Design Options for an Evaluation of Head Start Coaching

Design Report

OPRE Report #2014-82

July 2014

Head Start Professional Development: Design Options and Considerations for an Evaluation of Head Start Coaching

Design Report

July 29, 2014

Office of Planning, Research and Evaluation
Administration for Children and Families
U.S. Department of Health and Human Services
http://www.acf.hhs.gov/programs/opre
Wendy DeCourcey, Project Officer
Christine Fortunato, Project Specialist

American Institutes for Research
Eboni Howard, Project Director
Kathryn Drummond, Project Manager
Jonathan Farber
James Taylor

MDRC
Barbara Goldman
Marie-Andree Somers
Chrishana Lloyd

MEF Associates
Mike Fishman
Jessica Wille

Child Trends
Kathryn Tout

OPRE Report #2014-82

Suggested Citation
American Institutes for Research [AIR], MDRC, MEF Associates, and Child Trends (2014). *Head Start Professional Development: Design Options and Considerations for an Evaluation of Head Start Coaching*, E.C. Howard & K.V. Drummond (Eds.). Washington, DC: U.S. Department of Health and Human Services, Administration for Children and Families, Office of Planning, Research and Evaluation.

Disclaimer
This report and the findings within were prepared under Contract #HHSP23320095626W with the Administration for Children and Families, U.S. Department of Health and Human Services. The views expressed in this publication are those of the authors and do not necessarily reflect the views or policies of the Office of Planning, Research and Evaluation, the Administration for Children and Families, or the U.S. Department of Health and Human Services. This report and other reports are available from the Office of Planning, Research and Evaluation.

Contents

Acknowledgments

The completion of the *Head Start Professional Development: Developing the Evidence for Best Practices in Coaching* task order resulted from significant contributions and dedicated efforts of many people. This report represents a collaborative effort from researchers, early childhood care experts, and Head Start practitioners.

We are grateful to many research and administration staff at American Institutes for Research (AIR), MDRC, MEF Associates, and Child Trends who helped in numerous ways throughout the project, including quality review, editing, production, contract management, financial management, and information technology. We are particularly grateful for the multiple reviews, advice, and insights from Hans Bos (AIR), Mike Garet (AIR), Anja Kurki (AIR), Mengli Song (AIR), Howard Bloom (MDRC), JoAnn Hsueh (MDRC), Virginia Knox (MDRC), Shira Mattera (MDRC), Charles Michalopoulos (MDRC), Pamela Morris (MDRC), Michael Weiss (MDRC), and Marty Zaslow (Child Trends). Jessica Johnson (AIR) provided additional management support. Margaret Soli (Child Trends) and Eliza Brown (Child Trends) provided research support.

We also thank the expert consultants who served on the project team to provide insights about what is known about coaching and early childhood – Lindsey Allard Agnamba from School Readiness Consulting, Gaysha Beard from University of Delaware, Bridget Hamre from University of Virginia, Mary Louise Hemmeter from Vanderbilt University, Susan Landry from University of Texas, Douglass Powell from Purdue University, and Samantha Wulfsohn from New York Center for Child Development. In addition, Carol Vukelich from University of Delaware, Barbara Wasik from Temple University, and Betty Zan from University of Northern Iowa provided feedback on our review of coaching. We are deeply appreciative for the support that Linda Collins from Penn State provided throughout the project to devise, write, and review the extensive methodological approach. Her expertise on the multiphase optimization strategy framework was invaluable.

We greatly appreciate the leadership and availability of staff from the Office of Head Start who participated in Office of Planning, Research and Evaluation (OPRE) planning meetings and provided additional input along the way. We also appreciate the advice and information from the Head Start Center on Quality Teaching and Learning.

We also extend a very special thank-you to the Head Start grantee administrators, coaches, staff, and technical assistance providers who participated in stakeholder conversations. Their responsiveness and availability to provide reactions and feedback was exceptional and greatly appreciated. Without their gracious participation, the informative findings of this task order would not have been possible.

This report was completed in collaboration with Wendy DeCourcey and Christine Fortunato from OPRE. We are grateful for their genuine interest in and passion for this project and for always being available and willing to help it succeed. Jennifer Brooks provided supportive guidance and valuable insight throughout the entire project. We also thank their colleague Mary Bruce Webb for her review of the project's activities and reports.

Overview

There is a growing consensus in the early childhood education field that the provision of targeted high-quality professional development (PD) shows promise for improving teachers' practices classroom quality, and child outcomes. Coaching is a recommended PD practice which is increasingly widespread. Although the available evidence generally supports the positive effects of coaching overall, there are significant challenges with interpreting the evidence for the best combination of coaching practices.

The purpose of this report is to present design options for a study that will further investigate evidence for effective and efficient coaching practices. In particular, the proposed *Head Start Coaching Study* (hereafter called the *HS Coaching Study*) will evaluate specific dimensions of coaching that may impact teacher and classroom practices in Head Start (HS) and other early childhood settings. The resulting study of coaching intends to accomplish three goals:

1. Provide strong evidence for effective and efficient coaching practices of center-based teachers of three- to five-year-olds in HS programs.
2. Help HS programs make informed decisions about the allocation of PD resources when designing, implementing, and improving coaching programs.
3. Advance empirical knowledge about coaching within early childhood settings and set the stage for additional research about coaching as a PD strategy.

This report provides recommendations for the following aspects of the HS Coaching Study:

- The purpose of the study and the research questions
- The study design for testing the impact of coaching
- The implementation research component of the study
- The cost component of the study
- The measures
- The important logistical issues for this study

The report also provides information about the content of the coaching intervention and offers recommendations for selecting a PD developer to help implement the intervention. Descriptions of the process, criteria, and guiding principles are used throughout the report to support the design recommendations for the study.

The recommended HS Coaching Study was designed to help inform HS programs decisions about the allocation of their PD resources when developing and implementing coaching. In addition, the study aims to advance the research evidence about to what degree, dimensions of coaching, impact teacher practices, classroom quality, and child outcomes. Ideally, results from the HS Coaching Study will help in designing an optimal coaching intervention that will be the focus of additional research.

Executive Summary

The purpose of this report is to present design options for a study of the effectiveness of different coaching dimensions in Head Start (HS) programs. This design project was funded by the U.S. Department of Health and Human Services (HHS), Administration for Children and Families (ACF), Office of Planning, Research and Evaluation (OPRE).

Under the task order, *Head Start Professional Development: Developing the Evidence for Best Practices in Coaching*, a design team was formed of four research organizations (American Institutes for Research [AIR], MDRC, MEF Associates, and Child Trends), which developed the design options presented here with input from consultants and practitioners in the HS field. The resulting study of coaching intends to:

- Provide strong evidence for effective and efficient coaching practices of center-based teachers of three- to five-year-olds in HS programs.
- Help HS programs make informed decisions about the allocation of professional development (PD) resources when designing, implementing, and improving coaching programs.
- Advance the state of empirical knowledge about coaching within typical early childhood settings and set the stage for additional future research about coaching as a professional development strategy.

The work of the design task order included (1) examining the conceptual and theoretical frameworks for coaching in early childhood education settings, (2) determining the best methodology for rigorously evaluating the effectiveness of coaching dimensions, and (3) designing a study (hereafter called the *HS Coaching Study*) to evaluate specific dimensions of coaching that may impact teacher and classroom practices in HS and other early childhood settings. A dimension refers to a singular aspect or component of a coaching program (e.g., coach characteristics, type of coaching activity, dosage); the study will examine the effect of *varying the levels* of coaching dimensions.

This report provides recommendations for the following aspects of the HS Coaching Study:

- The purpose of the study
- The research questions
- The study design for testing the impact of coaching, including the following:
    - Application of the multiphase optimization strategy (MOST) framework
    - Systematic evaluation of three dimensions of coaching (dosage of coaching, recipient of coaching, and amount of coach training)
    - Use of a factorial design
    - Requirements for detecting effects and sample size
- The implementation research component of the study
- The cost component of the study
- The measures

- The important logistical issues for this study, such as participant recruitment, participant selection, the implementation monitoring, and the technical assistance that may be required

The report also provides information about the content of the coaching intervention and the standardized foundational coaching approach for the study. Although some approaches to coaching do not specify a particular content domain on which teachers and coaches will concentrate, we suggest that the goals of this study will be better met, and outcomes more precisely measured, by using a coaching approach with a specific content focus. After considering a number of content areas geared towards supporting various domains of early childhood development, we recommended that the HS Coaching Study focus on language development and the interactions between children and teachers that support that development. Language development is a critical domain of early child development, a well-established precursor to subsequent literacy skills that grow increasingly important as children approach entry to elementary school. It is one of the 11 domains within the HS Child Development and Early Learning Framework.

Descriptions of the process, criteria, and guiding principles are used throughout the report to support the design recommendations for the study. To help in planning for the HS Coaching Study, we provide estimates of the resources needed to conduct this study, suggested task structure, and a study timeline.

The Purpose of the HS Coaching Study

There is a growing consensus in the early childhood education (ECE) field that the provision of targeted high-quality professional development shows promise for improving teachers' practices, classroom quality, and child outcomes (Diamond & Powell, 2011; Dickinson & McCabe, 2001; Snyder, Hemmeter, & McLaughlin, 2011). Coaching is a recommended PD practice that is increasingly widespread. Although the available evidence generally supports the positive effects of coaching overall, there are significant challenges with interpreting the evidence for the effectiveness of coaching components.

Two extant literature reviews on coaching in ECE (Aikens & Akers, 2011 and Isner et al., 2011) noted a number of limitations. Most importantly, many studies did not provide detailed specifications about the coaching in their interventions. Overall, key limitations to extant coaching research are:

- Coaching is usually examined in combination with additional PD strategies; coaching is part of effective PD packages and is seldom studied on its own.
- Descriptions of coaching features (e.g., structure, process, and staffing aspects of coaching programs) lack sufficient detail.
- The most effective coaching actions and behaviors have not been identified through experimental methods. Coaching features are not examined separately in the extant literature. Little empirical support has been presented for the value from adding certain coaching strategies as part of a PD program (e.g., adding training for coaches).

- Few coaching studies have systematically examined the effectiveness of variations of coaching dimensions (e.g., how much training for coaches is most effective?).

There is a traditional PD paradigm for many evaluations—testing whole interventions rather than individual dimensions. In most evaluations of coaching, coaching *content* may be bundled, or combined, with delivery in a particular *format*, bundled with a particular *dosage* of the intervention, which is further bundled with delivery to a particular *recipient*. This combination of coaching features may then be combined with additional curriculum training and materials provided to teachers in a PD package. However, it leaves evaluators, policymakers, and program developers with an intervention "black box," for which it is hard to understand which individual dimensions influence outcomes.

The design for the HS Coaching Study aims to strengthen the research by evaluating coaching, as a stand-alone professional development component in the HS context and to examine the differential effects of several specific dimensions of coaching.

The Guiding Research Questions

Six research questions guided the design of the HS Coaching Study, two related to the impact of coaching dimensions, three related to the implementation of coaching dimensions, and one related to cost.

The key questions related to *impact* of the coaching dimensions are:

1. What is the effect of specific dimensions of coaching on teacher practices and classroom quality in HS programs?
2. Does the effect of one coaching dimension depend on the level of another coaching dimension?

The research questions related to the *implementation* of the coaching dimensions are:

3. Are the different coaching dimensions implemented with fidelity?[1]
4. What factors facilitate or challenge the fidelity of implementation of the different coaching variations?
5. How does implementation vary across grantees' program environments, populations, and other contextual program features?

The research question related to *cost* is:

6. What is the cost of implementing the different coaching dimension variations?

[1] Fidelity here refers to implementation of the coaching dimensions as designed. There may be several aspects of fidelity that may be of interest, including adherence, exposure, responsiveness, and quality.

Impact Component of the HS Coaching Study

The MOST Approach

As part of the task order, the design team wrote a review that outlined different possible design and methodology framework options for the study (Somers, Collins, Maier, 2013; http://www.acf.hhs.gov/programs/opre/research/project/head-start-coaching-study-design-phase). After reviewing a range of research methods for testing the effectiveness of coaching, the design team and OPRE staff members decided that the design for the HS Coaching Study should reflect the principles of the multiphase optimization strategy (MOST; Collins et al., 2005; 2009; in press).

The MOST framework is a staged and rigorous approach to developing and evaluating interventions.

- After a preparation phrase, an Optimization Phase is conducted, in which the relative effect of different intervention dimensions are assessed in a randomized screening experiment. Dimensions are selected for testing by examining the evidence base or, if the evidence base is weak, using strong theoretical support or recommendations from experienced practitioners and researchers.
- The results of this screening experiment are then used to build an optimal intervention model consisting of the selected dimensions that meet some minimum threshold for effect size, cost-effectiveness, and practical or theoretical importance.

In a second phase, the impact of this optimal model is evaluated in a standard two-group randomized experiment. The HS Coaching Study corresponds to the Optimization Phase of the MOST framework.

Systematic Evaluation of Coaching Dimensions

With the MOST approach as a guiding framework, we recommend that the HS Coaching Study examine the effect of three individual coaching dimensions:

(1) The amount or dosage of coaching (Dosage);

(2) The recipient of the coaching (Recipient; lead teacher only vs. teaching team); and

(3) The amount of coach training (Coach Training) or Delivery Mode (Mode; technologically-mediated vs. onsite)

Strictly speaking, we recommend that the study examine the effect of *varying the levels* of each of these coaching dimensions. For example, for Dosage, we suggest testing outcomes of having coaches meet with teachers on a bi-weekly vs. monthly basis.

Factorial Design for the Impact Study

To examine the dimensions, we suggest that a factorial design is the most suitable design for testing the effect of the three coaching dimensions. A factorial design is an experimental design in which the experimental conditions represent all possible combinations of the levels of the dimensions under investigation. Factorial experiments are well suited for building strong interventions in the Optimization Phase of the MOST framework (e.g., Collins et al., 2005; in press). Specifically, for three coaching dimensions, we recommend a factorial design with three factors and eight experimental conditions, as the table below illustrates.

Recommended 2^3 Factorial Design

Experimental Condition Number	Factors		
	Amount of Coaching (*DOSAGE*)	**Recipient of the Coaching (RECIPIENT)**	**Amount of Coach Training (TRAINING)***
1	Monthly	Lead teacher only	Orientation
2	Monthly	Lead teacher only	Ongoing
3	Monthly	Teaching team	Orientation
4	Monthly	Teaching team	Ongoing
5	Biweekly	Lead teacher only	Orientation
6	Biweekly	Lead teacher only	Ongoing
7	Biweekly	Teaching team	Orientation
8	Biweekly	Teaching team	Ongoing

Note. Unshaded cells represent the typical level (Level I) of the factor; shading denotes the enhanced level (Level II) of the factor.
*Or Mode, in which case the levels in the design would be remote coaching (in the unshaded cells) and in-person coaching (in the shaded cells).

Although factorial designs require more experimental conditions than other designs, a benefit is that they require a smaller sample size than other designs to statistically detect a dimension's effect of given magnitude. Another potential benefit of factorial designs is that they also account for—and provide information on—interaction effects between the dimensions that are being tested in the study. Thus, factorial designs make it possible to efficiently determine which particular components of an intervention are more important, as well as examine how these components interact with each other to produce the desired outcomes. For these reasons, factorial experiments provide findings that are useful for policymakers and practitioners who are creating or adapting interventions.

Minimum Detectable Effect Size and Sample Size

The report provides a full explanation of the power and sample plan for the HS Coaching study. The *minimum detectable effect size* (MDES) is a useful concept for making decisions about the sample size. Formally, MDES is the smallest true effect on the outcome of interest (scaled as an effect size) that can be detected with a reasonable degree of power. The recommendation is that the HS Coaching Study be able to detect a main effect on teacher and classroom outcomes of

0.20. The recommendation has two justifications. First, it seems reasonable to expect that the coaching dimensions in the study would have main effects of this size on teacher practices. Based on prior research, an additional 1.5 hours of coaching per month (which is the one of the variations that will be tested in the HS Coaching Study) could improve teacher practices by an effect size of about 0.09 to 0.26, with effects expected to larger for practices that teachers used less frequently at baseline.[2] Thus, it is reasonable to expect that the dimensions under study could have a main effect of 0.20 on teacher practices that are in greatest need of improvement. Second, it is probable that an effect size of 0.20 on teacher practices can also translate into a meaningful change in children's literacy-related outcomes. Even though child outcomes will not be measured in the HS Coaching Study, improving children's outcomes is one of the goals of coaching. An effect size of 0.20 on teacher and classroom outcomes translates into an effect of approximately 1.4 to 2.5 weeks of extra learning for children, or a 5 to 10 percent increase in children's literacy skills above and beyond what they would normally learn during the school year.[3]

We estimate that in the proposed factorial design approximately 248 centers across 31 HS grantees will be needed to detect an effect size of .20 if random assignment occurs at the center level. However, the final sample size will depend on (a) final decisions that OPRE and the study evaluation team make about the specifications of random assignment (whether dimensions are assigned at the coach level, and/or whether a HS grantee would allow the evaluators to randomly assign coaches to centers for the purposes of the study) and (b) how many classes exist per center for participating sites.

Implementation Component of the HS Coaching Study

Implementation research helps document the extent to which the intervention was implemented as intended. Implementation research identifies factors that may facilitate and challenge execution of the intervention that further contextualize the resulting impacts. For the HS Coaching Study, we recommend the following goals:

(1) To describe and assess the fidelity of implementation for the eight experimental coaching conditions in order to help interpret impacts.

(2) To inform future development of effective and feasible coaching models.

Documenting the foundational coaching model (including the implementation of the language content of the coaching) and the three systematically varied dimensions will be important to understand fidelity (i.e., the extent to which the coaches and teachers implement the levels of the targeted three dimensions—Dosage, Recipient, and Coaching Training—to which they were assigned) and the extent to which coaches and teachers adhere to the dimensions that are fixed and the natural variation across the teachers and coaches for other dimensions.

[2] This is based on a study conducted by Landry and her colleagues (2009), which found that four additional hours of coaching per month can improve teacher practices by an effect size of 0.23 to 0.70.
[3] Estimates of annual effect size gains are based on data from the CARES study (Mattera, Lloyd, Fishman, & Bangser, 2013).

Cost Component of the HS Coaching Study

If the evaluation team learns that particular coaching dimensions are effective, the total resources required to implement these dimensions will be important information for both planners within OHS and HS program directors. The cost aspect of the HS Coaching Study aims to accomplish the following goals:

(1) Provide information to HS grantees about the types of resources needed to develop and implement the targeted coaching dimensions within their programs.

(2) Gather information that can be used in a cost-effectiveness analysis.

Conducting this analysis would allow the evaluation team to determine the relative cost-effectiveness of each coaching dimension condition by comparing the financial resources required to implement a given level of a coaching dimension (e.g., low dosage of coaching or enhanced coach training) and its estimated effectiveness (effect size) when considered across all other dimension levels.

Measurement

The measurement approach, was designed to maximize study feasibility (conducting the study within the timeline and minimizing burden on participants) while simultaneously documenting the details and context of coaching with the necessary richness and specificity to answer the research questions.

Key constructs for the study were identified based on the research questions. We then provide details for specific recommended data collection tools for the impact, implementation, and cost research, including what they measure and their format, frequency, and specifications. Most suggested data collection tools serve multiple purposes in the HS Coaching Study. The measurement strategy is not simple. However, it is important to collect data with multiple respondents and at multiple levels to understand the complex practices that are part of the HS Coaching Study. We suggest six categories of data collection tools in addition to requesting program budgets. These are listed below:

(1) Implementation Contact, Time, and Attendance Logs
Participants: Coaches, PD trainers, teachers (using time sampling)
Purpose: Document and monitor attendance and details of coaching sessions, coach training, and teacher training

(2) Implementation Rating Logs
Participants: Coaches, PD trainers
Purpose: Document (a) coaches' report on utility and value of coach training; (b) coaches' and teachers' reports on utility and value of teacher training; (c) coaches' and teachers' reports on utility and value of coaching sessions; and (d) coaches' and trainers' reports of teachers using targeted strategies

(3) Participant Surveys
Participants: Center directors, coaches, teachers, PD trainers

Purpose: Gather data about participant characteristics, experiences, and perceptions of coaching

(4) Participant Interviews
Participants: Center directors, grantee liaisons; sample of coaches, teachers, PD trainers
Purpose: Gather data about how coaching was implemented, factors that facilitated or hindered implementation and fidelity

(5) Observations of Coaching Sessions and Coach Training
Participants: Coaches, teachers
Purpose: Assess key qualitative features of the coaching sessions

(6) Observations of Teacher Practices and Classroom Environment
Participants: Teachers
Purpose: Gather impact data about (a) classroom quality and (b) the specific language and teacher-child interaction practices that are targeted by the coaching

Conducting the Study

Relevance for the field

Aiming to design a study that is as relevant and compelling as possible for the HS field, as well as logistically feasible, as part of the design process we consulted with a limited number of stakeholders at OHS and in the HS practitioners. We spoke to stakeholders about either their experience with coaching programs or their opinion about coaching in general or in the context of the planned study. Feedback was gathered through individual calls, group webinar-format calls, and at an interactive conference presentation.

Logistical Issues

Problems related to implementation of the foundational coaching model and the eight coaching conditions could inevitably arise in a complex study in up to 31 grantees and 248 centers. Therefore, we recommend carefully explaining the study to potential participants, monitoring implementation, and providing assistance as necessary. The logistical issues examined by the coaching team include:

- Recommendations for participant recruitment and selection, including establishing partnership with OHS and HS grantees and consideration of funding the coaching efforts for each participating grantee.

- Monitoring of implementation and technical assistance, including establishing clear expectations, assigning an onsite liaison with each participating grantee to facilitate communication; and structured technical assistance. .

Conclusion

Using the first phase of the MOST framework to guide the HS Coaching Study design will allow for the systematic testing of the impact of coaching dimensions when controlling for all other variations studied. Certainly, adapting the MOST model to the complexities of the HS coaching interventions is not easy. However, the answers to the research questions for the HS Coaching

Study related to coaching impact, implementation, and costs will play an important role in informing HS programs decisions about the allocation of their PD resources when developing and implementing coaching approaches. In addition, the answers to these proposed research questions will advance the research evidence about coaching in early childhood settings. Ideally, results from the HS Coaching Study will help in designing an optimal coaching intervention that will be the focus of additional research.

I. Introduction

There is a growing consensus in the early childhood education (ECE) field that the provision of targeted, high-quality professional development shows promise for improving teachers' practices and, ultimately, child outcomes (Diamond & Powell, 2011; Dickinson & McCabe, 2001; Snyder, Hemmeter, & McLaughlin, 2011). Professional development (PD) refers to ongoing learning opportunities for the acquisition of skills and knowledge. Coaching, as a particular type of PD, is a recommended practice, which is increasingly widespread. The Office of Head Start (OHS) suggests that coaching is a flexible tool that can enhance teaching practices (Herren, 2009; National Center on Quality Teaching and Learning [NCQTL], 2012). Head Start (HS) coaching is one of the features of HS PD systems that also include national technical assistance (TA) centers, national centers to translate research to practice, and local PD resources.

Coaching in HS programs is usually supported by local grantee program funds, and as a result of this localized funding, its design and implementation vary widely in both form and content. Among HS grantees, coaches have many different roles and use different formats for interacting with staff. For example, in the Early Learning Mentor Coach (ELMC) descriptive study supported by the Office of Planning, Research and Evaluation (OPRE) of 121 HS grantees that received special funding to provide coaching, coaches were surveyed about their role. The 384 coaches, paid with ELMC funds, reported that they fulfilled an average of four roles (Howard et al., 2013). In terms of formats for interacting with staff, coaches may work with staff face-to-face (most common) or at a distance, using computer and video technology or via telephone (NCQTL, 2012; Howard et al., 2013).

Because coaching varies so readily from program to program, coaching details are defined in different ways in the ECE PD literature (Sheridan, Edwards, Marvin, & Knoche, 2009; Zaslow & Martinez-Beck, 2006). However, coaching is consistently distinguished from other forms of PD, such as formal coursework or group workshops, in that it incorporates an ongoing, individualized element that other PD lacks. The coach and teacher have regular interactions in which the teacher is advised and mentored by the coach. These interactions are intended to improve teacher practice and classroom quality—and, ultimately, child outcomes.

As coaching grows increasingly common in ECE, there is increasing interest in expanding coaching programs with guidance from empirical evidence about what are the most effective coaching practices to implement. Questions abound in the ECE field about what type of coaching roles, staff interactions, and other coaching elements are most effective for improving teacher practices and classroom outcomes; what the most rigorous evaluation approach to provide information about effectiveness is; and what information can inform decisions about the allocation of HS programmatic PD resources when developing and implementing coaching interventions.

Purpose of This Report

The purpose of this report is to present design options to examine select coaching components in HS programs that can address questions related to understanding the effectiveness of different coaching elements. The study design effort was funded by the U.S. Department of Health and

Human Services (HHS), Administration for Children and Families (ACF), and OPRE and was conducted in collaboration with OHS.

Under the task order *Head Start Professional Development: Developing the Evidence for Best Practices in Coaching*, a design team was formed of research organizations (i.e., American Institutes for Research [AIR], MDRC, MEF Associates, and Child Trends), with input from consultants and practitioners in the HS field. The project aims to develop design options for a study of coaching that will:

- Provide strong evidence for effective and efficient coaching practices of center-based teachers of three- to five-year-olds in HS programs
- Help HS programs make informed decisions about the allocation of PD resources when designing, implementing, and improving coaching programs
- Advance the state of empirical knowledge about coaching within typical early childhood settings and set the stage for additional future research about coaching as a PD strategy.

The work of the design task order included (1) examining frameworks about coaching, (2) determining the best methodologies for rigorously evaluating the effectiveness of coaching dimensions, and (3) designing a study (hereafter called the *HS Coaching Study*) to evaluate specific dimensions of coaching that may impact teacher and classroom practices in HS and other early childhood settings.

Organization of Report

This report begins by providing background about what is known about the effectiveness of coaching, the recommended evaluation approach for testing effectiveness for the HS Coaching Study, and the key research questions that frame the HS Coaching Study. In Section III, we describe the process and criteria used to choose the dimensions of coaching to investigate, and we describe the selected dimensions in depth. We also recommend aspects of the coaching intervention that will form the standardized, foundational approach for all study participants.

Section IV provides an overview of the recommended evaluation design, including guiding principles for the choice of that design. In the two subsequent sections, we present the study methodology in more detail. Section V presents the random assignment plan and sample size requirements. Section VI presents approaches to the impact analyses under particular models. In Section VII, we suggest two supplemental lines of research that can complement the main impact analyses; one focuses on an implementation study and the other on a cost study. Section VIII lays out key measures, data sources, and data collection methods. The last three sections focus on logistical issues critical for conducting the study. Section IX provides information on recruiting and selecting HS centers for the study. In Section X, we make suggestions for monitoring implementation to ensure fidelity of the intervention and the TA that may be required. Section XI provides resource estimates and lays out a suggested task structure and timeline for the study.

II. Evaluating Coaching Effectiveness

In this section, we discuss in brief the research that has examined coaching effectiveness in ECE. By examining two extensive literature reviews, along with other extant empirical research, we discuss the documented impact of coaching on teacher practice, classroom quality, and child outcomes. We then discuss some of the limitations of the available research. The section concludes with recommendations for the HS Coaching Study – including a suggested method for testing the effectiveness of coaching, the definition of coaching for this evaluation, and the six research questions.

What Is Known About Coaching Effectiveness?

Two systematic literature reviews completed in 2011 summarized extant evidence on ECE coaching (Aikens & Akers, 2011; Isner, et al., 2011). Both reviews included coaching studies that used experimental, quasi-experimental, and pre- and posttest designs, as well as descriptive studies. Isner et al. (2011) reviewed 44 quantitative coaching studies conducted between 1994 and 2010. Among the studies, 71 percent (31 studies) examined teacher practices and classroom quality; 48 percent (21 studies) included child outcomes. Aikens and Akers (2011) reviewed 72 studies conducted from 2000 to 2011 (14 overlapped with those in Isner et al. [2011]) to examine the relationship between coaching and several outcome areas. These reviews indicate that previous research generally supports the positive effects of coaching in three areas: (1) teacher practice, (2) observed classroom quality, and (3) child outcomes.

Teacher Practice

Teacher practice is defined as the strategies and activities the teacher uses with students in the classroom. In the Aikens and Akers (2011) review, of the 26 ECE studies that examined the relationship between coaching and classroom instruction, 22 found a positive relationship, indicating some type of improvement in a teacher's instruction. The coaching in these studies generally consisted of individual coaching sessions that focused on teachers' strengths and targeted areas of improvement through modeling and feedback. For example, Aikens and Akers (2011) cited Fiene's (2002) quasi-experimental study in their review, describing classroom observation measures showing that center-based providers who received four months of mentoring significantly improved in teacher sensitivity and effective discipline over the comparison group. As an additional example outside of the literature review, Wasik and Hindman (2011) used a randomized controlled study to test a nine-month intervention that focused on enhancing teachers' practices related to phonics, phonemic awareness, and oral language development. In classroom observations, intervention teachers demonstrated more language modeling and provided more linguistic feedback to children when compared with the control group.

Observed Classroom Quality

Coaching demonstrates positive effects across different measures of classroom quality. For example, in 27 of 31 reviewed ECE studies that included classroom measures, Isner et al. (2011) noted at least some positive impact on observed quality of the learning environment (e.g., Early Childhood Environment Rating Scale [ECERS; Harms, Clifford, & Cryer, 2004]; Classroom

Assessment Scoring System [CLASS; Pianta, La Paro, & Hamre, 2008]). Other studies have captured positive gains on content-specific classroom measures. For example, Neuman and Wright (2010; included in the review by Aikens & Akers [2011]), compared teachers who received coaching with those who received coursework (30 hours for each). Teachers who received coaching showed statistically significant improvements in classroom structural environment, as measured by the Early Language and Literacy Classroom Observation (ELLCO; Smith, Brady, & Anastasopoulos, 2008).

Child Outcomes

The most proximal outcomes of coaching are teacher pedagogy rather than child outcomes. However, improvement in child outcomes is the ultimate goal of coaching interventions. There is generally less evidence in the available literature that HS coaching improves child outcomes. After all, child outcomes are more distal from teacher coaching, which means that the effects on children may take more time to materialize and may be more difficult to measure reliably. In the Aikens and Akers (2011) review, of all studies reviewed, 35 (49 percent) examined child outcomes. Of these, 21 studies had positive findings. In some cases, the coaching was part of a broader PD intervention, so caution must be taken in attributing outcomes solely to coaching. For example, the review included an experimental study that tested the effects of in-person versus remote coaching for 15 weeks as part of a program to implement a new literacy curriculum (Powell, Diamond, Burchinal, & Koehler, 2010). Children in the two intervention groups outscored children in the control group on outcomes related to letter-sound skills and vocabulary. (There was no clear pattern of outcomes or superiority between remote and in-person coaching.) In the Aikens and Akers (2011) review, there were also examples in which coaching interventions did not lead to affect child outcomes. In a multigroup comparison study by Cusumano and colleagues (Cusumano, 2005; Cusumano, Armstrong, Cohen, & Todd, 2006), children whose teachers were part of the coaching intervention (coursework training and coaching) had similar growth in phonological awareness when compared with children whose teachers were in a second intervention condition (coursework training only) and the comparison condition (no training and no coaching).

Challenges With Interpreting the Evidence Base

Although the available evidence generally supports the positive effects of coaching on teacher practice, classroom quality, and child outcomes, there are significant challenges with interpreting the evidence for the effectiveness of coaching as an intervention strategy by itself. Both the Aikens and Akers (2011) and the Isner et al. (2011) reviews noted that many previous studies did not provide detailed specifications about coaching in their interventions. Investigators have not often provided full information about elements of the coaching model, such as how coaches were selected or trained or the specific structure of the coaching delivery itself. This makes it difficult to determine which coaching dimensions were bundled in the studies. In addition, Aikens and Aikers point that the frequency, duration, and nature of coaching vary across studies, making it difficult to determine the critical elements of coaching. Additionally, some studies included coaching as one feature of an intervention, but they do not specifically isolate and test coaching. When coaching is varied along with other features (e.g., curricula, additional workforce training), it is difficult to determine whether positive outcomes are due to coaching or other factors.

Overall, the key limitations of the available research to date are as follows:

- Descriptions of coaching features (e.g., structure, process, and staffing aspects of coaching programs) tend to be either unclear or not well specified.
- There was little empirical support presented for using certain coaching strategies as part of an intervention program (i.e., experimental methods have not yet shown which specific coaching actions and behaviors are most effective [Fixsen, Naoom, Friedman, & Wallace 2005]).
- Coaching, as a social intervention strategy by itself, is often examined in combination with additional intervention strategies. For example, coaching is included as an element to support a curriculum intervention, or coaching is included with other PD interventions, such as workshop institutes or other variations of training activities.

The design for the HS Coaching Study aims to provide options for evaluating coaching as a distinct PD social intervention in the HS context. The study also aims to address one other issue that has not often been evaluated in coaching studies: it will examine the differential effects of several specific dimensions of coaching. We discuss this in further depth below.

Evaluating Coaching Effectiveness

The design of the HS Coaching Study aims to provide options for evaluating coaching, as a PD intervention separate from other elements such as curriculum or teacher training, and to examine the differential effects of specific dimensions of coaching. Specifically, the study will examine the effect of *varying the levels* of coaching dimensions.

Social interventions often consist of multiple dimensions that are bundled or grouped together with the goal of improving targeted outcomes. In randomized controlled trials (RCTs) or evaluations that examine the impact[4] of social interventions, such bundling prevents researchers from understanding the individual contributing effects of the intervention's different dimensions (Collins, Murphy, & Strecher, 2007). When dimension effects are examined, it is typically done post hoc by using nonexperimental methods (Baker, Kupersmidet, Voegler-Lee, Arnold, & Willoughby, 2010; Collins, Murphy, Nair, & Strecher, 2005). For example, after the impact of a social intervention is estimated using an RCT, exploratory analyses examine whether intervention effects were associated with differential implementation of particular program features.[5] Or, in the context of a meta-analysis, multiple studies containing related interventions may be coded and analyzed according to their features and linked to outcomes accordingly (Dunst, Trivette, & Hamby, 2010). Because the effects of these dimensions are not directly studied, identified effects may be due to other causes.

[4] RCTs are often considered the gold standard for evaluating intervention effects. In this type of experimental test, participants are randomly assigned to either a treatment group that receives the intervention or to a control group that does not, and then average outcomes of the treatment and control groups are compared to evaluate the intervention's average effects.

[5] For example, the *Evaluation of Enhanced Academic Instruction in After-School Programs* (Black, Somers, Doolittle, Unterman, & Grossman, 2009) examined the extent to which the size of program impacts was correlated with program implementation characteristics (e.g., number of days the afterschool program was offered, attrition rates of program staff, and so on).

In general, there is little strong, empirical evidence about the effect of separate dimensions for social interventions (Green, Ha, & Bullock, 2010). This includes ECE coaching interventions that also contain multiple dimensions, which are rarely examined separately. For instance, coaching intervention content may be bundled with delivery in a particular format, bundled with a particular dosage of the intervention, which is further bundled with delivery to a particular Recipient. The traditional program evaluation paradigm—testing whole interventions rather than individual dimensions—leaves evaluators, policymakers, and program developers with an intervention "black box." The design for the HS Coaching Study aims to provide options to open this black box in order to explicitly examine the effects of different dimensions of ECE coaching.

The MOST Approach Recommended for the HS Coaching Study

After reviewing a range of research methods (see http://www.acf.hhs.gov/programs/opre/research/project/head-start-coaching-study-design-phase) for testing the effectiveness of coaching,[6] the design team and OPRE staff members decided that the design for the HS Coaching Study should reflect the principles of the multiphase optimization strategy (MOST; Collins, Dziak, & Li, 2009; Collins, Murphy, Nair, & Strecher, 2005; Collins, Murphy, & Strecher, 2007; Collins, Nahum-Shani, & Almirall, in press). The MOST framework is a staged and rigorous approach to developing and evaluating social interventions.

After an initial preparation phase, an optimization phase is conducted in which the relative effects of different intervention dimensions are assessed in a randomized screening experiment. Dimensions are selected for testing by examining the evidence base or, if the evidence base is weak, using strong theoretical support or recommendations from experienced practitioners and researchers. The results of this screening experiment are then used to build an optimal intervention model (Optimization Phase; L. Collins, personal communication, February 28, 2014) consisting of the selected dimensions that meet some minimum threshold for effect size, cost-effectiveness, and practical or theoretical importance. In a second phase, the impact of this optimal model is evaluated in a standard two-group randomized experiment. The MOST approach has most commonly been used for creating effective public health interventions (e.g., smoking cessation). Its use with social interventions, such as early childhood program training, is an innovative and cutting-edge strategy.

The HS Coaching Study corresponds to the Optimization Phase stage of the MOST framework. By carrying out the first phase of the MOST framework, the HS Coaching Study systematically selected dimensions and will test and determine the relative impact of coaching dimensions. However, adapting the MOST model to the complexities of HS coaching interventions will requires careful thought.

Testing the Relative Impact of Coaching Dimensions

As pointed out earlier, few coaching intervention studies have systematically compared the effectiveness of different variations of coaching dimensions. In other words, most studies on coaching interventions include different combinations of dimensions but do not systematically

[6] As part of the task order, the design team wrote a review that outlined different possible design and methodology framework options for the study. Appendix B provides a short overview of that report.

vary the use of these dimensions to determine the independent, additive, or interaction effects of a specific coaching dimension or of implementing different intensities of a particular dimension.

One of the few studies that did systematically compare different variations of coaching dimensions is from Landry, Anthony, Swank, and Monsegue-Bailey (2009), who used a multiple-conditions experiment design to study the impact of two coaching factors, including (1) one-on-one coaching (four hours per month versus no coaching) and (2) the type of student formative feedback used by teachers (detailed digital versions versus limited paper versions), across four conditions. Teachers in all four conditions received a small-group online training. The condition with the highest amount of in-class mentoring and the high level of feedback on children in addition to the online training yielded the most favorable outcomes for both teachers and students. In a subsequent study (Landry et al., 2011) to further test this optimal approach, the researchers assembled the three components to test in a larger scale RCT. This set of studies follows the MOST approach described earlier.

Beyond this one set of studies, studies examining individual dimensions of coaching do not appear in the published literature. Although some previous coaching studies used rigorous methodology, the unique effects of coaching dimensions on outcomes can rarely be determined because the entire intervention package was under investigation. In designing the HS Coaching Study, we aimed to make a unique contribution to the field by further building the evidence base for combinations and levels of coaching dimensions in HS settings.

Definition of Coaching for Evaluation

Because the HS Coaching Study will involve systematically varying and studying elements of coaching, we did not want to use a coaching definition that already established too many dimensions in our design work. Therefore, we purposely established a simple and straightforward definition of coaching.

We defined *coaching* as a capacity-building strategy that creates an ongoing partnership between an ECE expert (i.e., the coach) and an ECE staff member (i.e., the teacher) through the expert's provision of individualized support and guidance that strengthens the teacher's knowledge and practices to improve classroom quality.[7] Some approaches to coaching include coaching cycles involving different strategies. These cycles include phases when (a) the coach and teacher conduct planning, (b) the coach models and the teacher practices activities, (c) the coach then observes the teacher in the classroom, and (d) the two participate in subsequent reflection or feedback sessions (NCQTL, 2012; Snyder, Hemmeter, & McLaughlin, 2011). However, we purposely left out such specific strategies from our core definition. We wanted to make the explicit decision about each feature of coaching, and whether to:

- Systematically vary and evaluate the feature among participants,
- Standardize the feature across participants, or

[7] This report uses the term *coaching* but acknowledges that, in practice, the terms *mentor*, *consultation*, *facilitation*, and even the more broad terms of *on-site PD* and *technical assistance* can overlap with coaching (National Association for the Education of Young Children [NAEYC], 2013). There are distinctions among these terms (NAEYC, 2013); however, we explored all variations included within coaching.

- Allow the coaching feature to be left to the HS grantees discretion across participants.

Next we provide the research questions that the study aims to address.

Research Questions for Evaluation

Six research questions guided the design of the HS Coaching Study.

The key questions related to *impact* of the coaching dimensions are:

- What is the effect of specific dimensions of coaching on teacher practices and classroom quality in HS programs?
- Does the effect of one coaching dimension depend on the level of another coaching dimension?

The research questions related to the *implementation* of the coaching dimensions are:

- Are the different coaching dimensions implemented with fidelity?[8]
- What factors facilitate or challenge the fidelity of implementation of the different coaching variations? What types of TA and PD tools facilitate the implementation fidelity?
- How does implementation vary across grantees' program environments, populations, and other contextual program features?

The research question related to *cost* is:

- What is the cost of implementing the different coaching dimension variations?

The answers to these research questions will inform HS program decisions about the allocation of PD resources when programs develop and implement coaching interventions. In addition, the answers to these research questions will advance the research evidence about coaching in early childhood settings, will inform grantees and HS centers about more effective practices, and will inform the design of an optimal coaching model.

[8] Fidelity here refers to execution of the coaching dimensions as designed. There may be several aspects of fidelity that may be of interest, including adherence, exposure, responsiveness, and quality.

III. Dimensions of Coaching Interventions

Narrowing to Dimensions Suitable for Testing in the Evaluation

As mentioned earlier, the HS Coaching Study corresponds to the Optimization Phase of the MOST framework, in which the relative effects of intervention dimensions are tested using an experimental design. Dimensions can be selected for testing in this phase by examining the evidence base, by using strong theoretical support, and by utilizing recommendations from experienced practitioners and researchers (Collins, Dziak, & Li, 2009). A significant part of our design task was to develop a list of potential coaching dimensions and select from the list three dimensions for testing. In addition to selecting three dimensions, we also had to define two levels for each dimension that would provide clear contrast.[9] Later in the document, we explain our recommendation of a factorial design that results in eight experimental conditions representing all possible combinations of levels of the three factors. HS centers would be randomly assigned to one (and only one) of these experimental conditions (see Section IV). Because this factorial design involves combining the different levels for each coaching dimension selected, part of the process for determining the proposed coaching dimensions was to ensure that the dimensions could be successfully implemented in combination with other selected dimensions.

First, we examined extant literature to develop a broad list of possible dimensions and look at associated outcomes. Second, we held preliminary discussions among the design team, with OPRE staff, and with academic researchers who served as paid project consultants. Then, as the design team, we used multiple criteria to rank and prioritize the dimensions. Following the ranking and prioritization of the dimensions, we had in-depth conversations with OPRE about the ratings helped us to narrow down the list to a short, prioritized list of potential coaching dimensions. In addition, we also consulted with a limited number of stakeholders at OHS and in the HS field about either their experience with or their opinion about coaching in general or in the context of the planned study. Feedback was gathered either through individual calls, group webinar-format calls, or an interactive conference presentation. Finally, the final proposed dimensions for the coaching study were selected. In the sections that follow, we describe each of these five selection steps in greater detail.

Selection Step 1 – Review Extant Literature

As the first step in selecting dimensions to be tested for the study, the design team documented coaching dimensions, their variations, and the levels of intensity reported in extant research, wherever possible. The review of coaching dimensions considered a range of PD models, programs, and interventions that included coaching to support teacher practices and to promote positive outcomes for children in ECE and K-12 settings. Some researchers are beginning to explicitly describe the dimensions included in myriad coaching approaches (e.g., McGroder, Howard, Fishman, Rankin, & Helsel, 2012; Powell, Diamond, Burchinal, & Koehler, 2010; Taylor, 2008). However, many PD studies, which use coaching as a PD strategy, provide very few structural or process details regarding the coaching process itself.

[9] One level can be set to match practices considered common or typical in the field. The level represents a viable and still potentially effective option; it is not the same as a no-intervention or control condition on the dimension.

Based on the information gathered from the review of extant literature, the design team organized coaching dimensions into three broad types: structure, process, and staffing. For each dimension, we noted what levels or ranges have previously been implemented in other research studies, and also their reported effectiveness on various outcomes. Here we briefly list each dimension that we considered, by these three broad types. In Appendix A, we describe each dimension in more detail.

Structure. Structure dimensions involve the purpose for coaching as well as the organization of the coaching approach.

- Goals
- Recipient
- Dosage
- Format
- Additional PD coordinated with coaching
- Mode

Process. Process dimensions focus on the work between a coach and classroom staff.

- Teacher-coach relationship
- Use of tools
- Use of strategies
  - Planning
  - Modeling
  - Observing
  - Feedback

Staffing. Staffing dimensions include the selection and characteristics of coaches and managing their work.

- Coach selection
- Coach caseload
- Coach training
- Coach supervision

Selection Step 2: Phase 1 Stakeholder Input

As part of our design process, we asked different groups of stakeholders different types of questions about coaching and coaching models. To remain in compliance with the Paperwork Reduction Act (PRA) and approval process with the Office of Management and Budget (OMB) for federally sponsored data collection, we did not ask more than nine respondents to provide the same information within a 12-month period. We spoke individually to a limited number of HS directors or coaches from grantees that had previously used coaching. To help understand the HS

coaching context, we asked some participants to reflect on successes and challenges in implementing their own coaching programs. To validate our list of coaching dimensions, we asked another set of participants to provide feedback on coaching models more generally, including the structure, process, and staff dimensions listed earlier.

We also gathered feedback from an interactive round-table presentation at the HS 40th Annual National Research Conference in May 2013 (Howard & Drummond, 2013), which was attended by HS leaders. At this session, we asked the attendees to comment on what practitioners are interested in learning from a study on coaching, as well as give suggestions on the aspects of coaching they feel are important to test.

Selection Step 3: Prioritization Process

After preliminary discussions among the design team, with OPRE staff, and our paid ECE PD expert consultants, we developed criteria related to importance, design considerations, and implementation feasibility (Table 1 below). The criteria below were used to examine each of the structure, process, and staffing dimensions previously listed. After further conversations with OPRE, we established a small pool of top-priority dimensions that ranked at the highest level.

Table 1. Criteria Used for Prioritizing Coaching Dimensions for Testing

Key Areas for Consideration	Criteria for Selecting Coaching Dimensions
Importance	▪ Potential benefit to the field ▪ Evidence of variation in practice ▪ Evidence of effectiveness (previous research and theory) ▪ Strength of theory supporting dimension ▪ Relevance to HS grantees' interests
Design considerations	▪ Ability to determine two contrasting levels for the dimension ▪ Likelihood that the dimension could work with other dimensions ▪ Ability to identify dimension levels and implement them in in factorial design
Implementation feasibility	▪ Feasibility of implementing the dimension in HS ▪ Feasibility of coach to implement multiple levels ▪ Level of TA, monitoring, or piloting needed

Selection Step 4: Phase 2 Stakeholder Input

Step four of our dimension selection process involved a new set of stakeholder conversations with eight individuals from HS grantees—directors, coaches, or education managers.[10] Among these eight participants, five had developed coaching programs, as defined by having education specialists or coaches working with classroom teachers. The other three participants did not have a well-established coaching model (either no coaching or only a peer-mentor model, consisting of veteran teachers meeting with new staff). The individuals we spoke with varied in geographic location, urbanicity, and size. Depending on their role, stakeholders answered questions about

[10] Eight participants were selected to remain in compliance with the Paperwork Reduction Act (PRA) and approval process with the Office of Management and Budget (OMB) for federally sponsored data collection.

their grantee context and about current or desired coaching models or coaching roles. We asked these individuals to provide us with their feedback on our prioritized set of dimensions, including the definition of the dimension, its importance, and how potential levels of the dimension might look if implemented in HS grantees. We also asked a subset of stakeholder respondents to provide their opinions about logistics for a potential study. For instance, we asked about the kind of grantee that might be willing to participate in the study and what could influence a grantee's decision to participate. This informed our thinking about approaches to site selection criteria and recruitment (see Sections IV).

