The Role of Teacher-Child Interactions and Academic Content in Determining the Short-Term Impact of Head Start on Child’s Pre-academic and Social Skills

by
Laura R. Peck
Stephen H. Bell
Todd Grindal
Abt Associates Inc.

October 24, 2016
DRAFT

Acknowledgements
This paper was prepared for presentation at the Fall Conference of the Association for Public Policy Analysis and Management (APPAM), November 5, 2016 in Washington, DC. The authors appreciate input on an earlier draft of the paper from Jacob Klerman and Mike Lopez. The original analysis on which this draws was funded by the Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services (Jennifer Brooks, then project officer, now at the Gates Foundation) via subcontract to Westat (Camilla Heid, project director). In addition, we recognize our collaborators on that work, including Ronna Cook (Ronna Cook Associates), Frank Jenkins (Westat), Andrew Mashburn (Portland State University), and Jason Downer (University of Virginia).
The Role of Teacher-Child Interactions and Academic Content in Determining the Short Term Impact of Head Start on Child Development

As federal, state, and local governments seek to expand access to publicly funded preschool programs in the United States, policymakers and program leaders are eager to determine what program traits are associated with stronger outcomes for children. Two program traits—the degree to which programs focus on academic content and the quality of teacher-child interactions—represent malleable program components that might be related to larger program impacts. To date, there has been little empirical research on how variation in these program traits impacts the development of pre-academic and social skills among preschool-aged children.

This paper uses data from the Head Start Impact Study (HSIS), capitalizing on that evaluation’s experimental evaluation design, to examine the extent to which the level of teacher-child interactions or emphasis on academic content in Head Start classrooms impacts the development of children’s early skills. We follow children in the HSIS treatment group through the end of their Head Start experience—one year for the entering four-year-olds and two years for the three-year-olds—and control group children for the same duration. We then leverage the rich baseline data in the HSIS to identify the experimental sample members most likely to participate in Head Start programs with low or high levels of teacher-child interactions or in programs with less or more academic content. Within these statistically equivalent subgroups, we compare outcomes of treatment and control group children to estimate the impact of having had these types of experiences in Head Start classrooms.

We begin with providing the theoretical justification for why these traits of preschool programs might matter to children’s development, we discuss our methods—including data source, measurement, analytic approach and associated assumptions—before reporting our
findings. In brief, we find some evidence that Head Start’s impact varies by both teacher-child interaction levels and levels of academic content. Participating in classrooms with high levels of teacher-child interactions leads to generally stronger impacts on a range of pre-academic skills for both those children who enter Head Start at age three years and those who enter at age four years. Among children who enter Head Start at age 4 years, participating in classrooms with high levels of academic content produces stronger favorable impacts on their pre-academic skills but not their behavior. Among children who enter Head Start at age three years, participating in classrooms with high levels of academic content leads to less favorable impacts on children’s problem behaviors and pre-academic skill development. Although participating in classrooms with high levels of teacher-child interactions appear to be beneficial for both ages of children, high levels of academic content appear to be beneficial to older children but detrimental to younger children. These findings suggest that Head Start and other early childhood education programs should differentiate the curricular approaches and offer younger children a less academically oriented early childhood education experience.

BACKGROUND

Preschool programs represent a key developmental input for many young children in the United States (Pianta et al., 2009; Shonkoff, Boyce, & McEwen, 2009). Prior research suggests that participation in developmentally-focused early childhood programs generally confer substantial short and long-term benefits for young children (Duncan & Magnuson, 2013). Yet, much of the evidence on the favorable benefits of preschool comes with the caveat that only programs with certain characteristics, often referred to as “high quality” programs, yield meaningful benefits for children (Loeb, Fuller, Kagan, & Carrol, 2004; Mashburn et al., 2008).
As a result, substantial public funds have been dedicated to identifying and promoting the preschool program characteristics that may facilitate quality.

