THE ROLE OF PROGRAM QUALITY IN DETERMINING HEAD START'S IMPACT ON CHILD DEVELOPMENT

OPRE Report 2014-10

March 2014

Third Grade Follow-Up to the Head Start Impact Study

THE ROLE OF PROGRAM QUALITY IN DETERMINING HEAD START'S IMPACT ON CHILD DEVELOPMENT

OPRE Report 2014-10

March 2014

Contract Number: HHSP23320062929YC

Project Director:
Camilla Heid
Westat
1600 Research Boulevard
Rockville, MD 20850

Submitted to:
Jennifer Brooks, Project Officer
Office of Planning, Research and Evaluation
Administration for Children and Families
U.S. Department of Health
and Human Services

Submitted by:
Laura R. Peck
Stephen H. Bell
Social & Economic Policy

Abt Associates
4550 Montgomery Avenue
Suite 800 North
Bethesda, MD 20814

Suggested citation: Peck, Laura R., and Stephen H. Bell. (2014). The Role of Program Quality in Determining Head Start's Impact on Child Development. OPRE Report #2014-10, Washington, DC: Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.

This report and other reports sponsored by the Office of Planning, Research and Evaluation are available at http://www.acf.hhs.gov/programs/opre/index.html.

Overview

The Head Start Impact Study (HSIS) has shown that having access to Head Start improves children's preschool experiences and school readiness in certain areas, though few of those advantages persisting through third grade (Puma et al., 2012). Scholars and practitioners alike have wondered whether impacts might be larger or more persistent for children who participate in *high quality* Head Start as opposed to lower quality Head Start. In response, this report examines the vital policy question: To what extent does variation in the quality of children's Head Start experiences affect children's development? The HSIS experimental evaluation, which involved a nationally representative sample and included rich data at baseline, about programs and across several years of follow-up, provides an ideal source for analyzing the answer to this question.

Further informed by experts in the field, this report uses measures of quality based on the ECERS, Arnett, and teacher reports to capture three distinct dimensions of the Head Start setting: (1) "resources," which are the physical characteristics available in the program; (2) the "interactions" between teacher and child; and (3) children's "exposure" to academic activities in the classroom. Slightly less than three-fourths of the Head Start children in the study were in high quality classrooms for the resources and interactions quality measures, while on the exposure to academic activities measure, about one-fourth of the Head Start children were in high quality classrooms. Prior research posits that richer resources and more favorable interactions should be associated with better cognitive and social outcomes. The relationship of exposure to academic activities among children of this age is less clear, with some reason to think that too much such exposure may not necessarily benefit children.

We find little evidence that quality matters to impacts of Head Start using the available quality measures from the study across two age cohorts, three quality dimensions, five outcomes, and several years. The one exception is that for 3-year-old program entrants low exposure quality, defined as less exposure to academic activities during Head Start participation, produces better behavioral impacts in the short-run than more exposure to academic activities. Even so, there is no indication that either high quality Head Start or low quality Head Start in any dimension leads to program impacts lasting into third grade.

The analysis of quality makes use of the HSIS experimental evaluation design to capitalize on the fact that children were randomized into treatment and control groups, allowing any predicted quality subgroup of the treatment group to be matched to its counterpart in the control group. The analytic approach we take eliminates plausible rival explanations for observed impacts, an approach we advocate for future research that is otherwise challenged by potential selection bias on post-random assignment mediating factors, such as quality.

Section 1: Introduction

We know from the Head Start Impact Study (HSIS) that having access to Head Start moderately improves children's preschool experiences and school readiness in certain areas, with some of those advantages persisting through first grade but few lasting into third grade (Puma et al, 2012). Scholars and practitioners alike have wondered whether impacts might be larger for those who participate in high quality Head Start. To explore this, the current report considers the extent to which the *quality* of a child's Head Start experience affects children's development. To evaluate the impacts of Head Start quality on children's development, we ask: To what extent does variation in the quality of children's Head Start result in variation in impacts on children's development? In other words, does evidence of impact differ for children participating in high quality Head Start programs from that for Head Start participants as whole? To what extent might the main study findings understate what the program can accomplish in its strongest form?

Despite the importance of these questions to policy and practice, there are analytic challenges involved in addressing them. Among these challenges are: (1) defining the construct and potentially numerous dimensions of "quality" conceptually, (2) making "quality" as defined *measurable* with validity and reliability from the study's data, and (3) determining impacts by variation in Head Start quality given that Head Start quality is undefined in the control group who were not afforded access to Head Start. This report aims to tackle these main challenges and analyzes how the quality of children's Head Start experience influences the degree to which the program impacts their cognitive and social-emotional development.

In response to the first two challenges, we choose to operationalize quality in three distinct ways. For each of these quality constructs, we create index measures that collapse several variables, thereby potentially increasing the measures' reliability and validity. With regard to the third challenge, we know that (a) treatment group children who participate in high quality Head Start are likely to differ from treatment group children who participate in lower quality Head Start in ways that relate to their subsequent outcomes (e.g., social and intellectual development) independently of the Head Start program, and (b) the concept of Head Start quality is undefined in the control group. On the first of these two points, children who experience high quality Head Start certainly differ from those who do not in terms of where they live, and to which Head Start centers their families apply for admission to the program. It is also possible that more motivated and organized parents navigate the Head Start options in their communities more effectively (e.g., manage to get their children placed in classrooms with higher quality teachers) while at the same time doing more to *expand* Head Start's impact in the way they interact with their children at home.

Regarding the non-identifiability of control group children likely to participate in high and low quality programs—challenge (b) above—because of randomization we know that any subgroup that exists in one of the two randomly-divided experimental samples must have a counterpart in the other sample. Capitalizing on this property of the experimental design, we sort the HSIS control group on the same preexisting traits as characterize children in the Head Start treatment group who experienced low or high quality, and this serves as a benchmark for measuring the impact of Head Start at various quality levels through analytic procedures described in depth later in the report. To capitalize on having an experimental design, we use baseline characteristics to identify subgroups of treatment group children who do not participate in Head Start (i.e., no-shows) and who participate at varying levels of Head Start quality. We then apply a totally symmetric procedure to identify their counterparts in the control group (see Peck, 2003, 2013). This allows us to calculate impacts by Head Start quality subgroupings analogous to calculating impacts on subgroups of children defined (symmetrically in the treatment and control groups) by discrete, individual background characteristics such as sex or parental employment at enrollment.

In brief, we find little evidence that quality matters to impacts of Head Start on selected child outcomes, using the available quality measures from the study. The one exception is that for 3-year-old program entrants low exposure quality, defined as less exposure to academic activities during Head Start participation, produces better behavioral impacts in the short-run than more exposure to academic activities.

The remainder of this report proceeds as follows: Section 2 provides background on why we would expect variation in the quality of a child's preschool experience to be associated with variation in children's developmental outcomes and the trajectory of those outcomes over time. It also presents background on the HSIS's design, timing, and site coverage. Section 3 details the data that come from the HSIS and the particular measures we use in this research, including our measures of quality and selected outcomes of interest. Section 4 describes the methodological approach used to analyze the extent to which levels of Head Start quality result in variation in impacts on children's development. Section 5 presents the findings, and Section 6 concludes. An Appendix includes additional material on the analytic method, including details of alternative assumptions and results not otherwise presented in the main text.

Section 2: Background

This section discusses why and how the quality of Head Start should matter to children's cognitive and behavioral development. We then discuss the background of the Head Start Impact study, the source for this report's data.

Why & How Should Quality Matter?

As surveyed in Mashburn et al. (2008), the "quality" of early childhood education refers to a wide range of features that children experience in preschool classrooms and school settings that are presumed to impact their development. Definitions of high quality education may include the nature of children's experiences in classrooms (e.g., the furnishings and learning materials accessible to children, the frequency of instructional activities, the interactions between teachers and children), characteristics of teachers (e.g., level of education and field of study), and the nature of the interaction with those teachers (Mashburn et al., 2008).

In the context of the HSIS, higher Head Start quality experiences are hypothesized to affect children's cognitive, academic and social skill development differently from low quality experiences. For example, it is presumed that children will achieve better developmental outcomes—i.e., larger impacts—if they attend Head Start programs characterized by features such as: well-maintained furnishings, ample learning materials, instructionally-rich and emotionally-supportive interactions between teachers and children, teachers with bachelors and advanced degrees in child development or early childhood education, and exposure to proficient peers. If these quality factors matter during a child's Head Start experience, then they may provide longer-term advantages to developmental progress into kindergarten and beyond.

Research is mixed regarding which of these quality features influence which developmental outcomes, and for whom. Prior research establishes strong theoretical and empirical support that physical resources and social interactions are most consistently associated with children's development. In contrast, teacher proficiency and peer competence appear only indirectly related (Mashburn et al., 2008). Based on this past research, we expect that the quality of the learning environments that children in Head Start experience may affect the extent of program impacts, both at the end of their initial exposure to Head Start and as they continue into school.

The Head Start Impact Study

The Head Start Impact Study, congressionally-mandated, used a nationally representative sample of Head Start programs and newly entering 3- and 4-year-old children, and randomized children either to a Head Start group that had access to Head Start services in the initial year or to a control group that could receive any other non-Head Start services available in the community, chosen by their parents (e.g., Puma et al., 2005). About 60 percent of control group parents enrolled their children in some other type of preschool program in the first year. In addition, all children in the 3-year-old cohort could receive Head Start services in the second year. Under this randomized design, a simple comparison of outcomes for the two groups—treatment and control—yields an unbiased estimate of the impact of access to Head Start in the initial year on children's psychological development and school readiness (Puma et al., 2005).

This research design ensures that the two groups did not differ in any systematic or unmeasured way except through their access to Head Start services. It is important to note that, because the control group in the 3-year-old cohort was given access to Head Start in the second year, the findings for this age group reflect the added benefit of providing access to Head Start at age three, not the total benefit of having access to Head Start for two years. The study was designed to examine separately two cohorts of children, newly-entering 3- and newly-entering 4-year-olds. This design reflects the hypothesis that different program impacts may be associated with different ages of entry into Head Start.

In addition to random assignment, the HSIS is set apart from most program evaluations because it includes a nationally representative sample of programs and program participants, making its research findings generalizable to the national Head Start program as a whole as it existed in 2002-2003, not just to the studied sample of local programs and children. However, the study does not represent Head Start programs serving special populations such as tribal Head Start programs, programs serving migrant and seasonal farm workers and their families, or Early Head Start. Further, the study does not represent the 15 percent of Head Start programs in which the oversubscription for the available Head Start "slots" was too small to allow for an adequate-sized control group. The study sample, spread over 23 different states, consisted of a total of 84 randomly-selected local Head Start grantees/delegate agencies, 383 randomly-selected Head Start centers, and a total of 4,667 newly-entering children, including 2,559 3-year-olds and 2,108 4-year-olds.

At each of the included Head Start centers, program staff provided information about the study to parents at the time enrollment applications were distributed. Parents were told that enrollment procedures would be different for the 2002-03 Head Start year and that some decisions regarding enrollment would be made using a lottery-like process. Local agency staff implemented their typical process of reviewing enrollment applications and screening children for admission to Head Start based on criteria approved by their respective Policy Councils. No changes were made to these locally established ranking criteria for prioritizing which families to serve among a greater number of applicants than available, funded program slots.

The study collected information on all children determined to be eligible for enrollment in Fall 2002, and an average sample of 27 children per included center was selected from this pool: 16 who were assigned to the Head Start group and 11 who were assigned to the control group (in centers where fewer children than expected were actually available, a smaller sample of children was selected). The randomized children formed two study samples—newly-entering 3-year-olds (to be studied through two years of Head Start participation and beyond) and newly-entering 4-year-olds (to be studied through one year of Head Start participation and beyond).

Section 3: Data & Measures

This section identifies the source of our data and details the variables we use to measure Head Start quality and selected key child outcomes.