We also presented our study design to six individuals who serve HS grantees in support or TA capacities, through federal centers or local organizations, or as professional organization staff who worked for regional HS TA centers. We asked them to provide comments on the study design. A subset of this group was asked to provide their opinions about the kind of grantee that might be willing to participate in a potential study. In addition, OPRE staff also met with OHS staff within HHS to share with them the potential coaching dimensions under consideration to gather their feedback and input into the process.

Throughout this report, we refer to the opinions gathered during these interviews as *stakeholder* opinions. We use the term to refer to input gathered throughout the entire design process (i.e., selection steps 2 and 4), from stakeholders who work in or with HS grantees.

Selection Step 5: Final Selection of Dimensions for Coaching Study

Ultimately, on the basis of the information gathered during the previous four steps of the selection process, the design team recommended three dimensions for testing in the HS Coaching Study. These include the following:

- Dosage (how many hours of coaching HS teachers receive),
- Coach Training (whether and how coaches are trained), and
- Recipient (whether coaching is delivered individually or with classroom teaching teams).

Based on stakeholder opinions, in this final step of the selection process, we also recommend a fourth dimension—Mode—that could be considered by OPRE as an alternative dimension, given the high interest in conducting coaching at a distance by using technology.

Background on Coaching Dimensions to Vary Systematically

In this subsection, for each of these dimensions, we present the definition, level of theory and evidence, along with any feasibility issues related to testing the dimension. Later, after we present recommendations for the foundational coaching model, we will revisit these four coaching dimensions in order to describe the recommended levels for variation and testing in the HS Coaching Study.

Dosage

Definition. For purposes of the study, dosage is defined as the amount of individual coaching that an HS teacher receives. There are several ways to conceptualize dosage. It can be determined by

considering the frequency of coaching sessions, the length of each session, and the duration of the program.[11] The amount and frequency of coaching that is *intended* is the starting point for any coaching intervention. In practice, however, there is also the amount of coaching actually *offered* and the amount of coaching a recipient actually *received,* which can differ from what was intended (Howard et al., 2013; Wasik, Mattera, Lloyd, & Boller, 2013). We recommend that active steps be taken with the HS Coaching Study so that dosage actually offered to and received by participants is as close as possible to the intended dosage. The steps we recommend for monitoring and TA are aimed at maintaining intervention fidelity (described in more detail in Section X).

Theory. Coaching sessions need to be of sufficient frequency and duration to result in changes in teacher behavior (Halle, 2008). However, there is no accepted, established threshold, and several factors may affect what is sufficient (Wasik, Mattera, Lloyd, & Boller, 2013). Higher dosages of coaching provide teachers with more learning and problem-solving experiences and more chances to practice, build confidence, and develop mastery of new techniques. More time also allows the coach to have more opportunities to adapt to a teacher's unique needs. However, the complexity, breadth, or newness of the targeted behaviors should be considered in establishing optimal dosage (Halle, 2008; Halle et al., 2010). A coaching approach seeking to introduce or strengthen complex or new behaviors may require a larger dosage, whereas coaching programs that focus on specific, delimited behaviors could succeed with a smaller dosage.

Evidence. Among the dimensions, dosage has the greatest amount of empirical evidence for its effectiveness (Mashburn, Downer, Hamre, Justice, & Pianta, 2010; Powell, Diamond, & Burchinal, 2012; Shidler, 2009). However, coaching amounts vary greatly in the field. For instance, in one analysis of ECE coaching from 1995 to 2011, coaching programs ranged from one week (1 percent) to one year (22 percent) and usually occurred weekly (39 percent) or monthly (26 percent); however, of the 101 studies reviewed, 48 percent did not report the length of the teacher-coach sessions (NCQTL, 2012). Because there is no conclusive evidence on the threshold of coaching dosage that is necessary to change teacher practice, determining how coaching dosage affects teacher practice (along with what type of dosage—intended, offered, and received, as well as quality) can provide practical benefits to the field. Among HS stakeholders, there was strong consensus that dosage was one of the most important dimensions to test.

Feasibility. The same coach may implement different dosage levels systematically without extensive training, TA, or piloting; monitoring of systematic variation in dosage is also relatively straightforward. In addition, typical and enhanced levels of dosage can be combined successfully with other recommended dimension variations.

Coach Training

Definition. Coach training for the HS Coaching Study refers to the amount, content, and nature of preparation and ongoing training that the coach receives. Just as coaches provide job-embedded training to teachers, the coaches themselves may have a higher quality of coaching

[11] This represents one of several potential ways to conceptualize dosage (Berkel, Mauricio, Schoenfelder, & Sandler, 2010). Another way is the periodicity (frequency) of sessions, the duration of individual sessions, and the span or time period across which coaching sessions take place (Halle et al., 2010; Justice, Mashburn, Hamre, & Pianta, 2008).

behavior and support to the staff they work with when they receive job-embedded PD. The format of coach training may vary, including coaches preparing through self-study, attending training led by experts, peer coaching (observing or working with a mentor), or participation in a learning community of coaches. Such PD is designed for coaches to better understand the coaching content and practices and how best to scaffold and improve the practices of the adult learners (i.e., HS teachers).

Theory. More intensive coach training can affect the quality of coaching process dimensions (e.g., observation, modeling, planning, feedback, and teacher-coach relationship) and use of adult learning strategies that, in turn, could improve teacher learning, practice, and classroom outcomes. Training can ensure that coaches understand how to carry out their work effectively (Bryant et al., 2009).

Evidence. Little empirical research examines how the quantity or type of training provided to coaches makes a difference in coaching behavior and teacher-, classroom-, or child-level outcomes. In the design team's interviews with stakeholders, coach training was prioritized as a dimension that would be important to study and test.

Currently, levels or type of coach training appear to vary widely in practice. For example, among 50 ELMCs interviewed in a descriptive study of HS coaching programs grantees self-developed, a substantial number—34 percent—reported that they received no training or that they trained themselves. Another 16 percent reported that they received training on coaching. The remaining 50 percent received focused training on either assessments or on grantee or programmatic information (Howard et al., 2013). Most ECE coach programs in other studies involve initial trainings to orient participants to the project, to help participants learn about the curriculum or content focus, and to allow participants to discuss the coaching process and strategies; however, many studies do not provide sufficient detail on coach training (Isner et al., 2011).

Ongoing training, after initial training or orientation, for coaches seems to vary. A study on the use of coaching and teacher coursework to improve language and literacy development had a two-day orientation for coaches, followed by weekly debrief sessions (Neuman & Wright, 2010). Another study that used coaching to help improve classroom management had coaches meet for an initial three-day orientation (Head Start CARES demonstration; Morris, Raver, Millenky, Jones, & Lloyd, 2010). Representing a very high amount of coaching training, a study currently implementing the Building Blocks mathematics curriculum in New York City provides approximately six weeks of ongoing training to support coach knowledge and skill development (Making Pre-K Count study; MDRC, personal communication, March 12, 2014).

Feasibility. There are a few feasibility issues to consider with systematically testing coach training. It seems possible to set the two levels to provide a strong contrast in services and to successfully combine training with other dimension variations; however, the enhanced level would require financial resources and significant TA to implement. An additional methodological issue with training is that because one coach cannot simultaneously represent both levels of training, testing this dimension requires the random assignment of coaches (discussed in greater detail in Section V).

Recipient

Definition. The recipient is the staff member who receives coaching services. In current practice, ECE staff members who receive coaching may be selected based on their role (e.g., lead teachers), specific backgrounds (e.g., new teachers), or needs (e.g., teachers who have students with challenging behaviors).

We recommend testing two variations related to role—coaching only the lead teacher versus coaching the classroom teaching team. The lead teacher (sometimes referred to as the *teacher of record*) is the primary teacher responsible for planning and delivering instruction. The classroom teaching team includes an assistant teacher or aide in the classroom who works with the lead to support the classroom.[12]

Theory. Theoretically, coaching approaches that target different recipients have different strengths. Coaching only the lead teacher (1) makes it easier to individualize the coaching and (2) provides greater confidentiality, which could enable teachers to feel more open to sharing their challenges. Coaching the entire teaching team (1) allows evidence-based practices to be integrated into the classroom, (2) makes all teachers accountable for implementing evidence-based practices, and (3) provides the opportunity for teachers to collaborate and support each other within the classroom. There is theory to suggest the team coaching approach may produce more consistency and higher quality for children throughout the day, in activities conducted by the teacher or the assistant, and therefore may lead to better outcomes (Morris, Raver, Millenky, Jones, & Lloyd, 2010).

Evidence. The literature review for this study found no empirical studies that varied the recipient of coaching and examined resulting outcomes. Evidence shows that common practice with respect to coaching recipient varies, with a substantial portion including assistant teachers. The NCQTL (2012) summary of 101 ECE coaching studies from 1995 to 2011 found that, in 76 percent of the studies they reviewed, lead teachers were the recipients of the coaching. In the descriptive study of the HS ELMC initiative, many coaches in the study reported a wide variation in the recipient of coaching. In the ELMC study, coaches reported working with an average of 2.3 staff types, which could include any combination of lead teachers, assistant teachers, home visitors, family child care staff, administrators, supervisors, other administrators, or other. Many of the coaches (58.0 percent) reported providing coaching to 2 to 6 different staff types. Of the 381 coach ELMC survey respondents 19 percent focused on only the lead classroom teacher, and 38 percent focused on both the lead teacher and assistant teachers (Howard et al., 2013).[13] One major study incorporated the teaching team in coaching efforts to improve classroom quality. For example, in the HS Research-Based, Developmentally Informed (REDI) project, the teaching team was included in the workshop and coaching components (Domitrovich, Gest, Gill, Jones, & DeRousie, 2009a). Qualitative work suggests that teaching assistants feel more integrated into the teaching team when they attend and participate in the same PD programs as lead teachers (Morris, Raver, Millenky, Jones, & Lloyd, 2010).

[12] HS classrooms have different teacher structures: there may be colead teachers; there may be multiple assistants.
[13] The remaining coaches reported working with home visitors (19 percent), mostly along with classroom staff and sometimes with other program staff, or working with other combinations of classroom staff (24 percent), including administrators, supervisors, family child care staff, and other staff types, but excluding home visitors.

Reasons for limiting the recipient of coaching to only the lead teacher can vary. For instance, in terms of cost, not every site can afford the training time and materials required to coach multiple teachers per classroom. Logistically, sites may not be able to coordinate release time, establish appropriate substitute coverage, or fulfill union requirements for assistants to attend additional training and coach sessions. There was strong support among stakeholders that the recipient dimension was important to test so that HS grantees may determine whether it is worth the additional investment of coaching resources to include the teaching team.

Feasibility. Testing the recipient of coaching is feasible and can be done in conjunction with the other selected dimensions. One coach can be assigned to implement both levels simultaneously without the need for intensive TA. Those receiving the teaching team training will require extensive additional resources in terms of periodic substitute teacher coverage for one or more teachers per classroom to facilitate the coaching.

Alternative Dimension: Mode—Delivering Coaching Through Technology

Definition. Delivery mode involves the way coaching services are provided: in person or through technology. Although in-person coaching is currently more common, there is evidence supporting the positive effects of remote coaching, which can involve the combined use of exchanging videos, videoconferencing, telephone conversations, and e-mail exchange.

Theory. The theory of action for both modes of delivery, in person or through technology, requires targeted feedback linked directly to teachers' practices in classroom, such that the coach can adapt accurately to teacher needs and provide support and knowledge to facilitate changes in teacher practice. In-person interaction has traditionally been seen as a more potent way to develop trust and provide structure to engage the teacher in changing practice. However, because of the increase in technologically mediated coaching programs, the necessity of these components is coming into question. Remote coaching is seen as a way to deliver the same support, knowledge, and structure in an efficient manner that may save costs in terms of coach labor and expenses. Testing delivery mode in the HS Coaching Study provides potential benefits to the field related to long-term cost savings and coaching in isolated or rural areas.

Evidence. Previous empirical research (Landry, Anthony, Swank, & Monsegue-Bailey, 2009; Powell, Diamond, & Burchinal, 2012; Powell, Diamond, & Koehler, 2010) indicates that on-site and technologically mediated (remote) coaching approaches can be equally effective for changing teacher practice. However, this is a relatively limited body of research. There was consensus among HS grantee stakeholders that mode was an important dimension to test. Although interested in the possibility, the HS grantee members of our stakeholder group did not currently use remote coaching themselves. They had questions about technology infrastructure for some centers and wondered how well relationships could be developed through technology.

Feasibility. Creating the two contrasting levels for mode appears feasible, despite some challenges. Although implementing mode appears straightforward, establishing a purely in-person or purely technologically mediated condition is challenging; in-person coaches may occasionally use phone and e-mail, and technologically mediated coaches may be inclined to visit in person to establish rapport with the teacher. The main challenge is the substantial start-up costs of the technologically mediated mode. Supporting consistent implementation among

participants who may have varying levels of technology familiarity and experience may require substantial technical support, training, and assistance.

Pilot Testing. The PD developers or providers will have to adapt their materials and approach to be carried out remotely and actually develop and pilot-test an accessible, user-friendly technology platform (unless the selected provider already has technology-based infrastructure and routines). In addition, the study team should explore the technology capacity for potential participants.

Foundational Coaching Approach

This section provides recommendations for the foundational coaching approach that would support the test of the three systematically varied coaching dimensions (*Dosage*, *Recipient*, and *Coach Training* or, as an alternative, *Mode*). In order to isolate the effect of these dimensions, we strongly recommend the study establish and implement a foundational coach model that will provide the content, organization, and standardized aspects of the coaching intervention.

For each of the structural, process, and staffing dimensions not chosen for systematic variation and testing, we provide recommendations for how to structure the coaching element. In order to keep the foundational approach consistent across sites and conditions (the eight experimental conditions that represent all possible combinations of the levels of the dimensions under investigation), we make recommendations for the following:

- Goal and coaching content (including selecting a PD developer for the coaching content)
- Additional PD coordinated with coaching
- Coach roles and teacher-coach relationship
- Use of tools
- Coach strategies
- Coach selection
- Coach supervision
- Coach caseload

Specifically, we recommend that each coaching element satisfies at least one of three criteria: (1) the element is fixed, or held constant, for all participants; or (2) the element meets a minimum threshold; or (3) the element can be allowed to vary as it typically does among grantees (i.e., typical practice for the site or by individuals). In Section VIII (measures), we provide recommendations for how the study should collect data to document dimensions, including those that are tested and those elements of the foundational model.

Goal and Coaching Content

Definition and Background. The goal of coaching is the actual content and classroom practices on which coaches focus.

In developing our recommendations, the design team considered a number of learning domains and considered what classroom practices are of high priority for HS grantees and for children. According to the NCQTL (2012) review of 101 coaching studies, ECE coaching has been applied to preacademic skills (43 percent), social-emotional development (36 percent), communication skills (22 percent), and noncontent-specific instructional practices (25 percent). The systematic review of Aikens and Akers (2011) summarized that ECE coaches typically focus on classroom instruction, curriculum implementation, teacher-child interactions, and classroom environmental indicators. Some approaches to coaching do not specify particular content domains or sets of strategies on which teachers and coaches will focus, but rather they allow the partnership flexibility to set the goals and purpose of their work together based on identified needs (Boller, Blair, Del Grosso, & Paulsell, 2010; Koh & Neuman, 2009; Neuman & Cunningham, 2009).

Recommendations. The design team recommends that the HS Coaching Study focus on an important aspect of preschool classrooms—language development and the interactions between children and teachers that support that development. We suggest that choosing to focus on language development is valid because it is a critical domain of early child development and accepted as important in HS. It has been established that children learn about conversational and speech patterns from their caregivers; children's exposure to words correlates with socioeconomic status (SES), with children of lower SES having far less exposure (Hart & Risley, 2004). Language skills are a well-established precursor to subsequent literacy skills (Lonigan, 2006; Walker, Greenwood, Hart, & Carta, 1994; Whitehurst & Lonigan, 1998) that grow increasingly important as children approach entry to elementary school. Language activities are part of the accreditation process for the NAEYC (2013).

Language development is one of the 11 domains within the HS Child Development and Early Learning Framework, which includes elements for both receptive and expressive language (Fuentes, 2010). Constructs within the HS Learning Framework provide examples of goals for children's language skills:

- Attends to language during conversations, songs, stories, or other learning experiences
- Comprehends increasingly complex and varied vocabulary
- Comprehends different forms of language, such as questions or exclamations
- Comprehends different grammatical structures or rules for using language
- Engages in communication and conversation with others
- Uses language to express ideas and needs
- Uses increasingly complex and varied vocabulary
- Uses different forms of language
- Uses different grammatical structures for a variety of purposes
- Engages in storytelling
- Engages in conversations with peers and adults

In addition, CLASS (Pianta, La Paro, & Hamre, 2008), an observational protocol used by the OHS as part of the HS Designation Renewal System and therefore very important to HS grantees, has several scales that emphasize high-quality teacher-child interactions. The CLASS Instructional Support domain includes the following scales:

- The Concept Development scale emphasizes the use of instructional discussions and activities to promote a child's higher-order thinking skills and cognition, including discussions that require analysis and reasoning as well as integration of previous concepts.
- The Quality of Feedback scale includes a teacher's use of multiple strategies—scaffolding, feedback loops, prompting thought processes, providing information and encouragement—to provide feedback to children.
- In the Language Modeling scale, teachers are encouraged to promote frequent conversations with children, as well as between children, and teachers are to use strategies such as open-ended questions, extension of child responses, narration of the teacher's own actions, and advanced language to facilitate and stimulate the child's language use.

Selecting a PD Developer

As part of the preparation phase for the HS Coaching Study, we recommend that a PD developer be selected to adapt the established curriculum for this study. We recommend the PD developer be selected[14] for his or her capacity to work with coaches and teachers on helping HS children address key language development constructs. The selected PD approach should serve as a supplement to any existing curricula that an HS grantee might use. In other words, the set of strategies or practices should enhance the current instructional program.

To help students reach language developmental milestones, the selected PD developer could work with coaches to help teachers to increase the number of verbal interactions and opportunities for children to talk, guide teachers to more consistently engage in social interaction with children in their class, help teachers increase the quality of verbal interactions (such as asking open-ended, higher-order questions; using active listening; and making connections in conversation), and help teachers with their use of scaffolding conversations with children and encouragement of children's efforts to be involved in conversations (Justice, Pence, Beckman, Skibbe, & Wiggins, 2005; National Early Literacy Panel, 2008; National Institute for Literacy, 2009; Spencer, Goldstein, & Kaminski, 2012).

Because numerous coaching interventions and studies focus on language development (Diamond & Powell, 2011; Dickinson & McCabe, 2001; Landry, Swank, Smith, Assel, & Gunnewig, 2006; Neuman & Wright, 2010; Roskos, Christie, Vukelich, & Han, 2003; Wasik & Bond, 2001; Wasik, Bond, & Hindman, 2006; Wasik & Hindman, 2011; Zan & Donegan-Ritter, 2014; Zucker, Justice, Piasta, & Kaderavek, 2010), there are likely existing PD developers who could participate in the HS Coaching Study.

[14]A developer could be selected by the future HS Coaching Study contract team as part of the application process for the study or in conjunction with OPRE as part of the study-planning phase, in a selective application process that uses specific criteria.

The selected PD developer must be willing to adapt their extant curriculum resources (i.e., training and resource materials for teachers, materials for coaches) to fit the needs of the HS Coaching Study. For example, if the developer's PD typically applies to a broad array of language and literacy skills, the developer would narrow to the ones that fit just language content for this study.[15] Further, the selected developer would need to adapt their curriculum and resources to align with the variations of each of the tested dimensions and assure a strong enough contrast between these variations. The selected developer must also be able to train coaches on the use of needs assessments (see Use of Tools section later) to allow coaches to individualize which practice areas relating to language the coach and teacher will focus on.

Additional PD Coordinated With Coaching

Definition and Background. Coaching is often combined with other teacher PD activities that occur in a group setting and support the work of the coaches. Examples include a onetime workshop, a series of classes, a summer institute, or a professional learning community. If the PD is truly part of the coaching approach, it must be aligned and consistent with the topics covered as part of coaching. Thus, for the HS Coaching Study, the coordinated PD would focus on the language development content and strategies just discussed. We realize that HS had required trainings that all teachers regularly attend; we are not suggesting changes to that typical, business-as-usual training. We are only referring here to the training that is coordinated with the coaching as part of the HS Coaching Study.

Recommendations. We recommend that the HS Coaching Study fix the coordinated teacher PD (specifically the teacher training on the language content teachers receive in addition to coaching) for all conditions.

- All teachers (as well as the center directors and coaches) would attend an initial orientation (less than one day) on the targeted language practices.[16]
- Training of teachers, coaches, and center directors on the specific variations in the three dimensions they are to implement and other study requirements would occur at the same time.
- There should be no other study-provided coordinated group teacher PD (subject to final decisions with the provider or developer).

In recruiting sites, the study evaluation team should screen sites to ensure that potential participants do not have large extant PD efforts beyond required HS trainings and especially in the area of language that could interfere with estimating the impacts of coaching.

Coach Roles and Teacher-Coach Relationship

Definition and Background. This aspect refers to the wide range of roles that may be played by a coach (e.g., expert, friend, emotional supporter, or advocate; see Howard et al., 2013). It includes

[15] However, in conjunction with the PD developer, OPRE and the future evaluator should decide whether the content should involve a limited number of emergent literacy goals, in addition to the goals related to language development.

[16] During this orientation, participants would also receive an introduction to the overall study and its procedures.

the nature of the relationship between the coach and the teacher (e.g., the expectations about the supervisory relationship between the two and the power hierarchy).

Recommendations. We recommend fixing expectations for the coach role and the teacher-coach relationship for each participating grantee.

- Coaches may be hired by the grantee, but each grantee should use the same set of expectations in the hiring process:
  - The coach role will focus on collegial, collaborative, and cooperative expert support for the teacher.
  - The coach should serve in a nonsupervisory role.

The design team posits that role clarity, meaning how well a teacher understands the coach role, is necessary for effective coaching and is a prerequisite for any coaching approach. Therefore, the role and goal of the coach should be clearly articulated in all conditions. We acknowledge that exactly how and how well the coach develops relationships with individual teachers will typically vary.

Use of Tools

Definition and Background. This aspect focuses on the use of tools and data to guide and systematize a coach's work with teachers. Programs may use tools such as rubrics, checklists, or quality rating scales (e.g., Quality Rating and Improvement System [QRIS]) to select the focus of coaching activities. This information is commonly used to tailor the content and the delivery of coaching. Some developers of coaching interventions promote a data-driven approach for individualization (Hemmeter, Fox, & Snyder, 2014; Landry, Anthony, Swank, & Monsegue-Bailey, 2009). These approaches require coaches to collect data on teacher or child behavior so that coaches have concrete information on teacher needs and progress to create specific training plans with teachers (Domitrovich, Gest, Gill, Jones, & DeRousie, 2009a; Palsha & Wesley, 1998). For instance, an observational measure can help provide specific goals for quality improvement (Rubin, Sutterby, & Hoffman, 2011). Landry, Swank, Smith, Assel, and Gunnewig (2006) reported on a quasi-experimental intervention targeting preschool teachers' enhancement of children's language and literacy in which data collected with a teacher behavior rating scale was used to guide the coaching work with individual teachers.

Recommendations. We recommend that the HS Coaching Study fix the use of certain tools while allowing the use of other tools to vary.

- Coaches should receive from the PD developer common baseline training to provide a standard practice for the implementation of the needs assessments for teachers:
  - An initial needs assessment
  - A midyear follow-up needs assessment

  Because it is important for a coaching program to select a tool aligned with the overall goals of the program (Tout, Zaslow, Halle, & Forry, 2009), we recommend that final development and use of tools and data should occur in coordination with the selected developer.

- Beyond these guidelines, decisions about the use of tools will be left up to individual coaches and may vary.

Coach Strategies

Definition and Background. There are four strategies commonly used by coaches—planning, modeling, observing, and providing feedback—which are sometimes referred to collectively as a *coaching cycle* (Howard, et al., 2014; NCQTL, 2012; Snyder et al., 2012). *Planning* refers to preconferences during which the coach prompts the teacher to set goals and action steps in preparation for a lesson or observation. Coaches use *modeling* to demonstrate a teaching strategy (presumably one with empirical evidence or promise of effectiveness) to a teacher, with the goal of strengthening the teacher's understanding and confidence in using the technique. Coaches use *observation* to gather information about a teacher's classroom practice and provide specific and individualized *feedback* about the teacher's classroom practices.

The NAEYC (1993) provides a set of overarching principles for coaching ECE teachers that includes the following: "provide opportunities for application and reflection and allow for individuals to be observed and receive feedback on what has been learned" (p. 9). Moreover, the NCQTL (2012) review of coaching found that 72 percent of the studies reviewed use some type of feedback. In the ELMC descriptive study, observation, feedback, and discussion strategies that were commonly reported as strategies used by coaches with staff, however coaches reported that feedback was not as an effective strategy for changing staff practice as modeling or observation. For instance, only 19 percent of the 360 coaches responded that feedback was an effective strategy for changing teacher's practices, whereas 30 percent of the coaches cited that observation was effective and 65 percent of coaches cited modeling was the most effective strategy (Howard et al., 2013). Yet, despite evidence about coaches beliefs about feedback, observation, and modeling effectiveness, research does not yet indicate the amount of feedback that is necessary to effect teacher change, the mode of feedback that should be used, or when or where feedback is provided. In the ELMC descriptive study, coaches reported providing feedback based on live observations, reflective feedback, verbal feedback based on discussions with staff, and feedback resulting from live, on-site observation (Howard et al., 2013). Teachers themselves report on the value of coach feedback (Diamond & Powell, 2011; Morris et al., 2013), especially if the feedback is limited and concrete and involves additional exemplars and resources to avoid being overwhelming.

Recommendations. We recommend that the HS Coaching Study fix the provision of basic training on common coaching strategies.

- We recommend that the HS Coaching Study provide consistent common guidance to coaches on the use of the four strategies—planning, modeling, observing, providing feedback—as well as how to engage teaching staff in the coaching process.
- Beyond the common guidance, use of these strategies may be left to the HS grantee discretion as a function of assignment to specific conditions reflecting different variations of the levels of coach dimensions. For teachers who receive more Coach Training, there may be more strategy information exchanged and opportunity for coaches to learn about, discuss, and plan their use of the strategies. In addition, in the group receiving a higher coaching Dosage, there may be more opportunity to use these strategies with teachers.

Coach Selection

Definition and Background. Coach selection involves the process of identifying, recruiting, hiring, and matching candidates to fill coaching positions.

Recommendations. We recommend that the HS Coaching Study fix certain selection requirements for coaches and allow other requirements to vary.

- We recommend that study coaches should have at least a bachelor's degree; at least three years of ECE experience; interpersonal skills as documented in other HS coaching initiatives (Howard et al., 2013); and, depending on settings and population demographics for the study, possible bilingual proficiency.
- Prior coaching experience need not be required, which is fairly common in ECE coaching practice (Howard et al., 2013).
- Beyond these standard expectations for selecting and hiring coaches, grantees or centers can control the selection and hiring of the coaches locally.

Coach Supervision

Definition and Background. Coach supervision refers to the formalized process of providing oversight and support to coaches.

Recommendations. For the HS Coaching Study, we assume that supervisors of coaches can be staff within existing center or grantee organizational structures. We recommend fixing supervision of the coach expectations at a minimum standardized level, which some grantees may exceed, depending on their administrative structure. Supervision from HS programs can ensure study fidelity by encouraging coaches to deliver the expected coaching Dosage, attend the requisite meetings and trainings, and complete the required forms.

- The evaluator and PD developer would be responsible for standardizing expectations for supervision of the coaches in regards to implementing the foundational model.
- The content of the supervision would largely focus on administrative issues (related to logistics at the center and for the study, on the basis of input from evaluation and PD developer staff).
- Limitations would need to be set on how varied the supervision can be.
    - For instance, supervisors may not focus on providing additional training on the coaching content or strategies and will not be responsible for enhancing the quality of the coaching beyond ensuring a basic competent level of performance in order to avoid interfering with the test of the Coach Training dimension.
    - Supervisors could ask about general challenges or barriers at the center. However, it would be our expectation that content of coach-teacher sessions are kept confidential.

In order to ensure this fixed level of coach supervision, center administrators should sign a memorandum of agreement and statement of responsibilities as part of the study recruitment

process. Supervisors will be encouraged to attend teacher and coach orientations. Administrators and other grantee staff may also attend study orientations, teacher PD, and Coach Training.

Coach Caseload

Definition and Background. Coach caseload is the number of teachers or classrooms assigned to a single coach. In model or demonstration programs, coaches may have small caseloads (three to eight programs); statewide initiatives or QRIS coaching tend to have larger caseloads (up to 22; Isner et al., 2011). In the ELMC descriptive study of coaching in HS programs, coaches who reported working with both lead and assistant teachers together in a HS program reported a median caseload of 6 inclusive of these two staff types together (Howard et al., 2013).

Recommendations. Because the Dosage of coaching will vary across conditions, different assigned teacher-coach cases may represent more or less time commitment for the coach. For the HS Coaching Study, we recommend that caseload be within a fixed range. The range will avoid coaches having so high a caseload that they become overwhelmed or so low a caseload that grantees have difficulty recruiting for very limited part-time coaching positions. Our recommendation is in the middle of the caseload range reported in other coaching studies.

- We recommend setting the lower end of coach caseload at 12 classrooms, with a maximum of no more than 16 classrooms per coach.

Levels of Dimensions to Vary Systematically

The foundational coach model, composed of the elements just described, provides the standardized structure for the coaching intervention. Next, we will discuss the aspects of the coaching model that will be systematically varied: two levels for each of the three dimensions. In the recommended study design, there will be mixing and matching of the dimension levels across eight conditions to which centers (or coaches) will be randomly assigned. Through random assignment, one set of participants could experience a condition that is Level I for Dosage, Level II for Coaching Training, and Level I for Recipient. Participants in another condition could experience Level II for all three dimensions.

Dosage

Recommendations on Levels. We recommend testing whether providing a typical level of coaching Dosage (Level I) versus providing an enhanced level of coaching Dosage (Level II) has a differential impact on classroom teaching practices. Level I should still have the potential to impact teacher practice and represent a typical level in actual ECE coaching practice. The enhanced Level II should provide a program of coaching that is more intensive, but it should not be too much; it must be practical and sustainable in terms of cost and staffing and material resources.

We recommend setting the levels of Dosage as shown in Table 2. We suggest, however, that as part of the initial planning phase for the HS Coaching Study, the evaluator conduct additional exploration to confirm typical levels of coaching Dosage currently in practice at HS sites.

Table 2. Typical and Enhanced Levels of the Dosage Dimension

Dosage	
Level I	**Level II**
14 hours per classroom Delivered monthly 120-minute sessions	28 hours per classroom Delivered biweekly 120-minute sessions

We specify that both levels will implement 120-minute sessions that include teacher-coach meetings. Coaches' time to schedule, travel, participate in Coach Training, plan for the individual session, and complete study data collection should be allocated and occur outside of these coaching sessions.

We suggest allowing for some buffer to set up the coaching program and logistics at the beginning of the school year. The recommended time span for the coaching is seven months (this could be extended to eight months, depending on final decisions with the developer and provider of the program).

Coach Training

Recommendations on Levels. We recommend testing whether providing a typical level of Coach Training (Level I) versus providing an enhanced level of Coach Training (Level II) has a differential impact on classroom teaching practices. For Level I, we recommend a basic, initial training that orients coaches to their role and to the foundational coaching model. This appears to represent a typical level of Coach Training in ECE settings. For the contrasting Level II, we wanted a more intensive Coach Training and ongoing support, but at a level that would be practical and sustainable in terms of cost. Table 3 describes the recommended levels of Coach Training.

Table 3. Levels of the Coach Training Dimension

Coach Training*	
Level I	**Level II**
Two-day summer coach orientation Three quarterly one-hour large-group phone conferences with a coach trainer during school year.	Two-day summer coach orientation Three quarterly one-hour large-group phone conferences with a coach trainer during school year. Three additional days of summer training Weekly one-hour small-group phone conferences with a coach trainer during school year.

* This includes meetings facilitated by a coach trainer and coach trainer on-site visits.

As indicated in the table, all coaches will receive the following:

- A two-day summer orientation training on (a) what the coaching entails, (b) the targeted principles of adult learning for the targeted curriculum, (c) how to administer a short classroom needs assessment, and (d) an introduction to the study's procedures.
- Three large-group conference calls with coach trainers, covering both the language development content and the coaching process.

Coaches assigned to the more intensive Level II of training will receive two additional components:

- Additional training days include three additional summer training days over the summer.[17]
- Additional training also includes small-group conference calls with coach trainers on a weekly basis throughout the intervention. Academic consultants report that regular conference calls have been used in used in large PD studies (Neuman & Wright, 2010; Tout, Halle, Zaslow & Starr, 2009). Similarly, some stakeholders report that regular calls have been successful in supporting coaches in local coaching models. We recommend that the support calls involve facilitated discussion on the topics defined by the coaching curriculum, along with troubleshooting among coaches.

For teachers who receive more Coach Training, there will be more opportunity for coaches to learn about, discuss, and plan their use of the strategies discussed earlier—planning, modeling, observing, and providing feedback. Coaches assigned to the more intensive level of Coach Training would also receive more information on how to identify teacher needs by using data and how to actively engage teachers in the coaching process.

Regardless of which level coaches are assigned to, we recommend that all Coach Training focus on successful principles of adult learning, particularly coaching strategies that focus on promoting active engagement, sharing informative content, and discussion, as well as opportunities for practice, observation, self-reflection, and feedback (Birman, Desimone, Porter, & Garet, 2000; Bowman, Donovan, & Burns, 2000; Carlson et al., 2012; Dunst, Trivette, & Hamby, 2010; Wasik & Hindman, 2011).

Recipient

Recommendations on Levels. For the Recipient dimension, we recommend testing whether coaching only the lead teacher (Level I) versus coaching the classroom teaching team (e.g., lead teacher and assistant teacher or aide; Level II) has a differential impact on classroom teaching practices (Table 4).

Table 4. Individual and Teaching Team Levels of the Recipient Dimension

Recipient	
Level I: Individual	**Level II: Teaching team**
Lead teacher only	Classroom teaching team

[17] The precise mix of training activities must be finalized in conjunction with the PD developer.

The coaching recipients will have diverse backgrounds, including varied experiences, skills, and needs for information (e.g., working with more challenging populations). As in typical practice and reported in research (e.g., Howard et al., 2013), coaches may have to navigate the different needs and characteristics of teachers, whether working with the lead teacher only in the individual level (one-on-one) or the teaching team level (as a classroom team).

Alternative Dimension: Mode—Delivering Coaching Through Technology

Recommendations on Levels. If Mode is chosen as the third dimension tested, we recommend testing whether in-person coaching (Level 1I) versus remote, technologically mediated coaching (Level I) has a differential impact on classroom teaching practices (Table 5).

Table 5. Remote and On-Site Levels of the Delivery Mode Dimension

Delivery Mode	
Level I: Remote	**Level II: On-Site**
Remote, technologically mediated coaching involving phone, e-mail, and video, with one initial in-person coach visit	In-person coaching only

The technologically mediated Mode (Level I) may involve teachers submitting a videotape of their practice and coaches providing feedback either by e-mail or with a portable digital medium that pairs written feedback with selected segments of a teacher's videotaped practices. Use of videotape exemplars of high-quality practices associated with specific targeted teaching strategies may substitute for in-person coach modeling.[18]

Summary

Drawing on the design team's review of coaching and individual dimensions, we recommend the following approach and content for the HS Coaching Study:

- Testing of the suggested levels of three coaching dimensions: Dosage, Recipient, and Coach Training.
- Mode (remote versus on-site) could be considered as an option for an alternative coaching dimension to test.
- There would be a foundational coaching approach that would have a selected PD developer use or adapt the content and coaching approach focused on supporting the language development of children within HS classrooms.
- The foundational coaching approach is to be composed of coaching elements that are standardized, set at a minimum threshold, or left to program or coach discretion.

[18] The final details of delivering coaching via technology must be finalized in conjunction with the PD developer.

IV. Recommended Study Design: A Factorial Experiment

Given the recommendations of the previous chapter, the Head Start (HS) Coaching Study will examine the effect of three individual coaching dimensions: (1) the Dosage of coaching; (2) the Recipient of the coaching; and (3) the amount of Coach Training or, alternatively, the Mode of the coaching. Strictly speaking, the study will examine the effect of *varying the levels* of each of these coaching dimensions, but for simplicity in this section and those that follow, we will sometimes refer to "the effect of a coaching dimension".

A background paper for the HS Professional Development (PD) task order reviewed several experimental design options that could be used to examine the effect of varying dimensions of coaching in HS (Somers, Collins, & Maier, 2013). Five experimental designs were considered: factorial designs, comparative treatment designs, an individual experimental design, crossover designs, and adaptive clinical trials. These designs were compared in terms of how well they would be able to answer the study's research questions, their sample size requirements, the number of experimental conditions that would have to be implemented, and whether interactions between components could be estimated.

The main conclusion from this background paper is that a factorial design—specifically, a factorial experiment with three factors and eight experimental conditions—is the most suitable design for testing the effect of the three coaching dimensions.[19] A factorial design is an experimental design in which the experimental conditions represent all possible combinations of the levels of the dimensions under investigation. Although factorial designs require more experimental conditions than other designs, a benefit is that they require a smaller sample size than other designs to statistically detect a dimension's effect of given magnitude. The benefit of a smaller sample size can outweigh the disadvantage and cost of having to implement a larger number of conditions. Another potential benefit of factorial designs is that they also account for—and provide information on—interaction effects between the dimensions that are being tested in the study. (As will be discussed later in this section, interactions are less reliably estimated, but they can be examined for hypothesis-generating purposes.) Thus, factorial designs make it possible to get inside the black box of an intervention or program in order to efficiently determine which particular components of an intervention are more important, as well as examine how these components interact with each other to produce the desired outcomes. For these reasons, factorial experiments provide findings that are useful for policymakers and practitioners who are creating or adapting interventions. Given these desirable properties, factorial experiments are well suited for building strong interventions in the Optimization Phase of the multiphase optimization strategy (MOST) framework (Collins, Dziak, & Li, 2009: Collins, Murphy, Nair, & Strecher, 2005; Collins, Nahum-Shani, & Almirall, in press).[20]

[19] See Appendix B for a summary of the other experimental design options that the project team considered but that were ultimately deemed unsuitable for the HS Coaching Study.

[20] As described in Chapter 2, in the Optimization Phase of MOST, an experiment is used to test the effect of different components or features of a social intervention. The results from this experiment are then used to inform the design of an optimal intervention consisting of the components that meet some threshold for effectiveness or cost-effectiveness.

The next four sections of this report provide methodological details about the final recommended study design for the HS Coaching Study, along with the assumptions and rationales that led to the final recommended design. Section IV (the present section) provides an overview of the guiding principles that were used to select the design and a description of the 2^3 factorial design that is recommended for experimentally manipulating (and testing) the three coaching dimensions. Section V discusses the design team's recommendations related to the unit of random assignment (HS centers), the sample size requirements for the study, and the random assignment plan. Section VI describes the statistical model that can be used to estimate the effect of the coaching dimensions based on the final recommended design. Section VII describes the implementation research (IR) and cost study. Finally, Section VIII describes data collection and measurement options.

Guiding Principles for the Study Design

In selecting the most suitable final design for the HS Coaching Study, the design team established five guiding principles:

- *The study should be based on an experimental design*. The findings from the coaching study are intended to inform the resource allocation decisions of policymakers and practitioners in the HS context by providing reliable evidence on the causal effect of different aspects of coaching. Thus, the findings from the study should be rigorous and unbiased estimates of the causal relationship between aspects of coaching and a specified outcome. Accordingly, the design team also determined that random assignment should be used to assign sample members to experimental conditions.
- *The study design should provide estimates of the effect of varying each coaching dimension, as well as interactions, if feasible*. Given the study's research questions, the HS Coaching Study should provide estimates of the effect of each tested coaching dimension, on average, across the levels of the other dimensions of interest. If possible, the study should also provide estimates of the interaction between the dimensions of interest, in order to examine whether the effect of each coaching dimension depends on the level of the other dimensions being investigated. Interactions can inform policymakers and practitioners as to whether a coaching dimension is effective only when it is paired with other dimensions or whether its effect is robust regardless of what other aspects of coaching are being implemented.
- *The study should evaluate the effect of coaching dimensions on teacher and classroom outcomes*. Because its goal is to identify promising intervention components, the outcomes examined in the Optimization Phase of MOST are typically shorter-term proximal outcomes (such as teacher practices and teacher-child interactions), whereas impacts on longer-term outcomes (such as child outcomes) are typically examined in the Evaluation Phase of MOST. The HS Coaching Study can be viewed as the first phase of MOST, and therefore it will focus on teacher and classroom outcomes. Focusing on these shorter-term outcomes will have the further benefit of containing the data collection costs of the study.
- *The study design should minimize the number of sample members required to detect an effect of given magnitude*. A unique challenge with studies of intervention components or dimensions is that the expected effect of a single dimension is likely to be smaller in

magnitude than the effect of a complete intervention package. Prior studies have shown, for example, that the effect of entire teacher PD interventions on teacher practices and other classroom outcomes is approximately 0.40 to 0.90 standard deviations, which was generally large enough to also lead to effects on child outcomes.[21] Thus, we expect that the effect of an individual coaching dimension on teacher and classroom outcomes to be smaller than 0.40.[22] By extension, the HS Coaching Study *must be powered to detect smaller effect sizes* (and therefore will require *a larger sample size*) than an evaluation of a complete teacher PD or coaching package. This, in turn, may increase the cost of the HS Coaching Study relative to a standard impact evaluation – although we emphasize that the HS Coaching study will yield much more information than a standard impact evaluation. Thus, the final study design should minimize the number of sample members required to detect a given effect size.

- *The study design should examine no more than three dimensions and have no more than eight experimental conditions*. As noted earlier, it was determined that a factorial experiment is the most suitable design for the HS Coaching Study given its research questions. However, these designs can be challenging to implement and monitor in a reliable way because they have many experimental conditions. As will be shown later in this section, the experimental conditions in a factorial design represent every combination of the coaching dimensions' levels. Therefore, as the number of dimensions to be tested increases, so does the number of experimental conditions (if there are K dimensions, each with 2 Levels, then there are 2^K experimental conditions). The HS Coaching Study will be the first large-scale evaluation of coaching dimensions in early childhood settings, so it is important that the study be well implemented and that it provide a fair test of the dimensions under investigation.[23] On the basis of the design team's experience conducting random assignment studies, it was decided that the maximum number of coaching dimensions that can be well implemented and reliably monitored in the field given study resources is three.

Given these guiding principles, the final recommended study design is a 2^3 factorial experiment. This design will make it possible to examine the effect of three coaching dimensions (each with two levels) and their interactions. In total, the design will have eight experimental conditions. The details of the design and its characteristics are discussed in the remainder of this section. For simplicity, in this discussion we assume that HS centers will be the unit of random assignment (the recommended random assignment plan will be discussed in Section V).

[21] For example, the HS CARES examined the effect of three PD interventions (all of which included coaching) on teacher practices and found effect sizes ranging from 0.41 to 0.92 (Mattera, Lloyd, Fishman, & Bangser, 2013). This is in line with other prekindergarten studies. For instance, in the Foundations of Learning study, effects on teacher practice ranged from 0.46 to 0.90 (Morris, Raver, Millenky, Jones, & Lloyd, 2010). In the HS CSRP study, effects ranged from 0.52 to 0.89 (Raver et al., 2008). In the HS REDI study, effects ranged from 0.39 to 0.97 (Domitrovich et al., 2009b). These impacts on teacher practice were also sufficiently large to lead to impacts on child outcomes (Domitrovich et al., 2009; Morris, Raver, Millenky, Jones, & Lloyd, 2010).

[22] This assumes that no component of the PD intervention is reducing the overall effect size by exerting a negative effect (in which case, some components could actually be having an effect larger than 0.40). We think that this assumption is reasonable.

[23] Landry, Anthony, Swank, and Monsegue-Bailey (2009) used a factorial experiment to test the effect of two aspects of early childhood teacher PD (rather than aspects of coaching). Thus, this would be the first large-scale study to look at components of coaching more closely.

The Study Design: A 2^3 Factorial Experiment

The final recommended design for the HS Coaching Study is a 2^3 factorial experiment. In a factorial experiment, each of the coaching dimensions to be tested becomes an independent variable—called a *factor*—whose levels are manipulated by the evaluator.

As discussed in Section III, three coaching dimensions will be varied and tested: the amount of coaching (*DOSAGE*), the target of the coaching (*RECIPIENT*), and either the amount of Coach Training (*TRAINING*) or the Mode of delivery (*MODE*). These coaching dimensions can be operationalized into factors for the experiment, each with two levels: a typical level of the underlying dimension and an enhanced level of the underlying dimension.[24] Table 6 shows the definition of the levels for each coaching dimension or factor (i.e., *DOSAGE*, *RECIPIENT*, *TRAINING*, and *MODE*). Factor names are capitalized in the statistical literature, and therefore we will also use this convention in the technical sections of this report (sections IV, V, and VI).[25]

Table 6. Selected Coaching Dimensions as Factors in the Experiment

Factor	Description	Level	
		Level I	Level II
DOSAGE	Amount of time that the coach spends with a classroom's teaching staff	120-minute session once per month	120-minute session every other week
RECIPIENT	Recipient of the coaching	Lead teacher only	Classroom teaching team (lead and teaching assistant)
TRAINING	Amount of training that the coach receives	Summer orientation and quarterly large-group phone conferences	Summer orientation, summer training session, ongoing Coach Training during the year, and weekly small-group phone conferences
MODE	Mode of the coaching (remote vs. on-site)	Initial in-person meeting, with remainder of sessions conducted remotely (online)	Completely in-person coaching

Note. See Section III of this report for further information on the definition of these dimensions and their levels.

Table 7 illustrates the recommended factorial design based on these three coaching factors. Factorial designs are commonly described in terms of the number of factors in the experiment and the number of levels in each factor. Therefore, the design shown in Table 7 is a 2^3 factorial design because it has three factors each with two levels ($2 \times 2 \times 2$). As shown in Table 7, this factorial design results in eight experimental conditions, representing all possible combinations of levels of the three factors. HS centers would be randomly assigned to one (and only one) of

[24] As explained in Section III, we use the term *Level I* for the first level; level represents practices that are often typical in coaching programs and still potentially effective in and of themselves.

