B. COLLECTION OF INFORMATION USING STATISTICAL METHODS

B1. Sampling

The evaluation literature often discusses the appropriateness of the sample size for a study by focusing on the smallest program impacts that are likely to be detected with a specified level of confidence, assuming a sample of a given size and characteristics. These are usually called the program's "minimum detectable effects" (MDEs). Analysis of MDEs is also referred to as "power analysis," as it estimates the study's power to measure the effects it was designed to find.

Empirical information used. The next question to address is what sample sizes and configurations are required to achieve the preceding targets for precision (i.e., MDEs between 0.15 and 0.20 for student outcomes and, as presented below, between 0.30 and 0.60 for classroom or teacher outcomes) for the core study. To determine these requirements it was necessary to obtain empirical information on the variance components of the key outcome measures that will be used for the study. This information was obtained for roughly 150 student outcome measures and 50 classroom or teacher outcome measures both with and without statistical controls for baseline covariates (including pretests in most cases). These findings were computed using data from two national surveys of Head Start grantees/delegate agencies (i.e. Head Start Impact Study and FACES) and three group-randomized studies that focused on social and emotional outcomes for young children (two of which took place in Head Start centers). Results of these analyses were organized in an Excel spreadsheet program (the "EFFECT-O-SIZER") that was designed to facilitate real-time analyses of MDEs for alternative experimental designs given the full range of outcome measures in the relevant literature. A separate spreadsheet program was developed for student outcomes and classroom/teacher outcomes.

Precision of student outcomes. **Exhibit B1-1** (Appendix B) summarizes the likely MDEs of net impact estimates for a *single intervention* based on child-level outcome measures judged by the research team to be relevant for the present study. Exhibit B1-1 presents estimates of precision for sample sizes and the configuration needed in order to achieve our targeted precision for student outcomes, based on information from the EFFECT-O-SIZER, the parameters of which are listed in the top panel. This precision is based on a random-effects framework. All findings are based on the assumption that impacts will be estimated with regression adjustments for student baseline characteristics and a baseline measure of the outcome (pretest).

Exhibit B1-1 represents an experiment in which: 20 grantees/delegate agencies are sampled, 3 centers from each grantee/delegate agency (on average) are randomized to either a particular treatment group or the control group, there are 3 classrooms per center on average, and there are 8 four-year old students per classroom on average. For a pair-wise comparison of a single treatment group to the control group, randomizing 3 centers per grantee/delegate agency on average may be accomplished in a number of ways. Perhaps the best and simplest way is to randomize centers without replication in half of the grantees/delegate agencies (presumably the smaller ones) and randomize centers with replication in the other half of the grantees/delegate agencies (presumably the larger ones).

In addition to the design parameters shown in the column heading, the variance of true impacts across Head Start grantees/delegate agencies also is needed to compute minimum detectable effect sizes based on a random-effects framework. Because there are no readily available data on

this variance component, its value had to be assumed. The value was assumed to equal 0.01 for student outcomes and 0.0225 for classroom or teacher outcomes (in standardized effect size units squared).

The cells of Exhibit B1-1 show a simple mean of the MDEs estimates based on information from each study for which data were available. Some averages are based on numerous studies and others are based on only one study. The most frequent sources of this information are FACES 1997 and 2003—which is fortunate because they represent national samples. An "X" in a cell in Exhibit B1-1 indicates that the outcome measure was not collected by any of the five studies, that it was collected without a pretest, or that it is only available for a two-level model from one study.

The experimental design represented in Exhibit B1-1 provides MDEs near the upper end of our target range for child-level outcomes—that is, approximately 0.19 to 0.20 of a standard deviation.

Precision of classroom and teacher outcomes. Now consider our findings on precision for classroom or teacher outcomes. If the correlation between classroom or teacher outcomes and student outcomes is about 0.5 then to produce any given effect size on students requires producing an effect size of twice that on classrooms or students. So if our precision target range for student outcomes is about 0.15 to 0.20 standard deviation, this would imply a corresponding target range of about 0.30 to 0.40 standard deviation for classroom or teacher outcomes.

With this in mind, one option would be to specify sample design scenarios that produce MDEs between the 0.30 to 0.40 range for classroom or teacher outcomes. Another option, which we pursued, is to maintain the same design scenario specified in Exhibit B1-1 for student outcomes and present their MDES for classroom and teacher outcomes. **Exhibit B1-2** (Appendix B) presents these findings.

The layout of Exhibit B1-2 is identical to that of Exhibit B1-1. The column represents random-effects results for the same design scenario. Cell entries are the simple average of MDEs estimates for net impacts for each of the measures judged by the research team to be relevant for the present study. In most cases, only one or two estimates are available. (Information on classroom-level measures from FACES 2003 was not yet available.) As would be expected, the estimated MDEs for these classroom- and teacher-level outcome measures are larger than those for the child-level outcomes, equaling about 0.30 to 0.60 of a standard deviation.