Data Source

As noted above, the HSIS collected data from 383 randomly selected Head Start centers within 84 randomly selected Head Start grantee agencies, across 23 states. The respondents included parents, children, teachers, and other care providers. The resulting data set contains records for 4,667 newly entering children, which include 2,559 3-year-olds and 2,108 4-year-olds. The data includes follow-up records from the Head Start years (one year for 4-year-olds and two years for 3-year-olds) as well as kindergarten and first and third grade years.

The data set includes a rich set of baseline variables on the study's enrolled children, their families and the Head Start centers in which they enrolled as well as details on alternative care arrangements they might have had. The follow-up variables are similarly rich, including many measures of children's development in several domains, and parenting and family experiences. We discuss the specific variables that are relevant to our analysis next.

Quality Measurement

We posit that three main dimensions of a child's Head Start experience exist, one "structural" and two "process-related." The structural measure of quality considers the "resources" which are the physical characteristics of the setting. The process-related measures of quality consider the interactions between teacher and child and exposure to academic activities in the classroom. As informed by the study's expert panel on quality, we contend that these measures capture different dimensions of quality such that we are justified in using each of them, independently, to analyze something about that specific quality experience. We discuss the specific operationalization of each of these quality measures next.

Resources

We use a measure of resources that represents a facility's physical structure and its contents, using 17 of the items in the Early Childhood Environment Rating Scale (ECERS), those that form the subscale on materials. The measure includes 17 specific variables, each of which we coded to range from 1 to 7. Specific elements include characterizations about the indoor space and furnishings, space for both gross and fine motor play, private space, child-related display, and the availability of items relating to art, dramatization, nature/science and math/numbers. As an average of the 17 items, the resulting measure is also on a 1-to-7 scale, where we flagged those with an average score of 5 or greater as having "high" quality (recoded as 2) by this measure and those with a lower average score as having "low" quality (recoded as 1). Those treatment group members that were no-shows have a resulting score of zero.

Interactions

Our next quality measure aims to capture the quality of teacher-child interactions. It is an index computed from 31 variables, eight drawn from the ECERS and 23 from the Arnett Caregiver Interaction Scale. The eight ECERS elements include the following: encouraging children to communicate, developing reasoning skills, and staff-child interactions, for example. Each of these could range from 1 to 7 in value. The Arnett elements included the following characteristics of staff interactions: kneeling/bending to child's level, assisting children in making choices, exercising control over children, encouraging new experiences, being attentive when children speak, encouraging prosocial behavior, explaining reasons for child misbehavior, placing value on obedience, and speaking warmly to the children. Although the original values of these fell on a 1-to-4 scale, we recoded them so that they would have a 1-to-7 range and be comparable for averaging with the ECERS items. We defined "high" quality as an average score of 6 or higher and "low" quality as an average score below 6. As with the other two quality measures, zero is the classification for no-shows (and does not represent anything about quality experience).

Exposure

The process-related measure of exposure, as we have defined it, considers the frequency of academically-focused activities that children experience in the classroom. The measure contains 19 teacher-reported variables including the following: showing how to read a book, having child(ren) tell a story, discussing new words, learning names of letters, practicing letters' sounds, writing letters and one's own name, discussing calendar/days of week, counting, playing math games, working with rulers and measuring cups, for example. As with the resources quality measure, each of the items within this scale can range from 1 to 7, and our aggregate measure is an average of those. Those average scores of 6 or greater are identified as having "high" quality (recoded as 2) by this measure, and those with a lower average score are identified as having "low" quality (recoded as 1). Those treatment group members that were no-shows have a resulting score of zero. Some disagreement exists within the field regarding whether greater exposure to academic activities is age-appropriate and beneficial; nevertheless, we refer to those with higher scores as experiencing higher quality on the exposure measure.

Exhibit 3-1 summarizes these three quality measures separately for the 3- and 4-year old cohorts in the treatment group. As earlier noted, it shows that about 17 percent of the 3-year-old cohort and 23 percent of the 4-year-old cohort never participated in Head Start. This means that, despite having been randomized at the time of their Spring 2002 application to attend Head Start, by Spring 2003 those children had not attended, for even one day. Among those who did attend Head Start, 64 percent of the 3-year-old cohort and 73 percent of the 4-year-old cohort experienced high resource quality. Similarly, 72 percent and 79 percent of the two cohorts, respectively, experienced high quality interactions. A smaller proportion of each cohort—27 percent and 25 percent, respectively—experienced high quality exposure. As noted earlier, the field is conflicted on whether more exposure to academic activities is expected to be good for children's development.

Exhibit 3-1. Descriptive Statistics of Three Head Start Quality Measures among Treatment Group Members who Participated in Head Start, by Age Cohort

	3-Year Old Cohort		4-Year Old Cohort	
	Number	Percent	Number	Percent
Head Start Treatment Group				
Never participated in HS	243	16.6%	276	23.2%
Participated in HS	1,223	83.4%	915	76.8%
Among those who participated in HS…				
Resources (range = 1-7)				
High quality (5+)	684	64.2%	567	72.6%
Lower quality (<5)	382	35.8%	214	27.4%
Interactions (range = 1-7)				
High quality (6+)	764	71.7%	617	79.0%
Lower quality (<6)	302	28.3%	164	21.0%
Exposure (range = 1-7)				
High quality (6+)	278	27.4%	188	24.7%
Lower quality (<6)	735	72.6%	574	75.3%

Notes: Details of the elements comprising each measure appear in the narrative.

We chose the cutoffs we did—of 5 for resources and 6 for interactions and exposure—for the following reasons. For the resources measure, all items come from the ECERS, and it is common practice to use 5 as the threshold above which "high" quality is designated. The other two measures do not have a common convention. The interaction measure draws from a combination of Arnett and ECERS items, scaling them comparably and summing them. The choice of 6 out of 7 as the threshold for what designates "high" quality seems appropriate because of the distribution of resulting values on this measure: as Exhibit 3-1 shows, about 72 percent of the 3-year-old cohort in the treatment group and 79 percent of the 4-year-old cohort in the treatment group had high interactions quality. Since these percentages are already quite high, if we would have lowered the threshold to 5 points on the 7-point scale we would have less high-low variation to examine. As for exposure quality, this measure draws from teacher reports; like the interaction measure, it does not have a field-accepted designated threshold for what one might consider to be "high" quality. Furthermore, as noted above, whether more "exposure" to academic activities is helpful remains in debate. As a result, we chose the cut-point of 6 as our threshold in order to create a relatively high bar for designating "high" quality on this measure.

Outcome Measures

While the Head Start Impact Study explores many outcomes, for this analysis we examine five specific outcomes across two broad domains—cognitive and social-emotional. In the domain of cognitive outcomes, we include the PPVT, and the Woodcock-Johnson Letter Word Identification and Applied Problems variables. As key outcomes representing children's social-emotional outcomes, we include a measure of Social Skills and Positive Approaches to Learning (which we refer to as "social competence") and Total Child Behavior Problems. We choose these specific measures in part because of our interest in children's development across the domains listed but also because they are consistently measured across all points of HSIS follow-up. Therefore we are able to identify the extent to which Head Start quality has differential impacts not just by the end of the Head Start year but also over time. The current Head Start quality analysis considers the first point of follow-up, which is when we know Head Start's overall influence to be strongest and we therefore have the greatest chance of detecting different, stronger impacts from high quality Head Start. We also analyze the role of Head Start quality on children's outcomes in the following years of follow-up, which extends through the end of their third grade year. Each of the selected outcome variables is detailed below.

Cognitive Domain

Within the cognitive domain, "vocabulary knowledge" is a skill that represents children's oral language development, "pre-reading skills" focus on letter recognition, an important step toward reading proficiency, and "early math skills" include basic numeracy and math skills that are the foundation for more advanced quantitative development. These central cognitive outcomes are measured by the Peabody Picture Vocabulary Test (PPVT-III, adapted), and the Woodcock-Johnson III (WJ3) Letter-Word Identification and Applied Problems subsets, respectively. All three of these come from direct assessments of children and are available in each year of follow-up.

Social-Emotional Domain

In the social-emotional domain, we consider to variables: the extent to which children engage in an overall Social Skills and Positive Approaches to Learning measure (which we shorten as "social competence") as collected from interviews with parents at each of the follow-up points; and "total" child behaviors that are (1) aggressive or defiant, (2) inattentive or hyperactive, and (3) shy, withdrawn, or depressed. Each of these is described next.

Social Skills and Positive Approaches to Learning. Although many measures might represent the social-emotional domain of children's outcomes, we selected this measure, a composite of several elements as follows. Social skills focus on cooperative and empathic behavior, such as, "makes friends easily," "comforts or helps others," and "accepts friends' ideas in sharing and playing." Approaches to learning deal with curiosity, imagination, openness to new tasks and challenges, and having a positive attitude about gaining new knowledge and skills. Examples include, "enjoys learning," "likes to try new things," and "shows imagination in work and play." The seven items that comprise this scale came from parents' judgments whether the behavioral description was "not true," "sometimes true," or "very true" of the child. The scale's resulting scores can range from zero (meaning all the items were rated "not true" of the child) to 14 (meaning all the items were rated "very true" of the child).

Total Child Behavior Problems. Elements in the three subscales of this measure combine together to form the Total Child Behavior Problems scale that we use. Parents were asked to rate their children on items dealing with specific behaviors, and they did so on a three-point scale of "not true," "sometimes true," or "very true." Example items include the extent to which the child "hits and fights with others," "can't concentrate, can't pay attention" and "is unhappy, sad, or depressed." The 14 items in the scale result in the possible score ranging from zero (all items marked "not true") to 28 (all items marked "very true").

While the HSIS overall considers health and parenting domains as well, we focus this analysis of the role of Head Start quality specifically on this subset of outcomes in the cognitive and socio-emotional domains because prior theory and evidence indicate these are most proximally related to the quality of care and education.

Section 4: Analytic Methods

Comparing those children in high quality Head Start with those in low quality Head Start or with no Head Start exposure—all within the study's treatment group only—would result in impact estimates biased by selection. These samples would differ at baseline on unmeasured characteristics that lead to different outcomes independently of the effects of different quality Head Start experiences. So too would comparison of treatment group children exposed to a specific level of Head Start quality to the full control group, which is comprised of children who—had they been randomly assigned to the treatment group—would have, in distinct subgroups, not participated in Head Start, participated in low quality Head Start, or participated in high quality Head Start. Differences between these subgroups and the entire control group would reflect compositional distinctions, not simply the impact of Head Start.[1] To avoid these problems and capitalize on the experimental design of the HSIS, we use an approach established in Peck (2003) to create equivalent subgroups of treatment and control group members for separate analysis and that therefore results in internally valid (i.e., unbiased) estimates of Head Start's impact on that subgroup.[2] Because some misclassification of children is inevitable—e.g., predicting a child who actually receives low quality Head Start as likely to receive high quality Head Start—we convert results for predicted quality subgroups to results for actual quality subgroups under certain assumptions. This translates (subject to the validity of the assumptions) the internally valid impact estimates for predicted quality subgroups into externally valid—and more policy relevant—impact estimates for actual quality subgroups.

Description of Analytic Procedure

The technique we use identifies in identical fashion sample members from the treatment and control groups who are predicted to participate in high quality Head Start programs, then estimates impacts on that subpopulation as one would in any experimental subgroup analysis. The symmetry of the identification procedure ensures that equivalent subgroups are compared and guarantees that the impact estimates are free from differential selection bias or other sources of internal bias. Thus, the symmetric selection of treatment and control subgroup members within the experimental data ensures unbiasedness of the impact estimates generated for the subgroups examined. However, the subgroup for which the methodology produces unbiased impact estimates—children with the highest *predicted* probabilities of being in high quality Head Start programs, for example—is not necessarily the subgroup of policy interest—children who *actually* experience high quality Head Start. The predictive model, while symmetric for both treatment and control groups, is imperfect for both groups, potentially reducing the relevance (i.e., the external validity or generalizability) of the findings. This is why we develop and apply procedures to convert results so that they represent impacts on *actual* rather than *predicted* subgroup members, subject to certain assumptions.