[25] Although factors can have more than two levels, using more than two levels increases the sample size requirements for the study and its operational complexity, which in turn increases study costs. Therefore, for the HS Coaching Study, each factor should be limited to two levels.

these experimental conditions. For instance, in centers randomly assigned to Experimental Condition 1, lead teachers would receive coaching on a monthly basis, and their coach would have received less intensive training. In contrast, in centers randomly assigned to Experimental Condition 4, teachers and their assistants in each classroom would receive coaching on a monthly basis, and their coach would receive more intensive training.

Table 7. Recommended 2^3 Factorial Design

Experimental Condition Number	Factors		
	Amount of Coaching (*DOSAGE*)	**Recipient of the Coaching (RECIPIENT)**	**Amount of Coach Training (TRAINING)***
1	Monthly	Lead teacher only	Orientation
2	Monthly	Lead teacher only	Ongoing
3	Monthly	Teaching team	Orientation
4	Monthly	Teaching team	Ongoing
5	Biweekly	Lead teacher only	Orientation
6	Biweekly	Lead teacher only	Ongoing
7	Biweekly	Teaching team	Orientation
8	Biweekly	Teaching team	Ongoing

Note. Shading denotes the enhanced level (Level II) of the factor; unshaded cells represent the typical level (Level I) of the factor.
*Or Mode, in which case the levels in the design would be remote coaching (in the unshaded cells) and in-person coaching (in the shaded cells).

If the effect of Mode were tested—instead of the effect of Coach Training—then the third factor in the design would be *MODE* rather than *TRAINING*. The design would still have three factors and eight experimental conditions. HS centers assigned to the Level I condition of *MODE* would receive coaching primarily remotely or online, whereas centers assigned to the Level II conditions of this factor would receive in-person coaching only.

There are two features to note about this factorial design and its experimental conditions. First, every combination of dimension levels (experimental condition) must be possible to implement. Related to this point, there are potential challenges with testing delivery Mode (*MODE*) instead of Coach Training (*TRAINING*). In two of the experimental conditions (3 and 7), coaches would have to deliver coaching remotely to the entire teacher team (lead teacher plus teaching assistant). Although feasible, coaches would likely need additional training and technical support to deliver effective coaching to the entire team remotely. In Appendix D, we discuss another alternative design where *MODE* could replace *RECIPIENT* rather than *TRAINING*.

Second, notice that all eight experimental conditions in Table 7 include some coaching. This means that unlike an impact evaluation of a complete coaching intervention, the recommended design does not include a "no coaching" or traditional control condition. Instead, centers in the study will be asked to implement a particular combination of the levels of the coaching factors,

dependent on the experimental condition to which they are assigned.[26] Thus, the design in Table 7 will provide information on the effect of varying the level of a particular coaching dimension (which is the relevant question given the goals of the study), rather than the effect of coaching versus no coaching.

Types of Effect That Can Be Estimated

The factorial design that is represented in Table 7 can be used to estimate two types of effect. First, it can provide information about the main effect of each of the three coaching dimensions when considered across all levels of the other dimensions. Second, it can also provide an estimate of interaction effects, including the two-way interactions and the three-way interactions between these dimensions. This array of findings is likely to produce useful information for the field of ECE. These two different types of effect and their properties are discussed in greater detail here.

Main Effects

The *main effect* of a factor is the effect of that factor averaged across all the levels of all the other factors in the experiment. The main effect of a factor is obtained by comparing the mean outcomes of sample members assigned to the conditions in which the factor's level is set to the enhanced level (shaded cells for that factor in Table 7) against the mean outcomes of sample members assigned to the remaining half of the conditions in which the factor's level is set to the typical level (unshaded cells for that factor).

Table 8. How Main Effects Are Estimated Based on the 2^3 Factorial Experiment in Table 7

Main Effect of Factor	Compare Classroom Outcomes in Centers Assigned to the Following Two Sets of Experimental Conditions	
	Level I Conditions	Level II Conditions
Amount of coaching (*DOSAGE)*	1–4	5–8
Recipient of the coaching (*RECIPIENT*)	1, 2, 5, 6	3, 4, 7, 8
Amount of Coach Training (*TRAINING*) or Mode (*MODE*)	1, 3, 5, 7	2, 4, 6, 8

Note. The numbers in this table refer to the experimental conditions in Table 7. Level I conditions for a given factor are represented by unshaded cells for that factor in Table 7, whereas Level II conditions for a given factor are represented by shaded cells.

Table 8 shows which experimental conditions are compared to estimate the effect of each coaching dimension. For example, the main effect of *DOSAGE* can be obtained by comparing the mean teacher and classroom outcomes of HS centers assigned to conditions where *DOSAGE*

[26] Although there is not a single designated control group as with a typical RCT, this is still a controlled experiment; for more information, see Somers, Collins, and Maier (2013). It is also important to note that some factorial experiments *do* have a "no services" condition. Specifically, if the two levels of the factors are "on" versus "off" (as opposed to two different levels of intensity as in the HS Coaching Study), then one of the experimental conditions would receive no services because all factors would be turned "off".

is set to monthly (Conditions 1–4) with the mean outcomes of HS centers assigned to conditions where *DOSAGE* is set to biweekly (Conditions 5–8). Similarly, the main effect of *RECIPIENT* is the difference between mean outcomes for centers in the experimental conditions where *RECIPIENT* is set to lead teacher only (Conditions 1, 2, 5, and 6) and mean outcomes in the experimental conditions where *RECIPIENT* is set to teaching team (Conditions 3, 4, 7, and 8). The main effect of *TRAINING* (or *MODE*) is obtained in a similar fashion. There are three features of main effects that are worth noting.

First, the main effect of each coaching dimension is estimated by comparing the mean outcomes of *groups* of experimental conditions, not by directly comparing the mean outcomes of individual conditions. As illustrated in Table 8, the appropriate way to estimate the main effect of a dimension is to compare groups of experimental conditions. Individual experimental conditions are never directly compared with each other in a factorial analysis of variance (ANCOVA),[27] because the goal is not to provide a definitive answer to the question of which *single combination* of dimensions is best. Rather, as already noted, the goal is to estimate the main effect of each coaching dimension (or more specifically, the effect of varying the level of a coaching dimension).

Second, the main effect of each dimension is estimated using the entire sample. This feature is one of the most important advantages of the factorial experiment compared with other designs. The sample efficiency of a factorial experiment stems from the *balance property*: each level of each factor appears in half of the experimental conditions. For example, Table 7 shows that the monthly level of *DOSAGE* appears exactly two times at the lead-teacher-only level of *RECIPIENT* and exactly two times at the teaching-team level of *RECIPIENT*. This holds for every level of every factor at every level of every other factor. This balance property is what makes factorial experiments so sample efficient. The sample efficiency of the factorial design relative to other design options is discussed in greater detail in Somers, Collins, and Maier (2013).[28]

Third, the main effect of a dimension is different from a *simple effect*, which is the type of effect with which most education researchers are familiar. A simple effect is the effect of a factor *at a fixed level of the other dimensions in the study*. The reason that researchers are more familiar with simple effects is that they are the type of effect obtained from a regression analysis.[29] In the HS Coaching Study, for example, the simple effect of *DOSAGE* would be its effect when *RECIPIENT* is set to lead teacher and when *TRAINING* is set to orientation, or its effect when *RECIPIENT* is set to the entire classroom teaching team and when *TRAINING* is set to ongoing (both are examples of the simple effects of *DOSAGE*). If the effect of *DOSAGE* depends on the

[27] For example, the outcomes of HS centers in Condition 1 will not be compared directly to the outcomes of centers in Condition 2.

[28] The balance property also brings with it an important insurance policy: If, after random assignment, it is discovered that one of the factors cannot be varied—for example, Level II of *DOSAGE* cannot be implemented because of limited funds or poor implementation by coaches—then the factorial experiment can be analyzed as a 2^2 factorial experiment that excludes the problematic factor. This is a useful feature in a context in which there is uncertainty about funding or the feasibility of implementing certain factors.

[29] In the kinds of regression analyses typically used in education research, the regression coefficient associated with a binary (dichotomous) variable included on the right-hand side of the model is the effect at a fixed level of the other variables. This is due to the dummy coding of binary variables.

levels of the other dimensions, then these two simple effects are not equal to each other. For this reason, estimates of simple effects can be less useful from a policy or practical perspective, because they represent the effect of a dimension under a very particular set of circumstances that may not be broadly relevant.[30]

In contrast, the *main* effect of *DOSAGE* is its average effect across all levels of the *RECIPIENT* and *TRAINING* dimensions. Main effects are more useful from a policy perspective because they can help identify coaching dimensions whose effect is robust across the levels of other dimensions. An important advantage of the factorial experiment is that it provides estimates of main effects (as well as simple effects), while other design options provide only estimates of simple effects (see Somers, Collins, and Maier [2013] for a discussion).

Interaction Effects

The second type of effect that can be examined with a factorial experiment is an interaction effect. Two factors are said to interact when the size of the effect of one factor varies depending on the level at which another factor is set.[31] For example, there would be a *DOSAGE* × *RECIPIENT* interaction if the effect of *DOSAGE* when *RECIPIENT* is set to lead teacher only differs from its effect when *RECIPIENT* is set to teaching team. If the effect of *DOSAGE* is the same no matter what *RECIPIENT* is set to, then there is no *DOSAGE* × *RECIPIENT* interaction. The ability to examine interaction effects will be an important contribution of this study because not much is known about whether the effect of a particular aspect of coaching depends on the levels of the other features in the coaching model.

In this report, we define the two-way interaction between Dimension A and Dimension B as the effect of Dimension A *when Dimension B is set to its enhanced level* (Level II) minus the effect of Dimension A when *Dimension B is set to its typical level* (Level I):

$$(Effect_A|B = "Level\ II") - (Effect_A|B = "Level\ I")$$

This two-way interaction can be estimated by comparing the effect of Dimension A across the two subsets of the experimental conditions representing the levels of Dimension B. As an example, let's assume that we want to estimate the interaction between *DOSAGE* and *RECIPIENT*. This interaction would be estimated by comparing the effect of *DOSAGE* in the subset of experimental conditions where *RECIPIENT* is set to teaching team to the effect of *DOSAGE* in the subset of conditions where *RECIPIENT* is set to lead teacher only. The relevant experimental conditions (in Table 7) that would be compared are as follows:

- Effect of *DOSAGE* when *RECIPIENT* is set to teaching team: (7,8) *minus* (3,4)
- Effect of *DOSAGE* when *RECIPIENT* is set to lead teacher only: (5,6) *minus* (1,2)
- Difference of the two effects (interaction): [(7,8) *minus* (3,4)] *minus* [(5,6) *minus* (1,2)]

[30] When there are interactions between the dimensions, simple effects are not the same as main effects.
[31] Here, we are referring exclusively to interactions between factors in the experimental design, not between the factors and unmanipulated variables (for example, teacher experience).

Thus, to estimate interactions, one must split the sample into subgroups based on the levels of one of the dimensions. For this reason, the statistical power for an interaction effect is less than for a main effect. For example, if a main effect corresponding to a regression coefficient of 0.20 can be statistically detected given the sample size, a two-way interaction effect would have to be twice as large (i.e., 0.40) to be detected. Therefore, the analysis of interaction effects should be considered exploratory in the HS Coaching Study (statistical power will be discussed in greater detail in Section VI).[32]

[32] In some fields, interaction effects are defined as *half* of the difference in the effect of Dimension B across the two levels of A, i.e. (1/2)*$(Effect_A|B = "Level\ II") - (Effect_A|B = "Level\ I")$. Based on this definition, the statistical power for mains effects is the same as for interaction effects. However, this alternative definition of a two-way interaction is much less useful from a policy or evaluation perspective.

V. Random Assignment Plan and Sample Size Requirements

This section discusses the recommended random assignment plan and sample size requirements for the HS Coaching Study. As noted at the start of Section IV, the sample size for this study will be larger than for an impact evaluation of an entire coaching intervention package because the HS Coaching Study will examine the effect of individual coaching dimensions, and therefore it must be powered to detect effect sizes of smaller magnitude. However, the ability to detect smaller effect sizes must be balanced against the feasibility and additional cost of implementing a study with a large number of HS centers. Thus, the random assignment plan described in this section was chosen with the explicit goal of minimizing the sample size requirements while still ensuring that the study can be implemented in a field setting and be affordable.

In this section, we begin by discussing the effect size that was chosen for the purposes of powering the HS Coaching Study. Next, we describe the staffing structure of HS grantees and how this organizational structure affects the unit of random assignment and the minimum detectable effect size for each coaching dimension. Third, we present the recommended random assignment plan for the HS Coaching Study for two different scenarios that the evaluators may encounter in the field, and we discuss the sample size needed to detect effects of different magnitude under each plan. Finally, we conclude with a summary of our recommendations related to the random assignment plan and the number of centers to be recruited for the study, and we briefly discuss the implications of these recommendations for site recruitment and monitoring. Appendix C provides further technical detail on the sample size calculations presented in this section.

The Minimum Detectable Effect Size Used for Powering the Study and Parameter Assumptions

An important step in any study is to determine the sample size that is needed to detect an effect of meaningful magnitude. The *minimum detectable effect size* (MDES) is a useful concept for making decisions about the sample size. Formally, MDES is the smallest true effect on the outcome of interest (scaled as an effect size) that can be detected with a reasonable degree of power. A critical determinant of the MDES is the sample size: All else being equal, the larger the sample size, the smaller the effect size that the study will be able to detect. Conversely, the smaller the effect size that one would like to detect, the larger the sample size needs to be. This raises two important questions.

The first question relates to the *type* of effect that the HS Coaching Study should be powered to detect. On this issue, we recommend that the study should be powered to detect *main effects* as opposed to interaction effects. In theory, one could choose a sample size based on the desire to detect an interaction effect of given magnitude. However, this is not a desirable strategy in practice because: (1) choosing a target effect size for an interaction is more difficult than choosing a target effect size for a main effect because almost nothing is known about interaction effects; and (2) main effects are typically of greater scientific and policy interest, whereas interaction effects are most useful as a secondary source of information to help evaluators interpret main effects.

The second question relates to the *size* of main effect that the HS Coaching Study should be powered to detect. Prior studies of PD in early childhood settings have focused primarily on the impact of entire teacher PD packages. These packages typically include a combination of formal in-service teacher training, ongoing coaching for teachers during the school year, and other teacher supports. In the HS Coaching Study, the goal will be to test the effects if *individual* dimensions of coaching, which are likely to be smaller than the effect of an entire PD package. This means that the study should be powered to detect an effect size that is smaller than the effect of an entire PD intervention. Therefore, a larger sample size will be needed for HS Coaching Study as compared to what is required to evaluate the impact of an entire package.

We recommend that the HS Coaching Study be able to detect a main effect on teacher and classroom outcomes of 0.20, for two reasons. First, based on prior research, it seems reasonable to expect that the coaching dimensions in the study would have main effects of this size on teacher practices. A study conducted by Landry and her colleagues (2009) found that four hours of coaching per month can improve teacher practices by an effect size of 0.23 to 0.70, with effects being larger for practices that teachers used less frequently at baseline (such as practices related to phonological awareness).[33] This suggests that the main effect of *DOSAGE* on teacher practices in the HS Coaching Study (which is a test of the effect of an additional 1.5 hours of coaching per month) could be about 0.09 to 0.26.[34] Thus, it is reasonable to expect that the dimensions under study could have a main effect of 0.20 on teacher practices that are in greatest need of improvement. Second, an effect size of 0.20 on teacher practices can also translate into a meaningful change in children's literacy-related outcomes. Even though child outcomes will not be measured in the HS Coaching Study, improving children's outcomes is one of the goals of coaching. In general, effects on child outcomes are about 25 percent of the magnitude of effects on the teacher practices that target those child outcomes.[35] Thus, an effect size of 0.20 on teacher and classroom outcomes translates into an effect of about 0.05, which is equivalent to approximately 1.4 to 2.5 weeks of extra learning, or a 5 to 10 percent increase in children's literacy skills above and beyond what they would normally learn during the school year.[36]

In order to determine the sample size that is needed to detect a main effect of 0.20 on teacher practices, one must make assumptions about the unit of random assignment (which we assume

[33] This study, conducted by Landry and her colleagues (2009), used a 2x2 factorial design to test the effect of two dimensions of professional development for early childhood teachers: (1) one-on-one coaching (four hours per month versus no coaching) and (2) the type of student formative feedback used by teachers (digital versus paper).
[34] This was obtained by dividing the range of effect sizes from Susan Landry's study (0.23 to 0.7) by (1.5/4).
[35] This assumption is based on the results reported in Landry et al. (2009), studies published by MDRC and AIR (e.g., Garet et al., 2008), and from internal datasets. The exact percentage varies across outcomes (ranging from 15 to 35 percent), but on average it appears that impacts on child outcomes are about a quarter of the size of the impact on teacher outcomes.
[36] During their prekindergarten year, children's letter and word recognition (based on the Woodcock-Johnson test) are expected to grow by an effect size of about 0.92 standard deviations (or an effect of about 0.035 per week, assuming 26 weeks between the pretest and posttest). Effect size gains based on the Expressive Picture Vocabulary Test (EPVT) are 0.51 (about 0.02 per week). Thus, an effect size of 0.05 on child outcomes is equal to about 1.4 additional weeks of instruction based on the Woodcock-Johnson or 2.5 weeks based on the EPVT. Similarly, an effect size of 0.05 is equivalent to a 5 percent increase in learning (0.05 divided by annual gains of 0.92) based on the Woodcock-Johnson or a 10 percent increase based on the EPVT (0.05 divided by annual gains of 0.51). These annual effect size gains were estimated based on children in the control group centers in the CARES study (Mattera, Lloyd, Fishman, & Bangser, 2013).

will be HS centers or coaches, as discussed in the next section). Given that the primary unit of random assignment will be centers, assumptions must also be made about two key parameters: (1) the extent to which teacher and classroom outcomes vary *between* HS centers as opposed to *within* centers (this is also called the between-center intraclass correlation or ICC); and (2) the extent to which the baseline covariates predict between-center and within-center variation in the outcomes of interest.

In order to make informed assumptions about these parameters, we used data from the HS CARES study (Mattera, Lloyd, Fishman, & Bangser, 2013). The CARES study is an evaluation of the impact of three socio-emotional program enhancements on classroom and child outcomes in HS settings. Each enhancement being tested has intensive teacher training and coaching supports. Importantly, the HS centers that participated in the CARES study are similar to the types of centers that would be recruited for the HS Coaching Study: many of the CARES centers are located in urban areas[37] and these centers readily expressed willingness to participate in a large-scale study of teacher PD. In addition, teachers in the CARES study were assessed on practices that are similar to those that would be measured in the HS Coaching Study. For these reasons, parameters from the CARES data can provide reasonable assumptions about the structure of the data that we are likely to see in the HS Coaching Study.[38]

From the CARES data, we focused in particular on the center-level ICC and the between-center and within-center R^2 for the teacher practice measures that are most closely aligned with what would be measured in the HS Coaching Study (e.g., the Adapted Teaching Style Rating Scale and the CLASS instructional domain[39]). In order to examine the "stability" of the CARES parameters, we calculated the ICC and R^2 based on different subsamples of the CARES study (all centers, control group centers, and urban centers). We also examined whether the between-center and within-center R^2 are greater when the analysis controls for classroom-level baseline measures of the outcome of interest. Finally, we examined whether the between-center R^2 is similar when center-level mean CLASS scores are used as baseline covariates rather than classroom-level baseline measures.[40] There are two points worth highlighting in regards to the CARES parameters:

- The ICC and R^2 (between center and within center) are similar across the CARES outcomes and subsamples.[41] For the purposes of powering the HS Coaching Study and

[37] As will be discussed later in this section and in this report, recruitment for the HS Coaching Study will likely focus on large urban grantees in order to meet the sample size requirements for the study.

[38] We also collected parameter information from other studies, such as the FACES survey. Reassuringly, parameters from these studies were similar to what was obtained from the CARES data. These parameters—as well as power calculations based on these parameters—can be found in Appendix C.

[39] The Adapted TSRS is of interest because it is an adapted version of existing measure of teacher practices, which was modified for the purposes of the CARES study. The instructional domain of the CLASS is of interest because scores on this domain are lower, and therefore teachers would benefit most from being coached on the practices in this domain. The TSRS and the CLASS are discussed in Section VII of this report, as well as Appendix C.

[40] To examine the R^2 for center-level baseline pretests, we used the classroom-level baseline CLASS scores collected for the CARES study and aggregated them up to the center level. We then examined the extent to which these aggregate scores predict teacher practice outcomes (the Adapted TSRS).

[41] The CARES parameters are also similar to parameters we obtained from the FACES data, which confirms that parameters from the CARES data are a reasonable approximation for the purposes of calculating the MDES for the HS Coaching Study. The full set of ICC and R^2 from the CARES and FACES data—and sample size calculations based on each set of parameters—can be found in Appendix C.

making recommendations about the number of centers to recruit, we have used the most conservative set of CARES parameters (those that imply the largest sample size).[42]

- Controlling for grantee explains a substantial amount of the between-center variation in teacher practice outcomes. For this reason, we recommend that random assignment be blocked by grantee or by groups of similar grantees or centers. This approach will improve the precision of estimated effects (relative to not blocking), which will in turn reduce the sample size required to detect an effect of given magnitude. Blocking will be discussed in greater detail later in this section.
- Controlling for classroom-level baseline pretests of teacher practice does not appreciably increase the between-center and within-center R^2 over and above controlling for grantees. Moreover, at the center level, average CLASS scores are just as predictive of the teacher outcomes of interest as classroom-level pretests. As argued by Bloom, Richburg-Hayes and Black (2007), controlling for group-level administrative outcomes (such as school-level or center-level assessment scores from previous school years) can go a long way toward helping statistical power, and they are much less expensive to collect. Thus, for the purposes of powering this study, where relevant we assume that existing center-level CLASS scores from the fall of the previous school year will be used as baseline measures in the HS Coaching Study (rather than classroom-level pretests).

The MDES and sample size calculations presented in this report are based on a statistical significance level (Type I error rate) of 10 percent rather than the typical 5 percent level. Recent HS studies funded by HHS have reported statistical significance up to the 10 percent level (that is, effects with p-values less than or equal to 10 percent have been flagged with a star).[43] For the HS Coaching Study, we recommend using a 10 percent level as well. As explained earlier in this report, the HS Coaching Study can be viewed as the "Optimization Phase" of MOST, which is an exploratory phase whose purpose is to provide information that will lead to a rigorous evaluation of an optimal coaching intervention (the "Evaluation Phase" of MOST). Although a low Type I error rate (i.e., 5 percent) is desirable in the Evaluation Phase, it can be preferable to have a higher Type I error rate (10 percent) in the Optimization Phase given its purpose. In the HS Coaching Study, for instance, the goal will be to identify aspects of coaching that are effective. Given this objective, the cost of a Type II error (failing to detect a useful dimension) may be greater than the cost of a Type I error (falsely concluding that a dimension is effective).[44] Allowing a higher Type I error rate (10 percent) will reduce the risk of concluding that the effect of a promising dimension is not statistically significant.

Staffing Structure of HS Grantees: Implications for the Unit of Random Assignment, Blocking, and the MDES

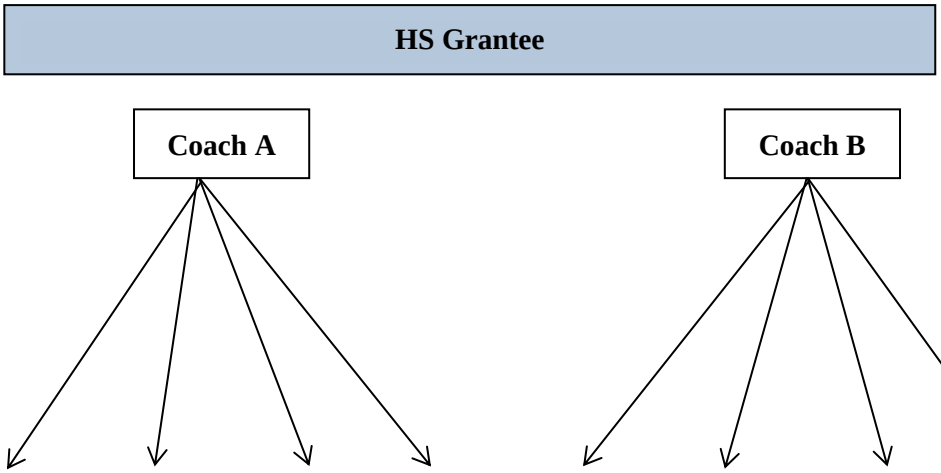
Table 9 shows the structure of a hypothetical HS grantee, based on our knowledge of coaching in the field.[45] As illustrated in this figure, there are usually multiple HS centers per grantee, and in

[42] These are the parameters based on the control group of centers only, with the adapted Teaching Style Rating Scale (TSRS) as the outcome of interest.
[43] For example, Mattera, Lloyd, Fishman, and Bangser (2013).
[44] Collins, Murphy, Nair, and Strecher (2005).
[45] This assumption is based on what we have seen when recruiting HS centers for other studies, as well as the review of coaching interventions prepared for this project (Taylor et al., 2013).

each HS center, there are about two to three classrooms on average, each with a lead teacher. Coaches are hired by the grantee. There are usually fewer coaches than centers, so coaches are often expected to work across multiple HS centers. Results from a recent study of coaches—the ELMC Study—showed that about three out of four coaches (73 percent) work across multiple centers.[46] In the illustrative example presented below for the HS Coaching Study, each coach (A and B) has a caseload of four centers with three teachers (classrooms) within each center.

Given the staffing structure of a HS grantee, there are three options for the unit of random assignment: coaches, centers, and teachers.[47] For the HS Coaching Study, we recommend that HS centers be the unit of random assignment rather than teachers, for several reasons.

Table 9. Staffing Structure of a Hypothetical HS Grantee



Center 1	Center 2	Center 3	Center 4	Center 5	Center 6	Center 7	Center 8
Classroom (Teacher)	Classroom (Teacher)	Classroom (Teacher)	Classroom (Teacher)	Classroom (Teacher)	Classroom (Teacher)	Classroom (Teacher)	Classroom (Teacher)
Classroom (Teacher)	Classroom (Teacher)	Classroom (Teacher)	Classroom (Teacher)	Classroom (Teacher)	Classroom (Teacher)	Classroom (Teacher)	Classroom (Teacher)
Classroom (Teacher)	Classroom (Teacher)	Classroom (Teacher)	Classroom (Teacher)	Classroom (Teacher)	Classroom (Teacher)	Classroom (Teacher)	Classroom (Teacher)

The first reason to make HS centers—rather than teachers—the designated unit of random assignment is to prevent spillover of the different coaching treatments across teachers. Because there are few teachers in each HS center (probably about two or three, based on the study team's experience), it is likely that teachers will interact with each other and talk about their experiences in the classroom. This means that if teachers were randomly assigned to receive different levels of the coaching dimensions, there could be spillover or contamination of the coaching treatments across teachers in a center. This spillover, in turn, would reduce the contrast between the

[46] Specifically, 27 percent of coaches worked in only one center, 39 percent worked in two to four centers, and about one third (34 percent) worked in five or more centers. On average, coaches worked across about four centers (Howard et al., 2013).

[47] Teachers are sometimes reassigned to different classrooms during the year, so the relevant unit of random assignment (and analysis) would be teachers rather than classrooms.

different types of coaching received by teachers and, therefore, the apparent effect of the coaching dimensions. If the reduction in the effect of the dimensions due to spillover is even small-to-moderate in size (between 0.01 and 0.06 standard deviations), then the sample size requirements would be the same for center-level random assignment and teacher-level random assignment.[48] Given that so little is known about the magnitude of spillover, especially in the context of a study of intervention, randomizing centers is preferable and will yield findings that are much more easily interpretable.

Second, randomizing centers (rather than teachers) is likely to be more acceptable to HS grantees. If teachers were the unit of random assignment, then the two or three teachers in a HS center could be assigned to receive different coaching treatments (e.g., different amounts of coaching). Because there are so few teachers per center, teachers would likely become aware of this difference in treatment, which could cause friction among staff.[49] Thus, randomizing centers (and having all teachers in a center receive the same treatment) could be more acceptable to HS grantees and facilitate site recruitment.

Third, randomizing centers (rather than teachers) will also reduce the TA and monitoring costs of the study. With center-level random assignment, all teachers in a HS center are assigned to receive the same treatment combination, which would make it easier to monitor whether coaches are providing the right services. In contrast, if teachers were the unit of random assignment, TA and monitoring efforts would be greater because the evaluators would have to ensure that *each teacher* in each HS center is getting the right level of the dimensions.

For all of these reasons, randomizing HS centers is recommended. As mentioned earlier, we are aware of only one other factorial experiment in the field of early childhood—a study by Landry and colleagues that used a 2x2 design to test of the effect of teacher PD components—and this study was able to successfully randomize 158 schools or centers to the conditions of the factorial experiment (Landry et al., 2009).

There is, however, one complication in the HS Coaching Study: The unit of random assignment for one of the three tested dimensions (*TRAINING*) will have to be coaches. Coach Training is a dimension that is delivered to the coach, which means that coaches will have to be randomly assigned to the levels of this coaching dimension (the unit of random assignment cannot be

[48] For example, let's assume that we want to power the study to detect an effect size of 0.20. If there were no spillover, then the sample size needed to achieve an MDES of 0.20 would be smaller if teachers were the unit of random assignment rather than centers. However, if there is spillover, then the sample size benefits of teacher-level random assignment are reduced or even eliminated. For example, let's assume that spillover is expected to reduce the effect of a dimension by 0.03 if teacher-level random assignment is used. This means that if teachers are the unit of random assignment, the study should actually be powered to achieve an MDES of 0.17 (= 0.20 − 0.03), which in turn will increase the sample size requirements compared to when there is no spillover. In contrast, if center-level random assignment is used, spillover is zero so the study can be powered to achieve an MDES of 0.20. Taking a step back, this means that if the spillover is large enough, the sample size needed under teacher-level random assignment could actually be the same (or larger) than under center-level random assignment. Thus, when comparing individual-level and cluster-level random assignment, it is useful to figure out how much spillover there would need to be for the sample size benefits of individual-level random assignment to cancel out (Rhoads, 2011). Based on the HS CARES parameter assumptions, center-level and teacher-level random assignment would "break even" if the spillover were 0.01 to 0.06 in magnitude.

[49] Shadish, Cook, and Campbell (2002) call this "resentful demoralization."

centers or teachers). In order to obtain an unbiased estimate of the effect of the Coach Training, it is important that coaches in Level I and Level II of *TRAINING* have similar characteristics. Randomizing coaches to Level I or Level II of *TRAINING* is the only way to ensure that this happens.

Thus, the unit of assignment will differ across the coaching dimensions: HS *centers* should be randomized to the levels of the *DOSAGE* and *RECIPIENT* dimensions, whereas *coaches* should be randomized to the levels of the *TRAINING* dimension. Because the unit of random assignment for the *TRAINING* dimension is at a higher level (coaches rather than centers), the MDES for this dimension will be larger than for the other two dimensions.[50] For this reason, in this section we assume that the study will be powered to detect an effect size of 0.20 on the *DOSAGE* and *RECIPIENT* dimensions rather than the *TRAINING* dimension. It is worth noting, however, that the random assignment plan that we recommend (described in the next section) also minimizes the MDES for the *TRAINING* dimension.

If Mode (*MODE*) were tested as a dimension instead of Coach Training (*TRAINING*), the unit of random assignment for this alternate dimension could be either centers or coaches. There are advantages and disadvantages to each type of unit:

- *Randomizing coaches to the levels of MODE:* In this scenario, half of the coaches would be randomly assigned to offer remote coaching with their centers, and the other half of coaches would be assigned to offer in-person coaching. The advantage of this approach is that only half of the coaches would have to be trained to deliver coaching remotely, which would make this dimension less costly to test. However, the disadvantage is that the MDES for the *MODE* dimension would be higher than for the other two dimensions given the sample size.
- *Randomizing centers to the levels of MODE:* In this scenario, the unit of random assignment for all three dimensions (*DOSAGE*, *RECIPIENT*, and *MODE*) would be centers. The advantages of this approach are that the MDES for the three coaching dimensions would be the same, and that the random assignment process would be much simpler. On the other hand, in this design, all coaches would have to be trained to deliver coaching remotely, which would increase the cost of the study.[51]

When discussing the random assignment plans in this next section, we further discuss the implications of using centers versus coaches as the unit of random assignment for *MODE*. Appendix D of this report also describes the implications of having *MODE* replace the *RECIPIENT* dimension rather than the *TRAINING* dimension.

[50] For a given sample size, the higher the level of random assignment, the greater is the MDES. See Appendix C for further discussion.

[51] If centers were randomized to delivery mode (*MODE*), then a given coach may have to implement "remote" coaching in some of the centers they serve, and "in person" coaching in their other centers, depending on the level of *MODE* to which a center has been assigned. This means that all coaches may have to implement remote coaching and be trained to deliver it.

Random Assignment Plans

In this section, we describe three random assignment (RA) plans. The choice of plan will depend on (a) whether one of the three dimensions is assigned at the coach level, and (b) whether a HS grantee would allow the evaluators to randomly assign (or reassign) its coaches to centers. As an overview, the three RA plans and their assumptions are as follows:

- **RA Plan 1:** This plan assumes that one of the three coaching dimensions is assigned at the coach level (whether *TRAINING* or *MODE*) and that a HS grantee would allow the evaluators to randomly assign (or reassign) its coaches to centers.
- **RA Plan 2:** This plan assumes that one of the three coaching dimensions is assigned at the coach level (whether *TRAINING* or *MODE*), but that coach assignments to centers would have to be taken as assigned by the grantee.

If one of the dimensions is assigned at the coach level, RA Plan 1 is preferred to RA Plan 2 (for reasons that will be discussed later). However, the study could include both types of grantee—some that allow the randomization of coaches (RA Plan 1) and some that do not (RA Plan 2).

- **RA Plan 3:** This plan assumes that none of the dimensions would be assigned at the coach level. This plan is only relevant if *MODE* is tested in the design instead of *TRAINING* and that the unit of random assignment for *MODE* is centers. As will be described later, the RA process would be much simpler in this situation.

For simplicity, in describing the three RA plans, we will refer to the hypothetical grantee in Table 9, which has two coaches who each serve four centers (for a total of eight centers in the grantee). However, the plans described in this section can be adapted to accommodate grantees with other staffing structures, such as a larger grantee or a grantee whose coaches have a smaller caseload. Random assignment under alternate staffing structures will be discussed in more detail later.

RA Plan 1: One of the Dimensions in the Design Is Assigned at the Coach Level and Coaches Can Be Randomized to Centers

Under RA Plan 1, the hypothetical HS grantee in Table 9 would allow the evaluators to change the set of centers that are assigned to Coach A and Coach B. Or similarly, the grantee would allow the study team to randomly assign newly hired coaches to centers. Random assignment would then proceed in two steps (shown in Table 10):

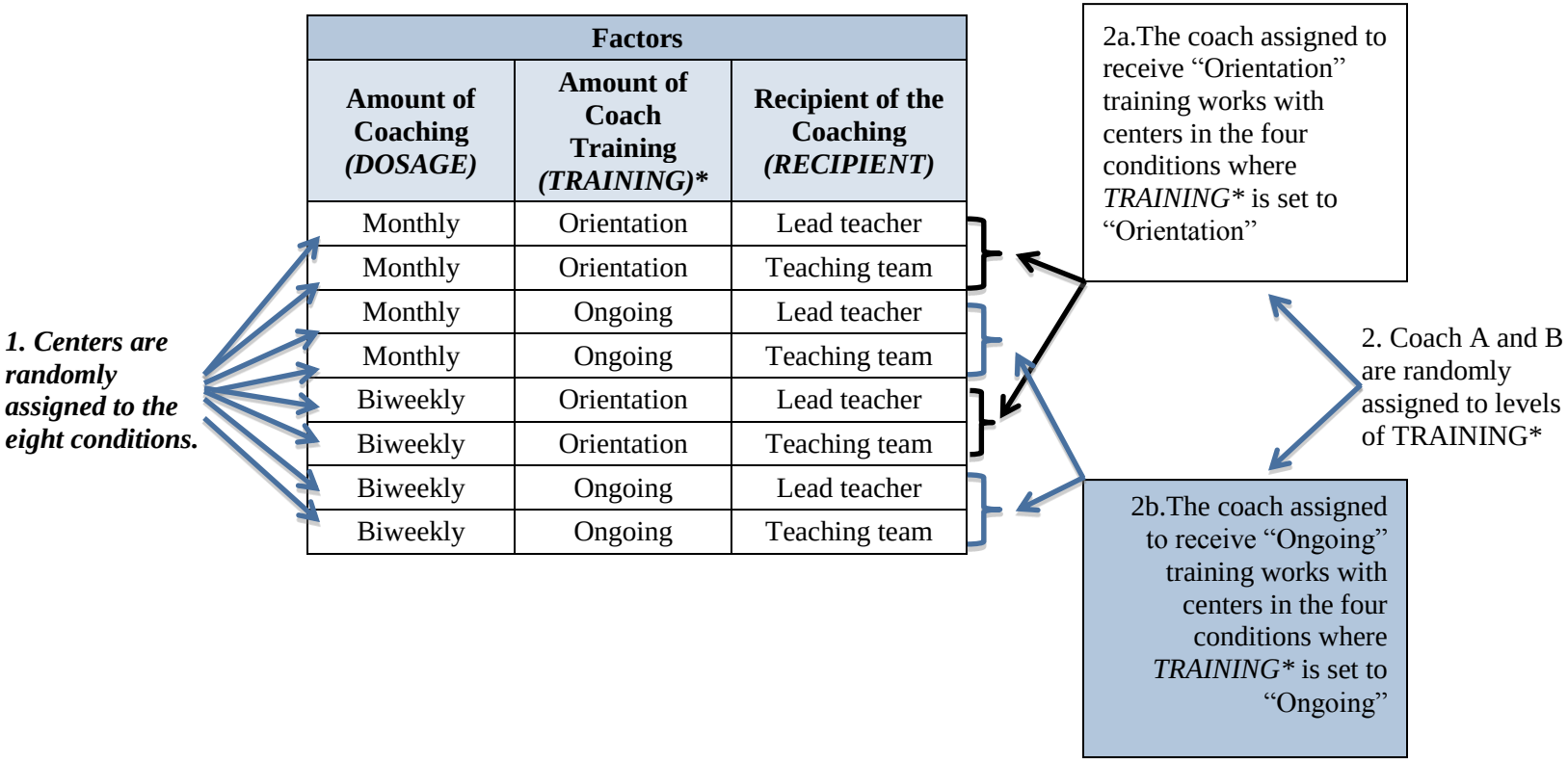
1. **HS centers would be randomly assigned to the eight possible combinations of the levels of *DOSAGE*, *RECIPIENT,* and *TRAINING* (or *MODE*).**

   For the hypothetical HS grantee in Table 9, for example, the grantee's eight centers would be randomly assigned to the eight experimental conditions representing all possible combinations of *DOSAGE*, *RECIPIENT*, and *TRAINING* or *MODE* (this design is shown in Table 10). One HS center would be assigned to each experimental condition. Based on the condition to which it is assigned, each center would receive a particular combination of the levels of *DOSAGE*, *RECIPIENT*, and *TRAINING* (or *MODE*).

2. **HS coaches would be randomly assigned to either Level I or Level II of the *TRAINING* (or *MODE*) dimension. Each coach would then work with the subset of HS centers that have been assigned to the experimental conditions where *TRAINING* (or *MODE*) is set to the same level.**

   For example, let's assume that *TRAINING* is the dimension being tested. For the hypothetical HS grantee in Table 10, Coach A and Coach B would be randomly assigned to receive either ongoing training or an orientation. The coach assigned to receive the orientation training (e.g., Coach B) would then work with the four centers assigned to the conditions where *TRAINING* is set to orientation, whereas the coach assigned to receive ongoing training (e.g., Coach A) would work with the four centers assigned to the conditions where *TRAINING* is set to ongoing. This is equivalent to having randomly assigned coaches to centers.

Table 10. RA Plan1: One of the Dimensions in the Design Is Assigned at the Coach Level and Coaches Can Be Randomized to Centers



* Or Mode (*MODE*), assuming that the unit of random assignment for this dimension is coaches. In this case, the levels to which coaches and centers would be assigned are "remote" or "in person."

Table 11 shows the number of HS centers needed to detect effect sizes of different magnitudes based on this random assignment plan. These calculations assume that grantees in the study would resemble the hypothetical grantee in Table 9 (two coaches per grantee, four centers per coach or eight centers per grantee, and two or three classrooms per center). The sample sizes in Table 11 are multiples of eight based on the assumption that centers would be randomized in blocks of eight centers. (See Table C5 in Appendix C for a more detailed table of MDES by sample size.) [52]

As shown in Table 11, if there were three classrooms per center and baseline outcomes data was not available, then a sample size of 248 centers would be needed to detect an effect size of 0.20 on the *DOSAGE* and *RECIPIENT* dimensions. If there were only two classrooms per center, then the sample needed to detect an effect size of 0.20 on the *DOSAGE* and *RECIPIENT* dimensions increases to 312 centers (an additional 64 centers). Table 11 also shows that the number of centers increases rapidly as the MDES decreases.[53]

Table 11. Sample Size Requirements Based on RA Plan 1

	No Baseline Outcomes Data				With Center-Level Mean CLASS Scores			
	DOSAGE and RECIPIENT		***TRAINING****		***DOSAGE and RECIPIENT***		***TRAINING****	
MDES	**(3)**	**(2)**	**(3)**	**(2)**	**(3)**	**(2)**	**(3)**	**(2)**
0.12	680	856	696	864	576	744	584	752
0.13	584	736	592	744	488	632	504	648
0.14	504	632	512	640	424	552	432	560
0.15	440	552	448	560	368	480	384	488
0.16	384	488	400	496	328	424	336	432
0.17	344	432	352	440	288	376	304	384
0.18	312	384	320	392	256	336	272	344
0.19	280	344	288	360	232	304	248	312
0.20	248	312	264	320	208	272	224	280

Note. These calculations are based on the ICC and R^2 from the HS CARES study (control group centers only, with an adapted version of the TSRS as the outcome measure). See Appendix C for parameter assumptions. The number in brackets (2 or 3) is the number of classrooms per center.
* Or *MODE*, assuming that the unit of random assignment for this dimension is coaches.

Importantly, under this random assignment plan, the MDES for the *TRAINING* (or *MODE*) dimension is only slightly higher than for the other two dimensions for a given number of centers. In some cases, the MDES for the two sets of dimensions is the same after rounding. For example, based on a sample of 248 centers and 3 classrooms per center, the MDES for the *TRAINING* or *MODE* dimension would also be 0.20 (the same as for the other two dimensions). Importantly, because *coaches are randomized to centers* in this plan, the MDES for the *TRAINING* (or *MODE)* dimension is minimized. As will be discussed later, if coaches cannot be

[52] The sample sizes in Table 11 are the smallest sample size needed to detect an effect size of given magnitude *rounded to the third decimal*.
[53] Because sample sizes are constrained to be multiples of eight centers, the increase in the sample size for a 0.01 decrease in the MDES is not always monotonic.

randomized to centers (and therefore coach assignments to centers must be taken as assigned by the grantee), then the MDES for the *TRAINING* (or *MODE*) dimension could be higher.

RA Plan 2: One of the Dimensions in the Design Is Assigned at the Coach Level but Coaches Cannot Be Randomized to Centers

In practice, some grantees may not agree to let the evaluators randomly reassign their coaches (or assign their newly hired coaches), yet it may be necessary to include these grantees in the study to meet the sample size targets. A grantee may not allow the random assignment of their coaches for several reasons. First, a grantee may serve a large geographic area, and therefore each coach can only travel to a specific subset of centers. Second, the grantee may feel that each coach has a working relationship with the centers that they already serve and that this relationship should not be disrupted. Third, grantees may want more experienced or skilled coaches to work with their weakest or most needy centers. For example, in Table 9, Coach A's four centers could be the strongest centers, and Coach B's centers could be the four weakest centers.

For grantees where coach assignments to centers must be taken as a given, the random assignment plan would be slightly more complicated (shown in Table 12):

1. **HS coaches would be randomly assigned to either Level I or Level II of the *TRAINING* (or *MODE*) dimension.**

   For example, if *TRAINING* were the dimension being tested, then Coach A and Coach B in Table 9 would be randomly assigned to receive either ongoing training or an orientation.

2. **Each coach's centers would then be randomly assigned to the four possible combinations of the levels of the *DOSAGE* and *RECIPIENT* dimensions.**

   Coach A's four centers would be randomly assigned to the four combinations of *DOSAGE* and *RECIPIENT*; the same process would then be repeated for Coach B.

As already explained, this random assignment plan is premised on the fact that grantees have particular preferences as to which coaches should be assigned to which centers. This raises two questions.

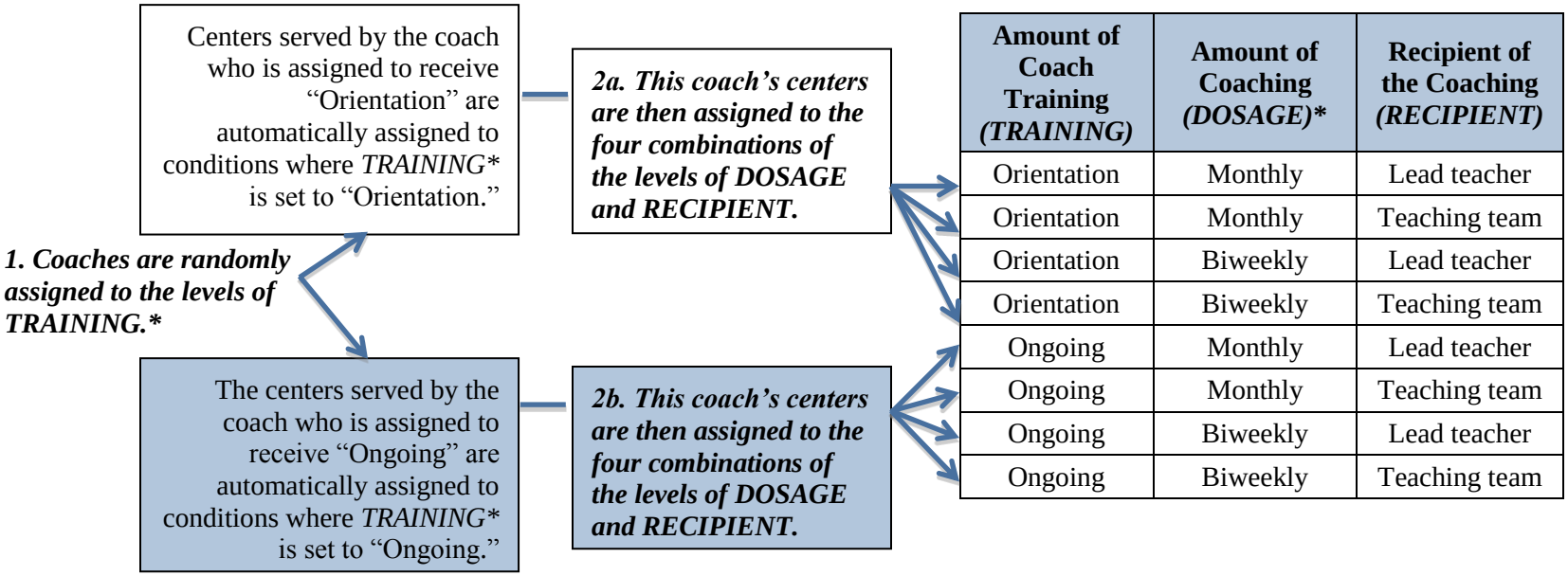
First, does the nonrandom assignment of coaches to centers compromise the internal validity of the estimated effect of the coaching dimensions? The answer is no, as long as grantees decide on their coach assignments prior to random assignment and do not change these assignments after random assignment. Under these conditions, random assignment should ensure that unobserved coach quality (as well as other center and teacher characteristics) is statistically equivalent in expectation at baseline across the two levels of each dimension and that estimated effects are internally valid.