Exhibit B1-3 (Appendix B) presents the total sample of the core study required by the three-treatment design for our design scenario. The top panel of the exhibit describes the scenario (for a single treatment group and control group). The first column of the exhibit represents a sample of 20 Head Start grantees/delegate agencies with 3 centers per grantee/delegate agency (on average) randomized to a single treatment group or a control group.[1] This scenario assumes 3 classrooms per center (on average) with 8 four-year-old students in each classroom (on average). The target minimum detectable effect sizes for student-level outcomes (using a random-effects

[1] [?] As noted earlier, this configuration comprises the average of a sample with half of its grantees randomizing two centers to the two experimental groups (one to each) and half randomizing four centers to the two experimental groups (two to each).

model) are 0.19. This scenario would require 20 grantees/delegate agencies with 120 centers, 360 classrooms and 2,880 four-year-old students.

As noted earlier, the three-treatment experimental design (with three treatment groups and a control group) will randomize Head Start centers within grantees/delegate agencies (blocks) and allocate the same number of centers to each treatment group and the control group (for a balanced design). The primary research questions addressed by the design will be:

- What are the *net impacts* of each intervention (treatment) on student-level outcomes relative to current Head Start practice for four-year old students (who are enrolled either in classrooms with only four-year olds or in mixed classrooms along with three-year-olds)?
- What are the *net impacts* of each intervention on classroom- or teacher-level outcomes relative to current Head Start practice?

Secondary research questions are:

- Overall, what are the *average net impacts* of the three interventions (combined) on student-level outcomes and classroom- or teacher-level outcomes relative to current Head Start practice?

The study's sample requirements are based on the desired precision of impact estimates that address its primary research questions. Thus they are driven by the desired precision of estimates of the net impacts of each treatment for four-year olds. This precision was summarized in Exhibit B1-1 for student outcomes and in Exhibit B1-2 for classroom or teacher outcomes.

Efficacy Study with 3-Year Olds

As described earlier in A1.2, we are planning to conduct an efficacy study on the three-year old children who will be in participating mixed-age classrooms. This component of the study will test the effects of social-emotional program enhancements pooled together (rather than the effect for any specific strategy) on outcomes for children.

We have little information to determine how many three-year old will be distributed across the classrooms, centers, and grantees that will be recruited for the core study. The challenge is that in a single grantee, we might have some centers (assigned to one of the treatment or control streams) with only four-year old classrooms. Because three-year olds may not be represented in all centers in that grantee, we would need to drop that grantee from any analysis of three-year olds. In larger grantees where we plan to test the effects across 8 centers (grantees with replication), our analysis may rely on only four rather than eight centers (in effect, losing replication in that grantee). In smaller grantees, we would have to eliminate a grantee completely with the loss of any one center in the analysis to maintain randomization.

By that same reasoning, we are also likely to have a lower number of classrooms within centers for an analysis of three- year old children (because some classrooms will serve only four-year olds and some will be mixed age). Again, we have little information in advance to inform our understanding about how many such classrooms would be appropriate for conducting our power analysis.

Both situations described here present a loss of power for this age group of children. This is why we have proposed assessing the effects for three-year olds by combining across program models, to test whether all of the social-emotional programs considered together, relative to a control condition, improves outcomes for classrooms and children.

Exhibit B1-4 (Appendix B) represents an add-on study that we think makes sense for the three-year olds embedded within the core study: 12 grantees/delegate agencies, 3 centers from each grantee/delegate agency (on average) are randomized to either a particular treatment group or the control group (assuming no replication), there are 2 classrooms per center on average, and there are 8 three-year old students per classroom on average.

In addition to the design parameters shown in the column heading, the variance of true impacts across Head Start grantees/delegate agencies also is needed to compute minimum detectable effect sizes based on a random-effects framework. Because there are no readily available data on this variance component, its value had to be assumed. The value was assumed to equal 0.01 for student outcomes and 0.0225 for classroom or teacher outcomes (in standardized effect size units squared).

The cells of Exhibit B1-4 show the range of MDEs for priority student outcomes from the FACES study (which included three-year olds) for our proposed unbalanced random assignment design. As expected, power is lower than the design for four-year olds. MDEs range from 0.24 to 0.38 of a standard deviation. However, power is consistent with the goal of this add-on component to be an efficacy trial for this age group of children, and is similar to the early efficacy trials conducted on four-year-old children that informed our core study here.

Exhibit B1-5 (Appendix B) presents our best estimate of the total sample of three-year olds for this embedded efficacy study. The top panel of the exhibit describes the scenario (for a pooled treatment group and control group). The first column of the exhibit represents a sample of 12 Head Start grantees/delegate agencies with 3 centers per grantee/delegate agency randomized to a single treatment group or a control group. This scenario assumes 2 classrooms per center (on average) with 8 three-year old students in each classroom (on average).