---

1 To check this proposition we conducted the described analysis, comparing low quality Head Start participants in the treatment group to the entire control group and then high quality Head Start participants in the treatment group to the entire control group. The results bore little resemblance to the main findings of our analysis described here, which are not subject to selection biases of this type.

2 Further discussion of this appears as a Method Note in Three Parts in Peck (2013), Bell and Peck (2013), and Harvill, Peck and Bell (2013).

The following steps—explained and justified next—are involved in carrying out this research approach:

1. Select random subsamples of the treatment group from which to predict the level of Head Start quality.
2. Using baseline characteristics, predict quality.
3. Use the resulting predicted quality variable to identify subgroups symmetrically in the treatment and control groups.
4. Analyze the impact of predicted quality by comparing mean outcomes between the symmetric treatment and control group subgroups created.
5. Convert results for predicted subgroups to represent impacts on actual subgroups under certain assumptions.

We also explore an alternative set of assumptions at Step 5 to examine the robustness of the findings to different conversion assumptions.

Step 1. Select random subsamples of the treatment group to predict Head Start Quality.

A key feature of this approach to subgroup analysis is *retaining the strength of the experimental design.* In order to do this, an important first step is to select a strategy for ensuring symmetric identification of subgroups. While prior work has used a single external "modeling" subsample to do so, the approach we take here is to choose several modeling subsamples for use in out-of-sample prediction. Through this process subgroups with equivalent predicted probabilities of participating in Head Start at a particular level of quality are identified in both treatment and control groups.

Using the entire treatment group for subgroup prediction at once *and* for impact analysis could introduce bias because of the better fit that is inevitable for the sample that is used for modeling. This has been referred to elsewhere as "overfitting bias" and can be avoided. To clarify, if the whole treatment group were used for prediction, then the model might more accurately identify the desired subgroup for treatment group cases than for predicted control group cases. This is because the prediction model would mold its parameters to the errors that exist in the outcome data due to random baseline variation between the groups. This would result in some unknown amount and direction of bias that is easily avoidable by keeping separate the predictive and impact estimation subsamples of the treatment group.[3] In this application, we select ten random 90-percent subsets of the treatment group from the combined 3-year-old and 4-year-old cohorts for predictive modeling,[4] as elaborated below.

Step 2. Using baseline characteristics, predict quality.

In this application, we create three distinct quality indicators for all members of the 3-year-old and 4-year-old treatment group cohorts, each with three levels: a value of 0 represents those who never participated in Head Start; a value of 1 represents "low quality" Head Start, among those who participated in the program; and a value of 2 represents "high quality" Head Start, also among those who participated in the program. The specific threshold for dividing high quality from low quality is measure-specific, as

---

3 Some have argued that the loss of sample size associated with choosing an external, modeling sample imposes too great a cost (e.g., Gibson, 2003); but the problem of potential overfitting bias diminishes as sample size increases, making the step of selecting a random subsample for modeling even more important in smaller samples (Harvill, Peck & Bell, 2013).

4 Initial examination of the predictions by cohort showed that the prediction rate was better for the pooled-cohort-prediction, which justifies our choice to pool.

defined in our measurement sub-section above. With this categorical quality measure as our dependent variable, we used a generalized logit procedure to predict no-show and quality status, with explanatory variables including center, family, and child characteristics as follows:

- Center Characteristics: center of random assignment (series of dummy variables, omitting the dummy for one center)
- Family Characteristics: home language, both bio-parents at home, primary caregiver's age, mother's education, bio-mother's recent immigrant status, mother's marital status, mother gave birth to study child as a teen
- Child Characteristics: sex, age, race, language

We expected that the center of random assignment would be the best predictor of the quality of Head Start; we further allow this to proxy other community characteristics that might be associated with higher quality.[5] Other family- and child-level characteristics might also be associated with the quality of Head Start that a child experiences. Rather than basing our decision for which predictor variables to include on arbitrary or theoretical factors, we follow the lead of propensity score methods (to which our treatment group predictive modeling procedure is closely akin) which advocate a "kitchen sink" approach for generating the greatest explanatory power and best correct prediction rate possible. We are uninterested in interpreting any of the coefficients on our explanatory variables from the prediction model but instead have as our goal the best "hit rate": correctly matching those predicted to be in each of our three subgroups with their actual subgroup experience.

With each of the ten 90-percent subsamples drawn in Step 1, we predict the quality experience of the remaining 10 percent of the sample, both within the treatment and the control group. This involves "out of sample" prediction for the entire sample, eliminating concerns about overfitting and ensuring symmetric prediction of the quality-related subgroups within treatment and control arms. Once we have replicated this process for the entire sample, we concatenate the subsamples together to maintain full use of the entire sample for analysis.

Step 3. Use resulting predicted quality variable to identify subgroups.

Within the sample, each individual is designated to a subgroup (no-shows, low quality and high quality) based on which category (0, 1 or 2) he or she has the highest probability of belonging to, given baseline characteristics.

Step 4. Analyze the impact of quality by comparing the treatment and control groups' mean outcomes, by subgroup.

Although this kind of analysis can involve a conventional split-sample subgroup analysis, we follow the HSIS's existing practice of pooling data and computing subgroups' impact estimates accordingly (see Puma et al., 2010b, for details).

---

[5] To gauge the extent to which our assertion that "the center of random assignment would be the best predictor" of Head Start quality we examined the correct prediction rates based on including only the center dummies and on adding the family and child characteristics to the center dummies. Our conclusion from this side analysis is that indeed the center dummies are the best predictors of quality. In fact, the family and child characteristics alone predict quality very poorly. The main reason to include the family and child characteristics in the model is not to distinguish further between levels of Head Start quality but instead to better identify those individuals who classify as no-shows.

Step 5. Convert impacts for predicted quality subgroups to impacts on actual quality subgroups.

This final step converts the impact estimates from Step 4, which represent impacts on predicted subgroups, to represent impacts on actual subgroups, under certain assumptions. Here we discuss our preferred assumptions, and the Appendix elaborates on two alternative sets of assumptions and their implications.

To design the conversion process, we begin with three equations that posit that the impact on each of the three predicted subgroups (non-participants—called "no-shows" from here on—low quality participants, and high quality participants, respectively) is a weighted sum of the impacts on actual subgroups, where the weights are the proportion of each subgroup that are correctly classified into that group.

$$I_N = s_N N_N + w_N L_N + g_N H_N$$

$$I_L = s_L N_L + w_L L_L + g_L H_L$$

$$I_H = s_H N_H + w_H L_H + g_H H_H$$

where the following notation applies:

I_N is the impact on predicted no-shows

I_L is the impact on predicted low quality participants

I_H is the impact on predicted high quality participants

N_N is the impact on predicted no-shows who are actual no-shows

N_L is the impact on predicted low quality participants who are actual no-shows

N_H is the impact on predicted high quality participants who are actual no-shows

L_N is the impact on predicted no-shows who are actual low quality participants

L_L is the impact on predicted low quality participants who are actual low quality participants

L_H is the impact on predicted high quality participants who are actual low quality participants

H_N is the impact on predicted no-shows who are actual high quality participants

H_L is the impact on predicted low quality participants who are actual high quality participants

H_H is the impact on predicted high quality participants who are actual high quality participants

s_N is the proportion of predicted no-shows who are actually no-<u>s</u>hows

s_L is the proportion of predicted low quality participants who are actually in the no-<u>s</u>how subgroup

s_H is the proportion of predicted high quality participants who are actually in the no-<u>s</u>how subgroup

w_N is the proportion of predicted no-shows who are actually in the lo<u>*w*</u> quality subgroup

w_L is the proportion of predicted low quality participants who are actually in the lo<u>*w*</u> quality subgroup

w_H is the proportion of predicted high quality participants who are actually in the lo<u>*w*</u> quality subgroup

g_N is the proportion of predicted no-shows who are actually in the high quality subgroup

g_L is the proportion of predicted low quality participants who are actually in the high quality subgroup

g_H is the proportion of predicted high quality participants who are actually in the high quality subgroup

This set of three equations contains nine unknowns, and so some (six) assumptions are necessary in order to solve the system. In this application, we make the following six assumptions:

(1) $N_N = 0$ – the impact on predicted no-shows who are actual no-shows is zero

(2) $N_L = 0$ – the impact on predicted low quality participants who are actual no-shows is zero

(3) $N_H = 0$ – the impact on predicted high quality participants who are actual no-shows is zero

(4) $L_H = L_L$ – the impacts on low quality participants are the same for children predicted to be high quality participants and children predicted to be low quality participants

(5) $H_H = H_L$ – the impacts on high quality participants are the same for children predicted to be high quality participants and children predicted to be low quality participants

(6) $H_N - L_N = H_L - L_L$ – the impact on high quality participants differs from impact on low quality participants by the same amount whether one looks at high and low quality cases predicted to be no-shows or high and low quality cases predicted to be low quality participants[6]

Ultimately, we must rearrange these equations, imposing our assumptions, to express the terms of interest—impacts on the actual subgroups—as a function of the elements that are known, the impacts on predicted subgroups and the relative proportions of those predicted to be in each group who are actually in each group. The resulting conversions are as follows:

$$L = \left(\frac{1-r}{w_N + g_N}\right) I_N - \left[\frac{(1-r)g_N(w_H + g_H) + rg_H(w_N + g_N)}{(w_H g_L - w_L g_H)(w_N + g_N)}\right] I_L + \left[\frac{(1-r)g_N(w_L + g_L) + rg_L(w_N + g_N)}{(w_H g_L - w_L g_H)(w_N + g_N)}\right] I_H$$

$$H = \left(\frac{1-p}{w_N + g_N}\right) I_N - \left[\frac{(1-p)w_N(w_H + g_H) + pw_H(w_N + g_N)}{(w_H g_L - w_L g_H)(w_N + g_N)}\right] I_L + \left[\frac{(1-p)w_N(w_L + g_L) + pw_L(w_N + g_N)}{(w_H g_L - w_L g_H)(w_N + g_N)}\right] I_H$$

where

1-r is the proportion of low quality participants who are predicted as no-shows; and

1-p is the proportion of high quality participants who are predicted as no shows.

The impact on the full *actual no-show* subgroup is a linear combination of N_N, N_L and N_H, all assumed to be zero, making the overall impact on the full no-show sample 0, consistent with the conventional Bloom

6 Or whether one looks at high and low quality cases predicted to be high quality participants, once one combines this final assumption with the previous two assumptions to derive $H_N - L_N = H_H - L_H$.

assumption (Bloom, 1984; Puma et al., 2005). The assumptions discussed here are our preferred assumptions and the ones that we use for our analysis. We discuss two alternative sets of assumptions in the Appendix and present results from those analyses.

Section 5: Findings

This section reports the estimated impacts of low and high quality Head Start services on the children who actually received those services, subject to the assumptions described earlier. By assumption, for children in the experimental treatment group who never participated in Head Start no impacts occurred: *N* = 0, as shown in Exhibits 5-1 through 5-5 for the 3-year-old cohort and in Exhibits 5-6 through 5-10 for the 4-year-old cohort. When the quality of Head Start services is divided between low and high for each of the quality dimensions examined, results get more interesting, as reported in the High, Low, and High-Low Difference rows of the exhibits. Like previous Head Start Impact Study reports involving subgroups, we confine discussion to measured impacts that we are confident (1) differ from zero *and* differ from impacts on a contrasting subgroup in the same division of the population, or (2) differ from zero in a consistent pattern across multiple years. We do not formally adjust for the increased potential for false positives that arises from conducting many hypothesis tests in exploratory research, but instead make an informal adjustment in our interpretation of results.