Second, how does the nonrandom assignment of coaches affect the precision and therefore the MDES of estimated effects? Unfortunately, this is the main disadvantage of allowing grantees to choose coach assignments. As already noted, one would expect the MDES for dimensions assigned at the coach level (*TRAINING* or *MODE*) to be larger than the MDES for the other two dimensions because its unit of random assignment is coaches. Under RA Plan 1, the MDES for the *TRAINING* (or *MODE*) dimension is actually only slightly larger because coaches are

randomly re-assigned to centers. However, if coach assignment must be taken as assigned by the grantee, the MDES for the *TRAINING* or *MODE* dimension could be noticeably larger than for the other two dimensions. The amount by which it is larger depends on the extent to which coaches are assigned to *centers based on centers' average classroom (teacher practice) outcomes*. The more coach assignments are associated with a center's outcomes (for example, if one coach works with all the strongest centers while the other coach works with all the weakest centers), the larger the MDES for the *TRAINING* or *MODE* dimension will be for a given number of centers.

Table 13 looks at the MDES for the *TRAINING* (or *MODE*) dimension under different assumptions about the extent to which grantees assign coaches to centers based on their outcomes. As discussed in the previous section (RA Plan 1), the MDES for the *TRAINING* or *MODE* dimension would be 0.20 if coaches were randomized to centers and the sample was 248 centers (this is shown in the second column of Table 13). However, for this same number of centers, the MDES for the *TRAINING* (or *MODE*) dimension would increase to 0.22 or 0.24 if coaches were not randomly assigned and coach assignments were weakly or moderately associated with centers' outcomes (RA Plan 2).

Table 12. RA Plan 2: One of the Dimensions in the Design Is Assigned at the Coach Level but Coaches Cannot Be Randomly Assigned to Centers



Amount of Coach Training *(TRAINING)*	Amount of Coaching *(DOSAGE)**	Recipient of the Coaching *(RECIPIENT)*
Orientation	Monthly	Lead teacher
Orientation	Monthly	Teaching team
Orientation	Biweekly	Lead teacher
Orientation	Biweekly	Teaching team
Ongoing	Monthly	Lead teacher
Ongoing	Monthly	Teaching team
Ongoing	Biweekly	Lead teacher
Ongoing	Biweekly	Teaching team

* Or Mode (*MODE*), assuming that the unit of random assignment for this dimension is coaches. In this case, the levels to which coaches and centers would be assigned are "remote" or "in person."

Table 13. MDES for the Main Effect of *TRAINING Under RA Plans 1 and 2 (Three Classrooms per Center)**

	No Baseline Outcomes Data			With Center-Level Mean CLASS Scores		
Number of Centers	Random Assignment of Coaches to Centers (RA Plan 1)	Assignment of Coaches to Centers Is Weakly Associated With Center-Level Outcomes (RA Plan 2)	Assignment of Coaches to Centers Is Moderately Associated With Center-Level Outcomes (RA Plan 2)	Random Assignment of Coaches to Centers (RA Plan 1)	Assignment of Coaches to Centers Is Weakly Associated With Center-Level Outcomes (RA Plan 2)	Assignment of Coaches to Centers Is Moderately Associated With Center-Level Outcomes (RA Plan 2)
200	0.23	0.25	0.27	0.21	0.22	0.24
208	0.22	0.24	0.26	0.21	0.22	0.23
216	0.22	0.24	0.26	0.20	0.21	0.23
224	0.22	0.23	0.25	0.20	0.21	0.23
232	0.21	0.23	0.25	0.19	0.21	0.22
240	0.21	0.22	0.24	0.19	0.20	0.22
248	0.20	0.22	0.24	0.19	0.20	0.21

Note. These calculations are based on the ICC and R^2 from the HS CARES study (control group centers only, with an adapted version of the TSRS as the outcome measure). See Appendix C for parameter assumptions. We further assume that there will be three classrooms per center and four centers per coach.
* Or *MODE*, assuming that the unit of random assignment for this dimension is coaches.

Thus, to minimize the MDES for dimensions assigned at the coach level (*TRAINING* or *MODE*), to the extent possible the study should include grantees that meet one of the two following criteria: (1) the grantee would allow the randomization of its coaches to centers or (2) coach assignments by the grantee appear to be uncorrelated with centers' outcomes (quasi-random assignment of coaches to centers). Existing center-level CLASS scores from prior school years could be used to determine whether a grantee meets the latter criterion.

Alternate Staffing Structures

The random assignment plans just described (RA Plans 1 and 2) can be adapted to staffing structures that are different from the one shown in Table 9:

- *Larger Grantees.* If a grantee has more than eight centers per grantee (say, 16 centers which is a multiple of 8), then there are two options with respect to the random assignment process. If the centers were similar in terms of their outcomes (for example, they had similar CLASS scores in the previous year), then the random assignment process would be similar to what has already been described, except that *two* centers would ultimately be assigned to each experimental condition. If the 16 centers are different from one another in terms of their outcomes, then the grantee's centers could be subdivided into two groups based on their prior outcomes (the eight strongest centers and

the eight weakest centers), and random assignment could be further blocked by these two subgroups.[54]

- *Smaller Grantees.* In practice, one could also block by a group of two similar small grantees. For example, it might be possible to recruit smaller grantees (say, two grantees with four centers and one coach each) and to combine them into a random assignment block, as long as the two grantees are similar with respect to their outcomes.[55]
- *Smaller Caseload*. The random assignment plans described earlier assume a caseload of four centers (that is, there are two coaches per grantee or block, and each coach works with four centers, for a total of eight centers, as in Table 9). However, the random assignment plans can be modified to accommodate a situation where a grantee's eight centers are distributed across a larger number of coaches (i.e., four coaches with a caseload of two centers each).[56]

Also, the study could include grantees that allow the randomization of coaches (RA Plan 1) and others that do not (RA Plan 2). In this mixed scenario, the MDES for the dimension assigned at the coach level (*TRAINING* or *MODE*) would be higher than the MDES under a pure RA Plan 1 scenario, but lower than the MDES in a pure RA Plan 2 scenario.

RA Plan 3: All Three Dimensions in the Design Are Assigned at the Center Level (Applicable for a Design in Which *MODE* is Tested Instead of *TRAINING*)

If Mode (*MODE*) were tested in the study rather than Coach Training (*TRAINING*)—and random assignment to *MODE* was at the center level—then the random assignment plan would be much simpler. In the first instance, random assignment would no longer be constrained by having to randomize coaches to the levels of one of the dimensions – HS centers would be the only unit of random assignment. Furthermore, it would no longer be as distinctly advantageous to recruit grantees that would allow the evaluators to randomize their coaches to centers (as in RA Plan 1), because none of the dimensions would be assigned at the coach level.[57]

Thus, there would only be one random assignment plan, and it would include only one step: A grantee's centers would be randomly assigned to one of the eight experimental conditions. In addition, coach assignments could be taken as assigned by the grantee. Each center would be randomly assigned to one of the eight experimental conditions, and the coach who serves that center would provide the center with the levels of the *DOSAGE*, *RECIPIENT*, and *MODE*

[54] This latter option would require that there also be at least four coaches per grantee (two coaches per smaller block), to ensure that in each block there is a coach assigned to each level of the *TRAINING* dimension.
[55] Under this approach, RA Plan 2 would have to be used because, by definition, it would not be possible to randomize coaches to HS centers (because a coach cannot typically work with centers that are located in another grantee). For this reason, it would be very important to combine only grantees that are similar to each other with respect to their average outcomes. If they are not similar, this could increase the MDES because it would approximate a situation in which coaches are assigned to centers based on centers' teacher practice outcomes (the two last columns of Table 13).
[56] This would simply add additional steps to the random assignment process. For example, RA Plan 1 would include an additional step that would randomize the four coaches to the eight centers given their assigned *TRAINING* level.
[57] As explained earlier in this section (under RA Plan 1 and 2), if one of the dimensions is assigned at the coach level, then randomizing coaches to centers (RA Plan 1) has the benefit of minimizing the MDES for that dimension. However, if all of the dimensions are assigned at the center level, then this is not relevant.

dimensions to which it has been assigned. To maintain the internal validity of the experimental design, however, grantees would have to agree to *not* change coach assignments after randomization.[58]

If the unit of random assignment for the *MODE* dimension were centers, the MDES for this dimension (as well as the *DOSAGE* and *RECIPIENT* dimensions) would be very similar to the MDES for the center-level dimensions in Table 11.[59]

Summary of Recommendations and Implications for Site Recruitment and Monitoring

On the basis of the issues just laid out, we can make the following general recommendations related to the random assignment plan and the sample size for the HS Coaching Study:

- *Unit of Random Assignment.* We recommend that cluster random assignment should be used for the HS Coaching Study, where the unit of random assignment is HS centers for the *DOSAGE* and *RECIPIENT* dimensions, coaches for the *TRAINING* dimension, and either centers or coaches for the *MODE* dimension.
- *Number of Classrooms per Center.* To minimize the number of centers needed to detect an effect size of 0.20, HS centers in the study should have at least three classrooms per center on average (based on the harmonic mean). To achieve this goal, one of the conditions for study participation could be that a center must have three classrooms at minimum. Alternatively, centers with only two classrooms could be allowed to participate, but a subset of recruited centers would then have to have more than three classrooms per center to make sure that the harmonic mean across the entire sample is three classrooms.
- *Recruitment and Random Assignment Plan.* If one of the dimensions in the study is assigned at the coach level (*TRAINING* or *MODE*), then the evaluators should recruit grantees that would allow them to randomly assign (or reassign) coaches to centers (RA Plan 1). If this is not feasible for all grantees, then the evaluators could also recruit grantees that appear to allocate coaches to centers quasi-randomly or based on some criterion that is not associated with teacher outcomes in the centers ("quasi-random" assignment) and use RA Plan 2 to assign coaches and centers to conditions.[60] (More generally, as part of the pilot work conducted for the study, the evaluators should examine the extent to which grantees would allow the randomization of coaches.) If none of the dimensions in the design are assigned at the coach level (RA Plan 3), then it would

[58] This random assignment plan can be adapted to alternate staffing structures (larger grantees or smaller grantees, and varying caseloads).

[59] In practice, it would slightly smaller because degrees of freedom would no longer need to be used by having to include coach random-effects in the model. However, the difference in the MDES would not be noticeable when rounded to the third decimal. For example, for 248 centers and three classrooms per center, the MDES would be 0.19989 rather than 0.20002. See Appendix C for further discussion.

[60] In order to recruit grantees where coach assignments are "quasi-random," the evaluators would need to talk to grantees about how they assign coaches to centers and obtain CLASS data from prior school years to confirm that coach assignment are not associated (or only weakly associated) with centers' outcomes. If the random assignment of coaches to centers were random or quasi-random, the MDES for a dimension assigned at the coach level (*TRAINING* or *MODE*) would be only slightly higher than the MDES for the two other coaching dimensions.

not be necessary to randomize coaches to centers or to recruit grantees where coach assignments are quasi-random.

- *Number of Centers.* The study should plan to recruit at least 248 centers. This would make it possible to detect a main effect of 0.20 on dimensions assigned at the center level (*DOSAGE*, *RECIPIENT*, and perhaps *MODE*) even if existing data on center-level CLASS scores from the prior school year were not available. With 248 centers, it would also be possible to detect a main effect of 0.20 or only slightly higher on dimensions assigned at the coach level (*TRAINING* or *MODE*).[61]
- *Blocking.* Random assignment should be blocked by grantee (or by groups of eight similar centers) because this will improve the precision of estimated effects and reduce the sample size requirements for the study, even if existing center-level CLASS scores are not available.

Following from these recommendations related to random assignment and the sample size, we can also make the following recommendations related to the analysis:

- *Interaction Effects*. When interaction effects are defined as in Section IV, the MDES for a two-way interaction is two times larger than the MDES for the main effect of a dimension. For example, if the MDES for a main effect is 0.20, then an interaction effect would have to be 0.40 to be statistically significant.[62] For this reason, the analysis of interactions should be considered exploratory.
- *Subgroup Effects*. The effect of the coaching dimensions for subgroups defined by center or grantee baseline characteristics (e.g., geographic location, organizational characteristics) will be difficult to detect in the HS Coaching Study unless they are larger in magnitude. The MDES for a subgroup based on approximately half of the centers in the sample (128 centers) would be 0.28 for the *DOSAGE* and *RECIPIENT* dimensions. For subgroups defined by classroom or teacher baseline characteristics, the MDES would be more acceptable: It would be 0.22 for a sample of two thirds of the classrooms (two classrooms per center in each of the 248 centers) and 0.28 for a sample of one third of the classrooms (one classroom per center in each of the 248 centers).[63] However, these MDES are "best case" scenarios because they assume that each center will have a classroom with the characteristic of interest. In reality, it is likely that some centers will *not* include a classroom with the characteristic of interest and that these centers will be excluded from the subgroup analysis. This, in turn, will reduce the number of centers in the analysis, which is a key factor for the MDES (the number of centers is more important than the number of classrooms per center). For these reasons, we recommend that any subgroup analyses, if conducted, should be considered exploratory.

[61] Because this is a one-year study, it is unlikely that centers would withdraw from the study. However, even if a grantee (eight centers) dropped out, the study would still be able to detect an effect of 0.204 on the *DOSAGE* and *RECIPIENT* dimensions and an effect of 0.19 if CLASS scores could be used as baseline measures.

[62] The MDES for a two-way interaction effect is twice as large because (a) the sample must be divided into two subgroups to estimate the effect of the first dimension at each of the two levels of the second dimension, and (b) the standard error of the *difference* in effects between these two subgroups (which represents the definition of an interaction) is larger than standard error of the effect for each subgroup.

[63] This assumes no pretests.

The process of random assignment will impose several constraints on-site recruitment. The first is a general constraint:

- *Number of Centers per Grantee.* There should be at least eight centers per random assignment block (per grantee or per group of similar centers), so that in each block there is at least one center in each of the eight experimental conditions.

The second constraint would be relevant if one of the dimensions in the study was assigned at the coach level (*TRAINING* or *MODE;* RA Plan 1 or 2):

- *Number of Coaches per Grantee.* There will need to be at least two coaches in each random assignment block (per grantee or per group of similar centers) so that at least one coach is assigned to each level of *TRAINING* in the random assignment block.

The third constraint would apply if the study included grantees where coach assignments to centers could not be changed (RA Plan 2):

- To maintain the internal validity of the experimental design, grantees would have to agree that they would *not* change coach assignments after randomization.

More generally, we assume that there will be no (or minimal) turnover of staff in the study, because the study is only one academic year. If a teacher left the study, the coach would work with the replacement teacher and the classroom would remain in the study. If a coach left the study, a replacement coach would have to be hired and given the right training. To verify the causal validity of the results and interpret the findings, it would be important to examine that staff attrition is statistically similar across the levels of each coaching dimension.

Finally, it is important to emphasize that strong TA and monitoring will be needed to maintain the integrity of the recommended random assignment plan. Specifically, it will be important to monitor the coaches' activities to make sure that the service contrast is being maintained. As already noted, coaches work across multiple centers. This means that each coach will have to implement both levels of the *DOSAGE* and *RECIPIENT* dimensions, based on the experimental condition to which a given center has been assigned. (If the third dimension was *MODE* and this dimension was assigned at the center level, then a coach would also have to offer both levels of the *MODE* dimension.) If a coach does not provide the right levels of these dimensions to teachers, this would create spillover across centers, which would dilute the service contrast and, by extension, the expected effect size.[64] Fortunately, we expect this type of spillover to be minimal because coaches should be able to vary the levels of *DOSAGE* and *RECIPIENT* (and *MODE*) with training and TA from the evaluators. In this regard, it will be extremely important to train the coaches carefully so that they understand the study design and how important it is to adhere to it. Monitoring and TA strategies are discussed in Section X of this report.

[64] This is different from spillover caused by *teachers* discussed earlier in this section, which would be a potential threat if the unit of random assignment were teachers rather than centers.

VI. Impact Analyses

In this section, we outline various issues to consider with respect to the estimation of effects and hypothesis testing, including: how to account for the features of the design when estimating effects, imputation and missing data, accounting for crossovers across experimental conditions, and multiple hypothesis testing.

Accounting for the Features of the Design

As explained in Section IV, the main effect of each dimension can be estimated simply by comparing the outcomes of HS centers and classrooms in different groupings of the experimental conditions (Table 8). Thus, the estimation of effects is simple in theory.

However, the estimation of effects must also account for the way in which random assignment was conducted, as well as the possible use of covariates:

- *Cluster Random Assignment.* As discussed in Section V, the random assignment plans for the HS Coaching Study assume that HS centers and coaches will be randomized to the eight experimental conditions. Thus, the analysis (and specifically, the standard errors used for hypothesis testing) must account for the fact that cluster random assignment was used. Specifically, the standard error of estimated effects for dimensions assigned at the coach level (*TRAINING* or *MODE*) should account for clustering at the center level and at the coach level, while the standard error of estimated effects for dimensions assigned at the center level should account for clustering at the center level. Otherwise, the standard errors from the analysis will be too small and one could conclude that an effect is statistically significant when in fact it is not.
- *Blocking.* As discussed in Section V, the precision of the estimated effects can be improved by blocking random assignment. For this reason, the sample size recommendations in Section V assume that random assignment will be blocked. Thus, in order to be able to detect effects of 0.20 with a sample of 248 centers, this feature of the design must be incorporated into the analysis.
- *Baseline Covariates.* As discussed in Section V, adjusting the estimated effects for random between-center differences in center-level CLASS scores in prior school years could further decrease the MDES (it would decrease to 0.19). Therefore, incorporating these scores into the analysis may be advantageous.

To account for these design features, the analysis of a factorial experiment can be set up as a regression model. For example, one could use a multilevel regression model to obtain main effects and interactions, which would make it possible to control for blocking and CLASS scores, while also providing the right standard errors. A second approach would be to use an ordinary least squares (OLS) regression, which would make it possible to control for blocking and CLASS scores. The OLS standard errors could then be corrected for clustering (for example, by using cluster-robust standard errors). Appendix E provides examples of regression models based on the first of these two approaches.

Testing for Baseline Equivalence

In a factorial experiment—as in other experimental designs—it is important to verify that random assignment resulted in experimental groups that are statistically equivalent in terms of their baseline characteristics. As noted earlier, in the HS Coaching Study, baseline characteristics will likely be measured at the center level—for example, center-level CLASS scores and other characteristics that are available from centers' administrative records. To test whether there is balance between experimental groups with respect to these center-level characteristics, one would use the same type of approach as is used for estimating main effects—that is, by comparing the characteristics of centers across different groups of experimental conditions (see Table 8). Ideally, the main "effect" of each dimension on each characteristic should be close to zero and not statistically significant, which would indicate that centers have similar characteristics on average across the experimental conditions.

Main Effect of the Treatment on the Treated (TOT)

As will be explained in Section X, the goal of monitoring and TA for the HS Coaching Study will be to make sure that HS centers and coaches are receiving the level of services to which they are randomized and that the contrast between the Level I and Level II of each dimension is strong. However, some centers may not receive the intended level of services; for example, a center assigned to receive biweekly coaching may actually receive less coaching because their coach was ill, or a center assigned to receive monthly coaching somehow receives more frequent coaching. In other words, some centers may take on characteristics of another experimental condition, thus weakening the intended contrast to be measured.

If this happens, then estimates of main effects from Model 1 will represent intent-to-treat (ITT) estimates of the effect of randomly assigning a center or coach to a particular level of that factor, rather than the effect of receiving the enhanced level (Level II) of that factor (also called treatment-on-the-treated [*TOT*]). In many ways, ITT estimates of main effects are more policy relevant because coaching services can be offered but their levels are difficult to enforce in a real-world setting. However, TOT estimates (the effect of receiving Level II of a coaching dimension) are more interesting from a practitioner's perspective. Under certain assumptions, TOT estimates can be obtained by adjusting the ITT estimates upward to account for the percentage of crossovers (Bloom, 1984). As an exploratory analysis, the study could present estimates of TOT main effects (assuming that all conditions for such an analysis are met).[65]

Missing Data and Imputation

There are multiple ways to handle missing data on the outcomes of interest. In the HS Coaching Study, for example, multiple imputation methods could be used to impute missing teacher and classroom outcomes. Alternatively, the analysis could be based on classrooms for which outcome data are available. There are pros and cons to these two approaches, both of which have been used in federally funded evaluations (see Puma et al., 2009, for an overview). For this

[65]Making crossover adjustments increases the magnitude of estimated effects, but it does not affect the *p* value of estimated effects because the standard error of the estimated effect is adjusted upward by the same amount. Therefore, the purpose of the crossover adjustment is simply to change the interpretation (and therefore magnitude) of the estimated effect.

reason, we do not make specific recommendations in this report and simply point out that the choice of approach needs to be carefully weighed prior to analysis.

However, the design team recommends that missing data on covariates—such as classroom and teacher characteristics—be imputed. In a random assignment study, the purpose of covariates is simply to improve precision (rather than to control for bias), so there are fewer (if any) drawbacks to imputation. For example, a method often used in randomized experiments is the dummy variable approach. This consists of imputing the missing values in a variable with a constant, such as the grand mean for the sample, and then including an indicator of missingness of this variable in statistical models that include the variable in question as a covariate. Puma, Olsen, Bell, and Price (2009) and internal studies by the design team have demonstrated that, with low rates of missing data, this approach yields unbiased estimates of program impacts in cluster randomized experiments. These findings should also apply to clustered factorial experiments. A more complex imputation method, such as multiple imputation, could also be used.

Multiple Hypothesis Testing

In the HS Coaching Study, the number of hypothesis tests to be conducted is large because (a) the study can look at the main effect of multiple dimensions and estimate interaction effects between these factors and (b) main effects and interactions can be estimated for multiple classroom and teacher outcomes. Conducting many hypothesis tests increases the probability of a Type I error (i.e., concluding that a coaching dimension improves teacher and classroom outcomes when in fact it does not). For example, let's assume that the Type I error rate used for hypothesis testing is 10 percent. If there are multiple estimates of effects, then the actual Type I error rate would be greater than 10 percent. In a two-group RCT, one approach for dealing with this issue is to adjust the Type I error rate downward (or the p-values upward) to account for multiple hypothesis testing.[66]

However, whether (and how) to make multiplicity adjustments in the Optimization Phase of the MOST framework (and the HS Coaching Study) is still an open question. In the HS Coaching Study, the goal will be to help practitioners and policymakers develop stronger coaching interventions. Given this exploratory objective, the Type I error rate (the risk that practitioners will use a dimension that is not effective) is perhaps less important that the Type II error rate (the risk that practitioners will not use a dimension that is effective). Viewed in this light, achieving a low Type I error rate in the Optimization Phase is perhaps not as crucial as it would be in the Evaluation Phase of MOST (where the impact of an optimal intervention is tested relative to a control condition). Therefore, having a Type I error rate that is somewhat higher than 10 percent may be acceptable, because the findings from the Optimization Phase can be considered exploratory.

For these reasons, we do not recommend using statistical adjustments to address the issue of multiple hypothesis testing in the HS Coaching Study. However, given the many outcomes and types of effects that will be estimated, we do recommend prioritizing the results and limiting the number of estimates that are used for decision-making purposes. Otherwise, the results will be

[66] See Schochet (2008) for a review of alternative methods of making these adjustments.

difficult to interpret and too overwhelming in number to utilize effectively. A useful strategy in this regard is to classify effect estimates as either primary or secondary.[67] In the HS Coaching Study, primary effects are those that would be used to make statements about whether a coaching dimension improves teacher and classroom outcomes, whereas secondary effects would be presented for descriptive or hypothesis-generating purposes only. As a concrete example, the main effect of the three dimensions could be considered primary, and interaction effects could be considered secondary.[68] Similarly, a subset of teacher and classroom outcomes could be designated as primary, whereas effects on other outcomes could be examined for descriptive purposes only.[69] Limiting the number of estimates that are used decision-making (versus descriptive purposes) will prevent the Type I error rate from being unacceptably high and will help policymakers and practitioner utilize the findings more effectively.[70]

[67] This classification would happen before conducting the analysis.

[68] The main effect of *TRAINING* (or *MODE*, if it was assigned at the coach level) could also be considered secondary if the evaluators cannot recruit grantees where coaches are sorted randomly—or close to randomly—across centers (in which case the MDES for the *TRAINING* or *MODE* dimension would be higher than for the other two dimensions).

[69] To choose the primary outcomes, a useful first step would be to start by grouping the outcomes of interest into domains. For example, measures of teacher practice and teacher-child interactions might naturally group into the following four domains: instructional support, emotional support, classroom organization, and specific language practices related to environmental supports and classroom practices and routines. One or two of these domains could be designated as the primary domains of interest. One outcome measure from each primary domain could then be designated as a primary outcome (which would result in one or two primary outcomes).

[70] A similar approach is used by the What Works Clearinghouse (WWC) at the Institute of Education Sciences (2014) to review (and rate) the rigor of impact evaluations in the field of education.

VII. Evaluation Components to Complement the Impact Study

In this section, we describe the design team's suggestions for the implementation research (IR) and cost study portions of the Head Start (HS) Coaching Study. For each of these components, we provide a brief overview of the goals, the research questions, and the approach. For the IR and cost study portions, we do not intend to provide the same in-depth level of preparation or analysis plan as we did for the impact study in previous sections. This is not, however, to deemphasize their importance because the design team strongly advocates for the inclusion of these two components to complement the impact findings. Section VIII (Measures) provides more detailed information about the specific proposed constructs and data sources that would be important to include in the evaluation IR and cost components of the study.

Implementation Research

Goals for the IR Agenda

IR helps document the extent to which the intervention—and, in this case, the multiple conditions of the coaching intervention—were implemented as intended. IR identifies factors that may facilitate and challenge execution of the intervention or evaluation design and that further contextualize the resulting impacts.

For the HS Coaching Study, we recommend IR goals to (1) describe and assess the fidelity of implementation for the eight experimental coaching conditions in order to help interpret impacts. The implementation data will be used to analyze the implementation of the foundational coaching approach; fidelity, content, and characteristics of the systematically varied coaching dimensions (including the strength of the contrast between levels); and the language content that coaches deliver to teachers. In addition, the IR will (2) inform future development of effective and feasible coaching models. IR will help identify common contextual factors across coaches, teachers, classrooms, sites, and dimensions that facilitate or impede effective implementation. It also includes examination of the nature and extent of professional development (PD) for the coaches as well as the technical assistance (TA) provided to the coaches, teachers, classrooms, and centers. This information will aid HS in formulating future coaching strategies.

These goals are aligned to the three study research questions related to understanding implementation of the HS Coaching Study, first presented in Section II:

3. Are the different coaching dimensions implemented with fidelity?
4. What factors facilitate or challenge the fidelity of implementation of the different coaching variations? What types of TA and PD tools facilitate the implementation fidelity?
5. How does implementation vary across grantees' program environments, populations, and other contextual program features?

The HS Coaching Study is designed to test the impact of varying coaching dimensions on teacher and classroom outcomes, within the context of implementing a foundational coaching model. As elaborated on in Section IV, the foundational coaching model includes the following:

- Dimensions that are specified as *fixed*
- Dimensions that must meet a *minimum threshold*
- Dimensions that can vary as they typically do among grantees

Documenting the foundational coaching model (including the implementation of the language content of the coaching) and the three systematically varied dimensions will be important for this study. The evaluator of the HS Coaching Study will first need to understand fidelity (i.e., the extent to which the coaches and teachers implement the levels of the targeted three dimensions—Dosage, Recipient, and Coaching Training—to which they were assigned). The evaluator will also need to document the extent to which coaches and teachers adhere to the dimensions that are fixed and the natural variation across the teachers and coaches for other dimensions. It is expected that the package of systematically varied, fixed, and varying (typical practice) dimensions will help explain the study's impacts and feasibility of implementation.

IR Construct Definitions and Illustrative Analyses

Fidelity of Coaching. Fidelity refers to whether the variations in the tested coaching dimensions are executed as designed and delivered in a clear and comprehensible manner. There may be several aspects of fidelity that may be of interest, including adherence, Dosage, exposure, and responsiveness.

We recommend that logs documenting Dosage, content, and implementation of coach and teacher practices be the primary data sources for investigating intervention fidelity. The logs will be completed by a variety of informants (see Measures, Section VIII) and will be housed in an online management information system (MIS). The log questions will be primarily closed ended and scaled in order to provide basic descriptive statistics of constructs and measures as well as comparisons across informants. For example, we recommend that fidelity of implementing the intended coaching models be assessed by coaches, and for a subset of coaching sessions, by both coaches and PD trainers.

Content and Characteristics of Coaching. A major responsibility of coaches is to encourage and support teachers' use of specific language practices and skills in their classrooms. As part of this work, coaches will engage in a variety of skills, strategies, and approaches with teachers, some of which are part of the foundational coach model and some of which will be systematically varied. We expect that changes in teacher practices are, in part, a result of how they implement these skills, strategies, and approaches.

Understanding the content and characteristics of coaching sessions will help us get inside the coaching "black box". While we expect to be able to document the content of the coaching sessions, measuring the quality of early childhood coaching is not very well developed. However, given that one goal of this study is investigating inside the black box of coaching, it is important to go beyond the logs. Both surveys and independent observations will be used to document the content and characteristics (including quality) of coaching. Surveys can measure

aspects such as the extent to which coaches and teachers liked training or whether they think they have learned a particular skill. Understanding the process of coaching is more complicated. For this, we propose using direct coaching observation protocols. When documenting the process of coaching, not only does the content need to be described (that is, what coaches are doing with teachers), but we also need to know how it is being done—behaviors that are best understood through observation.

Context of Coaching. Coaching will take place in a variety of settings and cannot be separated from the context in which it is embedded. Although the coaching intervention supports the implementation of teacher practices, the organizations and institutions where the coaches and teachers are housed, as well as individual characteristics of coaches, teachers, administrators, or other stakeholders, can affect the implementation of the intervention. Organizational contexts and individual characteristics can influence the rewards, sanctions, resources, and norms that support or hinder coaching implementation. In short, coaching and coaching activities and dimensions are embedded into a larger system that can impact intervention implementation.

The contextual moderators of coaching interact in a reciprocal manner, meaning they influence and impact each other. For instance, a setting with designated space and time set aside for coaching might facilitate more reflective coach-teacher interactions than a setting without these supports. In another example, coach characteristics such coach background—level of education, years of coaching experience, credentials, sense of efficacy, views about language practices, language knowledge, flexibility— might affect how well coaches respond to the Coach Training, and how well they implement their assigned coaching program model (i.e., experimental condition) in the face of challenging situations such as less than adequate space or time. Reporting on context and analyzing the relationships between contextual moderators of coaching should be approached with an eye toward documenting intervention patterns and variation across conditions, participants, or settings.

Table 14 lists illustrative examples of implementation features relating to fidelity, content and characteristics, and context of coaching that should be documented as part of the process to answer the research questions above.

Table 14. Example Implementation Features, by Coaching Dimension

Coaching Approach and Dimensions	Fidelity and Contrast Between Levels	Content	Characteristics	Context
Foundational Coaching Approach	▪ Were the predefined goals for the foundational coaching approach met? ▪ Did foundational coaching activities or content vary based on the coaching conditions assigned to the coach?	▪ What activities occurred, and what content was covered in the coaching sessions?	▪ Were the foundational strategies implemented with sufficient quality to result in teacher change?	▪ Did coach, teacher, organizational characteristics (e.g., type of supervision, teacher experience), or other factors (e.g., resources in classroom) affect the way that the foundational coaching approach was implemented?
Dimension: Dosage of Coaching	▪ Was the Dosage of coaching prescribed in the research design received by teachers? ▪ What was the variation in the Dosage? ▪ Was the Dosage of coaching received by teachers sufficient to ensure that there was a distinct contrast between Level I and Level II conditions?	▪ Did teachers in different Dosage conditions receive different content in their coaching sessions?	▪ Was there a difference in coaching strategies used under different Dosage conditions? ▪ Did teachers respond differently to coaching?	▪ What were the challenges (e.g., finding meeting time and place) and facilitators (e.g., support of administrators) related to Dosage receipt?
Dimension: Recipient Targeted by Coaching	▪ Did coaches meet with the lead teacher or with both the lead teacher and assistant teacher together, as appropriate to their assigned condition?	▪ What was the content of coaching received by the lead teacher or the lead and assistant teacher together? ▪ Was there variation in coaching content received by teachers by assigned dimension?	▪ What were the characteristics of the interactions between the lead teacher and the lead and assistant teacher together?	▪ What were the challenges and facilitators related to coaching teachers individually or together?
Dimension: Training of Coaches	▪ Was the Dosage of training prescribed in the research design received by coaches? ▪ What was the variation in the Dosage? ▪ Was the Dosage of training received by coaches sufficient to ensure that there was a distinct contrast between Level I and Level II conditions?	▪ What was the content of training received by coaches? ▪ Was there variation in training content received by coaches by assigned dimension?	▪ Did coaches with more or less training use different coaching strategies? ▪ Were there discernible differences in coaching characteristics between the two groups?	▪ What were the challenges and facilitators related to training coaches?

The next section will define what we mean by fidelity, content, characteristics, and context of the coaching intervention within the HS Coaching Study. We also provide some illustrative examples of how we might not only define but also consider analyzing these constructs within the study. For more in-depth detail about the specific instruments and measurement strategies that are being recommended for the IR, please refer to Section VIII.

Data Collection Approach

Our recommendation for the IR to collect important information relating to fidelity, content, characteristics, and context to answer the research questions listed earlier includes a fundamental set of principles to guide the work. These principals include the following:

- **Maximization of Staff and Fiscal Resources by Using Data Collected for Multiple Purposes.** To the extent possible, the IR component should use data that will already be collected as part of the impact study and as part of the TA and monitoring efforts that must be undertaken (see Section VI). The development of an MIS to store data will be an important resource for building in these efficiencies.
- **Use of Mixed-Methods Strategies.** Qualitative and quantitative data collection strategies should be used to facilitate the understanding of implementation. For example, quantitative data will provide measures of Dosage, topics covered, or use of different coaching strategies. Qualitative data will provide information on reactions of teachers and coaches to the coaching process and will identify processes that worked well and those that were challenging.
- **Documentation and Analysis of Critical Facets of Implementation.** Fidelity, characteristics (including quality), and content are key measures commonly used to assess implementation of interventions in early childhood. We recommend multiple data sources (MIS to log attendance, ratings, and content of sessions in addition to interviews and Web-based surveys) be used with multiple participants (coaches, teachers, grantee staff, and developers and trainers) involved in the study.

These data, when coupled with the factorial design, add a level of rigor and complexity that is not typical of many studies that involve PD. Section VIII (Measures) provides additional details about the specific data collection strategies and sources for the IR components of the HS Coaching Study.

Cost Study

We also have recommendations for the analyses of costs of the variations of each experimental dimension. The cost study would address one of the primary research questions: What is the cost of implementing the different coaching dimension variations?

There are two purposes behind the cost study. The first is to provide information to HS grantees about the types of resources needed to develop and implement these dimensions within their programs. The first purpose is critical; if the evaluation team learns that particular coaching models are effective, the total resources required to implement these models will be important information for both planners within the OHS and HS program directors. For example, the cost

study is expected to provide estimates of the average cost of providing Coach Training for different amounts of time per month. This information can help HS directors decide how best to use their available PD resources.

The second purpose of a cost study is to gather information that can be used in a cost-effectiveness analysis. Conducting this analysis would allow the evaluation team to determine the relative cost-effectiveness of each coaching dimension condition by comparing the financial resources required to implement a given level of a coaching dimension (e.g., Level I Dosage or Level II Coach Training) and its estimated effectiveness (effect size) when considered across all other dimension levels. The evaluation team of the HS Coaching Study can compare the marginal cost of moving from one level to another on any given dimension with the marginal change in effectiveness gained by changing the level. If enhanced dosage of coaching is found to be only marginally more effective in producing classroom outcomes than typical dosage, and the cost-effectiveness analysis reveals that enhanced coaching dosages are substantially more costly to implement, it may be more cost-efficient for HS grantees to adopt a coaching program with fewer coaching hours per classroom. In addition, if the evaluation team learns that two coaching dimensions have similar impacts on outcomes, but the second is more expensive than the first, HS programs might choose to invest in the first. Cost data can also be used to analyze differential efficiency in different programs and conditions, taking into account other variables such as CLASS scores. Overall, collecting cost information and conducting cost-effectiveness analyses can better support decision making about investments in dimensions of coaching interventions.

There are several different types of cost information that would be needed from HS programs participating in each experimental condition in order to conduct the recommended analyses. Resource information needed for the study includes both direct costs (e.g., materials, space) and staff time, which can be translated into dollars paid for staff compensation. Information the evaluators would need to collect for each study condition includes, but is not limited to, the following:

- Salaries of coaches (per hour)
- Hourly wages for substitutes to cover time of coaches while participating in out-of-classroom coaching activities
- Salaries of teachers and other staff participating in coaching
- Time coaches and teachers spend (weekly, monthly, or in total) participating in coaching (different by levels of the Dosage conditions)
- Hours coaches spend planning and preparing for coaching
- Salaries of HS administrators or others who supervise or provide administrative support (e.g., scheduling) for the coaches
- Time HS administrators spend supervising coaches and managing the program
- Cost of materials used in coaching, such as feedback forms or videos for demonstrating best practices (different by condition if technology-mediated condition is chosen)
- Time and cost of any training (e.g., trainers, materials) coaches receive (different by Level I and Level II conditions)

- Cost of TA and setting up and managing an MIS system

Data Sources

There are four key data sources from which cost information can be collected: (1) budgets, (2) audited financial statements, (3) time use data (from time logs or center director surveys), and (4) interviews with HS administrators. Budgets and audited financial statements track planned and actual expenditures. After weighing the pros and cons of each, we recommend collecting information for the cost study from the sources that follow, some of which are already are proposed for use in the impact study, the implementation study, or study monitoring activities:

- Program budgets
- Coaches time use logs and survey reports
- Time logs from coaches
- Interviews of center directors

Program Budgets. Program budgets are a good source of information about a coaching program's design and intent, and their advantage is that they are available before or at the start of the program. The disadvantage is that they provide a picture of *projected* spending, not actual expenditures. Although actual expenditures on nonpersonnel items (such as materials, consultants, and utilities) can vary from budgeted amounts, budgets typically provide fairly accurate information on salaries, which typically make up 70 percent or more of total program budgets. Financial statements are the most accurate representation of actual expenditures, but they are not available until several months after the fiscal year ends. Depending on study timing, it may be feasible to analyze actual audited expenditures when they are available; this would give a more accurate picture of actual resources used, when paired with teacher time logs, discussed later. Again, we recommend collecting data from program budgets and program administrator surveys in conjunction with teacher and coach time logs. However, if time logs are deemed too burdensome, budgets in conjunction with only a program director survey and coach reports of time spent with teachers would provide similar, but less detailed, information. Depending on study timing, it may be feasible to analyze actual audited expenditures when they are available; this would give a more accurate picture of actual resources used, when paired with teacher time logs, discussed below.

Time Use Logs and Time Survey Reports. Neither budgets nor financial statements provide information on how much time administrative leaders spend supervising coaches or how much time teachers spend with coaches, which are key elements in allocating salaries and understanding the true resources needed for a coaching model. For example, if a program director spends 5 percent of full-time employment supervising a coach, the evaluator should attribute 5 percent of his or her salary to the cost of implementing the coaching program. Similarly, if a teacher spends two hours per week working with a coach (outside of regular teaching time when the coach might be observing him or her), the evaluator should attribute 5 percent of his or her full-time salary to the cost of the coaching program.

We suggest that time should be collected from three sources:

- The time coaches spend on varying activities is also important to consider in estimating the FTEs of coaches needed for each dimension. As a part of the main impact study, coaches will enter into the MIS hours that they spend with each teacher (see Section VIII Measures, tool 1).
- It is also important to capture teachers' report of the time they spend on coaching activities to determine if it aligns with coach reports and to determine what coaching activities, if any, occur outside coach-teacher meetings. The evaluation team should consider asking teachers to keep track of time spent with coaches during a sample period of time (e.g., over a two week sample period). These data could be collected during two sample periods during the program year—one in fall and one in spring—to increase their reliability.[71]
- Time spent by program administrators can be captured in the proposed center director survey (see measures section, tool 3).

Interviews. Interviews with HS administrative leaders (see measures section, tool 4) constitute another source through which resource costs could be captured. Program directors to report the resources they have used, including but not limited to staff salaries, how much time teachers spend with coaches, how much time they spend in training, and materials costs. However, unlike surveys, some of this information is similar to what can be captured in surveys with administrators, however, interviewers would have the advantage of being able to gather the perspectives from program directors on the practicalities of coaching, additional cost elements they think should be considered in implementing coaching, and other challenges and considerations. Interviews can provide additional supporting evidence to help understand and expand on the quantitative information derived from the logs and surveys.

Sampling

If collecting cost data from all programs participating in the HS Coaching Study is too burdensome or expensive to implement, cost information could be collected from a stratified sample of programs within each experimental condition. This sample should intentionally include grantees of varying sizes, locations, and staffing arrangements to ensure that resulting cost information can be reasonably generalized to all HS centers. For example, including only large centers in the cost study could underestimate the true cost of a coaching program because large centers have a greater capacity to take advantage of economies of scale than do small centers.

Piloting

Before data collection begins, all protocols should be piloted with a small group of coaches, center directors, and teachers to ensure that survey questions, time log instructions, and online data collection fields are clear and understood as intended.

[71] However, Information that could be reported by teachers through the MIS reports that we suggest teachers would complete quarterly (see Measures section, tool 1).

Analysis of Cost Data

There are two key approaches to the analysis of the various types of cost data collected from budgets, audited financial statements, staff time use data, and interviews with HS directors: (1) cost calculations and (2) cost-effectiveness analysis.

Cost Calculations

Using data collected from program budgets and program staff allows estimation of the costs of each coaching dimension by using an "ingredients" approach to costing out (Levin & McEwan, 2001), in which the costs of interventions are built from the bottom up by listing the specific resources (personnel or nonpersonnel) needed to implement the intervention at each site. In the making of these estimations, costs should be calculated from the perspective of the programs—that is, what likely range of costs programs would incur to implement each coaching condition. Cost information should be collected from a sample of programs in each experimental condition. The evaluator will estimate the cost of implementing the Level I and Level II conditions of each dimension of the coaching program being explored. After these initial estimations, preliminary findings will be shared with a small group of research participants volunteering to participate in a focus group; this group can help the research team ensure that data are interpreted correctly. Marginal costs will then be estimated for the Level II condition of each dimension; in other words, the evaluator can estimate the cost of moving from the Level I condition to the Level II condition of each dimension (such as moving from typical Coach Training to enhanced Coach Training).

Cost-Effectiveness Analysis

For the recommended cost-effectiveness analysis,[72] the evaluation team should compare the marginal effect sizes with the marginal costs of the Level I and Level II conditions on each dimension in the study. They may also estimate the range of costs for various combinations of the Level I and Level II conditions of the three dimensions at different sizes of centers. For example, the evaluation team could compare the cost per standard deviation change in teacher effectiveness (based on one or more measures, described in the following section) for levels of Dosage.

[72] This is distinct from a cost-benefit analysis, which examines the monetary value of benefits in comparison with the cost of a program. Because it is difficult to estimate what costs might be saved in the future as a result of coaching, a cost-benefit approach is not very useful for translating benefits into dollar figures because these would be such rough estimates. Therefore, we do not recommend a cost-benefit approach for this study.

VIII. Measures

In this section, we provide a description of the key measures, data sources, and data collection strategies recommended for the HS Coaching Study and aligned to the research questions for the three core study components. As stated in the evaluation background (Section II), there are six research questions that guide the design of the HS Coaching Study. The research questions are organized by the three core components—the impact component (described in Sections IV–VI) and the IR component, and cost study component (described in Section VII).

- **The impact component.** The research questions relate to the impact of the coaching dimensions on classroom outcomes, including teachers' practices and the quality of the classroom environment. Data are needed to document the outcomes for each of the tested coaching dimensions.
- **The IR component.** The research questions relate to understanding the implementation fidelity of the tested coaching dimensions, the foundational coaching approach, and the variations in implementation of coaching dimensions across centers and classrooms that may be related to outcomes. Data collected for this component help interpret impacts by examining implementation quality, facilitators, barriers, challenges, TA, and HS contextual circumstances.
- **The cost study component.** The research questions relate to understanding the cost of implementing the tested coaching dimensions and their cost relative to each other and for Level I and Level II. Data collected for this component will help identify the most cost-effective interventions.

The data will also be used to monitor study activities, providing ongoing information about implementation fidelity and helping to identify areas of TA needed to support coaching and ensure fidelity, including adequate contrast between the variations being tested. More information about monitoring is described in Section X.

In this section, we describe the measurement approach in three parts. First, to provide an overview of the measurement strategy, we outline the key constructs for the study as implied by the research questions. Second, we provide details for specific recommended data collection tools for the impact and IR, including what they measure and their format, frequency, and specifications. We also provide a discussion about the respondents for each data collection tool and respondent burden. Finally, we conclude with a summary of measurement recommendations. We designed this measurement approach to maximize study feasibility (conducting the study within the timeline and minimizing burden on participants) while simultaneously documenting the details and context of coaching with the necessary richness and specificity to answer the research questions.

At several points within this section, we note that some decisions about measurement cannot be made until the specifics of the coaching content for the foundational model for the study (as described in Section III) have been established in partnership with the PD developer. For example, it is difficult to propose specific teacher practice measures without knowing the exact language strategies that will be targeted by the PD. The developers of the coaching content focused on language development will ultimately need to play a role in refining the measurement

plan. Here we specify the constructs and measures that will need customizing and also outline a recommended process to use in creating study-specific measures for teacher practice and classroom environment.

