



## APPENDIX 2

### I. Proposed Definitions Regarding Criteria for Evidence of Effectiveness

Nurse-Family Partnership offers the following definitions:

1. **“Significant, sustained, positive outcomes”** with respect to programs proven through randomized controlled research designs means statistically significant and meaningful health and developmental participant outcomes that are of sufficient magnitude and type as to positively impact child and parental conditions in the relevant benchmark area, and are sustained beyond the end of the home visitation program for a minimum of one year or until the child enters elementary school (depending upon the goals of the home visitation program). Evidence of sustainability comes from comparison of control and treatment groups utilizing appropriate statistical methods to account for missing data. Statistically significant positive outcomes must be demonstrated in multiple implementation sites, preferably across multiple research studies.

Areas of relevant outcomes include improvements in prenatal, maternal, newborn and child health (including prevention of child injuries and maltreatment), pregnancy outcomes, child development (including improvements in cognitive, language, social-emotional and physical developmental indicators), parenting skills, school readiness, child academic achievement, family economic self-sufficiency, and coordination of referrals for other community resources and supports as well as reductions in crime and domestic violence for home visitation program participants.

2. **“Significant, positive outcomes”** with respect to programs proven through quasi-experimental research designs means statistically significant and meaningful health and developmental participant outcomes that are of sufficient magnitude and type as to positively impact client conditions in the relevant benchmark area, as demonstrated through well-designed and implemented research studies reporting statistically significant positive effects.

Areas of relevant outcomes include improvements in prenatal, maternal, newborn and child health (including prevention of child injuries and maltreatment), pregnancy outcomes, child development (including improvements in cognitive, language, social-emotional and physical developmental indicators), parenting skills, school readiness, child academic achievement, family economic self-sufficiency, and coordination of referrals for other community resources and supports as well as reductions in crime and domestic violence for home visitation program participants. Statistically significant outcomes must be demonstrated across multiple research studies.

3. **“Well-designed and rigorous”** research designs include a well described theory for the impacts of the intervention, hypothesizing and testing outcomes of health and developmental significance, utilization of research methods that minimize bias (i.e., randomization of participants), demonstration of equivalency of comparison groups, minimizing attrition bias through utilizing an “intention to treat” approach in analyses,

demonstration of adequate statistical power through appropriate sample size, utilization of statistical techniques appropriate to the hypothesis in question, demonstration of generalizability (through multisite trials) and demonstration of sustainability of participant outcomes beyond a the end of the program.

## II. Discussion of Well-designed and Rigorous Randomized Controlled Trials and Quasi-Experimental Studies

### A. Rigor in the Conduct of Randomized Controlled Trials

Randomized controlled trials (RCTs) are considered the most rigorous research design for estimating intervention impact. This rigor is achieved because randomization of participants to intervention and control groups ensures that, within limits of probability; those groups are equivalent at the stage of randomization. Intervention – control differences observed after randomization are most likely due to intervention impact. Even RCTs, however, can produce questionable results if the following design features are not managed well:

1. **Post-Randomization Selection.** If all participants randomized to treatment conditions are not included in the calculation of program benefits, there is a possibility that participants in the intervention condition on whom follow-up assessments are completed will be different in some way from those in the control group, giving inaccurate estimates of intervention impact. One problem in the conduct of randomized controlled trials is for analyses to be conducted only on individuals who completed the program or met some threshold of intervention participation. This increases the likelihood that the program will appear to be more effective than it is. This problem is addressed by the conduct of “intention-to-treat” analyses, which call for analyzing all participants on whom post randomization assessments were conducted. If the sample has been stratified on the basis of baseline characteristics, it is acceptable to restrict the analysis within particular strata. Post-randomization selection is closely related to “attrition bias.”
2. **Attrition Bias.** One challenge with the conduct of randomized trials is retaining the sample. To the extent that the sample is reduced in size over time, statistical power is reduced, and questions can be raised about the extent to which the sample included at follow-ups is representative of that enrolled at baseline. This challenge is more problematic if those who drop out of the intervention group differ from those who drop out of the control group. Differential attrition by treatment condition is referred to as “attrition bias.” It can lead to the mistaken conclusion that the program is effective when it is not, or that the program is ineffective. Rigorous randomized controlled trials minimize attrition by instituting thorough-going retention strategies, setting in place methods to ensure equal follow-up by treatment condition, and examining the samples retained at various stages of follow-up to estimate potential attrition bias. Statistical controls may reduce attrition bias, but are unlikely to eliminate it if it is directly related to the intervention itself.
3. **Measuring Significant Health and Developmental Outcomes.** Rigorous trials choose outcomes of clear health and developmental significance, measured with the most valid and reliable methods available. So, if the outcome of interest is child language development, a design that measures directly the child’s language development is superior to one that measures language development on the basis of parent report (which is subject to reporting

bias). If one is interested in childhood injuries, measuring injuries from the review of medical records is superior to measurement by parent report. If the source of information is incomplete in characterizing the condition one wants to measure, such as child protective service records to measure child maltreatment (maltreatment is many fold more frequently occurring than is revealed in CPS records), use of such records can be misleading because of surveillance bias (increased detection of problems on the part of home-visited families). In these circumstances, it is best to have corroboration of findings from different sources of information.