As Exhibit 5-1 shows, there is no evidence of statistically significant differences in impact on PPVT score between high and low quality Head Start services for the 3-year-old cohort. A favorable impact of high resource quality exists in both the first and second years of Head Start for the 3-year-old cohort, though this was not significantly different from the impact for low resource quality in these years. The effect size that corresponds to these absolute impacts in PPVT test score units ranges from 0.16 to 0.31 (impact divided by the standard deviation of the control group mean).

Exhibit 5-1. Estimated Impacts on PPVT Scores for Actual Subgroups, by Quality Measure, by Follow-up Year, 3-Year-Old Cohort, Preferred Assumptions

	End of HS Year 1 2003	End of HS Year 2 2004	End of Kindergarten 2005	End of First Grade 2006	End of Third Grade 2008
Control Group Average	251.4	298.3	339.9	357.9	405.7
(standard deviation)	(34.3)	(36.6)	(28.4)	(30.1)	(29.5)
Resource					
No-shows	0.0 [a]	0.0	0.0	0.0	0.0
High	10.7 ***	5.8 *	1.0	2.1	4.1
Low	4.9	-4.4	-0.8	2.7	-0.9
High-Low Difference	5.8	10.2	1.8	-0.7	5.0
Interaction					
No-shows	0.0	0.0	0.0	0.0	0.0
High	7.6 **	4.8	0.7	4.5	3.5
Low	11.2	-4.3	-0.5	-4.2	-0.7
High-Low Difference	-3.5	9.0	1.2	8.7	4.2
Exposure					
No-shows	0.0	0.0	0.0	0.0	0.0
High	12.5	2.4	-3.9	5.0	-5.7
Low	7.1 *	2.6	1.5	1.3	5.2
High-Low Difference	5.4	-0.1	-5.4	3.6	-10.9

Notes: The impact is the regression-adjusted difference (impact) between the treatment and control groups in the number of points on each outcome measure. Some high-low differences appear not to sum because of rounding.
[a] No statistical significance noted because no-show impact estimates are derived by assumption to be zero.
*** statistically significant: p<0.01; ** statistically significant: p<0.05; * statistically significant: p<0.10

A similar set of findings appears in examining the WJ3 Letter-Word scores for the three-year-old cohort (Exhibit 5-2), including favorable impacts of high interaction quality and low exposure quality over the two pre-school years, the latter continuing into kindergarten. In two cases, a statistically significant *difference* in effectiveness between quality levels exists, favoring high interaction quality in the second Head Start year and low exposure quality in kindergarten. The latter finding suggests that greater exposure to academic activities, as reported by teachers, disadvantages children in the 3-year-old cohort, at least during their prekindergarten and kindergarten years. In contrast, the children with exposure to more academic activities in Head Start have largely unfavorable impacts on WJ3 Letter Word scores at the end of kindergarten. Effect sizes for the favorable impacts referenced here range from 0.21 to 0.35.

Exhibit 5-2. Estimated Impacts on WJ3 Letter-Word Scores for Actual Subgroups, by Quality Measure, by Follow-up Year, 3-Year-Old Cohort, Preferred Assumptions

	End of HS Year 1 2003	End of HS Year 2 2004	End of Kindergarten 2005	End of First Grade 2006	End of Third Grade 2008
Control Group Average	307.6	330.1	383.4	432.9	482.8
(standard deviation)	(27.4)	(27.6)	(31.6)	(35.3)	(29.8)
Resource					
No-shows	0.0 [a]	0.0	0.0	0.0	0.0
High	10.3 ***	5.2	1.0	1.5	3.9
Low	5.0	-0.9	-2.7	-1.8	-5.2
High-Low Difference	5.3	6.1	3.7	3.3	9.1
Interaction					
No-shows	0.0	0.0	0.0	0.0	0.0
High	9.7 **	7.5 **	2.2	3.8	3.7
Low	5.0	-8.8	-7.2	-8.6	-6.9
High-Low Difference	4.7	16.2 *	9.4	12.4	10.5
Exposure					
No-shows	0.0	0.0	0.0	0.0	0.0
High	7.2	-3.7	-24.9 **	-8.7	-10.5
Low	9.1 **	5.7 *	8.1 **	3.3	4.8
High-Low Difference	-1.9	-9.5	-33.0 *	-12.0	-15.3

Notes: The impact is the regression-adjusted difference (impact) between the treatment and control groups in the number of points on each outcome measure. Some high-low differences appear not to sum because of rounding.
[a] No statistical significance noted because no-show impact estimates are derived by assumption to be zero.
*** statistically significant: $p<0.01$; ** statistically significant: $p<0.05$; * statistically significant: $p<0.10$

Exhibit 5-3 shows no noteworthy impacts on WJ3 Applied Problems by quality level for the 3-year-old cohort. One significant difference in impacts arises for social competence, however (Exhibit 5-4). As with WJ3 Letter-Word scores, children exposed to relatively fewer academic activities as reported by teachers (i.e., lower exposure quality) appear to benefit more from Head Start participation than other children in terms of their social competence in kindergarten.

Exhibit 5-3. Estimated Impacts on WJ3 Applied Problems Scores for Actual Subgroups, by Quality Measure, by Follow-up Year, 3-Year-Old Cohort, Preferred Assumptions

	End of HS Year 1 2003	End of HS Year 2 2004	End of Kindergarten 2005	End of First Grade 2006	End of Third Grade 2008
Control Group Average	373.6	399.9	431.3	453.7	486.5
(standard deviation)	(27.4)	(21.8)	(21.2)	(20.6)	(22.8)
Resource					
No-shows	0.0 [a]	0.0	0.0	0.0	0.0
High	9.8 **	3.9	0.8	2.0	1.8
Low	-3.8	-3.2	-5.8	1.6	-2.9
High-Low Difference	13.6	7.1	6.5	0.4	4.7
Interaction					
No-shows	0.0	0.0	0.0	0.0	0.0
High	5.4 *	5.1	1.1	2.5	3.3
Low	4.8	-8.3	-7.9	0.2	-7.8
High-Low Difference	0.6	13.4	8.9	2.3	11.1
Exposure					
No-shows	0.0	0.0	0.0	0.0	0.0
High	2.5	6.6	-5.6	-0.5	-1.2
Low	6.2	-0.4	-0.2	2.5	0.9
High-Low Difference	-3.7	7.0	-5.4	-3.0	-2.0

Notes: The impact is the regression-adjusted difference (impact) between the treatment and control groups in the number of points on each outcome measure. Some high-low differences appear not to sum because of rounding.
[a] No statistical significance noted because no-show impact estimates are derived by assumption to be zero.
*** statistically significant: p<0.01; ** statistically significant: p<0.05; * statistically significant: p<0.10

Exhibit 5-4. Estimated Impacts on Social Competence for Actual Subgroups, by Quality Measure, by Follow-up Year, 3-Year-Old Cohort, Preferred Assumptions

	End of HS Year 1 2003	End of HS Year 2 2004	End of Kindergarten 2005	End of First Grade 2006	End of Third Grade 2008
Control Group Average	12.4	12.5	12.3	12.5	12.0
(standard deviation)	(1.8)	(1.8)	(1.8)	(1.7)	(1.9)
Resource					
No-shows	0.0 [a]	0.0	0.0	0.0	0.0
High	-0.2	0.4 *	0.2	0.3	0.4 **
Low	0.4	-0.1	0.4	-0.4	0.0
High-Low Difference	-0.6	0.5	-0.3	0.7	0.4
Interaction					
No-shows	0.0	0.0	0.0	0.0	0.0
High	0.0	0.4	0.3	0.2	0.3
Low	-0.1	-0.4	0.0	-0.5	0.2
High-Low Difference	0.1	0.8	0.3	0.7	0.1
Exposure					
No-shows	0.0	0.0	0.0	0.0	0.0
High	0.0	0.5	-0.7	-0.9	0.7
Low	-0.1	0.1	0.6 **	0.3	0.2
High-Low Difference	0.1	0.4	-1.3 *	-1.2	0.5

Notes: The impact is the regression-adjusted difference (impact) between the treatment and control groups in the number of points on each outcome measure. Some high-low differences appear not to sum because of rounding.
[a] No statistical significance noted because no-show impact estimates are derived by assumption to be zero.
*** statistically significant: p<0.01; ** statistically significant: p<0.05; * statistically significant: p<0.10

Impacts on problem behaviors (Exhibit 5-5) with negative signs signal desired reductions in poor behavior. These occur for certain high and low quality subgroups over the first three years of follow-up, consistent with the social competence findings but stronger. High resource quality is associated with favorable impacts over the first three years of follow-up, and these impacts are statistically significantly stronger than impacts on those experiencing low resource quality in two of those years. The reverse occurs for exposure quality: there are favorable impacts of low quality Head Start programs in the first year of the study; and in kindergarten, and those impacts are statistically significantly different from the unfavorable impacts of high exposure equality found in those same years. Effect sizes for statistically significant findings by subgroup range from 0.17 to 0.56 standard deviation units.

Exhibit 5-5. Estimated Impacts on Problem Behaviors for Actual Subgroups, by Quality Measure, by Follow-up Year, 3-Year-Old Cohort, Preferred Assumptions

	End of HS Year 1 2003	End of HS Year 2 2004	End of Kindergarten 2005	End of First Grade 2006	End of Third Grade 2008
Control Group Average	6.2	5.6	5.1	5.0	5.8
(standard deviation)	(3.6)	(3.8)	(3.9)	(3.9)	(4.4)
Resource					
No-shows	0.0 [a]	0.0	0.0	0.0	0.0
High	-1.0 **	-1.0 **	-1.0 **	-0.5	-0.3
Low	0.4	0.7	1.0 *	0.6	0.8
High-Low Difference	-1.3 *	-1.7	-2.0 ***	-1.2	-1.1
Interaction					
No-shows	0.0	0.0	0.0	0.0	0.0
High	-0.6 *	-0.9	-0.7	0.1	0.2
Low	-0.4	1.0	0.6	-0.8	-0.1
High-Low Difference	-0.1	-2.0	-1.3	0.8	0.3
Exposure					
No-shows	0.0	0.0	0.0	0.0	0.0
High	1.2 *	0.8	2.2 *	1.2	0.3
Low	-1.1 ***	-0.9	-1.2 **	-0.6	0.0
High-Low Difference	2.4 **	1.7	3.4 **	1.8	0.3

Notes: The impact is the regression-adjusted difference (impact) between the treatment and control groups in the number of points on each outcome measure. Some high-low differences appear not to sum because of rounding.
[a] No statistical significance noted because no-show impact estimates are derived by assumption to be zero.
*** statistically significant: p<0.01; ** statistically significant: p<0.05; * statistically significant: p<0.10

Considering the 3-year-old results across all quality dimensions, outcomes, and follow-up years, statistically significant differences in impact between high and low quality subgroups occur no more frequently than would be expected by chance when no true differences exist.[7] The strongest evidence concerns the benefit of less exposure to academic activities: low exposure quality surpasses high exposure quality in generating favorable impacts four times, especially for behavioral development outcomes.

[7] Among the 75 hypothesis tests of differences by quality level conducted on the 3-year-old cohort (three quality types, five follow up years, five outcome measures), about seven or eight are expected to be statistically significant by chance at the 0.10 significance level. In fact, seven are statistically significant.

Next, Exhibits 5-6 to 5-10 report the results for the 4-year-old cohort. As Exhibit 5-6 shows, there are no statistically significant differences in impacts on PPVT scores for 4-year-olds between high and low quality subgroups. Even so, in the first year all three high quality subgroups experienced favorable PPVT impacts, with that result echoed two years later, at the end of first grade, for resource and interaction quality. Effect sizes for significant findings range from 0.17 to 0.40.