Measurement Constructs

Measurement constructs for the HS Coaching Study can be described in five broad categories:

- **Category 1: Dimensions of Coaching That Are Systematically Varied.** Data collection needs to capture details of the three systematically varied coaching dimensions (Dosage, Coach Training, and Recipient) or, as an alternative, the dimension of Mode. Measures should provide information about the contrast between the different conditions specified by the factorial design and about costs.
- **Category 2: Elements of Coaching That Are Part of the Foundational Model.** The intent of documenting the foundational coaching approach is to assess the extent to which coaches are following the fixed components as well as the actual variation with those coaching components that the grantees were allowed to decide. Documenting how the underlying foundational model varies across centers and classrooms may provide an opportunity to explore how differences in delivery of coaching are related to outcomes.
- **Category 3: Fidelity, Content, and Characteristics of Coaching.** Outcomes of coaching for classrooms depend on a well-implemented coaching approach delivered with high quality and consistent with the conditions to which the coach is assigned. Coaching effectiveness also partially depends on, for example, the nature of the interactions between the teachers and coaches or the extent to which each coach individualized process and strategies to the needs of each teacher. Measures are needed to capture implementation fidelity and the degree to which coaches understood the different coaching dimensions. Measures also capture the bidirectional nature of coaching (for example, use of coaching strategies that are responsive to teachers' needs, teacher engagement in coaching) and the perceptions and attitudes of coaches and teachers that facilitate or impede the coaching (e.g., perceptions of productivity, usefulness, and sensitivity of the coach to the teacher's needs).
- **Category 4: Teacher and Classroom Outcomes.** The focal outcomes of the HS Coaching Study are teachers' language development strategies and practices within their classrooms and quality of classroom learning environments and teacher-child interactions that are closely aligned with the language content used by the coaches. Detection of impacts requires reliable and valid measures of teachers' practices, interactions with children, and the classroom language environment that are sensitive to the intervention.
- **Category 5: Contextual Factors.** Understanding contextual factors is an important measurement activity. Contextual factors include demographic characteristics (such as education, experience, attitude, and behavioral traits of teachers and coaches) as well as organizational and program-level features. A literature review of coaching found that teacher factors such as experience, mental health, and job stress are linked to various coaching outcomes through their influence on teacher engagement in coaching (Isner et al., 2011). Commitment to a profession in ECE, perceptions of organizational climate, and attitudes towards grantee leaders are also predictive of change. Readiness to change,

beliefs, and prior knowledge of classroom practices can make a difference in coach outcomes (Aikens & Akers, 2011). Similarly, coach attitudes, beliefs, and prior training and education may be linked to their effectiveness with teachers, though few studies examine this empirically.[73]

Table 15 provides details about the constructs within each of these five categories, with constructs to be measured organized by key feature.

Table 15. Overview of Key Constructs by Category

Feature	Constructs to Be Measured
Category 1: Dimensions of Coaching That Are Systematically Varied	
Dosage	▪ Time spent and frequency of coaching sessions ▪ Contrast between levels of Dosage specified by factorial design ▪ Coach salaries, full-time equivalent (FTE) spent coaching ▪ Perceptions of Dosage (opportunities and challenges)
Coach Training	▪ Content covered (specific language content, support for adult learners) ▪ Coach time and frequency of attendance ▪ Format (in person, remote technology, other) ▪ Training characteristics (teaching strategies used, coach ratings of trainings, research team ratings of trainings) ▪ Contrast between levels of Coach Training specified by factorial design ▪ Cost of trainers and training materials ▪ Content covered in training ▪ Peer network attendance (use and engagement) ▪ Perceptions, satisfaction (preferences, recommendations for improvement)
Recipient	▪ Attendance at initial training (lead or teaching team) ▪ Attendance at coaching sessions (lead or teaching team) ▪ Scheduling of coaching sessions (space, scheduling, teacher availability, substitute availability) ▪ Contrast between sessions with lead only compared with teaching team as specified by the factorial design ▪ Time or FTE teachers spent in coaching activities
Mode (Alternative)	▪ Coach attendance at training specific to remote delivery ▪ Context—availability and stability of technology for remote delivery; challenges of using technology ▪ Characteristics of coaching sessions

[73]Overall, it is expected that these characteristics will be balanced by random assignment; however, exploratory analysis of certain subgroups of teachers may be helpful to consider as covariates or for effects by subgroups. For example, teachers with less language and literacy knowledge may have different outcomes than those with more knowledge. Teachers who experience high job stress or who report mental health issues may have poorer outcomes than other teachers.

Feature	Constructs to Be Measured
	▪ Contrast between conditions as specified by the factorial design ▪ Content covered in coaching session ▪ Characteristics of interactions between coach and teacher(s) ▪ Cost of technology
Category 2: Elements of Coaching That Are Part of the Foundational Model[74]	
Structure of Coaching	
Content	▪ Language topics and teaching strategies (aligned with goals for children's development) that are addressed in training and coaching sessions
Coordinated Teacher PD	▪ Frequency, format, and content of PD sessions ▪ Attendance
Coach Role and Teacher-Coach Relationship	▪ Features of the teacher-coach relationship, including perception of trust, expectations, communication process, orientation to goals, and problem solving
Process of Coaching	
Coaching Tools	▪ Use of specific tools (assessments, checklists) to facilitate coaching by generating information to understand teachers' progress and guide selection of focal topics for coaching
Coaching Strategies	▪ Use of specific strategies during coaching sessions, including the following: • Planning—coach engages in planning with the teacher • Promoting reflection—coach provides opportunities for reviewing and assessing behavior and skills • Goal setting—coach works with the teacher to set specific goals • Modeling—coach models particular technique or teaching strategy • Observation—coach observes the teachers or teaching team practice in the classroom or in planning time • Providing feedback—coach provides specific and individualized feedback about a teacher's practice • Encouraging engagement
Staffing for Coaching	
Coach Selection	▪ The process of recruiting, hiring, and training coaches
Coach Supervision	▪ The structured, formalized process of providing support to coaches and attending to the improvement of coaching quality
Coach Caseload	▪ The number of teachers or classrooms assigned to individual coaches
Category 3: Fidelity and Characteristics of Coaching	
Fidelity of the Coach Model Variations	▪ Degree to which targeted language practices are the focus of coaching sessions ▪ Degree to which coaching strategies are delivered with appropriate frequency

[74] It is valuable to document features of the foundational model in order to understand how coaching elements work, beyond those that are systematically varied in the design.

Feature	Constructs to Be Measured
	▪ Degree to which conditions specified by the factorial design are implemented with the planned contrast
Characteristics of Coaching	▪ Quality of teacher-coach relationship ▪ Ability to manage time and interactions effectively ▪ Degree to which coach individualizes content and process on the basis of teacher characteristics
Category 4: Teacher and Classroom Outcomes	
Teacher Attitudes and Dispositions	▪ Attitudes about teaching, perception of the adult work environment, experience with coaching; readiness to change
Teacher Knowledge of Language Practices	▪ Knowledge inventory
Teacher Use of Language Practices	▪ Frequency and quality of engagement in targeted language practices, including provision of support for vocabulary and oral language
Quality of the Learning Environment	▪ Materials and activities designed to support language skills
Quality of Teacher-Child Interaction	▪ Instructional support, classroom organization, and social-emotional support
Category 5: Contextual Factors*	
Grantee Characteristics	▪ Location ▪ Size ▪ Available TA and training and coaching infrastructure
Program Characteristics	▪ Site size ▪ Staff characteristics ▪ Staff retention ▪ Characteristics of families served
Program Prior Experience With Coaching and Language Interventions	▪ Sites' previous use of coaches ▪ Content of prior coaching
Organizational Support for Coaching and Adult Work Environment	▪ Factors that impede or facilitate coaching in the classroom and site • Program culture and facilitation of PD for staff • Program schedule and availability of time and space for coaching ▪ Quality of the adult work environment • Program leadership • Promotion of adult well-being
Teacher Characteristics	▪ Education and training and other PD participation ▪ Prior experience with coaching ▪ Job stress ▪ Attitudes and beliefs ▪ Knowledge of language practices
Teacher Perceptions of Coaching	▪ Perceived value of coaching

Feature	Constructs to Be Measured
Coach Characteristics	▪ Education and training ▪ Experience with coaching
Coach Perceptions of Teacher Engagement	▪ Extent of teacher participation and engagement in the coaching

* Information on the program characteristics and characteristics of children and teachers at each center may be available from administrative records.

Data Collection Tools

To collect data across the construct categories in Table 15, we recommend that the HS Coaching Study use multiple data collection tools. We suggest six data collection tools. These are listed and numbered (Tools 1–6) in Table 16. For each tool, we outline our recommendations for which features (from the left-hand column in Table 15) can be documented. In some cases, the same features should be captured by multiple measures in order to triangulate information across respondents. Most tools serve multiple purposes—that is, they collect data for the impact, IR, and cost components of the study or for the purposes of monitoring study activities. As noted earlier, final decisions about data collection tools and features to measure will need to wait until the coaching content is selected.

Table 16. Features to Measure for Each of the Six Data Collection Tools

	Impact	IR	Cost	Monitor
Tool 1: Implementation Contact, Time, and Attendance Logs				
▪ *Category 1 features*: Dosage, Coach Training, Recipient, Mode ▪ *Category 2 features*: Content, Coordinated Teacher PD, Coaching Tools, Coaching Strategies, Coach Caseload ▪ *Category 3 features*: Fidelity of the Coach Model Variations	√	√	√	√
Tool 2: Implementation Rating Logs (Optional)				
▪ *Category 1 features*: Coach Training ▪ *Category 3 feature*: Fidelity of the Coach Model Variations ▪ *Category 5 feature*: Coach Perceptions of Teacher Engagement		√		√
Tool 3: Participant Surveys				
▪ *Category 1 features*: Recipient, Mode ▪ *Category 2 features*: Coordinated Teacher PD, Coach Role and Teacher-Coach Relationship, Coaching Strategies, Coach Selection, Coach Supervision ▪ *Category 4 features*: Teacher Attitudes and Dispositions, Teacher Knowledge of Language Practices ▪ *Category 5 features*: Grantee Characteristics, Program Characteristics, Program Prior Experience With Coaching and Language Interventions, Teacher Characteristics, Teacher Perceptions of Coaching, Coach Characteristics	√	√	√	

Tool 4: Participant Interviews				
▪ *Category 1 features*: Recipient, Mode ▪ *Category 2 feature*: Content, Coach Role and Teacher-Coach Relationship, Coaching Strategies, Coach Selection, Coach Supervision ▪ *Category 3 features*: Characteristics of Coaching ▪ *Category 5 features*: Grantee Characteristics, Program Prior Experience With Coaching and Language Interventions, Organizational Support for Coaching and Adult Work Environment	√	√		
Tool 5: Independent Observations of Coaching Sessions and Coach Training				
▪ *Category 1 features*: Coach Training, Recipient, Mode ▪ *Category 3 features*: Fidelity of the Coach Model Variations, Characteristics of Coaching		√		
Tool 6: Observations of Teacher Practices and Classroom Environment				
▪ *Category 4 features*: Teacher Use of Language Practices, Quality of the Learning Environment, Quality of Teacher-Child Interaction	√			

Next, we provide more details on the format, participants, and specifications for each of the six data collection tools.

Tool 1: Implementation Contact, Time, and Attendance Logs

The implementation contact and attendance logs will be used to document and monitor attendance and details of coaching sessions, Coach Training, and teacher training on language content teachers receive in addition to coaching (Table 17). The logs will measure all of the systematically varied dimensions (including contrast between each of the condition levels) as well as six other features of the coaching conditions as shown in Table 16.

Table 17. Specification of Tool 1: Implementation Contact, Time, and Attendance Logs

Format	Web-Based MIS
Participants	Coaches enter contact and attendance data biweekly or monthly (depending on experimental condition).
	Trainers enter contact and attendance data monthly.
	For two two-week sample periods, teachers enter time spent on coaching activities daily to estimate FTE and costs spent on coaching activities.
Specification of Measure	▪ The coaching log requires coaches to enter the time in minutes and hours that they spend with teachers overall as well as the time spent using particular coaching strategies (such as observation and modeling) and the content addressed in each coaching session.[75] Strategies and content are included in a checklist that facilitates easy reporting of the coaching session details. ▪ The trainer logs document attendance at Coach Training sessions (which is used to estimate Dosage) and content of training sessions and contacts with coaches (similar to the format used for coaches). ▪ Specific prompts and instructions for the log will need to be developed based on the language intervention that is selected for the study. ▪ Logs need clear instructions and anchor points for data entry. For example, data fields that require entry of hours and minutes are specific about what activities should be accounted for in the data entry.
Examples for the Measure	See Mattera, Lloyd, Fishman, and Bangser (2013) for an example of an MIS that captured similar data but for social-emotional curriculum in HS.

Note. We recommend a Web-based MIS to support data collection for the study. The MIS should be an accessible, user-friendly online data system that provides the opportunity to enter and monitor data in real time and collect data about coaching and training. An MIS is designed to reduce the burden of data entry on respondents by offering prepopulated data fields, drop-down menus, and automatic calculations and by checking for out-of-range values and missed fields. These simplify the experience of the respondent and reduce the time necessary on the back end of the system to clean and resolve data discrepancies. Respondents who do not have access to a Web-based system would be provided with alternative data collection tools to accommodate their needs. E-mail reminders with a link to the MIS could be sent to coaches following scheduled coaching visits and training to facilitate timely data collection.

Tool 2: Implementation Rating Logs (Optional)

Tracking the fidelity of coaching activities to support and implement the study is a critical study task. In addition to documenting quantitative features of these activities (Tool 1), it is necessary to assess qualitative features such as perceptions of utility, value, and adherence to process components of the training and the coaching model (use of strategies, delivery of content). It is also important to document the extent to which teachers practice new content and teaching strategies in their classrooms. Particularly for monitoring and technical assistance purposes, these features are best documented by those in the field, who are engaged daily with coaching and training. The rating logs will document (1) coaches' report on utility and value of training for the coaches (session productivity and clarity of content); (2) coaches' and teachers reports on

[75] Isner et al. (2011) cited several studies that used contact logs to document the time coaches spent with teachers and the content. The Impact of Two Professional Development Interventions on Early Reading Instruction and Achievement Study had coaches complete logs to record the hours spent with a teacher and the nature of each coach interaction with each teacher, including content (Garet et al., 2008).

utility and value of teacher PD on the language content; (3) coaches' and teachers' reports on utility and value of the coaching sessions; and (4) coaches' and trainers' reports of teacher adherence to process of teacher use of the targeted language strategies. Further details of the rating logs are provided in Table 18.

Table 18. Specification of Tool 2: Implementation Rating Logs

Format	Web-Based MIS Format*
Participants	Rating teacher practices in classroom: ▪ Coaches rate each classroom on a quarterly basis.[b] ▪ Trainers rate a subset of classrooms once (5 percent to 10 percent) to provide data on reliability of coach ratings at one or two points during the school year.
	Rating training: ▪ Coaches and teachers will rate trainings as they occur.
	Rating coaching: ▪ Coaches will rate coaching as they enter contact and attendance data. ▪ Teachers will rate coaching quarterly by using rating measures developed for the particular language PD.
Specification of Measure	Rating teacher practices in classroom: ▪ Classroom implementation and fidelity guidelines can be developed in partnership with the PD developers. These guidelines determine what will be monitored and assessed. ▪ Participants provided with descriptive anchor points for the rating at each level (e.g., 1–5) that accompanies the response options of "strongly disagree" to "strongly agree." These anchors are intended to increase rating reliability. Rating training: ▪ These ratings will collect input on the content covered by the training and perception of the quality and relevance of the training to implementation of the language intervention. Rating coaching: ▪ These ratings will request input on the degree to which individual coaching sessions (rated by coaches) or the overall experience of coaching (rated by teachers) met criteria for productivity, content, usefulness, and sensitivity to teachers' needs.
Examples for the Measure	In HS CARES (Mattera, Lloyd, Fishman, & Bangser, 2013), coaches rated classrooms on the degree to which teachers implemented specific teaching strategies and whether they were integrated throughout the day or used only sporadically. HS CARES also used ratings of training and quality of the coaching sessions.

*We recommend a Web-based MIS to support data collection for the study. See further information in the table note for Tool 1. [b]Although it would be preferable for trainers and their developers to conduct these ratings, it is likely not to be feasible, given probable staff limitations of the developers. On the basis of the design team's prior study experiences, the recommended strategy calls for the coaches to conduct ratings.

Tool 3: Participant Surveys

As shown in Table 16, participant surveys will be used to capture multiple features of the coaching model and context of coaching. Details about these Web-based surveys are provided in Table 19.

Table 19. Specification of Tool 3: Participant Surveys

Format	Web Based
Participants	Site directors and administrators, coaches, teachers, and trainers will complete a spring survey.
Specification of Measure	▪ Surveys gather data from participants about their characteristics, experiences, and perceptions and the context of coaching. Coach supervisors and administrative personnel estimate the FTE they spent on coach activities for the cost study.

Examples for the Measure	▪ Existing or adapted items or scales could be used for key constructs in the study. Examples include the following:
	• Teacher attitudes and dispositions
	  • Modernity scale (Schaefer & Edgerton, 1985)
	  • Child Care Worker Job Stress Inventory (Curbow, Spratt, Ungaretti, McDonnell, & Breckler, 2000)
	  • Psychological distress (Kessler et al., 2002)
	• Organization readiness for coaching
	  • Organizational Social Context (Glisson, 2007)—survey completed by staff or teacher to create scores on climate and culture at the organizational level and work attitudes at the individual level
	• Individual readiness for coaching
	  • Stages of Change Scale (Peterson, Baker, & Weber, 2010)—survey of teachers that provides a score from 1 to 5 indicating their readiness to implement new practices; coaches also can assess stages of change for teachers
	• Quality of teacher-coach relationship
	  • Working Alliance (Horvath & Greenberg, 1989)—survey completed by the teacher, coach, or external trained observer to assess relationship between coach and teacher
	• Adult work environment
	  • Supportive Environmental Quality Underlying Adult Learning (Whitebook, under development)—scales completed by teachers and directors to assess teaching supports, learning community, job crafting, adult well-being, and program leadership

Tool 4: Participant Interviews

Interviews with a subsample of participants will be used to collect in-depth details to answer research questions related to understanding how coaching was implemented, factors that facilitated or hindered implementation of the coaching conditions, types of TA and PD tools that facilitated implementation fidelity, and differences across different HS contexts. Qualitative data can provide insights into design features of the factorial experiment and the way they worked in

practice (see features of participant interviews listed in Table 16). Further details regarding recommended participant interviews are provided in Table 20.

Table 20. Specification of Tool 4: Participant Interviews

Format	Semistructured Interviews Conducted by Phone or in Person
Participants	All will participate in a spring interview.
	A sample of center directors, grantee liaisons coaches, teachers, and PD trainers will participate in a spring interview. The sample will correspond with study participants in the eight grantees included in the site visits.
Specification of Measure	▪ Grantee liaison and center director will provide their impressions concerning the language curriculum and implementation, the quality of the coaching and training, the organizational setting, and the sustainability of the practices targeted by the coaching. ▪ Coaches will provide their impressions of the coaching model and implementation, curriculum training, curriculum implementation, teacher-coach relationship and sessions, coaching sessions and the teacher dyad, informal interactions with peers to support practices, and the organizational setting. ▪ Teachers will reflect on their experience with the language curriculum and implementation, the coaching and training, their coteachers, and the organizational setting. ▪ Trainers will provide an additional perspective about coach quality and teacher implementation.
Examples for the Measure	Measurement items to include in the interviews should capture information about perceptions of coaching, training, challenges in adhering to the coaching conditions (coaches only), impressions of the effectiveness of the coaching conditions (coaches and trainers only), experiences with the language content, and experiences working with a teaching team compared with working alone (teacher only).

Note. We recommend that the research contractor conduct eight site visits to a sample of grantees participating in the HS Coaching Study. Depending on site visit schedules and travel expenses, interviews will be conducted in person whenever possible. However, if schedules do not permit visits with grantee liaisons, coaches, PD trainers, or site directors during the site visits, phone calls can be used to complete the interviews.

Tool 5: Independent Observations of Coaching Sessions and Coach Training

Structured observations of the coaching sessions will be conducted to assess key qualitative features such as type of questions, characteristics of feedback and reflection, engagement of teacher or teaching team, planning, goal development and assessment, active listening, and communication skills. Table 21 provides further details about the recommended observations.

Table 21. Specification of Tool 5: Independent Observations of Coaching Sessions and Coach Training

Format	On-Site
Participants	Four hundred observations of teacher-coach sessions, sampling all coaches and all experimental conditions (though not each coach implementing each condition)
	Training observations by the monitoring and TA team for a subset of teacher and Coach Trainings (including the initial coach and teacher trainings as well as trainings for coaches assigned to the enhanced level of Coach Training)
Specification of Measure	▪ A project-developed tool will be used by trained and reliable independent researchers to observe live coaching sessions. Alternatively, videos of sessions could be collected if on-site, live coding is not feasible. The tool should assess the context of the coaching session (challenges related to timing or facilities), teacher engagement, communication, specific coaching strategies used, content included, degree to which the coach individualizes the coaching session to the needs of the teacher or teaching team, and discussion of goals and progress. Tool development will need to focus on a concise set of items that capture key dimensions. ▪ A project-developed tool will be used by trained members of the monitoring and TA team to observe and record fidelity of training sessions.
Examples for the Measure	▪ In HS CARES, the trainers rated coach quality two to three times per year by using a set of 10 items assessing features such as the following: • Provision of feedback • Support for problem solving • Ability to motivate teachers This type of measure would need to be adapted for live coding by developing training and a manual with descriptions and examples of each practice and behavior. ▪ In the Impact of Two Professional Development Interventions on Early Reading Instruction and Achievement Study, observers from the project team completed a closed-ended observation fidelity form during each PD institute and seminar, documenting the time spent on the major topics and activities outlined in the syllabus for each day of PD. These data were used to measure the fidelity with which the intended PD program was delivered (see Appendix H of Garet et al. [2008]).

Tool 6: Observations of Teacher Practices and Classroom Environment

Structured observations of the classrooms and teaching practices will be used to gather information at the end of the year[76]about (1) classroom quality and (2) the specific language and teacher-child interaction practices that are targeted by the coaching as a primary outcome measure for the impact study (Table 22).

We do not recommend that the study conduct observations of teacher practices and classroom practices at baseline. As explained in Section V, data from prior a prior study conducted in HS centers (Mattera, Lloyd, Fishman, & Bangser, 2013) show that baseline measures of teacher practice do not substantially improve statistical power. Thus, we recommend using center-level CLASS scores from existing data sources to save on costs. These data can also be used to reduce the MDES and to test for baseline equivalence across experimental conditions (see Section V for further discussion).

[76] One of the conditions for study participation will be that the grantees and their centers be willing to provide administrative center-level data on extant CLASS scores at baseline.

Table 22. Specification of Tool 6: Observations of Teacher Practices and Classroom Environment

Format	On-Site
Participants	Teaching practices within classrooms will be observed by trained, reliable observers in the spring.
	Extant center-level CLASS data will be collected at baseline for the study.
Specification of Measure	▪ Likely two observation measures will be used—one for classroom quality (e.g., CLASS) and one for specific language practices.[77] ▪ Observation measures must be closely aligned with the coaching model. Because the language curriculum has not yet been decided, we do not propose a specific measurement protocol (though we suggest possible measures here). We propose the following seven steps for adapting an existing measure to align closely with the study constructs: 1. Identify the targeted teacher practices that the curriculum aims to change 2. Review relevant existing measures about that teacher practice 3. Select the base measure that best aligns with the curriculum's expected teacher practice changes and is most sensitive to hypothesized changes in teacher practices 4. Adapt the measure so that the questions are more highly aligned with teacher practices, adding questions as necessary (or create a measure if necessary) 5. Pilot measure 6. Create a manual and training and reliability standards 7. Collect data and analyze results ▪ For two measures, observations can to be conducted on separate days (with one observer) or on the same day (with two observers). Reliability checks will be conducted for approximately 10 percent of the observations.
Examples for Measure	The CLASS prekindergarten version (Pianta, La Paro, & Hamre, 2008) is recommended to assess the general quality of teacher-child interactions (capturing the domains of instructional support, emotional support, and classroom interaction). A language-and-literacy-focused tool will be used to assess the quality of specific language and literacy practices and features of the learning environment. Options to consider for this tool (or as a base tool that is adapted for the study) include the literacy subscale of the ECERS-Extension (Sylva, Siraj-Blatchford, & Taggart, 2010), the Supports for Early Literacy Assessment (Smith, Davidson, Weisenfeld, & Katsaros, 2001), and the ELLCO Pre-K (Smith, Brady, & Anastasopoulos, 2008). For an example of a tool that was adapted using this process, see the Adapted Teaching Style Rating Scale in the HS CARES study (Raver, Domitrovich, Greenberg, Morris, & Mattera, 2012).

[77] If necessary to contain study cost, the observations could be conducted with the language-and-literacy-specific measure only and not the CLASS. The CLASS is of great interest to HS grantees because of its use in the Designation Renewal System. Therefore, study buy-in may be increased if CLASS scores are included. However, because the dimensions of the CLASS will not be the direct focus of the PD, we may be less likely to see impacts on the practices assessed by the CLASS.

Respondent Burden

In addition to understanding the recommended data collection tools, it is important to specify what individual study respondents would be asked to contribute to the data collection efforts and when to assess respondent burden. Table 23 provides an overview of respondent burden.

Table 23. Data Collection Requests per Study Respondent

Respondent	Data Collection Tool	Periodicity	Estimated Time to Complete Tool
Center Director	Survey	Spring	30 minutes
	Interview (subsample)	Spring	60 minutes
	Request for program budgets	Spring	30 minutes
Coach	Contact, time, and attendance logs	Biweekly or monthly (depending on experimental condition)	15 minutes
	Implementation rating logs	Varied—depends on rating	15 minutes
	Survey	Spring	30 minutes
	Interview (subsample)	Spring	60 minutes
Teacher	Contact, time, and attendance logs	Daily, for two two-week sample periods in fall and spring	15 minutes
	Implementation rating log	Varied—depends on rating	15 minutes
	Survey	Spring	30 minutes
	Interview (subsample)	Spring	60 minutes
	Observation of classroom	Spring	No extra time
Curriculum Trainers	Contact, time, and attendance logs	Monthly MIS data entry	15 minutes
	Implementation rating logs (reliability checking)	10 percent sample of classrooms at one or two time points during school year, depending on number of classrooms	15 minutes
	Survey	Spring	30 minutes
	Interview (subsample)	Spring	60 minutes
Grantee Liaison	Interview	Spring	60 minutes

OMB Clearance and Institutional Review Board (IRB) Oversight. A federal OMB clearance package should be compiled at the initiation of the HS Coaching Study that describes the sampling, recruitment, and measurement plan for the study. IRB approval for the study must be sought from the contractors' IRBs with federalwide assurance. Additional IRB approval may be needed at the grantee level. These requirements can be determined during site selection.

Summary of Measurement Issues

The measurement strategy is not simple; it is important to collect data with multiple respondents and at multiple levels to understand the complex relationships and practices that are part of the HS Coaching Study. We recommend using distinct and complementary data collection and measurement methods to document the full range of implementation. However, our recommendations for measurement attempt to keep the data collection streamlined, while minimizing burden on any given participant. For instance, the same tool will be used across impact, implementation, and cost study components and to support monitoring and TA. In addition, a singular MIS platform can support multiple tools (i.e., both the contact, time, and attendance log and the rating log). Although there may be up-front costs for development, a Web-based MIS can reduce the burden of data collection and increase the efficiency of cleaning, compiling, and analyzing the data. The surveys, interviews, and study-administered classroom observations will be conducted only once a year. Program budgets for the cost study will also only be requested once.

Finally, we recommend that, after the PD approach and developer are selected, special care be taken with the additional planning of the independent classroom observations. These will probably use both the CLASS measure and another specific language observation measure that documents the frequency and quality of targeted language practices, teacher-child interactions, and classroom environment to support language development. The language-specific tool should be carefully aligned with the selected PD model so that the practices targeted in coaching are those assessed in the observation. We recommend that adaptations to existing measures be carefully documented and that adapted tools be piloted prior to use to ensure that they are feasible for data collection.

IX. HS Grantee Recruitment and Selection

Recruiting and selecting the appropriate centers among Head Start (HS) grantees[78] to participate in the HS Coaching Study aligned to the recommended design are critical to successfully examining the effects of the coaching dimensions that will be systematically varied. There are three design recommendations that have direct implications for center recruitment and selection: (1) cluster random assignment, in which HS centers are randomized to levels of the Dosage and Recipient dimensions, and in which coaches are randomized to the levels of the Training dimension; (2) the minimal number of centers and the average number of classrooms within centers needed to detect an effect size of 0.20; and (3) and the preference to recruit grantees that would allow the evaluators to randomly or quasi-randomly assign coaches to centers.

This section of the report provides recommendations about creating the sampling pool and recruiting the appropriate HS grantees and centers needed for the study, offering considerations for the process of grantee selection inclusion and exclusion criteria. This section also describes potential factors that may affect grantee participation.

Considerations for the HS Grantee and Site Selection Process

There are three steps that we recommend the evaluation team consider in selecting grantees for participation in the HS Coaching Study. Step one is identifying a broad pool of eligible grantees. Step two is specifying criteria for inclusion and exclusion. Step three is developing and applying criteria that potentially make one grantee a better candidate than another. Once grantees are selected, centers to be blocked for random assignment within each grantee can be identified.

Step one of the grantee selection process is identifying a broad eligible pool of HS grantees for the study that can maximize implementation efficiencies for the foundational coaching model and the evaluation design and that can minimize costs. Recall from the power analyses in Section V that a sample size of 248 centers would be needed to detect an effect size of 0.20 (Dosage and Recipient dimensions), if there were three classrooms per center. We suggest that the evaluation team consider 31 blocks (grantees or groups of similar grantees) with at least 8 centers per random assignment block, preferably with three classrooms within each center.[79] It may be possible to recruit some grantees that can provide two blocks (or 16 centers). If roughly half of the grantees provide two blocks, the study would need to recruit only 21 grantees (rather than 31).

In order to facilitate the efficient implementation of the coaching intervention, and the research design, it would be beneficial to consider a pool of grantees within geographic clusters around the country. An analysis of HS data (Administration for Children and Families & Early Childhood Learning and Knowledge Center) shows that current high concentrations of HS centers can be found in the following metropolitan areas:

- Chicago, Illinois

[78] The term *grantee* is used broadly to encompass both grantee and delegate agencies with operational responsibility for local HS programs.
[79] An alternative is to recruit 312 centers, if they have only two classrooms each.

- Los Angeles, California
- Miami-Dade, Florida
- New York, New York
- Philadelphia, Pennsylvania
- Sacramento, California
- San Antonio, Texas
- San Diego, California

Therefore, we recommend that the evaluation study team identify an eligible pool of HS grantees from or around major metropolitan geographic areas such as those listed above. Limiting the eligible pool of HS grantees to recruit for the study to certain geographic areas will significantly reduce the cost of implementing the coaching intervention and the research costs. Travel time and costs would be reduced for trainers, and data collection costs can be kept to a minimum when the study sites are concentrated in easily accessible locations.

Although the sample located in these geographic areas (approximately 12 percent of all HS centers) may not be representative of all HS centers across the country, it can provide geographic and regional diversity. Such diversity will address some concerns that the study results may not apply to the broad range of HS grantees across the country.

Once the broad pool of geographically clustered HS grantees is identified, step two of the site selection process is defining grantee inclusion and exclusion criteria—that is, criteria that determine whether a grantee is eligible for the sample. We recommend four criteria for a HS grantee to be eligible:

- Grantee must have at least eight HS centers (sites) with at least three classrooms within each center. It may be possible to include additional sets of sites in which two grantees have four centers each, if they are close enough to one another to share coaches.
- Grantees identified as poor performers by the OHS can be excluded. The study requires well-managed grantees that are not facing significant compliance issues with OHS.
- Migrant and Seasonal Head Start (MSHS) can be excluded, given the likely systematic differences in their operations (Fishman & Wille, forthcoming). MSHS grantees operate at different times of the year, often with varying lengths of program operation, making it difficult to pool them with regular HS programs.
- Three-year-old-only classrooms can be excluded if that the language content selected is not appropriate for those children.

Step three of the selection process specifies additional considerations in the final selection of grantees from the eligible pool. We recommend the following considerations:

- *Current grantee coaching practices.* Grantees vary in the existence and scope of current coaching efforts along a continuum from no current coaching program to a comprehensive effort. They can operate coaching programs in some centers and not in others (Howard et al., 2013). In order for grantees to be selected to participate in the HS

Coaching Study, they need to be willing to either develop a new coaching effort that meets study requirements or revise their current coaching approach to align with study requirements. Grantees who do not wish to change their current coaching practices should be removed from the recruitment pool.

- Stakeholder outreach conducted as part of the effort for the design of this study indicated that some grantees are unlikely to adjust existing coaching efforts, although some grantees whose current efforts are more minimal or could more easily align with the study conditions were more amenable to participation[80]. For some, the HS Coaching Study could likely enhance the Dosage and/or focus of their coaching efforts. Although it might make sense to recruit grantees with minimal or no coaching, these program partners may have additional challenges such as the need to provide resources to hire new staff as coaches and acclimating staff to the agency. Therefore, the extent to which grantees consider participating in the HS Coaching Study also depends on the resources that the study can make available to grantees (see below).
- *Ability to create a cluster with other nearby grantees.* The evaluation team may attempt to create geographic clusters that include at least four blocks of eight centers. This means the evaluation team would include no more than eight geographical grantee clusters.
- *Whether the grantee is in a monitoring year.* Grantees facing federal monitoring may experience additional pressures, requirements, and stress during the monitoring year. Although not a reason for exclusion, it is worth discussing with the grantee to ensure the grantee is comfortable conducting the study at that time.
- *Strength of grantee interest in study participation.* Although grantee participation is voluntary, the strength and depth of support for participation is a factor in grantee selection. This will be evidenced by the grantee's willingness to redeploy or contribute resources to implementing the coaching model and expressions of support at both the grantee and center levels, including support of teaching staff. All participating grantees will be required to agree to random assignment of centers to conditions and preferably random assignment of these coaches to centers. Also grantees and centers must agree to support teachers and coaches attending training and provide administrative support, coverage, training time and space for the coaches to meet with the teachers. All participating grantees should be required to sign a memorandum of understanding (MOU) stating their agreement to implement the programmatic and research requirements.

The final step in the selection process is identifying specific centers within each grantee for inclusion in a block of eight for random assignment to one of the coaching conditions. The selection of eight centers from a larger group of eligible centers can be customized to individual grantee circumstances. Centers could be selected randomly or based on specific criteria such as proximity to one another. Within blocks, these centers should be matched as closely as possible on ethnicity and urbanity.

[80] See section III for a more detailed on the results of the stakeholder outreach that was conducted.

Funding for Implementation

In order to encourage and in some cases enable grantee participation, the funders of the HS Coaching Study should consider the provision of funding to HS grantees to support the coaching efforts and to offset the costs of participating in the evaluation. As discussed earlier, it is easier for grantees with no current coaching effort to design a coaching program that aligns with the study design. However, those grantees likely need funding or assistance to build new coaching programs from scratch. We recommend that funds should be provided to grantees to hire and pay for coaches, for substitute teachers during coaching sessions, space for teacher and Coach Training, and about 10 percent of time from a grantee staff person to serve as liaison for the study.

Grantees with more extensive coaching programs may require less additional funding if they are willing and able to reallocate existing coaching resources to align with the requirements of the study design. However, even grantees with current coaching programs could also benefit from funds offered for study participation.

Grantee Recruitment Process

There are several ways to recruit grantees from the eligible sample pool. One approach is that grantees in the pool be recruited to participate from a targeted list. For example, grantees can first be identified for outreach by using the criteria outlined in step one and then be approached to participate via e-mail or phone calls. Another method is to use a volunteer approach, releasing an announcement with eligibility criteria and asking grantees to express interest prior to formal recruitment. Because this method does not involve sending e-mails and calling a set of grantees, the volunteer approach may be less labor-intensive. However, by not focusing resources on a pool of grantees that are likely to meet the minimum eligibility requirements, the evaluation team will have somewhat less control over the sample.

We recommend a hybrid approach to recruiting grantees. Using this approach, the evaluation team would reach out to the grantees in the eligible pool. Grantees contacted directly should be those in immediate proximity (50- to 100-mile radius) of selected metro areas. Then we recommend also publishing an announcement that enables other geographically-close grantees to indicate their interest in participation. This approach concentrates study team resources where they are most likely to bear fruit, while also providing an opportunity for participation to interested grantees that are not in the immediate metro areas but that otherwise meet the study criteria.

We recommend that Office of Planning, Research and Evaluation (OPRE) and the evaluation team use a combination of outreach strategies. A study e-mail announcement can be sent to targeted grantees. Outreach can also occur at any timely Office of Head Start (OHS) institutes, trainings or National Head Start Association meetings. We recommend telephone calls be used to further explain the study, ascertain study eligibility and fit, and assess grantee interest in participation. In addition, we recommend that the evaluation team conduct one to two site visits to the final interested candidate grantees that meet study criteria to further assess fit and explain the study to a broader audience for each grantee (including administrators and teachers) in order to build support and build center- and teacher-level buy-in.

In order to support grantee outreach efforts, we suggest that the evaluation team develop a set of recruitment and marketing materials to fulfill the following recommendations:

- Recruitment materials can be developed for grantee-level audiences. Grantee-level materials will describe the benefits of study participation; explain the study goals and methods; and outline study eligibility and selection criteria, the study recruitment and selection process, and the expectations of study participants.
- Materials can also be developed for center-level audiences. Materials for center directors and teachers can also review overall study goals and expectations of study participants. However, these materials will focus more on benefits and expectations at the center level. What can teachers and children gain through study participation? What does the coaching model mean for teachers' daily activities?
- In addition to providing basic information about eligibility and the study, marketing materials can include more in-depth supporting documents, such as a draft MOU and next steps for grantees interested in participating.

During the final stage of recruitment, we suggest that the evaluation team prepare a memo to OPRE with a list of recommended grantee participants. Depending on OPRE requirements and need, the memo may provide grantee inclusion recommendations, a discussion of how each grantee meets eligibility criteria, and a suggested process for approving grantee participation in the study and securing MOUs from grantees.

Recruitment Implications of Replacing Coach Training Dimension With Mode Dimension

If the Mode dimension is used for the study, the level of a grantee's technology capability should be given serious scrutiny during the grantee recruitment phase of the HS Coaching Study. Implementation of the Mode dimension will require that all grantees and half of all selected centers have access to broadband internet service and some threshold level of technical capacity.[81] Participating teachers will need access to video cameras with excellent sound capability, computers to securely transmit videos to their coaches, and computers and monitors to view video provided to them by their coaches. These can be provided either through the evaluation contract or through the grantee. It is preferable to provide some of the equipment through the evaluation contract to reduce the grantees' cost of participation. In addition, the research team should provide manuals and training for use of video and other computer equipment, and help desk available to coaches to deal with the inevitable technical problems.

If the Mode dimension is selected, there may also be interest in including more rural sites because this technologically-mediated coaching may be of special interest to grantees that find access to qualified on-site coaches challenging. That is likely to require the recruitment and selection of more grantees than assumed in the primary plan.

[81] Although the recommended study data collection plan relies on the use of technology (MIS; Web-based surveys), there are likely ways to alternatively collect data if a site has low-capacity technology or connectivity. For instance, coaches could move off-site to fill out the coaching log in a location that has internet connection or teachers could be offered paper-copy surveys. For the Mode dimension, there is no alternative to having high levels of technical capacity.

Summary

Selecting the appropriate centers among HS grantees aligned to the recommended design to participate in the HS Coaching Study is critical to successfully examining the effects of coaching dimensions. Selecting from an eligible pool of grantees in a fixed number of metro areas can create efficiencies for the recruitment process as well as for program implementation and data collection. Identifying and implementing explicit grantee selection criteria will ensure that selected centers meet minimum specifications and that conditions for implementing the coaching model are conducive to success. Finally, a focused, aggressive recruitment effort concentrated in a select number of metro areas will help produce a study sample that yields efficiencies in coaching model implementation and study data collection.

X. Program Monitoring and TA

This section of the report describes site monitoring to ensure fidelity of implementation of the foundational coaching model and the eight experimental coaching conditions and technical assistance (TA) to support grantee, coach, and developer efforts. Both rely heavily on the management information system (MIS) described in Section VI. As described in more detail in this section, the MIS will serve as the primary mechanism for identifying issues related to program implementation. We recommend the evaluation team monitor MIS information on an ongoing basis and identify issues that require further attention from the developers, grantees, or coaches. We also recommend establishing a TA plan to provide a foundational support to grantee and coach participants and to troubleshoot if additional fidelity training is needed during the course of the project.

Monitoring Implementation Fidelity

Monitoring implementation fidelity of both the foundational coaching model and the eight experimental coaching conditions is critical to the success of the study. This includes understanding the Dosage, content, and quality of implementation of the underlying coaching model, the systematically varied coaching dimensions (including the contrasting levels), and the language practices that underpin the eight treatment conditions to which the centers will be randomly assigned. This monitoring activity should occur across the course of implementation to ensure fidelity to the model across time. Effective monitoring involves both clear specification of implementation expectations and the means to collect information on fidelity to those expectations.

Goals for Monitoring

The content of monitoring focuses on two primary questions:

- Is the foundational coaching model being implemented as intended?
    - Do teachers participate in the content training?
    - Are the key language concepts the focus of coaching?
    - Do coaches observe teacher practice in the classroom?
    - Do coaches conduct feedback sessions with the teachers?
    - Are other fixed dimensions and those that vary among grantees within expected parameters?
- Are the levels of each dimension being implemented as planned, and is there an adequate contrast between the two conditions?
    - **Dosage.** Do coaches observe and conduct feedback sessions with teachers either biweekly or monthly, as appropriate to their assigned condition? Do they devote approximately two hours on each coaching interaction (combined observation and feedback session)?

- **Recipient.** Do coaches meet with just the lead teacher or with both the lead teacher and assistant teacher together, as appropriate to their assigned condition? Do they observe just the lead teacher or both the lead and assistant teacher?
- **Training.** Do coaches assigned to the intensive level participate in the additional training and support sessions designated for them? Do coaches assigned to the typical level participate in their assigned training? Do coaches in the typical level not participate in the training for coaches in the intensive level (or other professional development [PD] that serves as training)?

Use of MIS System for Monitoring

We recommend that monitoring data come primarily from the MIS. As outlined in Section VI, the MIS will provide real-time information on key aspects of the foundational coaching model and on the Dosage and Recipient dimensions. The management reports can be produced on a regular basis by MIS for use by the TA team to ensure that the MIS data are up-to-date for each coach, that all fields have been completed, and that entries align with study expectations. We suggest that the evaluation team staff e-mail coaches directly to reinforce appropriate entries or to indicate when irregularities are detected. If coaches do not respond appropriately, the evaluation team should provide additional intervention as needed to ensure fidelity to the model.

The MIS also provides an opportunity for the developer and trainers[82] of the coaches to document their contact with coaches. This will include documentation of coach support contacts (e.g., calls, e-mails) and assessment of coach practice during training visits. Even with the user-friendly design, it is necessary to provide training and ongoing TA to the study participants who will use the MIS.

Use of Liaisons for Monitoring

In addition, from the research side, we recommend that the evaluation team assign a site liaison to work with each grantee in order to provide for smooth and consistent communication with the grantee. From the program side, each grantee should be asked to assign a grantee liaison to work with the evaluation team. The grantee liaison may likely be the individual responsible for managing implementation of the coaching study for the grantee and will receive a stipend from the research contract for fulfilling these responsibilities. We recommend these site and program liaisons have scheduled monthly calls to discuss any issues identified by the evaluation team (using a standard agenda that includes key monitoring issues such as maintaining contrast across conditions for each dimension), provide the grantee an opportunity to raise issues (such as staffing changes that may affect program implementation), and identify any needs for TA.

TA for the Evaluation and the Intervention

We recommend that the evaluation team provide structured and consistent TA to the participating grantees and coaches to support study fidelity and data collection. TA is important to provide at several levels in a complex study such as the HS Coaching Study. The study goals

[82] The term *developer* is used in this section to capture the work of the program developer as well as program trainers who will work directly with coaches.

of understanding the impact of varying key coaching dimensions will not be successful if the foundational coaching model and the eight coaching conditions are not implemented with fidelity. The monitoring work described here enables the evaluation team to identify threats to fidelity. Problems could inevitably arise during implementation of a complex study in up to 31 grantees and 248 centers. TA is designed both to prevent problems before they arise and to address issues quickly once identified. In this section, we describe the type of TA we recommend as important for grantee administrators, centers and site administrators, classroom teachers, coaches, and trainers. We focus first on the TA to be provided prior to program implementation followed by ongoing assistance during implementation.

Evaluation Start-Up

Introduction for grantees and coaches. During the summer prior to study implementation, the evaluation team and developer should conduct a grantee and coach orientation. This is an opportunity to orient key grantee staff and all coaches to key aspects of the study. A group meeting at a central location for each grantee cluster (e.g., eight locations) provides an opportunity to deliver a consistent message to participants and respond to any questions or issues raised by grantee liaisons and coaches. The orientation is also an opportunity to build a sense of community and buy-in for the study. At this time, sessions that are specific to the needs of subsets of attendees will also be held. These include the following:

- Sessions held regionally (for multiple grantees) where grantee liaisons to provide an overview of coach content, coaching model, coaching dimensions and conditions being tested, and expectations for coach supervision
- Sessions for coaches to provide the core coaching model and the content (literacy and language) of the coaching. All coaches receive the same base level of Coach Training. Coaches in the intensive training condition will receive supplementary training at this session or at another time prior to start-up.
- Sessions for coaches on each of the coaching dimensions and conditions that they are expected to implement
- Sessions for coaches and PD trainers on the MIS that they are expected to complete.

The evaluation team can also develop a short video that can be shown at grantee-level orientations that grantee liaisons and coaches can retain, along with a sample PowerPoint presentation and sample questions and answers they could use to reinforce the information presented at the cluster trainings. At the conclusion of this orientation, grantee liaisons, coaches, developers, and evaluation team members will have a common set of expectations with regard to the foundational coaching model, the relevant experimental coaching conditions, and their respective roles and responsibilities for study implementation.

Introduction for center administrators and teachers. Center administrators and teachers will have received an initial orientation to the study during site recruitment in the spring. During teacher training on language content teachers receive in addition to coaching in the fall, the evaluation team should organize a section of the agenda to provide an overview of the study, roles and responsibilities of all parties, and a clear presentation of expectations with regard to participation in the study. The evaluation team would send the appropriate site liaison to make this

presentation in person at each of the fall trainings. Given the cluster grantee selection approach, there would be one training per cluster; attending eight of these initial sessions is relatively cost-efficient.

Manuals for the Evaluation

The evaluation team should prepare for each participating grantee liaison a manual that outlines the study objectives, procedures, and timelines. The manual would also spell out grantee roles and responsibilities for both implementing the coaching conditions and supporting study data collection. The manual can include copies of relevant documents, including the MOU, detailed implementation and data collection timelines, coach logs, and other data-collection instruments. The manuals can be handed out to and discussed with grantee liaisons during the orientation workshop.

