified gaps, and edited questions to elicit participants' perspectives on the reasons and mechanisms for results obtained through the surveys.	*Comment*: It is unclear whether replacement supervisors and home visitors will complete the interviews. *Response*: Replacement home visitors and supervisors will be eligible to participate in the interviews if they joined the study at least 6 months earlier.
Messages to home visiting program staff (28)	These messages thank staff for completing logs and remind staff to do so.	No changes	*Comments*: None

REFERENCES

Andreassen, C., and P. Fletcher. "*Early Childhood Longitudinal Study, Birth Cohort (ECLS-B) Methodology Report for the Nine-Month Data Collection (2001-02), Volume 1: Psychometric Characteristics*." Submitted to the U.S. Department of Education. Report No. (NCES 2005-100). Washington, DC: National Center for Education Statistics, 2005.

Aylward, G. P.. 1995. Bayley Infant Neurodevelopmental Screener. *San Antonio, TX: Psychological Corporation.*

Bayley, Nancy. Bayley Scales of Infant and Toddler Development-Third Edition: Administration and Technical Manual. San Antonio, TX: PsychCorp, 2006.

Bloom, Howard S., Carolyn J. Hill and James A. Riccio. 2003. "Linking Program Implementation and Effectiveness: Lessons from a Pooled Sample of Welfare-to-Work Experiments," *Journal of Policy Analysis and Management*, 22(4): 551 – 575.

Durlak, J. A., and E. P. DuPre. 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* 41, 3-4: 327-350.

Fagan, Joseph F. 2005. "The Fagan Test of Infant Intelligence-Manual." Website: http://infantest.com/ftii.pdf.

Filene, Jill H., James Bell, and Elliott G. Smith. 2011. National Cross-Site Evaluation of the Replication of Family Connections: Final Evaluation Report. Report submitted to the Administration for Children and Families.

Flanagan, D. P., & Harrison, P. L. (2005). Contemporary Intellectual Assessment: Theories, Tests, and Issues. (2nd Edition). New York, NY: The Guilford Press.

Flanagan, D. P., Ortiz, S. O., & Alfonso, V. C. (2007). Essentials of Cross-Battery Assessment. (2nd Edition). New Jersey: John Wiley & Sons, Inc.

Greenspan, S.I. *Greenspan Social-Emotional Growth Chart: A Screening Questionnaire for Infants and Young Children*. San Antonio, TX: Harcourt Assessment, 2004.

James, Tracy. 2001. "Results of the Wave 1 Incentive Experiment in the 1996 Survey of Income and Program Participation." *Proceedings of the Section of Survey Research Methods,* 834-839. Alexandria, VA: American Statistical Association.

Hack, Maureen, H. Gerry Taylor, Dennis Drotar, Mark Schluchter, Lydia Cartar, Deanne Wilson-Costello, Nancy Klein, Harriet Friedman, Nori Mercuri-Minich and Mary Morrow. 2005. "Poor Predictive Validity of the Bayley Scales of Infant Development for Cognitive Function of Extremely Low Birth Weight Children at School Age." *Pediatrics* 118, 2: 333-341.

Lewis, Michael and Harry McGurk. 1972. “Evaluation of infant intelligence: Infant intelligence scores--true or false?” *Science* 178:1174-1177.

Lewis, Michael and Harry McGurk. 1973. “Testing infant intelligence.” *Science* 182:737.

Mack, Stephen, Vicki Huggins, Donald Keathley, and Mahdi Sundukchi. 1998. “Do Monetary Incentives Improve Response Rates in the Survey of Income and Program Participation?” *Proceedings of the Section on Survey Research Methods,* 529-534. Alexandria, VA: American Statistical Association.

Martin, Elizabeth, Denise Abreu, and Franklin Winters. 2001. “Money and Motive: Effects of Incentives on Panel Attrition in the Survey of Income and Program Participation.” *Journal of Official Statistics* 17: 267-284.

Matheny, Adam P.. 1973. “Testing Infant Intelligence.” *Science* 182: 734.

McCall, Robert B.. 1981. “Early Predictors of Later IQ: The Search Continues.” *Intelligence* 5, 2: 141-147.

McGuigan, William M., Aphra R. Katzev, and Clara C. Pratt. 2003. “Multi-Level Determinants of Retention in a Home-Visiting Child Abuse Prevention Program.” *Child Abuse & Neglect 27:* 363-380.

Michalopoulos, Charles, Anne Duggan, Virginia Knox, Jill H. Filene, Erika Lundquist, Emily K. Snell, Phaedra S. Corso, Justin B. Ingels, Sue Kim, and Magdalena Mello, 2011. ACF-OPRE Report 2011-16. *Design Options for the Home Visiting Evaluation: Draft Final Report.* Administration for Children and Families, U.S. Department of Health and Human Services, Washington, DC.

Nápoles-Springer AM, Santoyo-Olsson J, O'Brien H, Stewart AL. (2006). Using Cognitive Interviews to Develop Surveys in Diverse Populations. Med Care, 44(Suppl 3):S21-S30.

Tulsky, D., Zhu, J. & Ledbetter, M. (Eds.). WAIS-III WMS-III Technical Manual (Wechsler Adult Intelligence Scale & Wechsler Memory Scale. (1997). Harcourt Brace & Company.

Wechsler, D. (1997). *Wechsler Adult Intelligence Scale–Third Edition.* San Antonio, TX: The Psychological Corporation.

Willis, Gordon B. **2005**. Cognitive Interviewing: A Tool for Improving Questionnaire Design. Thousand Oaks, CA: Sage.

Wilson, Ronald S.. 1983. “Testing Infant Intelligence.” *Science* 182: 734-736.