Exhibit 5-6. Estimated Impacts on PPVT Scores for Actual Subgroups, by Quality Measure, by Follow-up Year, 4-Year-Old Cohort, Preferred Assumptions

	End of HS Year 2003	End of Kindergarten 2004	End of First Grade 2005	End of Third Grade 2007
Control Group Average	290.3	331.9	363.1	405.7
(standard deviation)	(35.9)	(39.1)	(32.2)	(28.7)
Resource				
No-shows	0.0 [a]	0.0	0.0	0.0
High	6.2 *	1.4	7.7 ***	3.3
Low	8.5	11.1	0.1	2.2
High-Low Difference	-2.3	-9.6	7.6	1.1
Interaction				
No-shows	0.0	0.0	0.0	0.0
High	7.4 **	4.7	5.4 *	3.2
Low	3.8	2.0	4.1	1.9
High-Low Difference	3.5	2.7	1.3	1.4
Exposure				
No-shows	0.0	0.0	0.0	0.0
High	14.5 **	8.9	8.7	-2.5
Low	4.2	3.1	4.3	4.5
High-Low Difference	10.3	5.8	4.4	-7.0

Notes: The impact is the regression-adjusted difference (impact) between the treatment and control groups in the number of points on each outcome measure. Some high-low differences appear not to sum because of rounding.
[a] No statistical significance noted because no-show impact estimates are derived by assumption to be zero.
*** statistically significant: $p<0.01$; ** statistically significant: $p<0.05$; * statistically significant: $p<0.10$

Impacts on WJ3 Letter-Word (Exhibit 5-7) differ between high and low quality subgroups for resource and exposure quality in the Head Start year. Contrary to the 3-year-olds, for 4-year-olds more favorable impacts on Letter-Word scores occurred when children received low resource quality and high exposure to academic activities. A greater impact from high academic exposure also appears at the end of kindergarten. The findings here for individual subgroups are the most striking in magnitude among all results for both cohorts, ranging from 0.73 to 1.03 in effect size for statistically significant cases.

Almost none of the findings for WJ3 Applied Problems are noteworthy for the 4-year-old cohort (Exhibit 5-8). However, a statistically significant favorable effect for low resource quality occurs in kindergarten, an effect that is statistically significantly different from an *unfavorable* effect for high resource quality in that year. No meaningful impacts by quality level occur for the 4-year-old cohort on the two socio-emotional outcomes examined—social competence (Exhibit 5-9) and problem behaviors (Exhibit 5-10).

Exhibit 5-7. Estimated Impacts on WJ3 Letter-Word Scores for Actual Subgroups, by Quality Measure, by Follow-up Year, 4-Year-Old Cohort, Preferred Assumptions

	End of HS Year 2003	End of Kindergarten 2004	End of First Grade 2005	End of Third Grade 2007
Control Group Average	325.5	378.2	433.3	480.6
(standard deviation)	(28.5)	(31.6)	(36.5)	(28.7)
Resource				
No-shows	0.0 [a]	0.0	0.0	0.0
High	2.0	-4.3	0.1	2.1
Low	23.1 ***	13.2	6.0	4.7
High-Low Difference	-21.1 **	-17.5	-5.9	-2.6
Interaction				
No-shows	0.0	0.0	0.0	0.0
High	8.8 **	1.8	1.7	1.6
Low	7.0	-0.8	3.6	7.6
High-Low Difference	1.8	2.6	-1.9	-6.0
Exposure				
No-shows	0.0	0.0	0.0	0.0
High	29.4 ***	23.0 *	12.1	9.9
Low	1.6	-6.4	-1.2	0.6
High-Low Difference	27.8 ***	29.4 *	13.3	9.3

Notes: The impact is the regression-adjusted difference (impact) between the treatment and control groups in the number of points on each outcome measure. Some high-low differences appear not to sum because of rounding.
[a] No statistical significance noted because no-show impact estimates are derived by assumption to be zero.
*** statistically significant: p<0.01; ** statistically significant: p<0.05; * statistically significant: p<0.10

Exhibit 5-8. Estimated Impacts on WJ3 Applied Problems Scores for Actual Subgroups, by Quality Measure, by Follow-up Year, 4-Year-Old Cohort, Preferred Assumptions

	End of HS Year 2003	End of Kindergarten 2004	End of First Grade 2005	End of Third Grade 2007
Control Group Average	397.5	426.3	454.1	487.7
(standard deviation)	(24.0)	(21.9)	(19.8)	(19.4)
Resource				
No-shows	0.0 [a]	0.0	0.0	0.0
High	0.1	-5.4 *	0.1	-0.2
Low	13.3	13.1 *	2.9	-1.0
High-Low Difference	-13.1	-18.6 **	-2.7	0.8
Interaction				
No-shows	0.0	0.0	0.0	0.0
High	5.3	0.5	0.1	-1.9
Low	-0.3	-1.0	3.4	4.4
High-Low Difference	5.6	1.5	-3.3	-6.3
Exposure				
No-shows	0.0	0.0	0.0	0.0
High	-5.0	7.8	5.0	-1.4
Low	7.2	-2.3	-0.3	-0.1
High-Low Difference	-12.1	10.1	5.3	-1.3

Notes: The impact is the regression-adjusted difference (impact) between the treatment and control groups in the number of points on each outcome measure. Some high-low differences appear not to sum because of rounding.
[a] No statistical significance noted because no-show impact estimates are derived by assumption to be zero.
*** statistically significant: p<0.01; ** statistically significant: p<0.05; * statistically significant: p<0.10

Exhibit 5-9. Estimated Impacts on Social Competence for Actual Subgroups, by Quality Measure, by Follow-up Year, 4-Year-Old Cohort, Preferred Assumptions

	End of HS Year 2003	End of Kindergarten 2004	End of First Grade 2005	End of Third Grade 2007
Control Group Average	12.5	12.6	12.6	12.1
(standard deviation)	(1.8)	(1.5)	(1.6)	(1.9)
Resource				
No-shows	0.0 [a]	0.0	0.0	0.0
High	0.0	0.1	0.1	-0.4
Low	-0.3	0.0	-0.2	0.3
High-Low Difference	0.4	0.1	0.4	-0.7
Interaction				
No-shows	0.0	0.0	0.0	0.0
High	0.0	0.3	0.0	0.0
Low	-0.4	-0.4	0.0	-0.4
High-Low Difference	0.5	0.7	0.1	0.4
Exposure				
No-shows	0.0	0.0	0.0	0.0
High	-0.3	0.4	-0.6	-0.9
Low	0.0	0.0	0.2	0.1
High-Low Difference	-0.2	0.5	-0.8	-1.0

Notes: The impact is the regression-adjusted difference (impact) between the treatment and control groups in the number of points on each outcome measure. Some high-low differences appear not to sum because of rounding.
[a] No statistical significance noted because no-show impact estimates are derived by assumption to be zero.
*** statistically significant: p<0.01; ** statistically significant: p<0.05; * statistically significant: p<0.10

Exhibit 5-10. Estimated Impacts on Problem Behaviors for Actual Subgroups, by Quality Measure, by Follow-up Year, 4-Year-Old Cohort, Preferred Assumptions

	End of HS Year 2003	End of Kindergarten 2004	End of First Grade 2005	End of Third Grade 2007
Control Group Average	5.6	5.0	5.1	6.2
(standard deviation)	(3.8)	(3.3)	(3.8)	(4.2)
Resource				
No-shows	0.0 [a]	0.0	0.0	0.0
High	-0.6	0.4	0.1	0.1
Low	0.3	-0.4	-1.0	-2.2 *
High-Low Difference	-0.9	0.8	1.1	2.3
Interaction				
No-shows	0.0	0.0	0.0	0.0
High	-0.3	0.2	-0.5	-0.7
Low	-0.3	-0.2	0.4	-0.7
High-Low Difference	0.1	0.5	-0.9	0.1
Exposure				
No-shows	0.0	0.0	0.0	0.0
High	0.0	-0.7	-1.3	-1.8
Low	-0.4	0.4	0.1	-0.3
High-Low Difference	0.4	-1.1	-1.5	-1.5

Notes: The impact is the regression-adjusted difference (impact) between the treatment and control groups in the number of points on each outcome measure. Some high-low differences appear not to sum because of rounding.
[a] No statistical significance noted because no-show impact estimates are derived by assumption to be zero.
*** statistically significant: p<0.01; ** statistically significant: p<0.05; * statistically significant: p<0.10

Appraising the evidence for the 4-year-old cohort as a whole from Exhibit 5-6 through 5-10 combined, there is less evidence of impact differentials by quality level than for the 3-year-old cohort; indeed, fewer instances than expected by chance when no true differences exist.[8] Where the difference in impact is statistically significant, it is as likely to favor low quality in the area of resources and high quality in terms of greater exposure to academic activities. A developmentally-based explanation for these findings—including why they would differ for the 4-year-old cohort from the 3-year-old cohort—is unclear.

As elaborated in the Appendix, we estimated these impacts as of the end of the first year of follow-up using two sets of alternative assumptions. One of these alternative assumptions produces impact estimates that are substantially the same (identically so for the 4-year-old cohort) as those reported in the first columns of Exhibits 5-1 through 5-10, which are based on our preferred assumptions. The other alterative assumptions produce impact estimates that, for the 3-year-old cohort, strengthen somewhat the evidence that high resource and interaction quality Head Start can produce better short-run cognitive impacts than low quality Head Start, and that less exposure to academic activities can produce more favorable behavioral impacts than greater such exposure. However, these more favorable alternative results may be an artifact of including an assumption that impacts on actual high quality Head Start participants are twice as large as impacts on actual low quality Head Start participants (for those children predicted to be high quality participants).[9] Something of the same pattern is evident in the 4-year-old cohort—for potentially the same reason—but less noticeably.

8 With 60 tests of the hypothesis that the high and low subgroup impacts differ from each other, six are expected to be statistically significant by chance alone at the 0.10 significance level. We observe that four are.

9 Published standards in the scholarly literature indicate that where alternative assumptions are possible using the current technique and "the policy thrust of the findings varies across plausible scenarios…[the version of the findings to report]…should be based on the most plausible set of assumptions in the eyes of the researchers as declared prior to any analysis" (Bell & Peck, 2013).

Section 6: Conclusion

This report examines the influence of Head Start quality on children's selected developmental cognitive and social-emotional outcomes. Despite the importance to policy and practice of understanding the role of quality in influencing children's developmental progress the HSIS has not previously sought to address it primarily because of the analytic challenges involved in doing so. Among these challenges—each addressed here—are: (1) defining the construct and potentially numerous dimensions of "quality" conceptually, (2) making "quality" as defined *measurable* with validity and reliability from the study's data, and (3) determining impacts for varying levels of Head Start quality experience in the treatment group given that Head Start quality is undefined for the control group. We believe our methodology effectively addresses all of the challenges. That said, a point for future research is relevant: while we used expert panel input to determine the absolute threshold for designating a child's Head Start experience as "high" quality, in each of our three quality measures, other thresholds might be justified. Moreover, the measures[10] used in the HSIS's early 2000's were the best available at the time, but improvements in measuring quality have developed in the intervening decade, justifying alternative measurements of "quality" now than are possible with these data.

Applying these analytic innovations to the experimental HSIS evaluation data, we find little evidence that Head Start's impact varies systematically by the level of quality in the program for the available, limited quality measures. The frequency of statistically significant differences in impacts by quality levels is no greater than one would expect to observe by chance alone when no true differences exist. The one exception to this pattern is the discovery that, for 3-year-olds, lower exposure to academic activities is associated with more favorable short-run impacts on social development. There is almost no indication that either high or low quality Head Start in any dimension leads to Head Start impacts that last into third grade for either age cohort, consistent with the overall findings of the Head Start Impact Study not disaggregated by quality level.