Although there is enthusiasm in the preschool community regarding the potential benefits of large-scale quality improvement efforts, empirical evidence regarding the relationship between indicators of program quality and benefits for young children remains mixed. Some research suggests that children who participate in preschool programs that are rated as having higher overall quality, outperform children who attend preschool programs with lower quality ratings on a number of important cognitive and social emotional indicators (Burchinal, Vandergrift, Pianta, & Mashburn, 2010; Kontos et al., 1995; Lipsey, Hoffer, Dong, Farran, & Bilbrey, 2013). For example, a longitudinal analysis of U.S. children found that those who participated in early childhood programs that were rated as high quality based on the Observational Record of the Caregiving Environment or Home Observation for Measurement of the Environment (HOME) exhibited higher scores on measures of vocabulary development when compared to similar children who participated in programs rated as being lower quality (NICHD ECCRN, 2000). Research conducted outside of the U.S. has yielded similar results indicating benefits for children who participate in higher quality programs (Love et al., 2003).

This prior work linking program characteristics and child outcomes uses non-experimental methods, making it difficult to disentangle the effects of program features from child and family characteristics—such as maternal education and family income—associated with selection into particular types of preschool programs (Duncan, 2003). Recent efforts attempting to control for non-program confounding factors have identified only modest associations between overall program quality and cognitive and behavioral improvements for young children (Keys et al., 2013). A study by Sabol and colleagues (2013) examined data from nine states and found that preschool programs quality ratings, as determined by their state’s
quality rating and improvement system, were not consistently associated with children’s learning.

It is possible that current measures of preschool quality mask important traits of preschool programs that lead to larger impacts on outcomes for children. The instructional focus in a preschool program represents a potentially important dimension of preschool programs that might be related to child outcomes. Preschool programs vary substantially in the degree to which they explicitly focus instruction on the development of children’s pre-academic skills (Fulgini, et al., 2012). Although some prior work has indicated that attending a program with a high level of focus on academic skills is associated with stronger development of early mathematics skills (de Hahn et al., 2014), other research has suggested that such an academic focus can have negative consequences for children’s social-emotional development (Elkind, 2007).

The ways in which program content is delivered may also be associated with better or worse child outcomes. Neuroscience research indicates that reciprocal interactions between children and their caregivers are key to facilitating early development (Shonkoff & Phillips, 2000). Work by Mashburn and colleagues (2008) found that the level of preschool teachers’ instructional interactions predicted children’s academic and language skills even after adjusting for children’s prior skill levels and other relevant program and family characteristics. Non-experimental analysis by Burchinal and colleagues (2010) indicates that the level of teacher-child interactions is a strong predictor of children’s scores on measures of pre-academic skills and social behavior. This work has led Head Start, the largest provider of publicly funded preschool in the United States, to require that programs monitor the levels of teacher-child interactions and provide teachers with training to improve their interactions in the classroom.
THE PRESENT STUDY

In the current study we examine whether observed indicators of (1) teacher-child interactions and (2) the level of academic content in Head Start classrooms mediate the impact of the program on the development of children’s short-term pre-academic and behavioral skills. We focus on these dimensions of preschool programs, rather than the observed omnibus measures of “program quality” because teacher-child interactions and the level of academic content are measurable, policy malleable, and plausibly related to program impacts on children’s short-term developmental outcomes. Many broader measures of program quality incorporate measures of content and teacher-child interactions, but, unlike the work presented in this study, do not disaggregate them to permit analysis of the effects of each in its own right. In particular, we examine whether variation in these program traits affect Head Start’s impact on children’s development through the end of preschool. Although there is interest in the potential lasting effects of Head Start participation (see Lanham, 2016), we focus here on the short term to align with the program’s most central goal: to bring underprivileged children into Kindergarten on par with their more advantaged peers and ready to learn. The HSIS experimental evaluation, which involved a nationally representative sample and included rich data at baseline on children, families and programs, provides an ideal source for analyzing these questions. Impact evidence from that study (see Puma et al., 2012) shows that admission to Head Start on average moderately improved children’s preschool development and school readiness in certain areas, with some of those advantages persisting through first grade but few lasting into third grade.