The evaluation should also develop a manual for each coach that includes information on the foundational coaching model, the coaching conditions that each coach is responsible for implementing, screenshots from the coach logs, and instructions for completing the coach logs. In addition, the coach manuals should contain detailed guidance on completing the logs, including suggested anchors for use in their rating of classroom implementation quality.

Ongoing TA

Once the study period begins, there will be ongoing efforts to identify any TA needs. TA may be customized to the needs of grantee, coaches, and trainers. The evaluation team and the developer share responsibility for implementation support for the coaching model. Below, we provide suggestions for the type of supports PD developers, grantees, and coaches should receive.

Support for PD developers. The evaluation team should provide support and TA to the developers in the use of the contact logs, assessments of coach quality, and assessments of classroom implementation. We recommend the evaluation team schedule regular calls with the developers to discuss study progress and emerging implementation issues.

Support for grantees. As noted earlier, an evaluation team site liaison should be assigned to each grantee. We recommend that the evaluation team site liaison conduct monthly calls with the grantee liaison to discuss any challenges that have arisen during the prior month. These calls can also be used to discuss overall progress on program and research implementation. Should a matter requiring urgent attention emerge, it should be the responsibility of the research liaison to contact the grantee immediately to address the matter and may decide to make a site visit.

Support for coaches. The evaluation team has primary responsibility to ensure that coaches maintain the integrity of coaching conditions for each dimension and complete the coach logs. As the developer of the PD strategies and coaching oversees training and work of coaches in the field, issues may emerge related to the implementation of the coaching model, dimensions, or other conditions that need to be addressed. TA to coaches will focus on issues that emerge from regular review of coach logs, regular developer contact with coaches, and evaluation team contact with coaches and grantee liaisons. Coach support and TA should be in the service of ensuring that the coaches are maintaining assigned conditions (described in Section III) and not to replace or supplement planned levels of training.

On the basis of specific issues, the evaluation team and developers should develop specific TA strategies for working with grantees and coaches. Possible TA needs may include the following:

- Coach-specific issues, such as not meeting Dosage expectations, will be identified through coach log reviews or weekly discussions for coaches in the enhanced level between the developer and coach.
- Issues that are common across multiple coaches will be identified through coach log reviews.
- Issues with the coaching process or the content of the coaching will emerge from observation of teacher or Coach Training sessions.

For instance, if many coaches appear to be having a difficult time scheduling both biweekly and monthly coaching sessions with teachers, the developer can agree to put together a sample scheduling guide. If coaches are indicating they are having difficulty scheduling time for coaching sessions with teachers, we suggest the evaluation team put together a set of suggestions on ways center directors and grantee liaisons can support coaching and send them to all grantee liaisons and plan for site liaisons to discuss the issue at the next planned monthly call.

Monitoring and TA Implications of Replacing Coach Training Dimension With Mode Dimension

If the study involves remote versus on-site coaching, monitoring should ensure that coaching is being delivered in the proper Mode. The MIS can be adapted for the coach to indicate whether each coaching session was conducted in person or remotely to facilitate easy tracking. We recommend that the evaluation team monitor this on a weekly basis.

It is likely that substantial TA could be needed to implement a remote coaching model. In addition to training coaches in using a remote approach, teachers will need support acquiring access to the required technology and learning how to use the technology as well as ongoing support should they encounter difficulties using the technology. Some of this support could come from grantee administrative staff and some from the coaches. It may also be necessary to set up a help line to provide as-needed technology support to teachers.

Summary

Systematic monitoring of program implementation together with structured TA to ensure grantee and coach fidelity to the foundational coaching model and the experimental coaching conditions are critical to the success of the study. Monitoring will rely heavily on the MIS that can provide real-time information on key aspects of the foundational coaching model and on the Dosage and Recipient dimensions. We recommend the evaluation team have primary responsibility for identifying issues that emerge from the MIS. TA, delivered by the evaluation team and developers, will include up-front and ongoing training and support designed to anticipate grantee and coach needs as well as address issues that emerge during the course of the study. Developers should also provide ongoing coach-specific support, and the evaluation team, through its site liaisons, will provide similar support to grantee liaisons. Working together, the evaluation team

and the developers should use monitoring and TA to help grantees and coaches implement the foundational coaching model and experimental conditions with fidelity.

The process of monitoring and supporting grantees, teachers and teachers may yield lessons learned for the PD approach. The evaluation team and PD developers should plan to document any additional or alternative materials they would recommend, along with suggestions for ways to improve or streamline training and support for implementing this type of coaching in the future.

XI. Study Timeline and Resource Estimates

Next, we present an estimated timeline for the HS Coaching Study, assuming a start date of January 2016. We estimate that the study will require one and a half years of planning and preparation time. This will give time to select and finalize both the PD approach (set of language strategies and training for teachers and coaches) and measures for the study, as well as proceed through the Office of Management and Budget (OMB) clearance process before recruiting study participants. During the following year, the intervention will be implemented. Then data will be analyzed and findings reported.

- January 2016–July 2017: Study preparation
- August 2017–June 2018: Intervention and data collection
- August 2018–January 2020: Analysis and reporting

The following table lays out nine recommended tasks and associated deliverables with estimated due dates.

Table 24. Schedule of Deliverables

Deliverable	Date	Estimated Date Based on January 2016 Award
Task 1: Communication With Contracting Officer's Representative (COR)		
PD and key staff kickoff meeting with Health and Human Services	On award of contract	01/2016
1.2 Memo summarizing kickoff meeting	One week after meeting	01/2016
1.3 Biweekly COR meetings	Every two weeks following kickoff meeting	Ongoing
Task 2: Study Preparation and Measure Development		
2.1 Report on measures and piloting	12 weeks after contract awarded	04/2016
2.2 Submit draft OMB clearance package for recruitment materials for COR review	12 weeks after contract awarded	04/2016
2.3 Submit first draft OMB clearance package for recruitment materials for OMB	8 weeks after draft recruitment OMB package	06/2016
2.4 Submit revised draft OMB clearance package for recruitment materials for OMB	8 weeks after first draft recruitment OMB package	08/2016
2.5 Submit final recruitment OMB clearance package for OMB	8 weeks after revised draft recruitment OMB package	10/2016
2.6 Submit draft OMB addendum package materials for data collection forms for COR review	36 weeks after contract awarded	10/2016
2.7 Submit first draft OMB addendum package materials for data collection forms for OMB	12 weeks after draft data collection OMB package	01/2017

Deliverable	Date	Estimated Date Based on January 2016 Award
2.8 Submit revised draft OMB addendum package materials for data collection forms for OMB	8 weeks after first draft data collection OMB package	03/2017
2.9 Submit final OMB addendum package materials for data collection forms for OMB	8 weeks after revised draft OMB addendum package	05/2017
Task 3: Developing and Initiating the Intervention		
3.1 Materials used to select PD approach and developer	20 weeks after contract awarded	05/2016
3.2 Memo to name and describe selected PD approach and developer	32 weeks after contract awarded	08/2016
3.3 Interim progress report to describe adaptations to PD approach and site selection of coaches	44 weeks after contract awarded	11/2016
3.4 Manual and materials for training	64 weeks after contract awarded	4/2017
Task 4: Recruitment		
4.1 Biweekly progress reports on recruitment updates	Every month 01–5/2017	Ongoing
4.2 Memo to COR with recommended grantee participants	16 weeks after site recruitment begins	05/2017
4.3 MOUs for participating sites	20 weeks after site recruitment begins	06/2017
Task 5: Random Assignment to Conditions		
5.1 List of condition assignments	8 weeks after submitting MOU for participating sites	08/2017
Task 6: Impact Study		
6.1 Draft impact report	29 months after contract awarded	06/2018
6.2 Final impact report	16 weeks after draft impact report	10/2018
Task 7: Implementation Study		
7.1 Draft implementation study report	16 weeks after final impact report	02/2019
7.2 Final implementation study report	12 weeks after draft implementation report	05/2019
Task 8: Cost Study		
8.1 Draft cost study report	16 weeks after final implementation study report	09/2019

Deliverable	Date	Estimated Date Based on January 2016 Award
8.2 Final cost study report	16 weeks after draft cost study report	01/2020
Task 9: Monthly Reporting		
9.1 Monthly progress reports	Within 15 business days of the end of each calendar month	Ongoing

Next, we present the estimated cost ranges for each task of the study.

Task 1: Communication With COR

This task includes costs for the research contractor as laid out in Table 25.

Table 25. Assumptions for Resource Estimates Associated With Communication With COR

Task and Organization	Assumptions	Cost Range
Communication With COR ▪ Research contractor	Cost of staff time to: ▪ Plan for and attend kickoff meeting ▪ Meet biweekly with COR by phone during course of the four-year study Additional costs for: ▪ Telecommunications, printing ▪ Travel to kickoff meeting Deliverables: ▪ Subtask 1.1: Kickoff meeting with HHS ▪ Subtask 1.2: Memo summarizing the kickoff meeting ▪ Subtask 1.3: Biweekly COR meetings	$120,000–$160,000

Task 2: Study Preparation and Measure Development

This task includes costs for the research contractor (and any subcontractors) as laid out in Table 26.

Table 26. Assumptions for Resource Estimates Associated With Study Preparation and Measure Development

Task and Organization	Assumptions	Cost Range
Study Preparation (Measure Development) ▪ Research contractor	Cost of staff time to: ▪ Develop final measures • For the impact study • For the implementation study • For the cost study ▪ Pilot measures ▪ Develop the MIS system ▪ Produce progress reports ▪ Produce report on piloting measures Additional costs for: ▪ Telecommunications, printing Deliverable: ▪ Subtask 2.1: Report on measures and piloting	$450,000–$600,000
Study Preparation (OMB Package) ▪ Research contractor	Cost of staff time to: **Develop the OMB packages for recruitment and for data collection** Additional costs for: ▪ Telecommunications, printing Deliverables: ▪ Subtasks 2.2–2.5: OMB clearance package for recruitment materials—draft, revised, final ▪ Subtasks 2.6–2.9: OMB addendum package materials for data collection forms—draft, revised, final	$200,000–$250,000
	TOTAL	$650,000–$850,000

Task 3: Developing and Initiating the Intervention

This task includes costs for the research contractor and the PD developer, as laid out in Table 27.

Table 27. Assumptions for Resource Estimates Associated With Developing and Initiating the Intervention

Task and Organization	Assumptions	Cost Range
Developing and Initiating the Intervention ▪ Research contractor	Cost of staff time to: ▪ Develop materials to recruit or select the PD developer and approach (RFP, rubrics, and so on) ▪ Manage the process to select developer ▪ Monitor developer's progress Additional costs for: ▪ Telecommunications, printing Deliverables: ▪ Subtask 3.1: Materials used to select PD approach and developer ▪ Subtask 3.2: Memo to name and describe selected PD approach and developer ▪ Subtask 3.3: Interim progress report to describe adaptations to PD approach and site selection of coaches	$150,000–$180,000
Developing and Initiating the Intervention ▪ PD developer	Cost of staff time to: ▪ Modify PD approach, language intervention for teacher, and Coach Trainings, as needed, for the study ▪ Modify PD approach and trainings for Recipient dimension* ▪ Modify PD approach and trainings for Level II of Coach Trainings dimension* ▪ Modify PD approach and trainings for Mode dimension* ▪ Conduct teacher and Coach Training on PD approaches and intervention Additional costs for: ▪ Telecommunications, printing ▪ Travel to sites to train teachers Deliverables: ▪ Subtask 3.3: Interim progress report to describe adaptations to PD approach and site selection of coaches ▪ Subtask 3.4: Manual and materials for training	$200,000–$250,000
	TOTAL	$350,000–$430,000

* Indicates costs that will apply if dimension is selected.

Task 4: Recruitment

This task includes costs for the research contractor and participating sites as laid out in Table 28.

Table 28. Assumptions for Resource Estimates Associated With Recruitment

Task and Organization	Assumptions	Cost Range
Recruitment (Recruiting Sites) ▪ Research contractor	Cost of staff time to: ▪ Develop recruitment and other study materials (approximately 10 percent of task cost) ▪ Screen potential grantees for inclusion in the study, including phone interview screenings and site visits (approximately 30 percent of task cost) ▪ Recruit grantees, sites, and classrooms to participate in the study (approximately 60 percent of estimated cost) Additional costs for: ▪ Telecommunications, printing Deliverables: ▪ Subtask 4.1: Biweekly progress reports on recruitment updates ▪ Subtask 4.3: MOUs for participating sites	$1.0–$1.2 million
Recruitment (Hiring and Supporting Coaches) ▪ Participating sites	Includes costs for: ▪ Coach salaries ▪ Substitute pay for time teachers are out of class working with the coach ▪ Mileage reimbursement for coach travel between sites ▪ Coach equipment and supplies • Including video cameras and appropriate hardware and software for technology-mediated coaching* • Subtask 3.3: Interim progress report; contribute to describe site selection of coaches	$2.8–$3.3 million
	TOTAL	$3.8–$4.5 million

* Indicates costs that will apply if dimension is selected.

Task 5: Random Assignment to Conditions

This task includes costs for the research contractor as laid out in Table 29.

Table 29. Assumptions for Resource Estimates Associated With Random Assignment to Conditions

Task and Organization	Assumptions	Cost Range
Random Assignment to Conditions ▪ Research contractor	Cost of staff time to: ▪ Compile information about grantees and their classrooms (including location and teacher characteristics) ▪ Assign centers, teachers, and coaches to conditions Additional costs for: ▪ Telecommunications, printing Deliverable: ▪ Subtask 5: List of condition assignments	$60,000–$90,000

Task 6: Impact Study

This task includes costs for the research contractor (and any subcontractors) and the PD developer as laid out in Table 30.

Table 30. Assumptions for Resource Estimates Associated With Impact Study

Task and Organization	Assumptions	Cost Range
Impact Study (Staff Training and Ongoing Monitoring) ▪ Research contractor ▪ PD developer	Cost of research contractor staff time to: ▪ Prepare materials on study procedures ▪ Pay grantee liaison for study responsibilities ▪ Monitor intervention implementation on an ongoing basis Cost of PD developer staff time to: ▪ Monitor intervention implementation on an ongoing basis ▪ Conduct ongoing Coach Training (Level I and Level II follow-ups during the program year) Additional costs for: ▪ Telecommunications, printing ▪ Travel to sites to train staff and monitor implementation as needed	$2.5–$3.4 million
Impact Study (Survey Administration) ▪ Research contractor	Cost of staff time to: ▪ Manage survey administration ▪ Program survey questions into online survey software ▪ Follow up by e-mail and phone with nonrespondents to ensure high response rate Additional costs for: ▪ Telecommunications, printing	$260,000–$350,000

Task and Organization	Assumptions	Cost Range
Impact Study (Classroom Observations) ▪ Research contractor	Cost of staff time to: ▪ Manage observation administration and train observers (approximately 20 percent of cost) ▪ Schedule and conduct two-day classroom observations (approximately 80 percent of task cost) ▪ Enter data (less than 1 percent of task cost) Additional costs for: ▪ Telecommunications, printing ▪ Travel to sites to conduct observations	$3.1–$4.2 million
Impact Study (Analysis and Reporting) ▪ Research contractor	Cost of staff time to: ▪ Produce monthly progress reports ▪ Analyze data ▪ Produce draft report ▪ Respond to HHS comments to revise and produce final report Additional costs for: ▪ Telecommunications, printing ▪ Travel and fees for two conferences to present findings Deliverables: ▪ Subtask 6.1: Draft impact report ▪ Subtask 6.2: Final impact report	$325,000–$400,000
	TOTAL	$6.3–$8.4 million

Task 7: Implementation Study

This task includes costs for the research contractor (and any subcontractors) as laid out in Table 31.

Table 31. Assumptions for Resource Estimates Associated With Implementation Study

Task and Organization	Assumptions	Cost Range
Implementation Study (Independent Observation of Coach Sessions) ▪ Research contractor	Cost of staff time to: ▪ Manage observation administration and train observers ▪ Conduct independent observations of sample of coaching sessions Additional costs for: ▪ Telecommunications, printing ▪ Travel to sites to conduct observations	$1.4–$1.9 million
Implementation Study (Site Visits and Staff Interviews) ▪ Research contractor	Cost of staff time to: ▪ Manage interview administration and train interviewers ▪ Interview sample of key staff at grantee sites Additional costs for: ▪ Telecommunications, printing ▪ Travel to sites to conduct observations	$160,000–$250,000
Implementation Study (Analysis and Reporting) ▪ Research contractor	Cost of staff time to: ▪ Produce monthly progress reports ▪ Code interviews and analyze data ▪ Produce draft implementation report ▪ Respond to HHS comments to revise and produce final report Additional costs for: ▪ Telecommunications, printing Deliverables: ▪ Subtask 7.1: Draft implementation study report ▪ Subtask 7.2: Final implementation study report	$300,000–$350,000
	TOTAL	$1.9–$2.5 million

Task 8: Cost Study

This task includes costs for the research contractor (and any subcontractors) as laid out in Table 32.

Table 32. Assumptions for Resource Estimates Associated With Cost Study

Task and Organization	Assumptions	Cost Range
Cost Study ▪ Research contractor	Cost of staff time to: ▪ Conduct additional training and monitoring of sites for cost data collection ▪ Analyze data ▪ Produce draft cost study report ▪ Respond to HHS comments to revise and produce final report Additional costs for: ▪ Telecommunications, printing Deliverables: ▪ Subtask 8.1: Draft cost study report ▪ Subtask 8.2: Final cost study report	$300,000–$400,000

Task 9: Monthly Reporting

This task includes costs for the research contractor as laid out in Table 33.

Table 33. Assumptions for Resources Estimates Associated With Monthly Reporting

Task and Organization	Assumptions	Cost Range
Monthly Reporting ▪ Research contractor	Cost of staff time to: ▪ Produce monthly reports to the COR Additional costs for: ▪ Telecommunications, printing Deliverable: ▪ Subtask 9.1: Monthly reports submitted within 15 business days of the end of each calendar month	$70,000–$90,000

References

Administration for Children and Families Early Childhood Learning & Knowledge Center. (n.d.). *Head Start Center locations datasets*. Retrieved from http://eclkc.ohs.acf.hhs.gov/hslc/hs/directories/center-data

Aikens, N., & Akers, L. (2011). *Background review of existing literature on coaching.* Washington, DC: Mathematica Policy Research. Retrieved from http://www.first5la.org/files/07110_502.2CoachingLitRev_FINAL_07072011.pdf

Assel, M. A., Landry, S. H., Swank, P. R., & Gunnewig, S. (2007). An evaluation of curriculum, setting, and mentoring on the performance of children enrolled in pre-kindergarten. *Reading and Writing, 20*(5), 463–494.

Baker, C. N., Kupersmidt, J. B., Voegler-Lee, M. E., Arnold, D. H., & Willoughby, M. T. (2010). Predicting teacher participation in a classroom-based integrated preventive intervention for preschoolers. *Early Childhood Research Quarterly, 25*, 270–283.

Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory.* Englewood Cliffs, NJ: Prentice Hall.

Berkel, C., Mauricio, A. M., Schoenfelder, E., & Sandler, I. N. (2010). Putting the pieces together: An integrated model of program implementation. *Prevention Science, 12* 23–33.

Birman, B. F., Desimone, L., Porter, A. C., & Garet, M. S. (2000). Designing professional development that works. *Educational Leadership, 57*(8), 28–32.

Black, A. R., Somers, M.-A., Doolittle, F., Unterman, R., & Grossman, J. B. (2009). *The evaluation of enhanced academic instruction in after-school programs: Final report* (NCEE 2009-4077). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. Retrieved from http://ies.ed.gov/ncee/pubs/20094077/pdf/20094077.pdf

Bloom, H. S. (1984). Accounting for no-shows in experimental evaluation designs. *Evaluation Review, 8*(2): 225–246.

Bloom, H. S. (2005). Randomizing groups to evaluate place-based programs. In H. S. Bloom, *Learning more from social experiments: Evolving analytical approaches* (pp. 115–172). New York, NY: Russell Sage.

Bloom, H. S., Richburg- Hayes, L., & Black, A. R. (2007). Using covariates to improve precision for studies that randomize schools to evaluate educational interventions. *Educational Evaluation and Policy Analysis, 29*(1), 30–59.

Boller, K., Blair, R., Del Grosso, P., & Paulsell, D. (2010). *Better beginnings: The Seeds to Success modified field test: Impact evaluation findings.* Princeton, NJ: Mathematica Policy Research.

Bowman, B. T., Donovan, M. S., & Burns, M. S. (2000). *Eager to learn: Educating our preschoolers.* Washington, DC: National Academies Press.

Bryant, D. M., Wesley, P. W., Burchinal, M., Sideris, J., Taylor, K., Fenson, C., et al. (2009). *The QUINCE-PFI study: An evaluation of a promising model for child care provider training* (Final Report). Chapel Hill, NC: Frank Porter Graham Child Development Institute.

Carlson, J. S., Mackrain, M. A., Van Egren, L. A., Brophy-Herb, H., Kirk, R. H., Marciniak, D., et al. (2012). Implementing a statewide early childhood mental health consultation approach to preventing childcare expulsion. *Infant Mental Health Journal, 33*(3), 265–273.

Center for Advanced Study of Teaching and Learning. (2013). Curry School of Education.

Cohen, E., & Kaufmann, R. K. (2005). *Early childhood mental consultation* (DHHS CMHS-SVP0151). Rockville, MD: Center for Mental Health Services, Substance Abuse and Mental Health Services Administration.

Collins, L. M., Dziak, J. J., & Li, R. (2009). Design of experiments with multiple independent variables: A resource management perspective on complete and reduced factorial designs. *Psychological Methods, 14*(3), 202–224.

Collins, L. M., Murphy, S. A., Nair, V. N., & Strecher, V. J. (2005). A strategy for optimizing and evaluating behavioral interventions. *Annals of Behavioral Medicine, 30*(1), 65–73.

Collins, L. M., Murphy, S. A., & Strecher, V. (2007). The multiphase optimization strategy (MOST) and the sequential multiple assignment randomized trial (SMART): New methods for more potent eHealth interventions. *American Journal of Preventive Medicine, 32*(5), S112–S118.

Collins, L.M., Nahum-Shani, I., & Almirall, D. (in press). Optimization of behavioral dynamic treatment regimens based on the sequential, multiple assignment, randomized trial (SMART). *Clinical Trials*.

Curbow, B., Spratt, K., Ungaretti, A., McDonnell, K., & Breckler, S. (2000). Development of the child care worker job stress inventory. *Early Childhood Research Quarterly, 15*, 515–536.

Cusumano, D., Armstrong, K., Cohen, R., & Todd, M. (2006). Indirect impact: How early childhood educator training and coaching impacted the acquisition of literacy skills in preschool students. *Journal of Early Childhood Teacher Education, 27*, 363–377.

Cusumano, D. L. (2005). *Early learning experiences: Education with coaching and the effects on the acquisition of literacy skills in preschool children* (Doctoral dissertation). Retrieved from http://scholarcommons.usf.edu/etd/2843/

Department of Health and Human Services. (2011). 45 CFR Part 1307—Policies and procedures for designation renewal of Head Start and early Head Start grantees. *Federal Register, 76*(217), 70010–70032. Retrieved from http://eclkc.ohs.acf.hhs.gov/hslc/standards/Head%20Start%20Requirements/1307/Part%201307-FRNotice_2011-28880.pdf

Diamond, K. E., & Powell, D. R. (2011). An iterative approach to the development of a professional development intervention for Head Start teachers. *Journal of Early Intervention, 33*(1), 75–93.

Dickinson, D. K., & McCabe, A. (2001). Bringing it all together: The multiple origins, skills, and environmental supports of early literacy. *Learning Disabilities Research and Practice, 16*(4), 186–202.

Dickinson, D. K., McCabe, A., Anastasopoulos, L., Peisner-Feinberg, E. S., & Poe, M. D. (2003). The comprehensive language approach to early literacy: The interrelationships among vocabulary, phonological sensitivity, and print knowledge among preschool-aged children. *Journal of Educational Psychology, 95*(3), 465–481.

Domitrovich, C., Cortes, R. C., & Greenberg, M. T. (2000). The Teacher Style Rating Scale technical report. Unpublished manuscript, Pennsylvania State University.

Domitrovich, C. E., Gest, S. D., Gill, S., Jones, D., & DeRousie, R. S. (2009a). Individual factors associated with professional development training outcomes of the Head Start REDI program. *Early Education and Development, 20*(3), 402–430.

Domitrovich, C. E., Gest, S. D., Jones, D., Gill, S., & DeRousie, R. M. S. (2010). Implementation quality: Lessons learned in the context of the Head Start REDI trial. *Early Childhood Research Quarterly, 25*(3), 284–298.

Domitrovich, C. E., Gest, S. D., Sukhdeep, G., Bierman, K. L., Welsh, J. A., & Jones, D. (2009b). Fostering high-quality teaching with an enriched curriculum and professional development support: The Head Start REDI Program. *American Educational Research Journal*, 46(2), 567–597.

Dunst, C. J., Trivette, C. M., & Hamby, D. W. (2007). Meta-analysis of family-centered helpgiving practices research. *Mental Retardation and Developmental Disabilities Research Reviews, 13*(4), 370–378.

Dunst, C. J., Trivette, C. M., & Hamby, D. W. (2010). Meta-analysis of the effectiveness of four adult learning methods and strategies. *International Journal of Continuing Education and Lifelong Learning, 3*(1), 91–112.Epstein, A. S. (1993). *Training for quality: Improving early childhood programs through systematic inservice training.* Ypsilanti, MI: HighScope Press.

Fiene, R. (2002). Improving child care quality through an infant caregiver mentoring project. *Child and Youth Care Forum, 31*(2), 79–87.

Fishman, M., & Wille, J. *MSHS CARES report.* Manuscript submitted for publication.

Fixsen, D. L., Naoom, S. F., Blase, K. A., & Friedman, R. M. & Wallace, F.(2005). *Implementation research: A synthesis of the literature*.

Fuentes, Y. S. (2010). *The Head Start child development and early learning framework: Promoting positive outcomes in early childhood programs serving children 3-5 years old.* Washington, DC: U.S. Department of Health and Human Services, Administration for Children and Families, Office of Head Start. Retrieved from http://eclkc.ohs.acf.hhs.gov/hslc/tta-system/teaching/eecd/Assessment/Child%20Outcomes/HS_Revised_Child_Outcomes_Framework(rev-Sept2011).pdf

Garet, M. S., Cronen, S., Eaton, M., Kurki, A., Ludwig, M., Jones, W., et al. (2008). *The impact of two professional development interventions on early reading instruction and achievement* (NCEE 2008-4030). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Garet, M. S., Porter, A. C., Desimone, L., Birman, B. F., & Yoon, K. S. (2001). What makes professional development effective? Results from a national sample of teachers. *American Educational Research Journal, 38*(4), 915–945.

Glisson, C. (2007). Organizational climate, job satisfaction, and service outcomes in child welfare agencies. Manuscript submitted for publication.

Graham, J. W., Taylor, B. J., Olchowski, A. E., & Cumsille, P. E. (2006). Planned missing data designs in psychological research. *Psychological Methods, 11*(4), 323–343.

Green, D. P., Ha, S. E., & Bullock, J. G. (2010). Enough already about "black box" experiments: Studying mediation is more difficult than most scholars suppose. *Annals of the American Academy of Political and Social Science, 628*(1), 200–208.

Halle, T. (2008, October). *Dosage of professional development: What we have learned.* Paper presented at the Conference on Early Childhood Professional Development, Ann Arbor, MI.

Halle, T., Zaslow, M., Tout, K., Starr, R., Wessel, J., & McSwiggan, M. (2010). Beyond how much: What are we learning about structuring effective early childhood professional development? In S. B. Neuman & M. Kamil (Eds.), *Preparing teachers for the early childhood classroom: Proven models and key principles* (pp. 175–188). Baltimore, MD: Paul H. Brookes.

Harms, T., Clifford, R. M., & Cryer, D. (2004). *Early Childhood Environment Rating Scale Revised (ECERS-R).* New York, NY: Teachers College Press.

Hart, B., & Risley, T. R. (2004). The early catastrophe. *Education Review, 77*(1), 100–118.

Hemmeter, M. L., Fox, L., & Snyder, P. (2014). The Teaching Pyramid: A tiered model to address challenging behaviors and promote social-emotional development. In V. Buysse & E. Peisner-Feinberg (Eds.), *Handbook of response to intervention in early intervention* (pp. 85–103). Baltimore, MD: Paul H. Brookes.

Herren, J. K. (2009). *Being an effective mentor-coach* (Head Start Bulletin 80: Mental Health). Washington, DC: U.S. Department of Health and Human Services, Administration for Children and Families, Office of Planning, Research and Evaluation.

Hill, C., Bloom, H., Black, A. R., & Lipsey, M. W. (2007). *Empirical benchmarks for interpreting effect sizes in research*. New York, NY: MDRC.

Horvath, A. O., & Greenberg, L. S. (1989). Development and validation of the Working Alliance Inventory. *Journal of Counseling Psychology*, *36*(2), 223–233.

Howard, E., Allard Agnamba, L., Wessel, J., & Rankin, V. (in press). *Uses and definitions of implementation terms in early childhood education research*. Washington, DC: U.S. Department of Health and Human Services, Administration for Children and Families, Office of Planning, Research and Evaluation.

Howard, E., & Drummond, K. (2013). National Head Start Conference Presentation. Washington, DC.

Howard, E. C., Rankin, V. E., Fishman, M., Hawkinson, L. E., McGroder, S. M., Helsel, F. K., et al. (2013). *The descriptive study of the Head Start Early Learning Mentor Coach initiative*. Washington, DC: U.S. Department of Health and Human Services, Administration for Children and Families, Office of Planning, Research and Evaluation.

Isner, T., Tout, K., Zaslow, M., Soli, M., Quinn, K., Rothenberg, L., et al. (2011). *Coaching in early care and education programs and quality rating and improvement systems (QRIS): Identifying promising features.* Washington, DC: Child Trends.

Joyce, B., & Showers, B. (1996). The evolution of peer coaching. *Improving Professional Development Practice, 53*(6), 12–16.

Joyce, B., & Showers, B. (2002). *Student achievement through staff development* (3rd ed.). Alexandria, VA: Association for Supervision and Curriculum Development.

Justice, L. M., Mashburn, A., Hamre, B., & Pianta, R. (2008). Quality of language and literacy instruction in preschool classrooms serving at-risk pupils. *Early Childhood Research Quarterly, 23*, 51–68.

Justice, L. M., Pence, K. L., Beckman, A. R., Skibbe, L. E., & Wiggins, A. K. (2005). *Scaffolding with storybooks: A guide for enhancing young children's language and literacy achievement*. Newark, DE: International Reading Association.

Kessler, R. C., Andrews, G., Colpe, L. J., Hiripi, E., Mroczek, D. K., Normand, S. L., et al. (2002). Short screening scales to monitor population prevalences and trends in non-specific psychological distress. *Psychological Medicine, 32*(6), 959–976.

Kish, L. (1965). *Survey sampling*. New York, NY: John Wiley & Sons.

Koh, S., & Neuman, S. B. (2009). The impact of professional development in family child care: A practice-based approach. *Early Education and Development, 20*(3), 537–562.

Kugler, K. C., Trail, J. B., Dziak, J. J., & Collins, L. M. (2012). *Effect coding versus dummy coding in analysis of data from factorial experiments* (Technical Report 12-120). State College, PA: The Methodology Center, Pennsylvania State University.

Landry, S. H., Anthony, J. L., Swank, P. R., & Monsegue-Bailey, P. (2009). Effectiveness of comprehensive professional development for teachers of at-risk preschoolers. *Journal of Educational Psychology, 101,* 448–465.

Landry, S. H., Swank, P. R., Anthony, J. L., & Assel, M. A. (2011). An experimental study evaluating professional development activities within a state funded pre-kindergarten program. *Reading and Writing, 24*(8), 971–1010.

Landry, S. H., Swank, P. R., Smith, K. E., Assel, M. A., & Gunnewig, S. B. (2006). Enhancing early literacy skills for preschool children: Bringing a professional development model to scale. *Journal of Learning Disabilities, 39*(4), 306–324.

Levin, H. M., & McEwan, P. J. (2001). *Cost-effectiveness analysis: Methods and applications*. Thousand Oaks, CA: SAGE Publications.

Lloyd, C. M., & Bangser, M. (2009). *Promoting preschool quality through effective classroom management: Implementation lessons from the Foundations of Learning Demonstration.* New York, NY: MDRC.

Lloyd, C. M., & Modlin, E. L. (2012). *Coaching as a key component in teachers' professional development: Improving classroom practices in Head Start settings* (OPRE Report 2012-4). Washington, DC: U.S. Department of Health and Human Services, Administration for Children and Families, Office of Planning, Research and Evaluation.

Lonigan, C. J. (2006). Development, assessment, and promotion of pre-literacy skills. *Early Education and Development, 17*, 91–114.

Lonigan, C. J., & Whitehurst, G. J. (1998). Relative efficacy of parent and teacher involvement in a shared-reading intervention for preschool children from low-income backgrounds. *Early Childhood Research Quarterly, 13*(2), 263–290.

Mashburn, A. J., Downer, J. T., Hamre, B. K., Justice, L. M., & Pianta, R. C. (2010). Consultation for teachers and children's language and literacy development during pre-kindergarten. *Applied Developmental Science, 14*(4), 179–196.

Mattera, S., Lloyd, C. M., Fishman, M., & Bangser, M. (2013). *A first look at the HS CARES demonstration: Large-scale implementation of programs to improve children's social-emotional competence* (OPRE Report 2013-47). Washington, DC: U.S. Department of Health and Human Services, Administration for Children and Families, Office of Planning, Research and Evaluation.

McGroder, S. M., Howard, E. C., Fishman, M., Rankin, V. E., & Helsel, F. K. (2012). *Putting the pieces together: A program logic model for coaching in Head Start: From the descriptive study of the Head Start Early Learning Mentor Coach initiative.* Washington, DC: U.S. Department of Health and Human Services, Administration for Children and Families, Office of Planning, Research and Evaluation.

Morris, P., Lloyd, C. M., Millenky, M., Leacock, N., Raver, C., & Bangser, M. (2013). *Using classroom management to improve preschoolers' social and emotional skills*. New York, NY: MDRC.

Morris, P., Raver, C. C., Millenky, M., Jones, S., & Lloyd, C. M. (2010). *Making preschool more productive: How classroom management training can help teachers.* New York, NY: MDRC.

National Association for the Education of Young Children. (1993). *A conceptual framework for early childhood professional development: A position statement of the National Association for the Education of Young Children.* Washington, DC: Author.

National Association for the Education of Young Children. (2013). *NAEYC early childhood program standards and accreditation criteria and guidance for assessment.* Washington DC: Author.

National Center on Quality Teaching and Learning. (2012). *What do we know about coaching?* Washington, DC: Author.

National Early Literacy Panel. (2008). *Developing early literacy: report of the national early literacy panel.* Washington DC: National Institute for Literacy.

National Institute for Literacy. (2009). *Learning to talk and listen: An oral language resource for early childhood caregiver.* Washington, DC: Author.

National Professional Development Center on Inclusion. (2008). *What do we mean by professional development in the early childhood field?* Retrieved from http://npdci.fpg.unc.edu/sites/npdci.fpg.unc.edu/files/resources/NPDCI_ProfessionalDevelopmentInEC_03-04-08_0.pdf

Neuman, S. B., & Cunningham, L. (2009). The impact of professional development and coaching on early language and literacy instructional practices. *American Educational Research Journal, 46*(2), 532–566.

Neuman, S. B., & Wright, T. S. (2010). Promoting language and literacy development for early childhood educators: A mixed-methods study of coursework and coaching. *Elementary School Journal, 111,* 63–86.

Palsha, S. A., & Wesley, P. W. (1998). Improving quality in early childhood environments through on-site consultation. *Topics in Early Childhood Special Education, 18*(4), 243–253.

Peterson, S. M., Baker, A. C., & Weber, M. R. (2010). *Stage of change scale for early education and care 2.0: Mentor/coach form.* Rochester, NY: Children's Institute.

Pianta, R., La Paro, K., & Hamre, B. K. (2008). *Classroom Assessment Scoring System.* Baltimore, MD: Paul H. Brookes.

Powell, D. R., Diamond, K. E., & Burchinal, M. R. (2012). Using coaching-based professional development to improve Head Start teachers' support of children's oral language skills. In C. Howes, B. K. Hamre, & R. C. Pianta (Eds.), *Effective professional development in early childhood education: Improving teacher practice and child outcomes* (pp. 13–29). Baltimore, MD: Paul H. Brookes.

Powell, D. R., Diamond, K. E., Burchinal, M. R., & Koehler, M. J. (2010). Effects of an early literacy professional development intervention on Head Start teachers and children. *Journal of Educational Psychology, 102,* 299–312.

Powell, D. R., Diamond, K. E., & Koehler, M. J. (2010). Use of a case-based hypermedia resource in an early literacy coaching intervention with pre-kindergarten teachers. *Topics in Early Childhood Special Education, 29*(4), 239–249.

Powell, D. R., Steed, E. A., & Diamond, K. E. (2010). Dimensions of literacy coaching with Head Start teachers. *Topics in Early Childhood Special Education, 30,* 148–161.

Puma, M. J., Olsen, R. B., Bell, S. H., & Price, C. (2009). *What to do when data are missing in group randomized controlled trials* (NCEE 2009-0049). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Raudenbush, S. W. (2009). Adaptive centering with random effects: An alternative to the fixed effects model for studying time-varying treatments in school settings. *Journal of Education, Finance and Policy, 4*(4), 468–491.

Raver, C. C., Domitrovich, C., Greenberg, M., Morris, P. A., & Mattera, S. (2012). *Adapted Teaching Style Rating Scale*. MDRC: NY.

Raver, C. C., Jones, S. M., Li-Grining, M., Metzger, M., Champion, K. M., & Sardin, L. (2008). Improving preschool classroom processes: Preliminary findings from a randomized trial implemented in Head Start settings. *Early Childhood Research Quarterly, 63*(3), 253–255.

Raver, C. C., Jones, S. M., Li-Grining, C., Zhai, F., Bub, K., & Pressler, E. (2011). CSRP's impact on low-income preschoolers' preacademic skills: Self-regulation as a mediating mechanism. *Child Development, 82*(1), 362–378.

Rhoads, C. (2011). The implications of contamination for experimental design in education research. *Journal of Educational and Behavioral Statistics, 36*(1), 76–104.

Rodriguez, G., & Knuth, R. (2000). *Providing professional development for effective technology use.* Oak Brook, IL: North Central Regional Educational Laboratory. Retrieved from http://www.ncrel.org/sdrs/areas/issues/methods/technlgy/te1000.htm

Roskos, K. A., Christie, J., Vukelich, C., & Han, M. (2003). The effects of a well-designed literacy program on young children's language and literacy development (pp. 447–448). New York, NY: Mailman School of Public Health, Columbia University.

Rubin, R., Sutterby, J. A., & Hoffman, V. J. (2011). Professional development in culturally diverse settings. *Preparing teachers for the early childhood classroom: Proven models and key principles*, 163-172.

Schaefer, E. S., & Edgerton, M. (1985). Parent and child correlates of parental modernity. In I. E. Sigel (Ed.), *Parent belief systems: The psychological consequences for children* (pp. 287–318). Hillsdale, NJ: Erlbaum.

Schochet, P. Z. (2008). *Technical methods report*: *Guidelines for multiple testing in impact evaluations* (NCEE 2008-4018). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. New York, NY: Houghton Mifflin.

Sheridan, S. M., Edwards, C. P., Marvin, C. A., & Knoche, L. L. (2009). Professional development in early childhood programs: Process issues and research needs. *Early Education and Development, 20*(3), 377–401.

Shidler, L. (2009). The impact of time spent coaching for teacher efficacy on student achievement. *Early Childhood Education Journal, 36*(5), 453–460.

Smith, M., Brady, J., & Anastasopoulos, L. (2008). *User's guide to the Early Language & Literacy Classroom Observation: Pre-K tool.* Baltimore, MD: Paul H. Brookes.

Smith, S., Davidson, S., Weisenfeld, G., & Katsaros, S. (2001). *Supports for Early Literacy Assessment (SELA).* Unpublished instrument.

Snyder, P., Hemmeter, M. L., & McLaughlin, T. (2011). Professional development in early childhood intervention: Where we stand on the silver anniversary of PL 99-457. *Journal of Early Intervention, 33*(4), 357–370.

Snyder, P., Hemmeter, M. L., Meeker, K. A., Kinder, K., Pasia, C., & McLaughlin, T. (2012). Characterizing key features of the early childhood professional development literature. *Infants and Young Children*, *25*(3), 188–212.

Somers, M. A., Collins, L. M., & Maier, M. (2013). *Head Start professional development: Developing the evidence for best practices in coaching: Review of experimental designs for evaluating component effects in social interventions.* Washington, DC: U.S. Department of Health and Human Services, Administration for Children and Families, Office of Planning, Research and Evaluation.

Spencer, E.J., Goldstein, H., & Kaminski, R. (2012). Teaching vocabulary in storybooks: embedding explicit vocabulary instruction for young children. *Young Exceptional Children, 15*(1), 18–32.

Sylva, K., Siraj-Blatchford, I., & Taggart, B. (2010). *ECERS-E: The four curricular subscales extension to the Early Childhood Environment Rating Scale (ECERS)* (4th ed.). New York, NY: Teachers College Press.

Taylor, J. E. (2008). Instructional coaching: The state of the art. In M. M. Mangin & S. R. Stoelinga (Eds.), *Effective teacher leadership: Using research to inform and reform* (pp. 10–35). New York, NY: Teachers College Press.

Taylor, J. E., Lloyd, C. M., Tout, K., Powell, D., Zaslow, M., Agnamba, L. A., et al. (2013). *Head Start professional development: Developing the evidence for best practices in coaching: Review of coaching frameworks, components, and outcomes.* Washington, DC: U.S. Department of Health and Human Services, Administration for Children and Families, Office of Planning, Research and Evaluation.

Tout, K., Halle, T., Zaslow, M., & Starr, R. (2009). Evaluation of the early childhood educator professional development program: Final report. *Prepared for the US Department of Education, Policy and Programs Studies Service. Washington, DC, Child Trends*.

Tout, K., Zaslow, M., Halle, T., & Forry, N. (2009). Issues for the next decade of quality rating and improvement systems. *Washington, DC: Child Trends*.

Walker, D., Greenwood, C., Hart, B., & Carta, J. (1994). Predication of school outcomes based on early language production and socioeconomic factors. *Child Development, 65*(2), 606–621.

Wasik, B. A., & Bond, M. A. (2001). Beyond the pages of a book: Interactive book reading and language development in preschool classrooms. *Journal of Educational Psychology, 93*(2), 243–250.

Wasik, B. A., Bond, M. A., & Hindman, A. (2006). The effects of a language and literacy intervention on Head Start children and teachers. *Journal of Educational Psychology, 98*(1), 63–74.

Wasik, B. A., & Hindman, A. H. (2011). Improving vocabulary and pre-literacy skills of at-risk preschoolers through teacher professional development. *Journal of Educational Psychology, 103*(2), 455–469.

Wasik, B. A., Mattera, S. K., Lloyd, C. M., & Boller, K. (2013). *Intervention dosage in early childhood care and education: It's complicated* (Research Brief OPRE 2013-15). Washington, DC: Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services. Retrieved from https://www.acf.hhs.gov/sites/default/files/opre/dosage_brief_final_001_0.pdf

What Works Clearinghouse. (2014). *Procedures and standards handbook* (Version 3.0). Washington, DC: U.S. Department of Education.

Whitebook, M. (in press). Supportive Environmental Quality Underlying Adult Learning (SEQUAL). Berkeley, CA: Center for the Study of Child Care Employment, University of California.

Whitehurst, G. J., & Lonigan, C. J. (1998). Child development and emergent literacy. *Child Development, 69*, 848–872.

Zan, B., & Donegan-Ritter, M. (2014). Reflecting, coaching and mentoring to enhance teacher-child interactions in Head Start classrooms. *Early Childhood Education Journal, 42*(2), 93–104.

Zaslow, M., & Martinez-Beck, I. (Eds.). (2006). *Critical issues in early childhood professional development.* Baltimore, MD: Paul H. Brookes.

Zucker, T. A., Justice, L. M., Piasta, S. B., & Kaderavek, J. N. (2010). Preschool teachers' literal and inferential questions and children's responses during whole-class shared reading. *Early Childhood Research Quarterly, 25*(1), 65–83.

Appendix A. Other Coaching Dimensions Considered for Systematic Variation

Coaching Dimensions Considered for the HS Coaching Study

As discussed in Section III, the design team considered several coaching dimensions for possible inclusion and systematic variation for the HS Coaching Study. As explained in detail in Section III, the process of reviewing and selecting dimensions for inclusion into the HS Coaching Study was an iterative process that involved discussions among the design team, OPRE staff, and academic researchers, as well as consultation with stakeholders in the HS field. For each coaching dimension, we noted key design and implementation considerations and categorized each dimension to indicate a low, moderate, or high level for each criterion. Our literature review, along with stakeholder input that we received early in the design process, helped us to determine these ratings. Once we established a small pool of top-priority dimensions that ranked highest according to the criteria, we once again consulted with stakeholders to help us with the final selection process of three dimensions, and a fourth dimension, to be considered as an alternative.

The complete list of coaching dimensions considered for the study is as follows:

Structure:

- Goals
- Recipient
- Dosage
- Format
- Additional PD coordinated with coaching
- Mode

Process:

- Coach-teacher relationship
- Use of tools
- Use of strategies
    - Planning
    - Modeling
    - Observations
    - Feedback

Staffing:

- Coach selection
- Coach caseload
- Coach training
- Coach supervision

In Table A1, organized by dimension category, we provide a summary of our final rationale for inclusion or exclusion of the coaching dimensions that we considered.

Table A1. Coach Dimensions Considered for the HS Coaching Study

Coach Dimension	Definition	Rationale for Inclusion or Exclusion
Structure Dimensions		
Goals	Coaching approaches typically target key goals or interest areas. Coaching models vary in terms of the particular topics or goals on which models focus, the breadth or number of goals, and the degree of clarity or specificity of the goals.	**Excluded:** Varying this dimension would have made implementation of the HS Coaching Study highly complex and more costly. If multiple topics or sets of goals had been compared, likely multiple PD developers would have been needed in order to focus on the different topics. Multiple sets of resources, training plans, and trainers may have been required. For example, if one set of coaching goals is to focus on social-emotional development and another set of coaching goals is to focus on language, multiple PD developers would be needed. Creating a standard set of coaching goals, which we recommend in this design by having a foundational approach focused on language, will facilitate implementation and reduce costs by enabling the one PD developer and all coaches to focus on one topic area and one set of classroom practices.
Recipient	The staff members who receive coaching services.	**Included:** Theoretically, coaching approaches that target different Recipients have different strengths. For example, coaching only the lead teacher makes it easier to individualize coaching, whereas coaching the teaching team enables all teachers to be held accountable for implementing evidence-based practices in the classroom. The Recipient dimension was also considered one of the more feasible to implement because it could be done in conjunction with the other selected dimensions. Despite the theory that the team approach may produce higher quality for children throughout the day, there is a trade-off in terms of time and cost. Varying the Recipient of coaching will enable the development team to help programs determine how best to spend their limited resources in regard to which classroom staff should receive coaching.
Dosage	The amount of exposure to individual coaching that a Recipient receives. It can be determined by multiplying the frequency of coaching sessions, length of each session, and duration of the intervention.	**Included:** Of the numerous coaching dimensions examined, Dosage had the strongest theoretical rationale for positive effects and the greatest amount of empirical evidence for its effectiveness (Mashburn et al., 2010). It is also one of the most feasible dimensions to manipulate because the same coach can vary. Dosage systematically without extensive training, TA, or piloting. Varying the levels of Dosage teachers receive will enable the contract team to potentially learn whether enhanced levels of coaching make a greater difference in teacher outcomes than typical levels do.