[10] New measures, such as the CLASS (Classroom Assessment Scoring System) and ELLCO (Early Language and Literacy Classroom Observation), were not available at the time of the 2002 HSIS data collection.

Works Cited

Bell, Stephen H. & Laura R. Peck. (2013). "Using Symmetric Predication of Endogenous Subgroups for Causal Inferences about Program Effects under Robust Assumptions." *American Journal of Evaluation*, 24(3), 413-426. DOI: 10.1177/1098214013489338

Bloom, Howard S. (1984). Accounting for No-shows in Experimental Evaluation Designs. *Evaluation Review*, 8(2), 225–246. DOI:10.1177/0193841X8400800205

Downer, Jason & Andrew Mashburn. (2013). "Do School Experiences Play a Role in Hindering or Promoting the Persistence of Head Start Impacts on Cognitive and Social Outcomes during Elementary School?" Draft manuscript.

Gibson, Christina M. (2003). "Privileging the Participant: The Importance of Subgroup Analysis in Social Welfare Evaluations." *American Journal of Evaluation*, 24(4), 443-469. DOI: 10.1177/109821400302400403

Harvill, Eleanor, Laura R. Peck & Stephen H. Bell. (2013). "On Overfitting in Experimental Analysis Symmetrically Predicted Endogenous Subgroups from Randomized Experimental Samples: Part Three of a Method Note in Three Parts." *American Journal of Evaluation*, 34(4), 545-566. DOI: 10.1177/1098214013503201

Harvill, Eleanor & Laura R. Peck. (in progress). "Examining Prediction Quality Implications to Enhance the Social Impact Policy Pathfinder (SPI-Path)." Bethesda, MD: Abt Associates Inc. Unpublished manuscript.

Kemple, James J., & Jason C. Snipes. (2000). Career Academies: Impacts on Students' Engagement and Performance in High School. New York, NY: Manpower Demonstration Research Corporation.

Mashburn, Andrew J., Robert C. Pianta, Bridget K. Hamre, Jason T. Downer, Oscar A. Barbarin, Donna Bryant, Margaret Burchinal, Diane M. Early, & Carollee Howes. (2008). "Measures of Classroom Quality in Prekindergarten and Children's Development of Academic, Language, and Social Skills." *Child Development*, 79(3), 732-749.

Peck, Laura R. (2003). "Subgroup Analysis in Social Experiments: Measuring Program Impacts Based on Post-Treatment Choice." *American Journal of Evaluation*, 24(2), 157-187. DOI: 10.1016/S1098-2140(03)00031-6

Peck, Laura R. (2013). "On Analysis of Symmetrically-Predicted Endogenous Subgroups: Part One of a Method Note in Three Parts." *American Journal of Evaluation*, 34(2), 225-236. DOI: 10.1177/1098214013481666

Puma, Michael, Stephen Bell, Ronna Cook, Camilla Heid & Michael Lopez, et al. (2005). Head Start Impact Study: First Year Findings. Washington, DC: Department of Health and Human Services, Administration for Children and Families.

Puma, Michael, Stephen Bell, Ronna Cook, Camilla Heid, et al. (2010a). Head Start Impact Study Final Report. Washington, DC: U.S. Department for Health and Human Services, Administration for Children and Families.

Puma, Michael, Stephen Bell, Ronna Cook, Camilla Heid, et al. (2010b). Head Start Impact Study Technical Report. Washington, DC: U.S. Department for Health and Human Services, Administration for Children and Families.

Puma, Mike, Stephen Bell, Ronna Cook, Camilla Heid, Pam Broene, Frank Jenkins, Andrew Mashburn, and Jason Downer. (2012). Third Grade Follow-up to the Head Start Impact Study: Final Report. Washington, DC: Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services. OPRE Report 2012-45.

Appendix: Added Technical Details & Discussion of Alternative Assumptions

As discussed in the main body of this report, the technique we use for analyzing the influence of Head Start quality on children's developmental outcomes identifies predicted high quality sample members from the treatment group and the control group in symmetric fashion, and then estimates impacts on that subpopulation as one would in any experimental subgroup analysis. The symmetry of the selection procedure ensures that equivalent subgroups are compared and guarantees that the impact estimates are free from differential selection bias or any other sources of bias. Compared to conventional propensity score matching, for example, the symmetric selection of treatment and control subgroup members within experimental data ensures full internal validity—unbiasedness of the impact estimates generated for the subgroups examined. However, the subgroup for which the methodology produces unbiased impact estimates—children with the highest *predicted* probabilities of being in high quality Head Start programs—is not necessarily the subgroup of interest—children who *actually* experience high quality Head Start. The predictive model, while symmetric for both treatment and control groups, is imperfect for both groups, potentially reducing the relevance (i.e., external validity or generalizability) of the findings, which is why we ultimately convert the results so that they represent *actual* rather than *predicted* subgroup members. The body of this report discusses the five steps involved in carrying out this subgroup analysis of the effects of Head Start quality on children's outcomes. This Appendix provides some additional details about the results of our analytic process—including the correct prediction rate and the predicted subgroups' estimated impacts—and then elaborates on two possible alternative sets of assumptions, providing and analyzing results.

Results of Prediction Process

As noted in the text, the analysis starts by predicting which individuals would not participate in Head Start or would experience low or high quality Head Start. If there were perfect prediction, then the ultimate conversion step would be unnecessary. But, our prediction is not perfect. It is, however, better than random, and so we use this observation to justify using this approach.

As explained in Section 4, ten random subsets of the combined 3-year-old and 4-year-old treatment groups were used to develop a model predicting membership in the non-participant, low quality, and high quality subgroups. Because we observe both the predicted and actual quality measures within the treatment group, we can assess the predictive accuracy of the model. The following Exhibits present information on the accurate proportions of the predicted subgroups. We report this information for each of the three measures of quality that we use, following with an exhibit that presents the notation that identifies each of these elements for its use in our subsequent conversion process.

Exhibit A-1 cross-tabulates predicted quality subgroup membership in its rows by actual quality subgroup measurement in its columns. The following percentages appear:

- Row percentages that allocate members of a given predicted quality subgroup across actual quality categories (e.g., the top left entry in the exhibit indicates that 34.7 percent of predicted non-participants are actual non-participants);

- Column percentages that allocate members of a given actual quality subgroup across predicted quality categories (e.g., the top left bracketed entry indicates that 21.6 percent of actual non-participants are predicted as non-participants).

Exhibit A-1. Predicted by Actual Resource Quality

Predicted Resource Quality	Actual Resource Quality			Overall
	No-show	*Low*	*High*	
No-show	**34.7** [21.6]	23.0 [10.3]	42.3 [9.1]	11.8
Low	18.5 [27.2]	**62.3** [65.9]	19.3 [9.9]	28.0
High	16.2 [51.3]	10.4 [23.8]	**73.4** [81.0]	60.2
Overall	19.0	26.4	54.6	100.0

Notes: Diagonal elements in bold represent the correct placement of predicted within actual groups. The first numbers in each cell represent the proportion of the predicted that are in the actual group (the "row" percent). The numbers in brackets represent the proportion of the actual that are in the predicted group (the "column" percent). n=2,245

The numbers in brackets along the diagonal of the exhibit show that our model correctly predicted 21.6 percent of no-shows, 65.9 percent of low quality participants, and 81.0 percent of high quality participants. The "Overall" rows and columns indicate that the predicted distribution of cases among the three groups (12, 28 and 60 percent for each of the non-participant, low quality and high quality groups, respectively), is not wildly different from the actual distribution (of 19, 26 and 54 percent, respectively). These are unweighted numbers and reflect only the process of our analyzing the subset of cases that are relevant for this analysis and should not be construed as being nationally representative as weighted data would be. As Exhibits A-2 and A-3 show, the correct prediction rate for the high quality subgroup is 82.8 and 51.4 percent, respectively, for each of the interactions and exposure measures. Our correct placement rates are lower for the interactions and exposure measures than for the resources measure, but overall we conclude that the correct placement rates are acceptable for advancing this method of analyzing the effects of Head Start quality.

Another way to quantify the correct placement aggregates across the three groups. The overall hit rate that we achieve is 66 percent for resource quality, 64 percent for interactions quality, and 63 percent for exposure quality. In general in applying this analytic method, this rate should reflect that there is some useful prediction taking place: that is, the prediction should be better than a random sorting of the data into three groups, and ideally meaningfully better to instill confidence that the building blocks of the analysis—the experimentally-based impacts on predicted subgroups—are a reasonable starting point. We recognize that these terms—"meaningfully better" or "reasonable"—are subjective. In this case we reach the conclusion that the success of the prediction process is sufficient to warrant proceeding with the analysis. Current research is exploring how better to operationalize these constructs of "better" and "reasonable" to generate clear prescription for future applications (Harvill & Peck, in progress).

Exhibit A-2. Predicted by Actual Interactions Quality

Predicted Interactions Quality	Actual Interactions Quality			Overall
	No-show	*Low*	*High*	
No-show	**37.1** [21.6]	15.7 [8.5]	47.2 [8.6]	11.1
Low	13.5 [12.7]	**57.4** [49.7]	29.1 [8.6]	17.8
High	16.6 [65.1]	12.1 [41.9]	**70.3** [82.8]	71.2
Total	19.0	20.5	60.5	100.0

Notes: Diagonal elements in bold represent the correct placement of predicted within actual groups. The first numbers in each cell represent the proportion of the predicted that are in the actual group (the "row" percent). The numbers in brackets represent the proportion of the actual that are in the predicted group (the "column" percent). n=2,245

Exhibit A-3. Predicted by Actual Exposure Quality

Predicted Exposure Quality	Actual Exposure Quality			Overall
	No-show	*Low*	*High*	
No-show	**38.3** [22.7]	47.0 [9.2]	14.6 [8.0]	11.6
Low	17.5 [60.4]	**69.7** [79.7]	12.8 [40.6]	67.6
High	15.9 [16.9]	31.4 [11.0]	**52.7** [51.4]	20.8
Total	19.6	59.1	21.3	100.0

Notes: Diagonal elements in bold represent the correct placement of predicted within actual groups. The first numbers in each cell represent the proportion of the predicted that are in the actual group (the "row" percent). The numbers in brackets represent the proportion of the actual that are in the predicted group (the "column" percent). n=2,178

In addition to these placement percentages that result from our analysis, we report here the notation that we use in representing the conversion of results from predicted to actual subgroups. Readers should be able to use Exhibit A-4 to identify the elements from Exhibits A-1 through A-3 that are needed as inputs into the conversion formulae to compute the conversion factors themselves.

Exhibit A-4. Predicted by Actual Quality, Notational Information for Conversion

	Actual Quality		
Predicted Quality	*No-show*	*Low*	*High*
No-show	s_N	w_N $(1-r)$	g_N $(1-p)$
Low	s_L	w_L	g_L q
High	s_H	w_H	g_H

Notes: The first symbol in each cell represents the proportion of the predicted that are in the actual group (the "row" percent). The symbol in parentheses represents the proportion of the actual that are in the predicted group (the "column" percent).

Estimated Impacts on Predicted Subgroups

As noted in the report, we follow the HSIS existing practice to pool data and compute subgroups' impact estimates for our analysis of the effects of Head Start on these quality subgroups of interest. Though not of primary interest to most readers, the two exhibits below present the estimated impacts on each of the five outcomes under examination, by year, across the three quality measures, for each cohort. In essence, these are the "building blocks" for the estimates of impact on actual subgroups in a later subsection, using the formulas for *N*, *L*, and *H* given elsewhere. The estimates in the exhibits reflect comparison of symmetrically-selected subsamples of the treatment and control groups derived from baseline characteristics; hence, they are purely experimental and free from selection and other sources of bias. However, they do not fully reflect the non-participant and low and high quality subgroups their labels imply, because predicted members of a subgroup often are not actual members of that subgroup.