4. **Adequate Statistical Power.** Studies need to have sufficient numbers of participants to detect program impacts of reasonable magnitude on outcomes of health and developmental significance. Samples that are too small increase the likelihood that apparent program impacts are due to chance. Sometimes samples are so large that program effects may be statistically significant but so small that they are of limited health and developmental value.
5. **Calculation of Important Health and Developmental Impacts.** In the calculation and comparison of program impacts, use of conventional effect-sizes expressed in standard deviation units can be misleading unless one has an appreciation for the health and developmental, social, or economic meaning of the outcomes. A large effect on an outcome of questionable health and developmental importance (say self-report of parenting) is likely to be much less important than a smaller impact in standard deviation units on childhood injuries abstracted from medical records, death, or cost to government.
6. **Replication.** Program effects found in one well conducted randomized controlled trial are highly promising, but need to be replicated in different trials and with different populations in order to have increased confidence in the validity of program impact and the extent to which findings apply to different populations.
7. **Generalizability.** One of the concerns about use of randomized trials for guiding public policy is that small trials conducted under well controlled circumstances and the auspices of program developers (“efficacy trials”) may lead to findings that have limited applicability to other vulnerable populations served in more typical health and human service systems. It is important, therefore, when evaluating the results of randomized controlled trials, to determine the extent to which the results apply to a range of health and human service delivery systems and to populations that are going to be served under the “Maternal, Infant, and Early Childhood Home Visiting Program. Sometimes, this will require evaluation of results across trials.
8. **Sustained Effects.** Intervention effects may decay over time. In order for a program to meet the legislation’s requirement of “sustained positive outcomes,” studies must have found sustained effects beyond the end of the program. The period over which effects must be sustained may differ depending upon the home visiting program’s goals and when it ends, but a strong case can be made for expecting program effects to be sustained for at least a year or until the children enter elementary school, given legislative intent that home visiting programs affect children’s school readiness, achievement, and long-term family functioning. Given that outcomes of interest change as children grow and as families move into different developmental phases, it is important that program impacts on benchmark outcomes be present at least one year after end of the program. In comparing home visiting programs, it

will be important to note the duration of sustained program effects.

9. **Examination of Conditional Effects.** Program effects found in randomized controlled trials may be particularly pronounced for some subgroups and not effective at all for others. Moderated effects are to be expected; their reliable detection is helpful from the perspectives of science, public policy, and program delivery. The challenge from a research design perspective is that subgroup effects can be spurious due to the conduct of many statistical tests. Some moderated program effects reported in the literature are thus likely to be misleading. Greater confidence can be placed in findings of moderated program impact when they are based upon levels of stratification factors employed in the randomization, when there is a plausible theoretical rationale for intervention moderation, when they moderation were hypothesized, and most importantly, when they are replicated across trials.

## **B. Well-Designed and Rigorous Quasi-Experimental Studies**

It is much harder for quasi-experimental studies to be considered rigorous compared to RCTs because a larger number of alternative explanations can account for apparent program effects derived from quasi-experimental evaluations. Quasi-experimental studies therefore must invoke creative design features to be considered well-designed and rigorous, and even creative quasi-experimental designs are almost always less rigorous than RCTs. When quasi-experimental and experimental studies exist for a single program, the experimental studies are more valid unless they are obviously flawed or unless exceptionally creative features have been introduced into the quasi-experimental studies. Estimates of intervention impact derived from quasi-experimental studies are usually larger than those derived from randomized trials, because the lack of random assignment in a quasi-experimental design allows biases to operate that tend to inflate the observed impact in ways that are not due to the intervention's effects. This means that investments in programs justified with findings from quasi-experimental studies are less likely to produce the results seen in those studies. A well-conducted randomized controlled trial eliminates those biases and produces a more realistic and conservative estimate of effects that is more likely to be achieved in practice, assuming excellent implementation of the program.

All of the design and analysis challenges described above under standards of evidence for rigorous randomized controlled trials apply to quasi-experimental studies, only they tend to occur much more frequently. The most common threats to the validity of estimates derived from quasi-experimental designs are well described in the literature. The most commonly occurring problem undermining the rigor of quasi-experimental studies of home visiting programs is selection bias, and under most circumstances leads to falsely concluding that an intervention works.

In many quasi-experimental evaluations, the calculation of program benefits is based upon those who complete the program rather than those who enroll, or on those who meet certain thresholds of program participation. This leads to the selection of program participants who are better functioning in contrast to those who do not participate or those who drop; comparisons of "intervention" participants to those in a control group thus lead to inflated estimates of program benefit. Such studies can be improved if they are based upon "intention to treat" principles, but even then they will not rule out problems with differential selection of families into the intervention and control conditions to begin with.

Even if studies control statistically for the measured background characteristics of those who enroll and those who do not, they almost never are able to rule out motivational factors that lead some parents to accept the offer to participate compared to those who decline. Given that such programs require investments of time and commitment on the part of the families, those who participate usually are more motivated than those who decline. In some cases, however, families may decline because they correctly calculate that they do not need the program, and in others mothers may enroll because they correctly understand that they need help. Rigorous quasi-experimental design will employ creative methodologies to rule out these participant motivational factors. One such approach is called regression discontinuity (RD) analysis.

Regression discontinuity studies can be used to reduce sources of selection when there is a clear boundary that separates those who qualify and those that don't, such as children's achievement eligibility scores or their age when there is a clear cut-off date for program entry. This creative quasi-experimental design has been used in evaluations of center-based preschool programs. It is less clear how RD might be used to evaluate home visiting programs given typical enrollment procedures and limited pre-enrollment data available for analysis of home visiting programs, but such analyses may provide one way of addressing participant selection biases in estimates of home visiting program impact in the future. I know of no home visiting intervention that has been studied with RD methods. RD methods require assumptions that make the conclusions less tenable than those derived from well-designed, rigorous RCTs.

Studies that rely upon "propensity matching" designs can be biased by participant enrollment procedures that lead either to the recruitment of more competent participants in the intervention group, or to the recruitment of those that are worse off than those not enrolled. Statistical controls for health or socio-demographic factors that distinguish program participants from controls almost never account for differences in enrollment procedures or participant motivation. A rigorous propensity matching design would employ a creative methodology to rule out these forms of selection bias.