Overall Sample Requirements

Thus, the survey sample size for the baseline and follow-up lead teacher self-report surveys will be 360 respondents. The sample size for the teacher report on individual children, the parent survey will be 2,880 four-year old children and 768 three-year old children. Direct child assessment will only be completed with the 2,880 four-year old children. The trainer survey will be completed for 60 local coaches. Finally, it is estimated that implementation site visit interviews will be conducted with 60 local coaches, 360 lead and assistant teachers, 60 center directors, 180 center staff, 20 grantee/delegate agency directors, and, at most, 60 trainers. These sample estimates are based on the assumption that 80 percent of the research sample will be successfully interviewed. Sample sizes are depicted in **Exhibit B1-6** (Appendix B).

B2. Procedures for Collection of Information

The impact survey data and direct child assessments will be collected primarily through in-

person paper and pencil surveys, although in-person outreach and interviewing strategies will be used to maximize response rates. In some cases, surveys will be conducted over the phone or as mail-backs. MDRC will work with SRM to develop strategies that ensure an 80 percent response rate. All completed surveys will be reviewed to ensure all applicable fields are correctly completed and that all relevant interviewer notes are included in the data set. Any open ended and "other, please specify" items will be coded based on codes developed at SRM and approved by MDRC. Preliminary data files will be created and shared – with documentation – with MDRC on an agreed-upon schedule.

B2.1 Procedures for the surveys

Interviewer Selection. MDRC will work with SRM to ensure that the interviewers administering these surveys and assessments are professional interviewers, many of whom have worked on social research projects. Preference will be given to those who are multilingual, depending on the languages spoken by the research samples. Familiarity with the special requirements of interviewing low-income populations and children will be desirable. New personnel will be trained along with the seasoned interviewers.

Interviewer Training. MDRC will work with SRM to ensure sufficient interviewer training. In the past, this has typically involved 3-day trainings. For example, personnel who are new to interviewing were trained in general interviewing techniques and approaches in the first day of the session. Professional interviewers will be trained with the new recruits on project-specific material in the remaining 2 days. Some pre-training exercises are likely to be required, and the actual training will include an item-by-item review of the survey instruments, and practice interviews and critiques of those interviews. Direct assessments will first be practiced through role-playing, and later with actual children prior to entering the field.
Training will take place close to the time when the baseline assessments of Cohort 1 and Cohort 2 will begin. A booster training will also be conducted prior to the follow-up data collection for both cohorts.

All interviewers will sign a confidentiality pledge during training. They will be instructed on the importance of maintaining confidentiality and told that breaches of confidentiality will lead to dismissal.

MDRC will also work with SRM to establish procedures for monitoring early interviews for each interviewer (e.g., videotaping interviews), periodically monitoring interviews throughout the course of data collection, and procedures for offering feedback.

Conducting Interviews. In all cases, the interviewers will explain the purpose of the interview, and inform respondents that they will receive a small incentive for completing the survey. Each interviewer will be prepared to answer any questions about the study that sample members might have.

Interviewer Supervision. Interviewing field staff will be supervised directly by staff from SRM.

B3. Maximizing Response Rates

The goal will be to achieve an 80 percent response rate in each site. Procedures for obtaining the

maximum degree of cooperation include:

- Conveying the purposes of the survey to respondents so they will thoroughly understand the purposes of the survey and perceive that cooperating is worthwhile;
- Providing a toll-free number for respondents to use to ask questions about the survey and the survey firm's staff;
- Training site staff to be encouraging and supportive, and to provide assistance to respondents as needed;
- Hiring interviewers who have necessary skills for encouraging respondent cooperation;
- Training interviewers to maintain one-on-one personal rapport with respondent; and
- Offering appropriate payments to respondents.

Interviewers will also be trained to distinguish "soft" refusals from "hard" ones. Soft refusals often occur when the sample member has been reached at an inopportune time. In these cases, it is important to back off gracefully and to establish a convenient time to call or come back rather than to persist at the moment. Hard refusals do occur and must also be accepted gracefully by the interviewer.

B4. Pre-testing

Most of the questions proposed for this survey are either identical to questions used in prior MDRC evaluations or are similar, if not identical, to questions used in previous national surveys or major evaluations. Consequently, many of the items have been thoroughly tested on larger samples.

The CARES surveys have already undergone a number of revisions, following critiques by internal staff, by project consultants, and by staff at HHS. MDRC will also work closely with SRM's senior staff to conduct formal pre-tests of these surveys, using nine sample members with ample contact information to complete each survey in person.

B5. Consultants on Statistical Aspects of the Design

We consulted with an additional set of individuals outside of MDRC, in addition to Howard Bloom of MDRC (who is a lead member of the CARES project team), on the statistical aspects of the design and sampling, including: Carolyn Hill (Georgetown University); Stephanie Jones (Harvard University); Robert Nix (Pennsylvania State University); Mark Lipsey (Vanderbilt University); Stephen Raudenbush (University of Chicago); Tom Cook (Northwestern University); Jeff Smith (University of Michigan); Hendricks Brown (University of South Florida); Larry Hedges (Northwestern University).