Exhibit A-5. Estimated Impacts on PPVT Scores for Predicted Subgroups, by Quality Measure, by Follow-up Year, 3-Year-Old Cohort, Preferred Assumptions

	End of HS Year 1 2003	End of HS Year 2 2004	End of Kindergarten 2005	End of First Grade 2006	End of Third Grade 2008
Resource					
No-shows	13.4 *	-6.4	3.1	3.8	3.0
High	8.2 ***	4.7 **	0.3	1.5	2.6
Low	1.1	-0.6	-0.7	1.8	0.0
Interaction					
No-shows	15.3 **	-8.3	1.6	2.1	0.3
High	5.9 **	4.1 *	0.3	2.6	2.5
Low	6.2	0.2	-0.3	-1.2	0.7
Exposure					
No-shows	9.6	-6.2	5.1	-1.8	0.3
High	14.4 ***	3.1	-2.2	3.4	-1.2
Low	4.8 **	3.1	-0.1	2.0	3.1 *

Notes: The impact is the regression-adjusted difference (impact) between the treatment and control groups in the number of points on each outcome measure.

*** statistically significant: p<0.01; ** statistically significant: p<0.05; * statistically significant: p<0.10

Exhibit A-6. Estimated Impacts on WJ3 Letter-Word Scores for Predicted Subgroups, by Quality Measure, by Follow-up Year, 3-Year-Old Cohort, Preferred Assumptions

	End of HS Year 1 2003	End of HS Year 2 2004	End of Kindergarten 2005	End of First Grade 2006	End of Third Grade 2008
Resource					
No-shows	4.9	-0.4	3.6	2.4	2.1
High	8.6 ***	3.9	-0.1	0.6	2.1
Low	3.9	0.7	-2.0	-1.2	-2.7
Interaction					
No-shows	2.0	-6.8 **	1.2	3.1	-0.4
High	8.5 **	5.3 **	0.5	1.3	1.9
Low	5.3	-1.7	-3.7	-4.2	-2.8
Exposure					
No-shows	0.1	-2.7	-5.2	-4.7	-5.8
High	9.1 **	0.5	-9.9 **	-2.9	-3.2
Low	7.9 ***	4.2 **	3.2	1.9	2.9

Notes: The impact is the regression-adjusted difference (impact) between the treatment and control groups in the number of points on each outcome measure.
*** statistically significant: p<0.01; ** statistically significant: p<0.05; * statistically significant: p<0.10

Exhibit A-7. Estimated Impacts on WJ3 Applied Problems Scores for Predicted Subgroups, by Quality Measure, by Follow-up Year, 3-Year-Old Cohort, Preferred Assumptions

	End of HS Year 1 2003	End of HS Year 2 2004	End of Kindergarten 2005	End of First Grade 2006	End of Third Grade 2008
Resource					
No-shows	3.9	4.2	2.0	4.8	-2.0
High	7.9 **	2.0	-0.4	1.2	1.2
Low	-0.8	-1.7	-3.9	0.9	-1.2
Interaction					
No-shows	2.7	5.9	1.5	6.4	-1.1
High	6.9 **	2.0	-0.5	1.2	1.5
Low	2.2	-3.9	-4.5	0.2	-3.4
Exposure					
No-shows	3.7	0.9	1.7	3.0	-4.9
High	9.9 **	3.3	-3.3	0.3	0.3
Low	4.6 *	0.5	-1.2	1.4	1.1

Notes: The impact is the regression-adjusted difference (impact) between the treatment and control groups in the number of points on each outcome measure.
*** statistically significant: p<0.01; ** statistically significant: p<0.05; * statistically significant: p<0.10

Exhibit A-8. Estimated Impacts on Social Competence for Predicted Subgroups, by Quality Measure, by Follow-up Year, 3-Year-Old Cohort, Preferred Assumptions

	End of HS Year 1 2003	End of HS Year 2 2004	End of Kindergarten 2005	End of First Grade 2006	End of Third Grade 2008
Resource					
No-shows	0.0	0.2	0.2	0.2	-0.1
High	0.0	0.2 *	0.2	0.1	0.4 **
Low	0.0	0.0	0.3	-0.2	0.1
Interaction					
No-shows	0.0	0.0	0.4	0.1	0.0
High	0.0	0.3 **	0.2	0.1	0.3 **
Low	0.0	-0.1	0.1	-0.2	0.3
Exposure					
No-shows	-0.3	0.2	0.2	0.0	-0.1
High	0.2	0.3	-0.2	-0.4	0.5 **
Low	0.0	0.1	0.3 **	0.1	0.3 *

Notes: The impact is the regression-adjusted difference (impact) between the treatment and control groups in the number of points on each outcome measure.
*** statistically significant: p<0.01; ** statistically significant: p<0.05; * statistically significant: p<0.10

Exhibit A-9. Estimated Impacts on Problem Behaviors for Predicted Subgroups, by Quality Measure, by Follow-up Year, 3-Year-Old Cohort, Preferred Assumptions

	End of HS Year 1 2003	End of HS Year 2 2004	End of Kindergarten 2005	End of First Grade 2006	End of Third Grade 2008
Resource					
No-shows	-0.6	0.1	-0.1	-0.6	0.6
High	-1.2 ***	-0.7 *	-0.6 **	-0.3	-0.2
Low	0.0	0.2	0.5	0.4	0.4
Interaction					
No-shows	-0.5	0.1	-0.2	-0.9	0.2
High	-0.9 ***	-0.6 *	-0.4	0.1	0.1
Low	-0.9	0.3	0.2	-0.3	0.0
Exposure					
No-shows	-0.6	0.2	-0.2 *	-0.5	0.8
High	-0.1	0.1	0.8	0.5	0.1
Low	-1.0 ***	-0.6	-0.5	-0.2	-0.1

Notes: The impact is the regression-adjusted difference (impact) between the treatment and control groups in the number of points on each outcome measure.
*** statistically significant: p<0.01; ** statistically significant: p<0.05; * statistically significant: p<0.10

Exhibit A-10. Estimated Impacts on PPVT Scores for Predicted Subgroups, by Quality Measure, by Follow-up Year, 4-Year-Old Cohort, Preferred Assumptions

	End of HS Year 2003	End of Kindergarten 2004	End of First Grade 2005	End of Third Grade 2007
Resource				
No-shows	1.3	-4.6	3.2	-3.6
High	5.2 **	3.2	5.5 ***	3.3
Low	5.8 *	8.2 *	1.5	2.8
Interaction				
No-shows	3.8	2.5	7.6 *	-2.0
High	6.1 ***	3.5	3.8 *	3.0 *
Low	2.3	2.5	3.4	2.5
Exposure				
No-shows	3.2	-2.2	3.8	0.0
High	4.2	6.3	5.9	0.4
Low	5.9 ***	3.9 *	4.0 *	3.1

Notes: The impact is the regression-adjusted difference (impact) between the treatment and control groups in the number of points on each outcome measure.
*** statistically significant: p<0.01; ** statistically significant: p<0.05; * statistically significant: p<0.10

Exhibit A-11. Estimated Impacts on WJ3 Letter-Word Scores for Predicted Subgroups, by Quality Measure, by Follow-up Year, 4-Year-Old Cohort, Preferred Assumptions

	End of HS Year 2003	End of Kindergarten 2004	End of First Grade 2005	End of Third Grade 2007
Resource				
No-shows	2.9	-2.5	-6.7	-3.6
High	5.5 *	-1.2	1.8	2.7
Low	14.4 ***	7.9 *	4.9	4.1
Interaction				
No-shows	4.7	1.3	-3.7	-1.0
High	8.0 **	1.1	2.3	2.4
Low	7.5	0.0	3.2	5.2
Exposure				
No-shows	7.5	2.1	-5.4	-2.3
High	7.7	9.9 *	6.9	5.9
Low	7.4 **	-1.7	1.6	2.2

Notes: The impact is the regression-adjusted difference (impact) between the treatment and control groups in the number of points on each outcome measure.
*** statistically significant: p<0.01; ** statistically significant: p<0.05; * statistically significant: p<0.10

Exhibit A-12. Estimated Impacts on WJ3 Applied Problems Scores for Predicted Subgroups, by Quality Measure, by Follow-up Year, 4-Year-Old Cohort, Preferred Assumptions

	End of HS Year 2003	End of Kindergarten 2004	End of First Grade 2005	End of Third Grade 2007
Resource				
No-shows	3.7	-4.2 *	0.5	-1.7
High	1.9	-1.9 *	0.5	-0.1
Low	8.0 *	7.8	1.9	-0.5
Interaction				
No-shows	6.8 *	-2.8	3.0	-2.3
High	3.2	0.6	0.2	-0.5
Low	2.7	0.0	1.7	2.3
Exposure				
No-shows	7.1 *	-2.2	1.1	-1.8
High	-1.6	3.7	2.5	-0.6
Low	3.9	-0.3	0.4	0.0

Notes: The impact is the regression-adjusted difference (impact) between the treatment and control groups in the number of points on each outcome measure.
*** statistically significant: p<0.01; ** statistically significant: p<0.05; * statistically significant: p<0.10

Exhibit A-13. Estimated Impacts on Social Competence for Predicted Subgroups, by Quality Measure, by Follow-up Year, 4-Year-Old Cohort, Preferred Assumptions

	End of HS Year 2003	End of Kindergarten 2004	End of First Grade 2005	End of Third Grade 2007
Resource				
No-shows	0.0	-0.1	-0.1	0.1
High	0.0	0.1	0.1	-0.2
Low	-0.2	0.1	-0.1	0.1
Interaction				
No-shows	0.1	-0.1	-0.2	0.1
High	0.1	0.2	0.1	-0.1
Low	-0.7 **	-0.1	0.0	-0.3
Exposure				
No-shows	0.2	0.2	0.1	0.4
High	0.2	0.2	-0.3	-0.5
Low	-0.2	0.0	0.1	-0.1

Notes: The impact is the regression-adjusted difference (impact) between the treatment and control groups in the number of points on each outcome measure.
*** statistically significant: p<0.01; ** statistically significant: p<0.05; * statistically significant: p<0.10

Exhibit A-14. Estimated Impacts on Problem Behaviors for Predicted Subgroups, by Quality Measure, by Follow-up Year, 4-Year-Old Cohort, Preferred Assumptions

	End of HS Year 2003	End of Kindergarten 2004	End of First Grade 2005	End of Third Grade 2007
Resource				
No-shows	1.0 *	0.6	0.5	-0.6
High	0.0	0.2	-0.1	-0.2
Low	-0.4	-0.3	-0.7	-1.3 *
Interaction				
No-shows	1.1 **	0.1	0.9	-0.3
High	-0.1	0.1	-0.4	-0.6
Low	-0.2	-0.1	0.0	-0.6
Exposure				
No-shows	0.8	0.6	0.3	-0.8
High	-0.7	-0.3	-0.7	-1.0
Low	0.0	0.1	-0.1	-0.4

Notes: The impact is the regression-adjusted difference (impact) between the treatment and control groups in the number of points on each outcome measure.
*** statistically significant: p<0.01; ** statistically significant: p<0.05; * statistically significant: p<0.10

It is with the information embedded in Exhibits A-5 through A-14 that we impose our conversion factors in order to "reallocate" the estimated impacts from the predicted to the actual subgroups.[11] We elaborate next on the implications of imposing alternative assumptions for a single year of outcomes, the first year of follow-up.

[11] Some may find it interesting to note that there is about the same frequency of statistically significant effects observed in the predicted subgroups as reported for the converted, actual impacts: 35 of the 225 tests for the 3-year-old cohort and 23 of the 180 tests for the 4-year-old cohort are statistically significant.