Coach Dimension	Definition	Rationale for Inclusion or Exclusion
Format	*Format* refers to the personnel involved in coaching sessions. Coaches often meet with a single teacher or a teaching pair, or it is also possible for coaching to occur in group sessions with three or more teachers simultaneously.	**Excluded:** When PD is defined as coaching, it most often occurs at the individual level, according to both extant research and HS stakeholders. We felt that providing coaching to a group of three or more individuals could be confused with other types of group professional training and development. We also felt that a group coaching dimension could be confounded with the Recipient dimension. Both group coaching and the Recipient dimension provide staff other than the lead teacher with peer collaboration for implementing best practices. Using a consistent individualized format, which we recommend in this design, will enable coaches to conduct classroom observations of teachers and provide opportunities for the coaches and teachers to meet to reflect on experiences, provide feedback, and model best practices.
Additional PD coordinated with coaching	Using coaching with other teacher PD activities (e.g., workshops, coursework, professional learning communities) that relate to coaching goals.	**Excluded:** Additional PD to supplement coaching is confounded with other dimensions, such as Dosage and delivery Mode. For example, the amount of time a teacher would spend in additional trainings would impact the total Dosage for this PD effort. In addition, other PD in combination with delivery Mode (i.e., the technologically mediated condition) would (a) create a mixed-Mode condition, in which some PD is in person and other PD is via technology, or (b) necessitate the development of additional technologically mediated trainings and meetings. Just providing initial training and coaching sessions, which we recommend, does not provide as many opportunities for teachers to share knowledge with each other. If this did occur, it would be difficult to determine what the causes of any change in teacher practice were.
Mode	The way coaching is conducted and coaching services are provided: in person or via technology.	**Included as alternative:** Testing Mode provides potential benefits to the field related to long-term cost savings and coaching in isolated or rural areas. Although some studies have found both on-site and technologically mediated coaching approaches to be equally effective for changing teacher practice, the body of research is limited. Although in-person coaching is currently more common, there is increasing interest and some evidence supporting the positive effects of remote coaching. Systematically varying the Mode of coaching will enable the contract team to determine each Mode's impact on teacher practice.

Coach Dimension	Definition	Rationale for Inclusion or Exclusion
Process Dimensions		
Coach-teacher relationship	A coach's role may vary from expert, friend, or emotional supporter to advocate (Howard et al., 2013). The nature of the relationship between the coach and the teacher may also vary between a supervisory relationship and nonsupervisory support.	**Excluded:** Some aspects of teacher-coach relationships are difficult to specify or to vary (e.g., telling the coach to act as a warm, supportive "friend"). We considered testing the variation in the coach's supervisorial role. However, the design team was concerned that varying the supervisory relationship could become a barrier to potential participants because the study could either (a) change the nature of the current center administrators' responsibilities and duties by asking them to coach or (b) change the nature of center administration by requiring that a coach join that structure. Systematically varying the supervisory relationship between the coach and the teacher also is confounded with a dimension that was highly prioritized (i.e., Coach Training). For example, coaches would need additional training to be in an administrative or supervisory role. Having coaches not take on the additional role of becoming supervisors enables them to prioritize other focal areas when participating in Coach Training. In addition, in typical practice, coaches are not formal supervisors of program staff (Howard et al., 2013).
Use of Tools	The types of tools and data used to guide and systematize a coach's work with teachers.	**Excluded:** Although some basic set of coaching tools may be required to structure the work of effective coaching, the development of multiple coaching tools requires resource-intensive effort. The type of training coaches would receive would be impacted by the use of the many tools they could be using. For example, coaches would need initial training on how to use each specific tool and support on how to share what they are learning from using these tools with teachers to improve teacher practice. Establishing a plan for the use of certain tools enables all coaches to receive the same common baseline training to support the use of specific needs assessments, which reduces the amount of time coaches need to participate in training to learn various tools, leading to a reduction in associated training costs.
Planning	The pre-conferences or initial meetings between the coach and teacher during which the coach prompts the teacher to set goals and action steps in preparation for a lesson or observation.	**Excluded:** Systematically manipulating the level of planning a coach does with a teacher would be very difficult to implement. We felt that planning could be only somewhat systematically varied through Coach Training because coaches in the enhanced condition are intended to receive more coaching strategies to implement systematic planning than are those in the typical condition. We recommend providing coaches only with common guidance on using planning because the implementation of this coaching strategy will be left to the HS grantees discretion.

Coach Dimension	Definition	Rationale for Inclusion or Exclusion
Modeling	The coach demonstrating a technique or teaching strategy (presumably one with empirical evidence or promise of effectiveness) to the teacher, with the goal of strengthening the teacher's use of the technique.	**Excluded:** Systematically manipulating the level of modeling a coach provides for a teacher would be very difficult to implement. We felt that modeling could be only somewhat systematically varied through Coach Training because coaches in the enhanced condition are intended to receive more coaching strategies to implement systematic modeling than are those in the typical condition. We recommend providing coaches only with common guidance on modeling because the implementation of this coaching strategy will be left to the HS grantees discretion..
Observation	The process used by the coach to gather information about a teacher's classroom practice and provide feedback to give specific and individualized information to a teacher about the teacher's classroom practices.	**Excluded:** Systematically manipulating the level of observation a coach conducts with a teacher would be very difficult to implement. The design team felt that observation could be only somewhat systematically varied through Coach Training because coaches in the enhanced condition are intended to receive more coaching strategies to implement systematic observation than are those in the typical condition. We recommend providing coaches only with common guidance on conducting observations because the implementation of this coaching strategy will be left to the HS grantees discretion.
Feedback	The provision of specific and individualized information from a coach to a teacher about the teacher's classroom practices.	**Excluded:** Systematically manipulating the level of feedback a coach provides for a teacher would be very difficult to implement. The design team felt that feedback could be only somewhat systematically varied through Coach Training because coaches in the enhanced condition are intended to receive more coaching strategies to implement systematic feedback than are those in the typical condition. We recommend providing coaches only with common guidance on providing feedback because the implementation of this coaching strategy will be left to the HS grantees discretion.
Staffing Dimensions		
Coach Selection	The process of identifying, recruiting, hiring, and matching candidates to fill coaching positions.	**Excluded:** We felt it would be very difficult to control the type of coaches who were hired beyond a certain minimum required level. Creating certain minimum requirements that all coaches must possess, such as a bachelor's degree, enables the grantees or centers, for the most part, to control the selection and hiring of the coaches, meaning the coaches' qualifications will typically vary somewhat.

Coach Dimension	Definition	Rationale for Inclusion or Exclusion
Coach Caseload	The number of teachers or classrooms assigned to a single coach.	**Excluded:** In the study design, the Dosage of coaching will vary across conditions, which impacts the coach workload. We felt that these two dimensions were similar enough that only Dosage needed to be tested. Setting a fixed caseload for coaches, which we recommend, enables the design team to study the impact that coaching Dosage has on teacher practice.
Coach Training	The amount, content, and nature of preparation and ongoing PD that a coach receives.	**Included:** Training coaches can affect the quality of several coaching process dimensions (e.g., observation, modeling, planning, feedback, and teacher-coach relationships) that could potentially influence a teacher's practice and classroom outcomes. The levels of Coach Training currently vary widely in practice, placing more importance on the need to test this dimension. Varying the level of training that coaches receive will enable the contract team to potentially learn whether Coach Training that is focused on the coaches' understanding of the content and practices they are to coach has an impact on the practices of the adult learners they coach.
Coach Supervision	The formalized process of providing oversight and support to coaches.	**Excluded:** As with the roles and relationships dimension, the design team was concerned that this dimension could change (a) the current center administrators' responsibilities and duties by asking them to supervise coaches or (b) change the nature of center administration by requiring that a supervisor of coaches join that structure. Systematically varying the supervisory relationship between the coach and a supervisor would impact the costs associated with this study design. For example, the contractor would have to provide a greater amount of TA and training to the staff member on how to become a coach supervisor. Limiting the level of grantee supervision across all grantees enables the contractor to spend the money that would be required for coach supervision training on other areas such as travel for coaches to attend trainings.

Appendix B. Other Experimental Designs That Were Considered but Rejected

The study design used in the optimization phase of the MOST framework does not necessarily need to be a factorial experiment—any other appropriate experimental design can be used. Thus, the design team considered several other experimental designs for the HS Coaching Study.

The alternative designs that we considered differed from the recommended factorial experiment in one of two ways: They either included more coaching dimensions (and therefore estimated more effects and generated more information about coaching) or they had fewer experimental conditions (and were therefore less complex from an operational perspective).

First, we considered using a 2^{4-1} fractional factorial design. A fractional factorial design is a factorial design in which a carefully selected subset (fraction) of experimental conditions from the complete factorial design is retained in the experiment. For example, a 2^{4-1} fractional factorial design would include four factors, but it would include only eight of the experimental conditions from the 16 conditions of a complete (2^4) factorial experiment (see Somers, Collins, and Maier [2013] for a discussion of fractional designs). The sample size required to detect a given effect size by using this design would be the same as for the recommended factorial experiment. However, it was decided that this design should not be used, given the constraint that no more than three coaching dimensions can be well implemented and monitored given study resources.[83]

Second, we considered using a 2^{3-1} fractional factorial design. This is a design that would examine the effect of the three factors in Table 7 but that would retain only four of the eight experimental conditions. The sample size required to detect a given effect size by using this design would be the same as for the recommended factorial experiment. However, this alternative design was deemed unsuitable because dropping experimental conditions in a design with only three factors would make it impossible to disentangle main effects from two-way interactions. More technically, main effects and two-way interactions would be aliased; see Somers, Collins, and Maier (2013) for a discussion.

Third, we considered using a comparative treatment (CT) design, also known as a *multigroup* or *multiarm* experiment. In this design, there would be only four experimental conditions: In the first condition, centers would receive the Level I of each dimension; in the three other experimental conditions, one of the three dimensions would be set to Level II and all others would be set to Level I.[84] The effect of each dimension would then be estimated by comparing the mean outcomes of centers assigned to the condition for which that dimension is set to Level II with the mean outcomes of centers assigned to the first condition (for which all dimensions are set to Level I). The advantage of the CT design is that it would require fewer experimental conditions than the recommended factorial design (four rather than eight). On the other hand, there are several disadvantages to this design. First, the CT design would require a larger

[83] Adding a fourth dimension would increase study costs because Coach Training and TA on this additional dimension would have to be provided by the evaluators. The dimension's implementation would also have to be monitored, which would increase data collection costs.

[84] There are also other ways to set up the conditions of the CT design. See Somers, Collins, and Maier (2013) for more information.

sample—two times as many centers as the factorial experiment. Second, the CT design cannot be used to examine interactions. And third, the CT design does not provide estimates of main effects or the average effect of a dimension across all levels of the other dimensions being tested. Instead, the CT design would provide estimates of the effect of each dimension at a specific level of the other two dimensions. In the HS Coaching Study, for example, it would not be possible to obtain an estimate of the average effect of *DOSAGE* across all levels of *TRAINING* and *RECIPIENT* (the main effect). Rather, the CT design would provide an estimate of the effect of *DOSAGE* when *RECIPIENT* is the lead teacher only and *TRAINING* for the coach is a summer orientation.[85] This type of effect is not as useful as for practical and policy purposes because the effect of *DOSAGE* then becomes very context specific. For these reasons, the factorial experiment better meets the objectives of the HS Coaching Study (for further discussion of the CT design, see Somers, Collins, and Maier [2013]).

[85] This is just one example of a simple effect (or the effect of *DOSAGE* at some other fixed level of the other two dimensions).

Appendix C. MDES and Sample Size

In this appendix, we provide a discussion of how the MDES for each of the dimensions were calculated. As explained in Section V, the design team recommends that the random assignment unit for the HS Coaching Study should be clusters: the unit of random assignment should be HS *centers* for the *DOSAGE* and *RECIPIENT* dimensions, *coaches* for the *TRAINING* dimension, and *coaches or centers* for the *MODE* dimension. Because the unit of random assignment differs across dimensions, so does the MDES for their main effect. For this reason, we begin by discussing the MDES for the main effect of dimensions assigned at the center level (*DOSAGE*, *RECIPIENT*, and perhaps *MODE*). We then discuss the MDES for the main effects of dimensions assigned at the coach level (*TRAINING* and perhaps *MODE*). We conclude with a brief discussion of the MDES for interaction effects.

MDES for the Main Effect of Dimensions Assigned at the Center Level

For the *DOSAGE* and *RECIPIENT* dimensions (and perhaps *MODE*), the unit of random assignment is centers, so the MDES for the main effect of these two dimensions was calculated using the following formula:

$$M_{J-K}\sqrt{\frac{\rho_{CEN}(1-R^2_{CEN})}{P(1-P)J}+\frac{(1-\rho_{CEN})(1-R^2_{CL})}{P(1-P)Jn}} \qquad (1)$$

where

- J is the total number of centers in the study.
- n is the number of classrooms per center (and therefore JCn is the total number of classrooms in the study).
- P is the proportion of centers assigned to each level of the dimension (always = 0.5 in this report)
- ρ_{CEN} is the between-center ICC (proportion of the total variation in teacher and classroom outcomes that is between centers)[86]
- R^2_{CEN} is the proportion of the between-center variation in teacher and classroom outcomes that is explained by center or teacher baseline characteristics
- R^2_{CL} is the proportion of the within-center (or between-classroom) variation in teacher and classroom outcomes that is explained by teacher baseline characteristics

[86] If one of the three dimensions in the study design were randomly assigned at the coach level, then strictly speaking, the relevant ICC here would be the proportion of between-center variation in outcomes *net of the between-coach variation in outcomes*. However, if coaches are randomized to centers (as in Random Assignment Plan 1)—or if the assignment of coaches to centers is quasi-random—then the between-coach variation in outcomes would be equal to or near zero. By extension, the between-center variation in outcomes *net of between-coach variation* (which is equal to zero) would be equal to the between-center variation in outcomes. For this reason, we simply use the total between-center variation in outcomes when calculating the MDES for the dimensions assigned at the center level (*DOSAGE*, *RECIPIENT* and *MODE*).

- K is the number of center-level covariates[87]
- $M_{J\text{-}K}$ is a multiple of the standard error of the impact estimate (the degrees of freedom multiplier)

If all three dimensions in the study design were assigned at the center level (that is, if *MODE* were tested rather than *TRAINING* and *MODE* were assigned at the center level), then *K* would be smaller and the degrees of freedom would be larger, which would in turn reduce the multiplier and the MDES.[88] However, the amount by which the MDES would be reduced would be very small, so in this report, we present the MDES for the more conservative scenario (where *K* is equal to its value under the assumption that one of the dimensions being tested is assigned at the coach level).

More generally, notice that Equation (1) has two terms: a term for the variation in teacher and classroom outcomes between centers and a term for the variation in outcomes between classrooms within a center. This reflects the two sources of error in a cluster-randomized experiment where the unit of random assignment is centers. In general, the greater is the between-center ICC, the larger is the MDES for a given sample size. However, if a large proportion of the between-center variation in outcomes can be explained by the covariates and baseline measures (i.e., R^2_{CEN} is high), then the MDES can be reduced relative to what it would be without these covariates.

As explained in Section V, we used data from the CARES study to obtain estimates of the parameters needed to calculate the MDES.[89] The CARES study is the most relevant dataset for making assumptions about the HS Coaching Study because it includes several large urban grantees with the capacity to participate in a large-scale study of PD. Using these data, we estimated the total between-center variation in the data (ρ_{CEN}); the proportion of between-center variation that is explained by grantee fixed effects; and the proportion of between-center and within-center variation explained by grantee fixed effects and classroom-level pretest measures of the outcome (R^2_{CEN} and R^2_{CL}). These parameters were calculated for different subsamples of centers (all centers, control group centers, and urban centers), and for the types of classroom and teacher outcomes that are most similar to the measures that would be used in the HS Coaching

[87] *K* is equal to two main effects at the center level (*DOSAGE* and *RECIPIENT*) + three two-way interaction terms + one three-way interaction term + three center-level baseline CLASS subdomain scores (for MDES that adjusted for these scores) + the number of coaches (= J/4, based on the assumption that each coach serves four centers). The number of coaches uses degrees of freedom because coach random effects must be included in the analysis to account for that level of clustering (see the statistical model in Section VI).

[88] *K* would be smaller because the analysis would no longer have to account for the clustering of centers within coaches, which means that coach random-effects would no longer be included in the analysis (nor would they reduce the degrees of freedom). Specifically, *K* would be equal to two main effects at the center level (*DOSAGE* and *RECIPIENT*) + three two-way interaction terms + one three-way interaction term + the number of random assignment blocks (=J/8, based on the assumption that there are eight centers per grantee) + three center-level baseline CLASS subdomain scores (for MDES that are adjusted for these scores).

[89] For information on the study, see Mattera, Lloyd, Fishman, and Bangser (2013).

Study (the adapted Teaching Style Rating Scale[90] and the CLASS instructional support domain[91]).

Having obtained parameter estimates from the CARES data, we then looked at ICCs and R^2 from other ECE studies, to make sure that the CARES parameters generally align with parameters from other datasets. We looked in particular for studies that met the following two criteria: (1) the study collected data on teacher practice measures that are similar to what we would be using in the HS Coaching Study (i.e., the TSRS, the Arnett, the COEMET, or the CLASS instructional support domain), and (2) the study included centers in several urban areas (since the HS Coaching Study will likely recruit large urban grantees). Only one other study met these two criteria: the FACES survey.[92] Table C1 shows the ICC and R^2 parameter estimates from the CARES and FACES survey. For the FACES, we focused on the Arnett as an outcome measure, because this measure is most similar to the types of practices that would be assessed in the HS Coaching Study.

Table C1. Intraclass Correlations and R^2 From the CARES and FACES Data

				Explained by Grantee or Center Characteristics*		Explained by Grantee or Center Characteristics* Plus Pretests	
Study/ Data	Outcome	Subsample	Between Center ICC (ρ_{CEN})	R^2 b/w Centers (R^2_{CEN})	R^2 Within Centers (R^2_{CL})	R^2 b/w Centers (R^2_{CEN})	R^2 Within Centers (R^2_{CL})
CARES	CLASS instr. sup.	All centers	0.238	0.866	0	0.866	0.063
CARES	Adapted TSRS	All centers	0.277	0.567	0	0.567	0.025
CARES	Adapted TSRS	Control group	0.387	0.500	0	0.648	0.032
CARES	Adapted TSRS	Urban centers	0.162	0.829	0	0.861	0.006
FACES	Arnett—Lead Teacher	4-year-old classes	0.100	0.145	0	0.223	0.029
FACES	Arnett—Ass't Teacher	4-year-old classes	0.209	0.132	0	0.177	0.015
FACES	Arnett—Lead Teacher	3-year-old classes	0.149	0.048	0	0.129	0.023
FACES	Arnett—Ass't Teacher	3-year-old classes	0.208	0.049	0	0.080	0.022

* The CARES R^2 includes grantee fixed effects as covariates. It is not known whether the FACES R^2 also account for grantees. Adapted TSRS = Adapted Teaching Style Rating Scale.

[90] The TSRS was used in the REDI study (Domitrovich, Cortes, & Greenberg, 2000). The Adapted TSRS was created by Raver, Domitrovich, Greenberg, Morris, and Mattera as part of the HS CARES demonstration.

[91] We focus on the CLASS instructional domain because scores on this domain are lower, and therefore teachers would benefit most from being coached on the practices in this domain.

[92] We also considered three other studies but decided to exclude them because they did not meet our two criteria. The REDI study was conducted in central Pennsylvania (not large urban sites). The 4R study measured classroom outcomes using the CLASS only (so there are fewer teacher practice measures). The CSRP was conducted in an urban area but randomization was blocked by matched pairs so the reported R^2 for this study may be larger than what we would see in the HS Coaching Study (which would block by grantee or groups of grantees).

As can be seen in Table C1, the between-center ICCs from the CARES data (range = 0.162 to 0.387) are similar to the ICCs from the FACES data (0.100 to 0.209). The between-center R^2 (R^2_{CEN}) is larger for the CARES data than the FACES data, but this is likely due to the fact that the R^2_{CEN} for the FACES data may not control for grantee fixed effects (in which case, the R^2_{CEN} from FACES are lower than what would be expected in the HS Coaching Study).[93] Although the FACES parameters are useful for validating the general range of ICCs that we see in the CARES data, in general the parameters from the latter study are more relevant because centers in the CARES study are similar in profile to the types of centers that would be recruited for the HS Coaching Study.

Table C1 also shows that classroom-level pretest measures of the outcomes of interest do not appreciably increase the between-center and the within-center R^2 (R^2_{CEN} and R^2_{CL}) beyond what is explained using center characteristics and grantee fixed effects. For example, in the first set of parameters from the CARES data, the between-center R^2 (R^2_{CEN}) is 0.866 when controlling for grantee fixed effects, and the value of this parameter is unchanged when pretests are also included as control variables. Similarly, the pretest measures explain very little of the variation in outcomes within centers: The R^2_{CL} values are no higher than 0.063 in either the CARES or FACES data. Thus, in terms of minimizing the MDES, there is little gain to collecting baseline measures of teacher practice. Instead, we recommend using existing data on center-level mean CLASS scores as covariates in the analysis.[94]

Thus, we calculated the MDES under the assumption that baseline outcomes data would not be available ("No Baseline Outcomes Data") and under the assumption that existing CLASS data would be used as baseline measures ("With Center-Level Mean CLASS score"). The parameters used for each scenario are as follows:

- "No Baseline Outcomes Data": The MDES are calculated using the R^2 in the "Explained by Grantee or Center Characteristics" columns
- "With Center-Level Mean CLASS scores": The MDES are calculated using the R^2_{CEN} in the "Explained by Grantee or Center Characteristics Plus Pretests" columns but we set R^2_{CL} to 0.

At the end of this appendix, we present MDES tables for each set of assumptions in Table C1. For all MDES calculations, we assume that there are eight centers per grantee or per random assignment block, that each coach serves four centers, and that there are two to three classrooms per center (n). The statistical significance level (alpha) is set at 10 percent and power at 80 percent. As shown in these tables, the MDES for a given sample size is very similar across assumptions. As explained in Section V, to recommend a sample size for the HS Coaching Study, we used the set of parameters based on CARES control group centers—which are the most conservative of the CARES parameters (i.e., the lead to the largest MDES for a given number of centers).

[93] Because the FACES parameters are from published reports, it is not possible to determine whether the R^2 are calculated based on having controlled for grantees.

[94] Using the CARES data, we estimated the value of R^2_{CEN} with *center-level* mean CLASS scores as the baseline covariate (defined here as the center-level averages of the baseline CLASS scores collected for the study). We found that the value of R^2_{CEN} in this analysis is very similar to its value when classroom-level pretests are used, even when the outcome measure is the TSRS.

MDES for the Main Effect of Dimensions Assigned at the Coach Level

Recall that the unit of random assignment for the *TRAINING* (and perhaps *MODE*) dimension is coaches, which means that the MDES must account for an additional level of clustering:

$$M_{C-Z}\sqrt{\frac{\rho_{COA}(1-R^2_{COA})}{P(1-P)C}+\frac{(\rho_{CEN}-\rho_{COA})(1-R^2_{CEN})}{P(1-P)J}+\frac{(1-\rho_{CEN})(1-R^2_{CL})}{P(1-P)Jn}} \qquad (2)$$

where the parameters are defined in Equation (1) and

- *C* is the total number of coaches (=J/4, based on the assumption that each coach serves four centers)
- ρ_{COA} is the between-coach ICC (proportion of the total variation in teacher and classroom outcomes that is between coaches)
- $(\rho_{CEN}-\rho_{COA})$ is the proportion of the total variation in teacher and classroom outcomes that is between centers *within coaches* (as opposed to the total variation in classrooms outcomes between all study centers, which is ρ_{CEN})
- R^2_{COA} is the proportion of the between-center variation in teacher and classroom outcomes that is explained by coach, center, or teacher baseline characteristics
- *Z* is the number of coach-level covariates[95]
- $M_{C\text{-}Z}$ is a multiple of the standard error of the impact estimate (the degrees of freedom multiplier)

Notice that in Equation (2), there are three terms: a term for the variation in outcomes between coaches, a term for the variation in outcomes between centers within coaches, and a term for the variation in outcomes between classrooms within centers. Because of this extra term, the MDES for the *TRAINING* (or *MODE*) dimension is larger than the MDES for the two other dimensions [this can be seen by comparing Equation (1) and (2)]. In general, the greater is the between-coach ICC, the greater will be the MDES for dimensions assigned at the coach level (*TRAINING* or *MODE*), both in absolute terms and relative to the MDES for dimensions assigned at the center level (*DOSAGE* and *RECIPIENT*).

The value of ρ_{COA}—and by extension the MDES for dimensions assigned at the coach level such as *TRAINING*—depends on how coaches are assigned to centers. At one extreme, if coach assignments are random, the outcomes of centers served by each coach are the same on average, and therefore the between-coach variation in outcomes is zero ($\rho_{COA}=0$). Thus, under Random Assignment Plan 1 in Section V, we can assume that between-coach variation is zero.

However, in practice, coach assignments are unlikely to be random and may be related to centers' outcomes. First, grantees may explicitly assign coaches to centers based on center-level

[95] *Z* is equal to the number of grantees or blocks (= JC/8 based on the assumption that there are eight centers per block) + one main effect at the coach-level (for *TRAINING* or *MODE*). The main effects of *DOSAGE* and *RECIPIENT* and interaction terms do not use degrees of freedom because these variables are measured at a level below the coach. The same applies to center-level mean CLASS scores.

outcomes. For example, a more experienced coach may be assigned to work with all the "weakest" centers, while a less experienced coach may be assigned to with all the "strongest" centers, or vice versa. Second, grantees may assign coaches based on geography to minimize coach travel time. If the grantee is large, for example, one coach may be assigned to work with all centers in the East, and another coach may get all centers in the West. However, if there are large regional differences in teacher quality (for example, the West has stronger centers), then the assignment of coaches by region will indirectly lead to one coach being assigned to the strongest centers and the other coach working with the weaker centers. This means that under Random Assignment Plan 2 (where coaches cannot be randomize to centers), ρ_{COA} is likely to be greater than zero.

For this reason, we calculated the MDES for the *TRAINING* dimension under different scenarios about the value of ρ_{COA}. From the CARES data, we know that the total between-center variation in teacher outcomes across study centers (ρ_{CEN} in Table C1) is 0.387 based on the most conservative CARES assumptions (control group centers). To obtain the MDES for the *TRAINING* dimension, we made different assumptions about how much of this total variation is due to variation between coaches (ρ_{COA}) as opposed to variation between the centers served by a coach ($\rho_{CEN} - \rho_{COA}$).[96] In Section V, we show the MDES for two scenarios representing two different ways to partition this variation:

- For the scenario where coach assignments are *not* associated with center outcomes (i.e., where coaches are randomized to centers as in RA Plan 1), we assume that 0 percent of ρ_{CEN} is due to variation between coaches $\rho_{COA} = 0$, and that 100 percent is due to variation between centers within coaches ($\rho_{CEN} - \rho_{COA} = 0.387$).
- For the scenario where coach assignments are *weakly* associated with center outcomes, we assume that 10 percent of ρ_{CEN} is due to variation between coaches $\rho_{COA} = 0.10 * 0.387 = 0.0387$, and that 90 percent is due to variation between centers within coaches ($\rho_{CEN} - \rho_{COA} = 0.90 * 0.387 = 0.3483$).
- For the scenario where coach assignments are *moderately* associated with center outcomes, we assume that 25 percent of ρ_{CEN} is due to variation between coaches $\rho_{COA} = 0.25 * 0.387 = 0.0968$, and that 75 percent is due to variation between centers within coaches ($\rho_{CEN} - \rho_{COA} = 0.75 * 0.387 = 0.2903$).

In all scenarios, we further assumed that R^2_{COA} is equal to R^2_{CEN} (in other words, that the between-coach variation explained by grantees and CLASS mean scores is the same as the between-center variation explained by these characteristics). These parameter values were then applied to Equation (2) to obtain the MDES for the *TRAINING* dimension under different scenarios in Table 13. In the latter table, for RA Plan 1 we assume that coach assignments are *not* associated with center outcomes, or that $\rho_{COA} = 0$. We also make this assumption in the tables included at the end of this appendix.

[96] Unfortunately, it is not possible to determine how large this parameter is in practice using the CARES data because coach identifiers were not available.

MDES for Main Effects for Each Set of Parameter Assumptions

The end of this appendix includes a set of tables that present the MDES for the main effect of each coaching dimension, for each set of assumptions in Table C1. The sample sizes in these tables are multiples of eight based on the assumption that centers would be randomized in blocks of eight centers. The tables further assume that the three tested dimensions in the design are *DOSAGE*, *RECIPIENT*, and *TRAINING* and that coach assignments are *not* associated with center outcomes ($\rho_{COA} = 0$, or RA Plan 1 in Section V). If *MODE* were tested instead of *TRAINING*—and this dimension was randomly assigned at the coach level— then the MDES for *MODE* would be the same as the MDES in the "TRAINING" columns of these tables. If *MODE* were assigned at the center level, then the MDES for *MODE* (as well as the *DOSAGE* and *RECIPIENT* dimensions) would be very similar to the values in the "DOSAGE & RECIPIENT" columns of these tables.[97]

Table C2 summarizes the tables by showing the MDES based on a sample of 248 centers, which is the sample size needed to detect an effect size of 0.20 on the *DOSAGE* and *RECIPIENT* dimensions based under to most conservative set of CARES assumptions ("CARES TSRS CONTROL"). As this table illustrates, the MDES for a given sample size is similar across assumptions.

Table C2. MDES With a Sample of 248 Centers, by Parameter Assumptions

	No Baseline Outcomes Data				With Center-Level Mean CLASS Scores			
	DOSAGE & RECIPIENT		TRAINING		DOSAGE & RECIPIENT		TRAINING	
Parameter Assumptions	(3)	(2)	(3)	(2)	(3)	(2)	(3)	(2)
CARES CLASS ALL	0.17	0.20	0.17	0.21	0.17	0.20	0.17	0.21
CARES TSRS ALL	0.19	0.22	0.19	0.22	0.19	0.22	0.19	0.22
CARES TSRS CONTROL	0.20	0.22	0.20	0.23	0.18	0.21	0.19	0.21
CARES TSRS URBAN	0.18	0.21	0.18	0.22	0.17	0.21	0.18	0.21
FACES ARNETT LEAD 4 YR	0.20	0.23	0.20	0.24	0.19	0.20	0.20	0.23
FACES ARNETT ASS'T 4 YR	0.21	0.24	0.22	0.25	0.21	0.21	0.21	0.24
FACES ARNETT LEAD 3 YR	0.21	0.24	0.21	0.24	0.20	0.21	0.21	0.24
FACES ARNETT ASS'T 3 YR	0.22	0.24	0.22	0.25	0.21	0.22	0.22	0.25

Note. The number in brackets (2 or 3) is the number of classrooms per center.

MDES for Interaction Effects

The MDES for a two-way interaction effect is about two times larger than the MDES for a main effect. The MDES for an interaction is larger for two reasons. First, as explained in Section IV, to estimate the interaction between Dimension A and Dimension B, the effect of Dimension A

[97] As explained earlier, it would not be exactly the same because the degrees of freedom used by the analysis would be slightly different in a design where all three dimensions are assigned at the center level.

must be estimated *for each level of Dimension B*. This is similar to estimating a subgroup effect, where the two subgroups in this case are defined based on the levels of Dimension B. Because the sample size is being subdivided, the MDES for the effect of Dimension A at a given level of Dimension B is larger than the MDES for the main effect of Dimension A (since the latter is based on the entire sample). Second, as explained in Section IV, an interaction is the *difference* in the effect of Dimension A between the two levels of Dimension B. As a general rule, the MDES for the difference in effects between two subgroups is larger than the MDES for the effect of each subgroup, because differences in effects are less reliably estimated. Taken together, these two factors double the MDES for an interaction effect relative to a main effect.

Table C3. MDES Based on ICC and R^2 From CARES (All Centers, CLASS as Outcome)

Number of Centers	No Baseline Outcomes Data				With Center-Level Mean CLASS Scores			
	DOSAGE & RECIPIENT		TRAINING		DOSAGE & RECIPIENT		TRAINING	
	(2)	(3)	(2)	(3)	(2)	(3)	(2)	(3)
96	0.330	0.275	0.350	0.292	0.330	0.275	0.350	0.292
104	0.317	0.264	0.335	0.278	0.317	0.264	0.335	0.278
112	0.305	0.254	0.321	0.267	0.305	0.254	0.321	0.267
120	0.294	0.245	0.308	0.257	0.294	0.245	0.308	0.257
128	0.285	0.237	0.298	0.248	0.285	0.237	0.298	0.248
136	0.276	0.230	0.288	0.239	0.276	0.230	0.288	0.239
144	0.268	0.223	0.279	0.232	0.268	0.223	0.279	0.232
152	0.261	0.217	0.271	0.225	0.261	0.217	0.271	0.225
160	0.254	0.212	0.263	0.219	0.254	0.212	0.263	0.219
168	0.248	0.206	0.256	0.213	0.248	0.206	0.256	0.213
176	0.242	0.202	0.250	0.208	0.242	0.202	0.250	0.208
184	0.237	0.197	0.244	0.203	0.237	0.197	0.244	0.203
192	0.232	0.193	0.238	0.198	0.232	0.193	0.238	0.198
200	0.227	0.189	0.233	0.194	0.227	0.189	0.233	0.194
208	0.223	0.185	0.229	0.190	0.223	0.185	0.229	0.190
216	0.218	0.182	0.224	0.186	0.219	0.182	0.224	0.186
224	0.215	0.178	0.220	0.183	0.215	0.179	0.220	0.183
232	0.211	0.175	0.216	0.179	0.211	0.175	0.216	0.179
240	0.207	0.172	0.212	0.176	0.207	0.172	0.212	0.176
248	0.204	0.170	0.208	0.173	0.204	0.170	0.208	0.173
256	0.201	0.167	0.205	0.170	0.201	0.167	0.205	0.170
264	0.197	0.164	0.201	0.168	0.197	0.164	0.201	0.168
272	0.194	0.162	0.198	0.165	0.195	0.162	0.198	0.165
280	0.192	0.159	0.195	0.163	0.192	0.160	0.195	0.163
288	0.189	0.157	0.192	0.160	0.189	0.157	0.192	0.160
296	0.186	0.155	0.190	0.158	0.186	0.155	0.190	0.158
304	0.184	0.153	0.187	0.156	0.184	0.153	0.187	0.156
312	0.182	0.151	0.185	0.154	0.182	0.151	0.185	0.154
320	0.179	0.149	0.182	0.152	0.179	0.149	0.182	0.152
328	0.177	0.147	0.180	0.150	0.177	0.147	0.180	0.150
336	0.175	0.146	0.178	0.148	0.175	0.146	0.178	0.148
344	0.173	0.144	0.175	0.146	0.173	0.144	0.175	0.146
352	0.171	0.142	0.173	0.144	0.171	0.142	0.173	0.144
360	0.169	0.141	0.171	0.143	0.169	0.141	0.171	0.143
368	0.167	0.139	0.169	0.141	0.167	0.139	0.169	0.141
376	0.165	0.138	0.168	0.139	0.165	0.138	0.168	0.139
384	0.164	0.136	0.166	0.138	0.164	0.136	0.166	0.138
392	0.162	0.135	0.164	0.136	0.162	0.135	0.164	0.136
400	0.160	0.133	0.162	0.135	0.160	0.133	0.162	0.135

Note. The number in brackets (2 or 3) is the number of classrooms per center. The MDES in this table are for main effects.

Table C4. MDES Based on ICC and R^2 From CARES (All Centers, TSRS as Outcome)

Number of Centers	No Baseline Outcomes Data				With Center-Level Mean CLASS Scores			
	DOSAGE & RECIPIENT		TRAINING		DOSAGE & RECIPIENT		TRAINING	
	(2)	(3)	(2)	(3)	(2)	(3)	(2)	(3)
96	0.356	0.308	0.378	0.328	0.353	0.306	0.375	0.325
104	0.342	0.296	0.361	0.313	0.339	0.294	0.358	0.310
112	0.329	0.285	0.346	0.300	0.326	0.283	0.343	0.297
120	0.318	0.275	0.333	0.288	0.315	0.273	0.330	0.286
128	0.308	0.266	0.321	0.278	0.305	0.264	0.318	0.276
136	0.298	0.258	0.311	0.269	0.295	0.256	0.308	0.267
144	0.290	0.251	0.301	0.261	0.287	0.249	0.298	0.258
152	0.282	0.244	0.292	0.253	0.279	0.242	0.289	0.251
160	0.275	0.238	0.284	0.246	0.272	0.236	0.281	0.244
168	0.268	0.232	0.277	0.240	0.265	0.230	0.274	0.238
176	0.262	0.227	0.270	0.234	0.259	0.225	0.267	0.232
184	0.256	0.222	0.263	0.228	0.253	0.220	0.261	0.226
192	0.250	0.217	0.258	0.223	0.248	0.215	0.255	0.221
200	0.245	0.212	0.252	0.218	0.243	0.211	0.250	0.216
208	0.240	0.208	0.247	0.214	0.238	0.206	0.244	0.212
216	0.236	0.204	0.242	0.209	0.234	0.203	0.240	0.208
224	0.232	0.201	0.237	0.205	0.229	0.199	0.235	0.204
232	0.228	0.197	0.233	0.202	0.225	0.195	0.231	0.200
240	0.224	0.194	0.229	0.198	0.222	0.192	0.227	0.196
248	0.220	0.191	0.225	0.195	0.218	0.189	0.223	0.193
256	0.217	0.187	0.221	0.191	0.215	0.186	0.219	0.190
264	0.213	0.185	0.218	0.188	0.211	0.183	0.215	0.187
272	0.210	0.182	0.214	0.185	0.208	0.180	0.212	0.184
280	0.207	0.179	0.211	0.183	0.205	0.178	0.209	0.181
288	0.204	0.177	0.208	0.180	0.202	0.175	0.206	0.178
296	0.201	0.174	0.205	0.177	0.199	0.173	0.203	0.176
304	0.199	0.172	0.202	0.175	0.197	0.171	0.200	0.173
312	0.196	0.170	0.199	0.173	0.194	0.168	0.197	0.171
320	0.194	0.168	0.197	0.170	0.192	0.166	0.195	0.169
328	0.191	0.165	0.194	0.168	0.189	0.164	0.192	0.167
336	0.189	0.163	0.192	0.166	0.187	0.162	0.190	0.165
344	0.187	0.162	0.189	0.164	0.185	0.160	0.188	0.163
352	0.184	0.160	0.187	0.162	0.183	0.158	0.185	0.161
360	0.182	0.158	0.185	0.160	0.181	0.157	0.183	0.159
368	0.180	0.156	0.183	0.158	0.179	0.155	0.181	0.157
376	0.178	0.155	0.181	0.157	0.177	0.153	0.179	0.155
384	0.177	0.153	0.179	0.155	0.175	0.152	0.177	0.154
392	0.175	0.151	0.177	0.153	0.173	0.150	0.175	0.152
400	0.173	0.150	0.175	0.152	0.171	0.149	0.174	0.150

Note. The number in brackets (2 or 3) is the number of classrooms per center. The MDES in this table are for main effects.

Table C5. MDES Based on ICC and R^2 From CARES (Control Group Centers, TSRS as Outcome)

Number of Centers	No Baseline Outcomes Data				With Center-Level Mean CLASS Scores			
	DOSAGE & RECIPIENT		TRAINING		DOSAGE & RECIPIENT		TRAINING	
	(2)	(3)	(2)	(3)	(2)	(3)	(2)	(3)
96	0.363	0.324	0.386	0.344	0.338	0.297	0.359	0.315
104	0.348	0.311	0.368	0.328	0.324	0.285	0.343	0.301
112	0.336	0.299	0.353	0.315	0.312	0.274	0.328	0.288
120	0.324	0.289	0.339	0.303	0.302	0.265	0.316	0.277
128	0.313	0.280	0.327	0.292	0.292	0.256	0.305	0.268
136	0.304	0.271	0.317	0.282	0.283	0.248	0.295	0.259
144	0.295	0.263	0.307	0.274	0.275	0.241	0.285	0.251
152	0.287	0.256	0.298	0.266	0.267	0.235	0.277	0.243
160	0.280	0.250	0.290	0.258	0.260	0.229	0.269	0.237
168	0.273	0.244	0.282	0.252	0.254	0.223	0.262	0.230
176	0.267	0.238	0.275	0.245	0.248	0.218	0.256	0.225
184	0.261	0.233	0.269	0.240	0.243	0.213	0.250	0.219
192	0.255	0.228	0.262	0.234	0.237	0.209	0.244	0.215
200	0.250	0.223	0.257	0.229	0.233	0.204	0.239	0.210
208	0.245	0.219	0.251	0.224	0.228	0.200	0.234	0.206
216	0.240	0.214	0.246	0.220	0.224	0.197	0.229	0.201
224	0.236	0.211	0.242	0.216	0.220	0.193	0.225	0.198
232	0.232	0.207	0.237	0.212	0.216	0.190	0.221	0.194
240	0.228	0.203	0.233	0.208	0.212	0.186	0.217	0.191
248	0.224	0.200	0.229	0.204	0.209	0.183	0.213	0.187
256	0.221	0.197	0.225	0.201	0.205	0.180	0.210	0.184
264	0.217	0.194	0.222	0.198	0.202	0.178	0.206	0.181
272	0.214	0.191	0.218	0.195	0.199	0.175	0.203	0.178
280	0.211	0.188	0.215	0.192	0.196	0.172	0.200	0.176
288	0.208	0.185	0.212	0.189	0.194	0.170	0.197	0.173
296	0.205	0.183	0.209	0.186	0.191	0.168	0.194	0.171
304	0.202	0.181	0.206	0.184	0.188	0.165	0.192	0.168
312	0.200	0.178	0.203	0.181	0.186	0.163	0.189	0.166
320	0.197	0.176	0.200	0.179	0.184	0.161	0.187	0.164
328	0.195	0.174	0.198	0.177	0.181	0.159	0.184	0.162
336	0.192	0.172	0.195	0.174	0.179	0.157	0.182	0.160
344	0.190	0.170	0.193	0.172	0.177	0.155	0.180	0.158
352	0.188	0.168	0.191	0.170	0.175	0.154	0.178	0.156
360	0.186	0.166	0.189	0.168	0.173	0.152	0.175	0.154
368	0.184	0.164	0.186	0.166	0.171	0.150	0.173	0.152
376	0.182	0.162	0.184	0.164	0.169	0.149	0.172	0.151
384	0.180	0.160	0.182	0.163	0.167	0.147	0.170	0.149
392	0.178	0.159	0.180	0.161	0.166	0.146	0.168	0.147
400	0.176	0.157	0.179	0.159	0.164	0.144	0.166	0.146

Note. The number in brackets (2 or 3) is the number of classrooms per center. The MDES in this table are for main effects.

Table C6. MDES Based on ICC and R^2 From CARES (Urban Centers, TSRS as Outcome)

Number of Centers	No Baseline Outcomes Data				With Center-Level Mean CLASS Scores			
	DOSAGE & RECIPIENT		TRAINING		DOSAGE & RECIPIENT		TRAINING	
	(2)	(3)	(2)	(3)	(2)	(3)	(2)	(3)
96	0.343	0.285	0.364	0.302	0.340	0.281	0.361	0.299
104	0.329	0.273	0.348	0.289	0.327	0.270	0.345	0.285
112	0.317	0.263	0.334	0.277	0.315	0.260	0.331	0.273
120	0.306	0.254	0.321	0.266	0.304	0.251	0.318	0.263
128	0.296	0.246	0.309	0.257	0.294	0.243	0.307	0.254
136	0.287	0.238	0.299	0.248	0.285	0.236	0.297	0.245
144	0.279	0.231	0.290	0.240	0.277	0.229	0.287	0.238
152	0.271	0.225	0.281	0.233	0.269	0.223	0.279	0.231
160	0.265	0.219	0.274	0.227	0.262	0.217	0.271	0.224
168	0.258	0.214	0.267	0.221	0.256	0.212	0.264	0.219
176	0.252	0.209	0.260	0.216	0.250	0.207	0.258	0.213
184	0.246	0.204	0.254	0.210	0.244	0.202	0.252	0.208
192	0.241	0.200	0.248	0.206	0.239	0.198	0.246	0.203
200	0.236	0.196	0.243	0.201	0.234	0.194	0.241	0.199
208	0.232	0.192	0.238	0.197	0.230	0.190	0.236	0.195
216	0.227	0.188	0.233	0.193	0.225	0.186	0.231	0.191
224	0.223	0.185	0.229	0.189	0.221	0.183	0.227	0.187
232	0.219	0.182	0.224	0.186	0.217	0.180	0.222	0.184
240	0.215	0.179	0.220	0.183	0.214	0.177	0.218	0.181
248	0.212	0.176	0.217	0.180	0.210	0.174	0.215	0.178
256	0.209	0.173	0.213	0.177	0.207	0.171	0.211	0.175
264	0.205	0.170	0.210	0.174	0.204	0.168	0.208	0.172
272	0.202	0.168	0.206	0.171	0.201	0.166	0.204	0.169
280	0.199	0.165	0.203	0.168	0.198	0.163	0.201	0.167
288	0.197	0.163	0.200	0.166	0.195	0.161	0.198	0.164
296	0.194	0.161	0.197	0.164	0.192	0.159	0.196	0.162
304	0.191	0.159	0.195	0.161	0.190	0.157	0.193	0.160
312	0.189	0.157	0.192	0.159	0.187	0.155	0.190	0.157
320	0.186	0.155	0.189	0.157	0.185	0.153	0.188	0.155
328	0.184	0.153	0.187	0.155	0.183	0.151	0.185	0.153
336	0.182	0.151	0.185	0.153	0.180	0.149	0.183	0.151
344	0.180	0.149	0.182	0.151	0.178	0.147	0.181	0.150
352	0.178	0.147	0.180	0.150	0.176	0.146	0.179	0.148
360	0.176	0.146	0.178	0.148	0.174	0.144	0.177	0.146
368	0.174	0.144	0.176	0.146	0.172	0.142	0.175	0.144
376	0.172	0.142	0.174	0.144	0.170	0.141	0.173	0.143
384	0.170	0.141	0.172	0.143	0.169	0.139	0.171	0.141
392	0.168	0.140	0.171	0.141	0.167	0.138	0.169	0.140
400	0.167	0.138	0.169	0.140	0.165	0.137	0.167	0.138

Note. The number in brackets (2 or 3) is the number of classrooms per center. The MDES in this table are for main effects.