Discussion of Alternative Assumptions

We chose the preferred set of assumptions discussed in the main body of this report because we believe them to be reasonable. Nevertheless, we recognize the possibility that results on the impact of actual Head Start quality level may vary under other assumptions.[12] To assess the robustness of the results to the assumptions, we apply two alternative sets of assumptions and re-compute Step 5 of the procedure. In both instances, we retain the first three assumptions from above as non-controversial:

(1') (1") $N_N = 0$ – The impact on predicted no-shows who are actual no-shows is zero.

(2') (2") $N_L = 0$ – The impact on predicted low quality participants who are actual no-shows is zero.

(3') (3") $N_H = 0$ – The impact on predicted high quality participants who are actual no-shows is zero.

The fourth assumption from before is also retained:

(4') (4") $L_H = L_L$ – The impacts on low quality participants are the same for children predicted to be high quality participants and children predicted to be low quality participants.

This leaves two assumptions to reconsider. In the first alternative scenario, our goal is to adopt assumptions that are reasonable in our view but sufficiently different from the original final two assumptions to provide a strong *contrast* in the sensitivity analysis:

(5') $L_L = L_N$ – The impacts on low quality participants are the same for children predicted to be low quality participants and children predicted to be no-shows.

This assumption, in conjunction with (4'), postulates that low quality Head Start has the same impact on three types of children with different propensities to participate in it: predicted no-shows, predicted low quality, and predicted high quality. It seems more reasonable to suppose a relatively weak version of the program has a uniform (and possibly smaller) impact of this sort than that high quality Head Start does. Indeed, with the original assumption (5) replaced by (5'), no assumptions about homogeneous impacts from high quality Head Start participation are made in this scenario.

(6') $H_H = 2L_H$ – The impact on predicted high quality participants who are actual high quality participants is two times the impact on predicted high quality participants who are actual low quality participants.

Assuming a magnitude relationship of this sort, as opposed to strict equality, puts a new twist into the first alternative scenario. It is in fact no more exacting an assumption than that of pure equality made for the low quality participants, and it is the least extreme simple multiplicative relationship. It involves children of similar background characteristics (for the characteristics that predict high quality participation) but for otherwise similar children we assume here that actual high quality services have a larger impact than actual low quality services.

[12] Bell and Peck (2013) consider how the validity of the assumptions can be improved through strategic choices of background variables to include in the quality prediction model. This work argues that the best choice of predictors are those exogenous variables that most strongly predict membership in the endogenous subgroup of interest (here, either high quality Head Start participation or low quality participation) but that are otherwise unrelated to program impact magnitude.

Inserting these new assumptions into the derivation results in the following formulas for impacts on the *actual low quality* subgroup (L) and the *actual high quality* subgroup (H):

$$L = \frac{I_H}{w_H + 2g_H}$$

$$H = \frac{(1-p)}{g_N} I_N + \frac{q}{g_L} I_L + \left[\frac{2(p-q)g_N g_L g_H - (1-p)w_N g_L - q w_L g_N}{g_N g_L (w_H + 2g_H)}\right] I_H$$

where

q is the proportion of high quality participants predicted as low quality participants.

As before, $N = 0$.

The final alternative scenario examines the sensitivity of the findings in the range between the other two scenarios. Here, we return to the original fifth assumption:

(5”) $H_H = H_L$ – The impacts on high quality participants are the same for children predicted to be high quality participants and children predicted to be low quality participants, and we add to it an assumption from the second scenario:

(6”) $L_L = L_N$ – The impacts on low quality participants are the same for children predicted to be low quality participants and children predicted to be no-shows.

This set of assumptions leads to the following formulas for impacts on the actual low quality subgroup (L) and the actual high quality subgroup (H):

$$L = -\left[\frac{g_H}{(w_H g_L - w_L g_H)}\right] I_L + \left[\frac{g_L}{(w_H g_L - w_L g_H)}\right] I_H$$

$$H = \left(\frac{1-p}{g_N}\right) I_N + \left[\frac{(1-p)w_N(g_H) + p w_H(g_N)}{(w_H g_L - w_L g_H)(g_N)}\right] I_L - \left[\frac{(1-p)w_N(g_L) + p w_L(g_N)}{(w_H g_L - w_L g_H)(g_N)}\right] I_H$$

As before, $N = 0$.

Sensitivity to Assumptions

Exhibits A-15 and A-16 report the results from imposing the first set of alternative assumptions. Examining the differences and similarities between these results and those in the main text for the preferred set of assumptions, we make the following observations. For the 3-year-old cohort, the alternative assumptions strengthen the case that low quality Head Start can have favorable effects on child development in the first Head Start year. But they also add to the evidence of statistically significant *differences* in effectiveness by quality level favoring high quality Head Start. The same general pattern is evident in the 4-year-old cohort under these assumptions, but less noticeably so.

Exhibit A-15. Estimated Impacts for Actual Subgroups, by Quality Measure and Outcome, at the end of the first Head Start year (2003), 3-Year-Old Cohort, First Alternative Assumptions

	Outcomes				
	Cognitive			Social-Emotional	
Actual Quality	**PPVT**	**WJ3 Letter-Word**	**WJ3 Applied Problems**	**Social Competence**	**Problem Behaviors**
Resource					
No-shows	0.0 [a]	0.0	0.0	0.0	0.0
High	9.0 ***	8.4 ***	4.2 **	0.1	-0.4 *
Low	4.4 ***	4.7 ***	4.0 **	-0.1	-0.4 ***
High-Low Difference	4.6 **	3.7 *	0.2	0.1	0.0
Interaction					
No-shows	0.0	0.0	0.0	0.0	0.0
High	8.2 ***	7.3 ***	4.8 **	-0.0	-0.5 **
Low	3.9 ***	4.9 ***	2.7 *	-0.0	-0.3 *
High-Low Difference	4.3 **	2.5 *	2.1	0.0	-0.2
Exposure					
No-shows	0.0	0.0	0.0	0.0	0.0
High	13.9	15.8	12.0	-0.1	-2.6 ***
Low	5.7 *	5.1 **	2.3	-0.0	-0.2
High-Low Difference	8.2	10.8	9.7	-0.1	-2.8 ***

Notes: The impact is the regression-adjusted difference (impact) between the treatment and control groups in the number of points on each outcome measure.
[a] No statistical significance noted because no-show impact estimates are derived by assumption to be zero.
*** statistically significant: p<0.01; ** statistically significant: p<0.05; * statistically significant: p<0.10

Exhibit A-16. Estimated Impacts for Actual Subgroups, by Quality Measure and Outcome, at the end of the Head Start year (2003), 4-Year-Old Cohort, First Alternative Assumptions

	Outcomes				
	Cognitive			Social-Emotional	
Actual Quality	**PPVT**	**WJ3 Letter-Word**	**WJ3 Applied Problems**	**Social Competence**	**Problem Behaviors**
Resource					
No-shows	0.0 [a]	0.0	0.0	0.0	0.0
High	7.0 ***	10.4 ***	5.3	-0.1	0.1
Low	3.6 **	3.1 *	1.5	0.0	0.3 *
High-Low Difference	3.4	7.3 ***	3.8	-0.1	0.2
Interaction					
No-shows	0.0	0.0	0.0	0.0	0.0
High	5.9 ***	7.2 ***	3.2	-0.1	-0.2
Low	3.3 ***	4.7 **	2.5	0.0	-0.2
High-Low Difference	2.6	2.5 *	0.7	-0.1	0.0
Exposure					
No-shows	0.0	0.0	0.0	0.0	0.0
High	5.0	-5.3	14.6	0.0	-0.6
Low	6.5 ***	12.0 ***	0.1	-0.1	-0.2
High-Low Difference	-1.5	-17.2	14.4	0.1	-0.5

Notes: The impact is the regression-adjusted difference (impact) between the treatment and control groups in the number of points on each outcome measure.
[a] No statistical significance noted because no-show impact estimates are derived by assumption to be zero.
*** statistically significant: p<0.01; ** statistically significant: p<0.05; * statistically significant: p<0.10

Next, we report the results from imposing the second set of alternative assumptions in Exhibits A-17 and A-18. While we might characterize the second alternative assumptions as being a middle ground between our preferred assumptions and the first alternative assumptions, the results from imposing these assumptions are very similar to the results estimated by using our preferred assumptions—almost identical in the case of the 4-year-old cohort. This is the case even though the conversion factors themselves appear to be quite different in their structure. In turn, an interpretation of the differences between the results generated by imposing the first and second sets of alternative assumptions is likewise identical to the discussion of the differences between the results generated by imposing the preferred assumptions and the first set of alternative assumptions. Overall, the meaning of our conclusions does not differ when we impose these alternative assumptions. Given that one alternative set of assumptions provides modestly different results and the other alternative set provides largely identical results, we feel justified basing our conclusions regarding the impact of being in Head Start by quality subgroup on the preferred assumptions.

Exhibit A-17. Estimated Impacts for Actual Subgroups, by Quality Measure and Outcome, at the end of the first Head Start year (2003), 3-Year-Old Cohort, Second Alternative Assumptions

	Outcomes				
	Cognitive			Social-Emotional	
Actual Quality	**PPVT**	**WJ3 Letter-Word**	**WJ3 Applied Problems**	**Social Competence**	**Problem Behaviors**
Resource					
No-shows	0.0 [a]	0.0	0.0	0.0	0.0
High	11.5 ***	10.7 ***	10.0 **	-0.2	-1.0 **
Low	3.1	4.2	-4.3	0.3	0.4
High-Low Difference	8.4	6.4	14.2	-0.5	-1.3
Interaction					
No-shows	0.0	0.0	0.0	0.0	0.0
High	8.0 **	9.7 **	5.5	0.0	-0.6 *
Low	10.8	5.7	4.7	-0.2	-0.5
High-Low Difference	-2.8	4.0	0.8	0.2	-0.1
Exposure					
No-shows	0.0	0.0	0.0	0.0	0.0
High	16.5	6.2	2.6	0.1	1.3
Low	5.7	9.4 **	6.2	-0.1	-1.1 ***
High-Low Difference	10.9	-3.2	-3.6	0.1	2.4 **

Notes: The impact is the regression-adjusted difference (impact) between the treatment and control groups in the number of points on each outcome measure.

[a] No statistical significance noted because no-show impact estimates are derived by assumption to be zero.

*** statistically significant: p<0.01; ** statistically significant: p<0.05; * statistically significant: p<0.10

Exhibit A-18. Estimated Impacts for Actual Subgroups, by Quality Measure and Outcome, at the end of the Head Start year (2003) 4-Year-Old Cohort, Second Alternative Assumptions

	Outcomes				
	Cognitive			Social-Emotional	
Actual Quality	PPVT	WJ3 Letter-Word	WJ3 Applied Problems	Social Competence	Problem Behaviors
Resource					
No-shows	0.0 [a]	0.0	0.0	0.0	0.0
High	6.0 *	1.3	-0.4	0.0	-0.5
Low	8.9	24.5 ***	14.3	-0.3	0.1
High-Low Difference	-3.0	-23.2 ***	-14.7	0.4	-0.6
Interaction					
No-shows	0.0	0.0	0.0	0.0	0.0
High	7.6 ***	8.7 **	5.3	0.0	-0.2
Low	3.5	7.9	0.1	-0.5	-0.4
High-Low Difference	4.1	0.9	5.2	0.5	0.2
Exposure					
No-shows	0.0	0.0	0.0	0.0	0.0
High	14.8 *	28.1 ***	-6.9	-0.3	0.3
Low	4.0	2.1	7.9 *	0.0	-0.5
High-Low Difference	10.8	26.0 **	-14.8	-0.2	-0.8

Notes: The impact is the regression-adjusted difference (impact) between the treatment and control groups in the number of points on each outcome measure.

[a] No statistical significance noted because no-show impact estimates are derived by assumption to be zero.

*** statistically significant: p<0.01; ** statistically significant: p<0.05; * statistically significant: p<0.10