Table C7. MDES Based on ICC and R^2 From FACES (Lead Teachers, Four-Year-Olds, Arnett as Outcome)

Number of Centers	No Baseline Outcomes Data				With Center-Level Mean CLASS Scores			
	DOSAGE & RECIPIENT		TRAINING		DOSAGE & RECIPIENT		TRAINING	
	(2)	(3)	(2)	(3)	(2)	(3)	(2)	(3)
96	0.376	0.319	0.399	0.339	0.369	0.312	0.391	0.331
104	0.361	0.306	0.381	0.323	0.354	0.300	0.374	0.316
112	0.347	0.295	0.365	0.310	0.341	0.288	0.358	0.303
120	0.335	0.284	0.351	0.298	0.329	0.278	0.344	0.292
128	0.324	0.275	0.339	0.287	0.318	0.269	0.332	0.281
136	0.315	0.267	0.328	0.278	0.308	0.261	0.321	0.272
144	0.306	0.259	0.317	0.269	0.300	0.254	0.311	0.264
152	0.297	0.252	0.308	0.261	0.291	0.247	0.302	0.256
160	0.290	0.246	0.300	0.254	0.284	0.240	0.294	0.249
168	0.283	0.240	0.292	0.248	0.277	0.235	0.286	0.242
176	0.276	0.234	0.285	0.241	0.271	0.229	0.279	0.236
184	0.270	0.229	0.278	0.236	0.265	0.224	0.272	0.231
192	0.264	0.224	0.272	0.230	0.259	0.219	0.266	0.225
200	0.259	0.219	0.266	0.225	0.254	0.215	0.261	0.221
208	0.254	0.215	0.260	0.221	0.249	0.211	0.255	0.216
216	0.249	0.211	0.255	0.216	0.244	0.207	0.250	0.212
224	0.244	0.207	0.250	0.212	0.240	0.203	0.245	0.208
232	0.240	0.204	0.246	0.208	0.235	0.199	0.241	0.204
240	0.236	0.200	0.241	0.205	0.231	0.196	0.236	0.200
248	0.232	0.197	0.237	0.201	0.228	0.193	0.232	0.197
256	0.228	0.194	0.233	0.198	0.224	0.190	0.229	0.194
264	0.225	0.191	0.229	0.195	0.220	0.187	0.225	0.190
272	0.222	0.188	0.226	0.192	0.217	0.184	0.221	0.187
280	0.218	0.185	0.222	0.189	0.214	0.181	0.218	0.185
288	0.215	0.183	0.219	0.186	0.211	0.179	0.215	0.182
296	0.212	0.180	0.216	0.183	0.208	0.176	0.212	0.179
304	0.209	0.178	0.213	0.181	0.205	0.174	0.209	0.177
312	0.207	0.175	0.210	0.178	0.203	0.172	0.206	0.175
320	0.204	0.173	0.207	0.176	0.200	0.169	0.203	0.172
328	0.202	0.171	0.205	0.174	0.198	0.167	0.201	0.170
336	0.199	0.169	0.202	0.172	0.195	0.165	0.198	0.168
344	0.197	0.167	0.200	0.170	0.193	0.163	0.196	0.166
352	0.195	0.165	0.197	0.168	0.191	0.161	0.194	0.164
360	0.192	0.163	0.195	0.166	0.189	0.160	0.191	0.162
368	0.190	0.161	0.193	0.164	0.187	0.158	0.189	0.160
376	0.188	0.160	0.191	0.162	0.184	0.156	0.187	0.158
384	0.186	0.158	0.189	0.160	0.183	0.155	0.185	0.157
392	0.184	0.156	0.187	0.158	0.181	0.153	0.183	0.155
400	0.182	0.155	0.185	0.157	0.179	0.151	0.181	0.153

Note. The number in brackets (2 or 3) is the number of classrooms per center. The MDES in this table are for main effects.

Table C8. MDES Based on ICC and R^2 From FACES (Assistant Teachers, Four-Year-Olds, Arnett as Outcome)

Number of Centers	No Baseline Outcomes Data				With Center-Level Mean CLASS Scores			
	DOSAGE & RECIPIENT		TRAINING		DOSAGE & RECIPIENT		TRAINING	
	(2)	(3)	(2)	(3)	(2)	(3)	(2)	(3)
96	0.390	0.343	0.414	0.364	0.385	0.338	0.409	0.358
104	0.374	0.329	0.395	0.347	0.369	0.324	0.390	0.342
112	0.360	0.317	0.379	0.333	0.356	0.312	0.374	0.328
120	0.348	0.306	0.365	0.320	0.343	0.301	0.360	0.315
128	0.337	0.296	0.352	0.309	0.332	0.291	0.347	0.304
136	0.326	0.287	0.340	0.299	0.322	0.282	0.336	0.294
144	0.317	0.279	0.330	0.289	0.313	0.274	0.325	0.285
152	0.309	0.271	0.320	0.281	0.304	0.267	0.316	0.277
160	0.301	0.264	0.311	0.273	0.297	0.260	0.307	0.269
168	0.293	0.258	0.303	0.266	0.289	0.254	0.299	0.262
176	0.286	0.252	0.295	0.259	0.283	0.248	0.291	0.256
184	0.280	0.246	0.288	0.253	0.276	0.242	0.285	0.249
192	0.274	0.241	0.282	0.248	0.270	0.237	0.278	0.244
200	0.269	0.236	0.276	0.242	0.265	0.232	0.272	0.239
208	0.263	0.231	0.270	0.237	0.260	0.228	0.266	0.234
216	0.258	0.227	0.265	0.233	0.255	0.223	0.261	0.229
224	0.254	0.223	0.260	0.228	0.250	0.219	0.256	0.225
232	0.249	0.219	0.255	0.224	0.246	0.216	0.251	0.221
240	0.245	0.215	0.250	0.220	0.242	0.212	0.247	0.217
248	0.241	0.212	0.246	0.216	0.238	0.208	0.243	0.213
256	0.237	0.208	0.242	0.213	0.234	0.205	0.239	0.209
264	0.233	0.205	0.238	0.209	0.230	0.202	0.235	0.206
272	0.230	0.202	0.234	0.206	0.227	0.199	0.231	0.203
280	0.227	0.199	0.231	0.203	0.224	0.196	0.228	0.200
288	0.223	0.196	0.228	0.200	0.220	0.193	0.224	0.197
296	0.220	0.194	0.224	0.197	0.217	0.191	0.221	0.194
304	0.217	0.191	0.221	0.194	0.214	0.188	0.218	0.191
312	0.215	0.188	0.218	0.192	0.212	0.186	0.215	0.189
320	0.212	0.186	0.215	0.189	0.209	0.183	0.212	0.186
328	0.209	0.184	0.213	0.187	0.206	0.181	0.210	0.184
336	0.207	0.182	0.210	0.184	0.204	0.179	0.207	0.182
344	0.204	0.179	0.207	0.182	0.202	0.177	0.205	0.179
352	0.202	0.177	0.205	0.180	0.199	0.175	0.202	0.177
360	0.200	0.175	0.203	0.178	0.197	0.173	0.200	0.175
368	0.197	0.173	0.200	0.176	0.195	0.171	0.198	0.173
376	0.195	0.172	0.198	0.174	0.193	0.169	0.195	0.171
384	0.193	0.170	0.196	0.172	0.191	0.167	0.193	0.169
392	0.191	0.168	0.194	0.170	0.189	0.165	0.191	0.168
400	0.189	0.166	0.192	0.168	0.187	0.164	0.189	0.166

Note. The number in brackets (2 or 3) is the number of classrooms per center. The MDES in this table are for main effects.

Table C9. MDES Based on ICC and R^2 From FACES (Lead Teachers, Three-Year-Olds, Arnett as Outcome)

Number of Centers	No Baseline Outcomes Data				With Center-Level Mean CLASS Scores			
	DOSAGE & RECIPIENT		TRAINING		DOSAGE & RECIPIENT		TRAINING	
	(2)	(3)	(2)	(3)	(2)	(3)	(2)	(3)
96	0.387	0.335	0.411	0.356	0.379	0.328	0.403	0.348
104	0.371	0.321	0.392	0.340	0.364	0.315	0.385	0.332
112	0.357	0.310	0.376	0.326	0.351	0.303	0.369	0.318
120	0.345	0.299	0.362	0.313	0.338	0.292	0.355	0.306
128	0.334	0.289	0.349	0.302	0.327	0.283	0.342	0.295
136	0.324	0.280	0.337	0.292	0.318	0.274	0.331	0.286
144	0.314	0.272	0.327	0.283	0.308	0.266	0.320	0.277
152	0.306	0.265	0.317	0.275	0.300	0.259	0.311	0.269
160	0.298	0.258	0.308	0.267	0.292	0.253	0.302	0.261
168	0.291	0.252	0.300	0.260	0.285	0.246	0.295	0.254
176	0.284	0.246	0.293	0.254	0.279	0.241	0.287	0.248
184	0.278	0.241	0.286	0.248	0.272	0.235	0.280	0.242
192	0.272	0.235	0.280	0.242	0.267	0.230	0.274	0.237
200	0.266	0.231	0.274	0.237	0.261	0.226	0.268	0.232
208	0.261	0.226	0.268	0.232	0.256	0.221	0.263	0.227
216	0.256	0.222	0.263	0.227	0.251	0.217	0.257	0.222
224	0.251	0.218	0.258	0.223	0.247	0.213	0.253	0.218
232	0.247	0.214	0.253	0.219	0.242	0.209	0.248	0.214
240	0.243	0.210	0.248	0.215	0.238	0.206	0.243	0.210
248	0.239	0.207	0.244	0.211	0.234	0.202	0.239	0.207
256	0.235	0.204	0.240	0.208	0.231	0.199	0.235	0.203
264	0.231	0.200	0.236	0.205	0.227	0.196	0.232	0.200
272	0.228	0.197	0.232	0.201	0.224	0.193	0.228	0.197
280	0.225	0.195	0.229	0.198	0.220	0.190	0.225	0.194
288	0.222	0.192	0.226	0.195	0.217	0.188	0.221	0.191
296	0.218	0.189	0.222	0.193	0.214	0.185	0.218	0.188
304	0.216	0.187	0.219	0.190	0.211	0.183	0.215	0.186
312	0.213	0.184	0.216	0.187	0.209	0.180	0.212	0.183
320	0.210	0.182	0.214	0.185	0.206	0.178	0.209	0.181
328	0.207	0.180	0.211	0.183	0.203	0.176	0.207	0.179
336	0.205	0.178	0.208	0.180	0.201	0.174	0.204	0.176
344	0.203	0.175	0.206	0.178	0.199	0.172	0.202	0.174
352	0.200	0.173	0.203	0.176	0.196	0.170	0.199	0.172
360	0.198	0.171	0.201	0.174	0.194	0.168	0.197	0.170
368	0.196	0.170	0.199	0.172	0.192	0.166	0.195	0.168
376	0.194	0.168	0.196	0.170	0.190	0.164	0.193	0.166
384	0.192	0.166	0.194	0.168	0.188	0.162	0.191	0.165
392	0.190	0.164	0.192	0.166	0.186	0.161	0.188	0.163
400	0.188	0.163	0.190	0.165	0.184	0.159	0.187	0.161

Note. The number in brackets (2 or 3) is the number of classrooms per center. The MDES in this table are for main effects.

Table C10. MDES Based on ICC and R^2 From FACES (Assistant Teachers, Three-Year-Olds, Arnett as Outcome)

Number of Centers	No Baseline Outcomes Data				With Center-Level Mean CLASS Scores			
	DOSAGE & RECIPIENT		TRAINING		DOSAGE & RECIPIENT		TRAINING	
	(2)	(3)	(2)	(3)	(2)	(3)	(2)	(3)
96	0.396	0.349	0.420	0.371	0.391	0.344	0.415	0.366
104	0.380	0.335	0.401	0.354	0.375	0.331	0.396	0.349
112	0.366	0.322	0.385	0.339	0.361	0.318	0.380	0.335
120	0.353	0.311	0.370	0.326	0.349	0.307	0.365	0.322
128	0.342	0.301	0.357	0.315	0.337	0.297	0.352	0.310
136	0.331	0.292	0.345	0.304	0.327	0.288	0.341	0.300
144	0.322	0.284	0.334	0.295	0.318	0.280	0.330	0.291
152	0.313	0.276	0.325	0.286	0.309	0.272	0.320	0.282
160	0.305	0.269	0.316	0.278	0.301	0.265	0.312	0.275
168	0.298	0.262	0.307	0.271	0.294	0.259	0.303	0.267
176	0.291	0.256	0.300	0.264	0.287	0.253	0.296	0.261
184	0.284	0.251	0.293	0.258	0.281	0.247	0.289	0.255
192	0.278	0.245	0.286	0.252	0.275	0.242	0.282	0.249
200	0.272	0.240	0.280	0.247	0.269	0.237	0.276	0.243
208	0.267	0.236	0.274	0.242	0.264	0.232	0.271	0.238
216	0.262	0.231	0.269	0.237	0.259	0.228	0.265	0.234
224	0.257	0.227	0.263	0.232	0.254	0.224	0.260	0.229
232	0.253	0.223	0.259	0.228	0.250	0.220	0.255	0.225
240	0.248	0.219	0.254	0.224	0.245	0.216	0.251	0.221
248	0.244	0.216	0.250	0.220	0.241	0.213	0.246	0.217
256	0.240	0.212	0.246	0.217	0.237	0.209	0.242	0.214
264	0.237	0.209	0.242	0.213	0.234	0.206	0.238	0.210
272	0.233	0.206	0.238	0.210	0.230	0.203	0.235	0.207
280	0.230	0.203	0.234	0.207	0.227	0.200	0.231	0.204
288	0.227	0.200	0.231	0.204	0.224	0.197	0.228	0.201
296	0.224	0.197	0.228	0.201	0.221	0.194	0.225	0.198
304	0.221	0.194	0.224	0.198	0.218	0.192	0.222	0.195
312	0.218	0.192	0.221	0.195	0.215	0.189	0.219	0.193
320	0.215	0.190	0.218	0.193	0.212	0.187	0.216	0.190
328	0.212	0.187	0.216	0.190	0.210	0.185	0.213	0.188
336	0.210	0.185	0.213	0.188	0.207	0.182	0.210	0.185
344	0.207	0.183	0.210	0.186	0.205	0.180	0.208	0.183
352	0.205	0.181	0.208	0.183	0.202	0.178	0.205	0.181
360	0.203	0.179	0.206	0.181	0.200	0.176	0.203	0.179
368	0.200	0.177	0.203	0.179	0.198	0.174	0.201	0.177
376	0.198	0.175	0.201	0.177	0.196	0.172	0.198	0.175
384	0.196	0.173	0.199	0.175	0.194	0.171	0.196	0.173
392	0.194	0.171	0.197	0.173	0.192	0.169	0.194	0.171
400	0.396	0.349	0.420	0.371	0.391	0.344	0.415	0.366

Note. The number in brackets (2 or 3) is the number of classrooms per center. The MDES in this table are for main effects.

Appendix D. Alternative Design Option: Testing the Effect of Coaching Dosage, Delivery Model, and Coach Training

In this appendix, we discuss the random assignment plan, sample size requirements, and analysis for an alternative design that would examine the effect of the following three coaching dimensions: coaching Dosage (*DOSAGE*), Coach Training (*TRAINING*), and Mode of the coaching (*MODE*). Table D1 shows the eight experimental conditions in this alternative study design.

Table D1. Alternative 2^3 Factorial Design: *DOSAGE*, *MODE*, and *TRAINING*

Experimental Condition Number	Factors		
	Amount of Coaching (*DOSAGE*)	Delivery Mode (*MODE*)	Amount of Coach Training (*TRAINING*)
1	Monthly	Remote	Orientation
2	Monthly	Remote	Ongoing
3	Monthly	In person	Orientation
4	Monthly	In person	Ongoing
5	Biweekly	Remote	Orientation
6	Biweekly	Remote	Ongoing
7	Biweekly	In person	Orientation
8	Biweekly	In person	Ongoing

Note. Shading denotes Level II of the factor; unshaded cells represent Level I of the factor.

As in Section V of the report, we assume that the unit of random assignment would be HS centers for the *DOSAGE* dimension and coaches for the *TRAINING* dimension. As for the *MODE* dimension, we continue to assume that the unit of random assignment for this dimension could be either centers or coaches. The RA plan and analysis under these two scenarios for the *MODE* dimension are described below.

If the unit of random assignment for MODE is centers

If the unit of random assignment for *MODE* were centers, then two dimensions in the alternative design would be assigned at the center level (*MODE* and *DOSAGE*) and one dimension would be assigned at the coach level (*TRAINING*). Thus, the relevant random assignment plan would be RA Plan 1 or 2 in Section V ("*One of the Dimensions in the Design is Assigned at the Coach Level*"), except that the three coaching dimensions in the design would be *DOSAGE*, *MODE* and *TRAINING* (rather than *DOSAGE*, *RECIPIENT* and *TRAINING*). This means that for a sample of 248 centers, the MDES for the *DOSAGE* and *MODE* dimensions would be 0.20. The MDES for the *TRAINING* dimension would depend on whether coaches could be randomized to centers (RA Plan 1) or not (RA Plan 2). In terms of the statistical analysis, effects could be estimated using Model 1 in Appendix E of this report, but *MODE* would be included in the model rather than *RECIPIENT*.

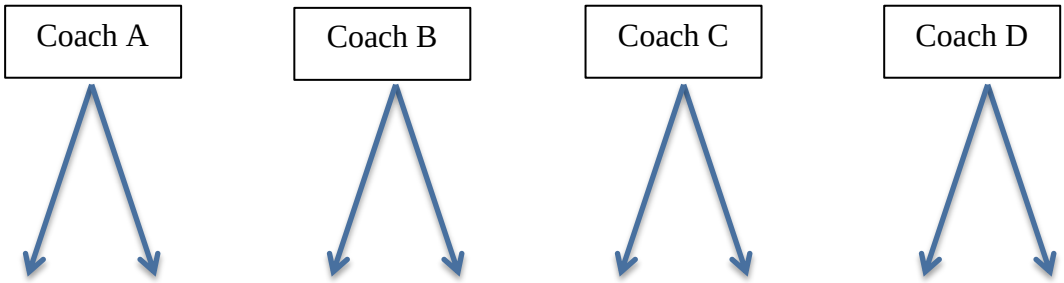
If the unit of random assignment for MODE is coaches

If coaches (rather than centers) were randomly assigned to the levels of *MODE*, then two of the dimensions would be assigned at the coach level (*MODE* and *TRAINING*) and *DOSAGE* would be assigned at the center level.

In this scenario, there would need to be 4 coaches per random assignment block—because there are four different combinations of the levels of the two dimensions assigned at the coach level (*TRAINING* and *MODE*). Table D2 shows the structure of a hypothetical grantee (or random assignment block) meeting this requirement. In this example, there are four coaches per grantee, and each coach works with four centers. (For simplicity, the figure does not show classrooms within centers, but we continue to assume that there would be approximately 3 classrooms per center.) Thus, this alternative design may require recruiting larger grantees compared to the other designs presented in the report, since each grantee would need 4 coaches. Or, it may be necessary define RA blocks by combining two similar grantees into one block.

Like the designs discussed in Section V of the report, the specific RA plan depends on whether grantees would allow the evaluators to reassign coaches to centers, or whether coach assignments would have to be taken as assigned by the grantee. The RA plans under each scenario are described below.

Table D2. Hypothetical Staffing Structure in Each Random Assignment Block (Alternative Design)



Center 1	**Center 2**	**Center 5**	**Center 6**	**Center 9**	**Center 10**	**Center 13**	**Center 14**
Center 3	**Center 4**	**Center 7**	**Center 8**	**Center 11**	**Center 12**	**Center 15**	**Center 16**

RA Plan 4: Two Dimensions are Assigned at the Coach Level and Coaches Can Be Randomized to Centers

Under this plan, random assignment would proceed in two steps:

1. **Centers would be randomly assigned to the eight possible combinations of the levels of *DOSAGE*, *MODE*, and *TRAINING*.**

   For the hypothetical HS grantee in Table D2, the grantee's centers would be randomly assigned to the eight experimental conditions representing all possible combinations of *DOSAGE*, *MODE*, and *TRAINING* (these are the eight conditions in the design shown in

Table D1). In this example, two centers would be assigned to each experimental condition.

2. **Coaches would be randomly assigned to the four possible combinations of the levels of the *TRAINING and MODE* dimensions. Each coach would then work with the centers that have been assigned (in Step 1) to the experimental conditions where *TRAINING* and *MODE* are set to the same level.**

   For example, for the hypothetical grantee in Table D2, Coaches A, B, C, and D would be randomly assigned to receive one of the four combinations of *TRAINING* and *MODE* shown in Table D3. Let's assume that Coach A is assigned to receive an orientation and to deliver remote coaching. Coach A would then work with the four centers that have been assigned to the conditions where *TRAINING* is set to orientation and *MODE* is set to remote (Conditions 1 and 4). The same principle would apply for the other three coaches. As explained in Section V, this is equivalent to randomly assigning coaches to centers.

Note that if blocks are constructed by combining two or more grantees (in order to meet the requirement of four coaches per block), then randomizing coaches to centers would probably not be possible, because coaches cannot typically work across grantees. In this case, RA Plan 5 (described later in this appendix) would have to be used.

Table D3. Combinations of *MODE* and *TRAINING* to Which Coaches Are Assigned

Experimental Condition Number	Factors	
	Delivery Mode (*MODE*)	Amount of Coach Training (*TRAINING*)
1 & 4	Remote	Orientation
2 & 5	Remote	Ongoing
3 & 6	In person	Orientation
4 & 8	In person	Ongoing

Note. Shading denotes Level II of the factor; unshaded cells represent Level I of the factor.

The minimum detectable effect size (MDES) based on different sample sizes (number of centers) are shown in Table D3.[98] As shown in this table, the sample size requirements are essentially the same as for the other designs discussed in this report. Specifically, if we assume that no baseline outcomes data was available, a sample of 256 centers would be needed to detect an effect size of 0.20 for the *DOSAGE*, *MODE*, and *TRAINING* dimensions (compared to 248 in the main design described in the report). The sample size is slightly larger for the alternative design because it was assumed that the sample size should be a multiple of 16 centers (248 is *not* a multiple of 8, so the sample size here is slightly higher, 256).[99] More generally, the sample sizes in Table D4

[98] The sample sizes in Table D3 are the smallest sample size needed to detect an effect size of given magnitude *rounded to the third decimal*.

[99] These MDES are based on the formulas in Appendix B using the CARES (control group) parameter assumptions. We focused on sample sizes that are multiples of 16 centers (number of centers per block, as shown in Figure D-1), rather than multiples of 8 centers as in other parts of the report.

are either the same as in Table 11 (when the sample size in Table 11 is a multiple of 16) or eight centers larger than in Table 11 (when the sample size in Table 11 is not a multiple of 16).[100]

Table D4. Sample Size Requirements Based on RA Plan 4

	No Baseline Outcomes Data				**With Center-Level Mean CLASS Scores**			
	DOSAGE		***TRAINING & MODE***		***DOSAGE***		***TRAINING & MODE***	
MDES	**(3)**	**(2)**	**(3)**	**(2)**	**(3)**	**(2)**	**(3)**	**(2)**
0.12	688	864	688	864	576	752	592	752
0.13	592	736	592	736	496	640	496	640
0.14	512	640	512	640	432	560	432	560
0.15	448	560	448	560	368	480	384	496
0.16	400	496	400	496	336	432	336	432
0.17	352	432	352	448	288	384	304	384
0.18	320	384	320	400	256	336	272	352
0.19	288	352	288	352	240	304	240	304
0.20	256	320	256	320	208	272	224	288

Note. These calculations are based on the ICC and R^2 from the HS CARES study (control group centers only, with an adapted version of the TSRS as the outcome measure). See Appendix B for parameter assumptions. The number in brackets (2 or 3) is the number of classrooms per center.

RA Plan 5: Two Dimension are Assigned at the Coach Level but Coaches Cannot Be Randomized to Centers

For grantees where coach assignments to centers must be taken as assigned by grantees, the random assignment plan would be as follows:

1. **Coaches would be randomly assigned to the four possible combinations of the levels of the *TRAINING* and *MODE* dimensions.**

   Coaches A, B, C and D would each be assigned to one of the combinations of levels in Table D2.

2. **Each coach's centers would then be randomly assigned to one of the levels of the *DOSAGE* dimension.**

   For example, two of Coach A's four centers would be randomly assigned to receive monthly coaching while the other two centers would receive biweekly coaching. This process would be repeated for each coach's centers.

As explained in Section V, if grantees' decisions about coach assignments are related to the outcomes of centers, then this would increase the MDES for the coach-level dimensions (*TRAINING* and *MODE*). Table D5 looks at the MDES for the *TRAINING* and *MODE* dimension under different assumptions about the extent to which grantees assign coaches to centers based on their outcomes. As discussed in the previous section (RA Plan 4), the MDES for

[100] Because sample sizes are constrained to be multiples of 16 centers, the increase in the sample size for a 0.01 decrease in the MDES is not always monotonic.

the *TRAINING* or *MODE* dimension would be 0.20 if coaches were randomized to centers and the sample was 256 centers (this is shown in the second column of Table D5). However, for this same number of centers, the MDES for the *TRAINING* and *MODE* dimensions would increase to 0.21 or 0.23 if coaches were not randomly assigned and coach assignments were weakly or moderately associated with centers' outcomes (RA Plan 5). More generally, for a given number of centers, some of the MDES in Table D5are slightly smaller than in Table 13, because fewer degrees of freedom are used by for the alternative model.[101]

Table D5. MDES for the Main Effect of *TRAINING* and *MODE* Under RA Plans 4 and 5 (Three Classrooms per Center)

	No Baseline Outcomes Data			With Center-Level Mean CLASS Scores		
Number of Centers	Random Assignment of Coaches to Centers (RA Plan 3)	Assignment of Coaches to Centers Is Weakly Associated With Center-Level Outcomes (RA Plan 4)	Assignment of Coaches to Centers Is Moderately Associated With Center-Level Outcomes (RA Plan 5)	Random Assignment of Coaches to Centers (RA Plan 3)	Assignment of Coaches to Centers Is Weakly Associated With Center-Level Outcomes (RA Plan 4)	Assignment of Coaches to Centers Is Moderately Associated With Center-Level Outcomes (RA Plan 5)
192	0.23	0.25	0.27	0.21	0.22	0.24
208	0.22	0.24	0.26	0.20	0.22	0.23
224	0.21	0.23	0.25	0.20	0.21	0.22
240	0.21	0.22	0.24	0.19	0.20	0.22
256	0.20	0.21	0.23	0.18	0.19	0.21
272	0.19	0.21	0.23	0.18	0.19	0.20
288	0.19	0.20	0.22	0.17	0.18	0.20

Note. These calculations are based on the ICC and R^2 from the HS CARES study (control group centers only, with an adapted version of the TSRS as the outcome measure). See Appendix B for parameter assumptions. We further assume that there will be three classrooms per center and four centers per coach.

Statistical Model

If *TRAINING* and *MODE* were both assigned at the coach level, then the estimation of main effects would be as discussed in Section VI (i.e., by comparing the groups of centers in different experimental conditions). Like the *TRAINING* dimension, the standard error of the *MODE* dimension would have to account for two levels of clustering (at the center level and at the coach level). This could be accomplished using a multilevel model, or by using a correction to adjust the OLS standard error upwards.[102]

[101] This is because for a given number of centers, there are fewer random assignment blocks (since there must be 16 centers per block for the alternative design rather than 8 centers per block). Thus, the model uses fewer degrees of freedom.

[102] If using a multilevel model (see Appendix E), one would include both *MODE* and *TRAINING* as independent variables in Equation 1c of Model 1.

Appendix E. Statistical Models

This section describes the multilevel models that could be used to estimate the main effect of each coaching dimension on teacher and classroom outcome, as well as interaction effects between these dimensions. Multilevel models make it possible to account for the features of the design—such as cluster RA, blocking, and the use of covariates to improve precision. (At the end of this appendix, we discuss the limitations of this approach and alternatives.)

The unit of RA affects the structure of the multilevel model. Therefore, this appendix presents models under two scenarios:

- One of the three dimensions is randomly assigned at the coach level (*TRAINING* or *MODE*) [RA Plan 1 or 2 in Section V]
- All three dimensions are assigned at the center level (*DOSAGE*, *RECIPIENT*, and *MODE*) [RA Plan 3 in Section V)

In all models, we assume that the baseline measures used in the analysis will be measured at the center level (e.g., using center-level CLASS scores from existing data sources) rather than by collecting new baseline data at the classroom level.[103] For simplicity, the presentation of the models assumes that the outcomes of interest will be classroom-level measures of teacher practice and child-teacher relationships. However, we will also discuss the implications for the analysis of having multiple measures of an outcome per classroom (for instance, survey responses from both the lead teacher and the teaching assistant).

Model 1: One of the Dimensions Is Randomly Assigned at the Coach Level (RA Plan 1 or 2)

If one of the dimensions is assigned at the coach level (*TRAINING* or *MODE*), it will be important to account for the fact that RA is being conducted at two levels: at the center level and at the coach level. This can be accomplished by using a multilevel model with three levels:

Level 1 (classrooms within HS centers):

$$Y_{ijc} = \alpha_{0jc} + \varepsilon_{ijc} \quad \textbf{(1a)}$$

Level 2 (HS centers within coaches):

$$\alpha_{0jc} = \gamma_{0c} + \gamma_{1c} DOSAGE_{jc} + \gamma_{2c} RECIPIENT_{jc} + \gamma_{3c} DOSAGE_{jc} \times RECIPIENT_{jc} + \sum_k \lambda_k W_{kjc} + u_{jc} \quad \textbf{(1b)}$$

[103] As discussed in Section V, this assumption is motivated by the fact that in prior studies conducted in HS centers, classroom-level measures of baseline teacher practice do not substantially improve the MDES above and beyond controlling for grantees.

Level 3 (coaches):

$$\gamma_{0c} = \beta_0 + \beta_2 TRAINING_c + \sum_d \rho_d B_{dc} + \omega_c \quad \textbf{(1c)}$$
$$\gamma_{1c} = \beta_1 + \beta_4 TRAINING_c$$
$$\gamma_{2c} = \beta_3 + \beta_6 TRAINING_c$$
$$\gamma_{3c} = \beta_5 + \beta_7 TRAINING_c$$

where *i* denotes classrooms, *j* denotes HS centers, and *c* denotes coaches. The variables in this model are defined as follows:

Y_{ijc}	=	Classroom outcome for classroom i in center j served by coach c
$DOSAGE_{jc}$	=	Dichotomous indicator for the level of *DOSAGE* that center j was randomly assigned to receive (+1/2 if biweekly coaching; -1/2 if monthly coaching)
$RECIPIENT_{jc}$	=	Dichotomous indicator for the level of *RECIPIENT* that center j was randomly assigned to receive (+1/2 if entire teaching team; -1/2 if lead teacher only)
$TRAINING_c$	=	Dichotomous indicator for the level of *TRAINING* that coach c is assigned to receive (+1/2 if ongoing sessions; -1/2 if orientation)
B_{dc}	=	A set of D dichotomous indicators for RA block (=1 if center j and coach c are in a particular RA block; =0 otherwise)[104]
W_{kjc}	=	A set of K baseline characteristics for center j served by coach c (for example, center-level mean CLASS scores)
ω_c	=	Between-coach random variation in outcome Y
u_{jc}	=	Between-school (within-coach) random variation in outcome Y
ε_{ijc}	=	Between-classroom (within-center) random variation in outcome Y

In the multilevel model shown here, note that *DOSAGE* and *RECIPIENT* appear at level 2 because the level of RA for these dimensions is HS centers, whereas *TRAINING* (or *MODE*) appears at level 3 because the unit of RA for this dimension is coaches.[105]

[104] As explained in Section V, blocks will be grantees or subgroups of similar centers within grantees.

[105] Note that in this model, the highest level of clustering—coaches clustered within HS *grantees*—is accounted for by including a set of indicator variables for RA blocks in the model, which in this study will be grantees or groups of similar centers within grantees. As will be explained, these indicator variables also improve the precision of estimated effects.

Combining the three levels of the model (1a, 1b, and 1c) allows the statistical model to be collapsed to one equation as follows:

$$Y_{ijc} = \beta_0 + \beta_1 DOSAGE_{jc} + \beta_2 TRAINING_c + \beta_3 RECIPIENT_{jc} + \beta_4 DOSAGE_{jc} \times TRAINING_c + \beta_5 DOSAGE_{jc} \times RECIPIENT_{jc} + \beta_6 TRAINING_c \times RECIPIENT_{jc} + \beta_7 DOSAGE_{jc} \times RECIPIENT_{jc} \times TRAINING_c + \sum_d \rho_d B_{djc} + + \sum_k \lambda_k W_{kjc} + \sum_s \theta_s X_{sijc} + \omega_c + u_{jc} + \varepsilon_{ijc} \quad \textbf{(1)}$$

If the effect of Mode (*MODE*) was tested instead of the effect of the *TRAINING* dimension—and *MODE* were assigned at the coach level—the analysis would be the same, except that *MODE* would be substituted for *TRAINING* in the model.

The main effects and interaction effects of interest—as defined in Section IV of this report—are simply the regression coefficients from Model 1:

β_1 = Main effect of *DOSAGE*

β_2 = Main effect of *TRAINING* (or *MODE*)

β_3 = Main effect of *RECIPIENT*

β_4 = Interaction effect of *DOSAGE* × *TRAINING* (or *MODE*)

β_5 = Interaction effect of *DOSAGE* × *RECIPIENT*

β_6 = Interaction effect of *TRAINING* (or *MODE*) × *RECIPIENT*

β_7 = Interaction effect of *DOSAGE* × *RECIPIENT* × *TRAINING* (or *MODE*)

Because a multilevel model is used, the standard errors for the regression coefficients (main effects and interactions) appropriately account for clustering. Specifically, the standard error for the main effects of *DOSAGE* and *RECIPIENT* will account for clustering at the center level, and the standard error for the *TRAINING* dimension will account for clustering at the center level and coach level. As explained in Section V, given the goals of the HS Coaching Study, we recommend that the statistical significance level used for hypothesis testing should be 10 percent.

Several features of the model are worth noting.

- *Coding of Factor Levels.* The levels of the factors are coded as +1/2 and -1/2, as opposed to dummy coded (i.e., coded as 1 and 0). The purpose of this alternative coding is to be able to estimate main effects more easily. For reasons discussed in Section III, the goal in this study will be to estimate the *main effect* of each dimension—or its average effect across all levels of the other dimensions being tested. However, if dummy coding were used to code the factors, the regression coefficients in Model 1 would represent the *simple effect* of each dimension—that is, its effect at a given level of the other dimensions. However, by coding the levels as +1/2 and -1/2, the regression coefficients in Model 1 become *main effects* as defined in Section III. (This can be shown using simple algebra.) Although the two approaches to coding produce the same overall test of the

model (omnibus *F*), the advantage of the alternative coding is that it more easily produces main effects and interactions that can be interpreted as described in Section IV.[106] Another advantage is that the regression coefficients are uncorrelated (whereas with dummy coding, regression coefficients are correlated). Thus, alternative coding (+1/2 and -1/2) is recommended for factorial experiments that are being conducted for the purpose of testing intervention components. For a technical discussion of the difference between dummy coding and other types of coding, see Kugler, Trail, Dziak, and Collins (2012).[107]

- *Center Characteristics.* Model 1 includes a set of baseline center-level characteristics (*W*). The primary covariates here would be centers' average CLASS scores (for each subdomain) from prior school years.[108] Strictly speaking, it is not necessary to control for baseline characteristics because, on average, classrooms and centers in the eight experimental conditions should have similar characteristics at baseline because of RA. However, as discussed in Section V, controlling for these characteristics may slightly improve the precision of estimated effects, which in turn could reduce the MDES (by about 0.01 standard deviations, on the basis of assumptions from prior studies).[109]
- *RA Blocks.* As recommended, HS centers will be randomized by grantee or by groups of similar centers within grantees. Model 1 includes a set of indicators for these RA blocks for several reasons. First, controlling for blocking will improve the precision of estimated effects.[110] Second, including RA blocks in the model will account for the nesting of HS centers and coaches within grantees, which is important for obtaining the correct standard error of estimated effects. Finally, controlling for blocks accounts for possible differences in the RA ratio across blocks.[111]
- *Fixed-Effects Approach.* Estimated main effects and interaction effects from Model 1 are fixed-effects estimates because the effects of interest are not allowed to vary across RA

[106] If dummy coding is used, the main effect of a factor can still be obtained, but it is a more complicated linear combination of not only that factor's r regression coefficient, but also the regression coefficients for the other factors being tested.

[107] Rather than coding the levels as +1/2 and -1/2, one could instead code the levels as +1 and -1 (this is called "effect coding"). If effect coding is used, then the main effect of a factor would be 2 times its regression coefficient, a two-way interaction effect (as defined in Section IV) would be 4 times the associated regression coefficient, and the three-way interaction effect would be 8 times the associated regression coefficient.

[108] One could also control for other center characteristics such as staff-child ratios, and the average demographic characteristics of children at each center, which are potentially available from administrative records. However, because center-level characteristics reduce the degrees of freedom for the analysis, it is best to include only characteristics that are predictive of the outcome so that the loss of degrees of freedom is balanced by more precise effect estimates.

[109] In addition to center-level data, centers may be able to provide information on the characteristics of teachers from their roster data. If these data were available, they would be included as covariates in level 1 of the model.

[110] As discussed in Section V, we found that blocking by grantees in the CARES study substantially contributed to reducing the MDES.

[111] If the number of participating centers from a grantee is a multiple of eight, the random assignment ratio would be balanced (the same number of centers per experimental condition). However, if the number of participating centers is not a multiple of eight, then the random assignment ratio may be imbalanced (there can be more centers in some conditions than others). Such differences in the random assignment ratio must be accounted for to obtain an unbiased estimate of impacts. There are several ways to account for variation in the random assignment ratio. The two most common are to (a) "block-mean" center the covariates on the right-hand side of the model or (b) include block fixed-effects in the model. Raudenbush (2009) shows that these two methods produce the same impact estimate. Model 1 is based on the latter approach.

blocks (grantees). An alternative approach would be to model the variation in effects across blocks by adding a fourth level to the multilevel model (coaches within grantees) and allowing the regression coefficients for the main effects to vary randomly across blocks. This approach would yield random-effects estimates of main effects. The advantage of this approach is that the variation in main effects across grantees is part of the model, and, therefore, it becomes possible to examine whether the magnitude of main effects is associated with grantee characteristics.[112] Using a random-effects approach would also generalize the inference of the estimated effects to a broader population of grantees and centers. However, the precision of main effects is reduced when using a random-effects model, and, therefore, the sample size needed to detect a main effect of given magnitude would need to be even larger than reported in Section V. Thus, we recommend that a random-effects approach not be used.

- *Pooled Effects.* In Model 1, the estimated effect of each dimension (factor) is its pooled effect across all RA blocks (grantees), with block-specific effects weighted based on their precision. This means that the estimated effects from larger grantees (those with more centers or classrooms) may have a greater weight in the pooled effect. This self-weighting approach maximizes the precision of estimated effects and, therefore, minimizes the MDES. An alternative option would be to reweight the sample so that blocks (grantees) are weighted equally in the estimation of the average effect; however, reweighting can appreciably increase the standard error of estimated effects and, by extension, the MDES. Therefore, we do not recommend reweighting the sample so that grantees are equally weighted. If equal weighting is an important goal—for example, because larger effects are expected in some grantees than others—then rather than reweighting the analysis, it would be preferable to recruit the same (or a similar) number of centers per grantee so that each grantee can implicitly have a similar weight in the analysis (hence obviating the need to reweight the analysis). Because the study sample may likely include large HS grantees, the number of centers in each grantee may be quite similar across grantees by virtue of the study design and recruitment plan.

Model 2: All Three Dimensions Are Randomly Assigned at the Center Level (RA Plan 3)

If all three dimensions in the design are assigned at the center level (*DOSAGE*, *RECIPIENT*, and *MODE*), then the analysis must account for RA at the center level. This can be accomplished using a simplified version of Model 1 that has two levels, with classrooms nested within centers:

Level 1 (classrooms within HS centers):

$$Y_{ij} = \psi_{0j} + \varepsilon_{ij} \qquad \textbf{(2a)}$$

[112] This type of question could be examined by expanding Model 1 to include interactions between the factors and a variable representing a particular grantee characteristic.

Level 2 (HS centers):

$$\psi_{0j} = \phi_0 + \phi_1 DOSAGE_j + \phi_2 MODE_j + \phi_3 RECIPIENT_j + \phi_4 DOSAGE_j \times MODE_j + \phi_5 DOSAGE_j \times RECIPIENT_j + \phi_6 MODE_j \times RECIPIENT_j + \phi_7 DOSAGE_j \times RECIPIENT_j \times MODE_j + \sum_k \lambda_k W_{kj} + \eta_j \quad \textbf{(2b)}$$

where *i* denotes classrooms and *j* denotes HS centers. The new variables in this model are defined as follows:

$MODE_j$	=	Dichotomous indicator for the level of MODE that center j was randomly assigned to receive (+1/2 if in-person coaching; -1/2 if remote coaching)
η_j	=	Between-school random variation in outcome Y
ε_{ij}	=	Between-classroom (within-center) random variation in outcome Y

The collapsed version of this model would be the following:

$$Y_{ij} = \phi_0 + \phi_1 DOSAGE_j + \phi_2 MODE_j + \phi_3 RECIPIENT_j + \phi_4 DOSAGE_j \times MODE_j + \phi_5 DOSAGE_j \times RECIPIENT_j + \phi_6 MODE_j \times RECIPIENT_j + \phi_7 DOSAGE_j \times RECIPIENT_j \times MODE_j + \sum_k \lambda_k W_{kj} + \eta_j + \varepsilon_{ij} \quad \textbf{(2)}$$

Because of the effect coding of the factors, the main effects and interactions are simply the regression coefficients from the model:

ϕ_1	=	Main effect of *DOSAGE*
ϕ_2	=	Main effect of *MODE*
ϕ_3	=	Main effect of *RECIPIENT*
ϕ_4	=	Interaction effect of *DOSAGE* × *MODE*
ϕ_5	=	Interaction effect of *DOSAGE* × *RECIPIENT*
ϕ_6	=	Interaction effect of *MODE* × *RECIPIENT*
ϕ_7	=	Interaction effect of *DOSAGE* × *RECIPIENT* × *MODE*

The general features of Model 2 (such as blocking and controlling for center-level characteristics) are the same as discussed in the context of Model 1.

Testing for Baseline Equivalence

In a factorial experiment—as in other experimental designs—it is important to verify that RA resulted in experimental groups that are statistically equivalent in terms of their baseline

characteristics. As noted earlier, in the HS Coaching Study, baseline characteristics (*W*) will likely be measured at the center level—for example, center-level CLASS scores and other characteristics available from centers' administrative records. To test whether there is balance between experimental groups with respect to these center-level characteristics, one would use an adapted version of Models 1 and 2 in which the lowest level (classrooms within centers) is omitted and in which the outcome variable (*Y*) is replaced by the center-level characteristic of interest.[113] Ideally, the main effects of each dimension on the center-level baseline characteristics from these regressions should be close to zero and not statistically significant.

Cautions and Extensions

More generally, it is worth noting that although multilevel models are the most common analytical approach in cluster RA studies in the field of education, their estimation can become problematic when there are very few observations in a cluster. For instance, we highlight here the fact that there may be only about two or three classrooms per HS center and that some centers may have outcomes data for only one classroom (if, for example, classroom observations cannot be conducted in some classes on the day of the site visits). If a substantial number of centers in the study sample have only one classroom data point, then the multilevel model described here can be difficult to estimate (specifically, it may not converge).[114] In this situation, one possible option would be to aggregate the classroom data to the center level (that is, calculate center-level means of the variables) and then fit a two-level model (centers within coaches, or levels 2 and 3 in Model 1) to the aggregated data. Another option would be to forgo the use of a multilevel model altogether and instead use an OLS regression with cluster-robust standard errors. The most suitable approach may depend on the structure of the data in the actual sample, and one would likely want to try several analytical approaches in order to test the sensitivity of the results.

On a related point, it is also worth noting that some of the data collected for the HS Coaching Study could be at the teacher level—for instance, there may be survey data from both the lead teacher and the assistant teacher. For such outcomes, there would be an additional level of clustering (teachers within classrooms). To handle this additional level of clustering, another level could be added to Models 1 and 2 (teachers within classrooms). However, as just discussed, multilevel models can be difficult to estimate if some clusters have only one observation (for instance, if some classrooms had only one teacher who responded to the survey but others had two teacher responses). Similar to the earlier discussion of this issue, one approach would be to aggregate the teacher data to the classroom level and to use Models 1 or 2 as specified, or one could use an OLS approach with the teacher-level data and use cluster-adjusted standard errors. Again, the choice may depend on the data structure, so, for now, we simply highlight this as an issue for further consideration once the HS Coaching Study's sample has been recruited and the data have been collected.

[113] The analysis would exclude center-level characteristics (*W*) from the right-hand side of the model.

[114] This is because in the subset of the sample in which there is only one classroom per center, there would be one less level of clustering. For example, in Model 1, for centers with one classroom, there would be two levels of clustering (centers within coaches), whereas in the remainder of centers (those with two or more classrooms), there would be the usual three levels of clustering.

ABOUT AMERICAN INSTITUTES FOR RESEARCH

Established in 1946, with headquarters in Washington, D.C., American Institutes for Research (AIR) is an independent, nonpartisan, not-for-profit organization that conducts behavioral and social science research and delivers technical assistance both domestically and internationally. As one of the largest behavioral and social science research organizations in the world, AIR is committed to empowering communities and institutions with innovative solutions to the most critical challenges in education, health, workforce, and international development.

LOCATIONS

Domestic

Washington, D.C.

Atlanta, GA

Baltimore, MD

Chapel Hill, NC

Chicago, IL

Columbus, OH

Frederick, MD

Honolulu, HI

Indianapolis, IN

Naperville, IL

New York, NY

Portland, OR

Sacramento, CA

San Mateo, CA

Silver Spring, MD

Waltham, MA

International

Egypt

Honduras

Ivory Coast

Kenya

Liberia

Malawi

Pakistan

South Africa

Zambia

1000 Thomas Jefferson Street NW
Washington, DC 20007-3835
202.403.5000 | TTY 877.334.3499

www.air.org

Making Research Relevant

2226_06/14