The analysis of interest here is challenging for two main reasons: the first reason—having to do with quality measurement of what goes on in Head Start classrooms—is overcome by the HSIS’s valid measurement of the classroom activities of interest, teacher-child interactions and the level of academic content provided in Head Start classrooms. The second challenge involves
how analytically to examine the relationship between children’s participation in programs that possess particular traits (low or high levels of teacher-child interactions or of academic content) and their outcomes. We will elaborate on four aspects of this methodological challenge in order to motivate our choice of method. The four aspects are: those with high (or low) levels of exposure are selected; the treatment-group inducement of high or low levels of exposure is a concept that is not defined in the control group; use of propensity score matching alone to identify high or low level control group counterparts will introduce overfitting bias into impact estimates; and principal strata identification is relevant when the endogenous experience is usefully leveraged from within both experimental arms, whereas this analysis is interested in the experiences of treatment group members alone. We discuss each of these in turn.

First, comparing children with high levels of exposure to the focal measures to those with low levels of exposure is not useful because the two groups (those with high and those with low exposure) select into their relative exposure levels. This selection can involve a family’s personal characteristics, where parents choose a perceived “better” center over another in which to enroll their child. Or, it can involve the general community availability of centers whose traits are influenced accordingly by those community traits. Further, it can involve the center’s administration and their choices for how to configure and operate their programs. Regardless which of these mechanisms is at play—family, community or centers’ administrative traits—comparing those children who experience low levels of a given Head Start program trait to those children who experience high levels is riddled with selection bias. Although using multiple regression analysis can adjust for some traits that might influence children’s outcomes, it cannot overcome the problem of unobserved measures associated with those selection forces (Burchinal, et al., 2010). Even more advanced analytic techniques, such as instrumental variable estimation, cannot fully overcome this selection bias problem (Duncan, 2003).
With experimental data from the HSIS where we are not necessarily limited to comparing children with high levels of experience to certain program traits to children with low levels of exposure. Randomization ensures that the treatment and control groups are alike in all ways—both measurable and unmeasureable—such that the only systematic difference between them is that the treatment group gained access to Head Start. Despite this advantage, the second methodological challenge is that of the construct of interest—treatment-induced experience of high (or low) levels of certain program traits—is undefined in the control group. If this is thought of as an omitted variable problem (that the concept of a treatment program characteristic is missing in the control group), then one could imagine using instrumental variables as a potential solution, where the instrument is that of random assignment. Indeed, IV is commonly used with experimental evaluation data, and it is fitting to do so when the analysis’ main assumption—that of the “exclusion restriction,” which implies that the impact’s lone pathway is through randomization—is credible. This is reasonable in the situation of adjusting for no-shows and cross-overs (e.g., Angrist, Imbens & Rubin, 1996), but it is less reasonable when the endogenous factor of interest is more complex than simple treatment take-up.

This observation leads to our third methodological challenge: that of symmetric identification of the subgroups of interest, and the related, narrower issue of overfitting bias. A proposed solution to the problem of the endogenous factor being unobserved in the control group is that of using propensity score matching to model the treatment group’s endogenous traits and construct a propensity score that permit identifying control group counterparts that share the same profile of observed predictors of those endogenous traits. Example applications of this approach appear in Harknett (2006) and Schochet and Burghardt (2007). Those analysts predict subgroup membership among control group members and analyze the impact of being in that subgroup as the difference between treatment group members and their predicted control group
counterparts. The main problem with this approach is that of comparing treatment group “actuals” to control group “predicted.” The two groups are similar but—as is the case in any propensity score analysis—may differ on unobservable characteristics. Moreover, the group used for predicting (the treatment group in the examples noted above) will have a better model fit than the group within which the propensity score is needed (the control group, above). So, even if comparing predicted subgroups to one another, without engaging in out-of-sample prediction, overfitting bias in some unknown direction and magnitude will arise (Abadie, Chingos & West, 2014; Harvill, Peck & Bell, 2013; Kemple & Snipes, 2000); and it is avoidable.

Conducting an analysis within the framework of principal stratification (Frangakis & Rubin, 2002) is a desirable solution to the challenge of endogenous subgroups, but, as noted above, it is more useful when the endogenous experience is observed in both the treatment and control arms. In that case, the potential treated outcomes of those in the control group who experience the post-randomization event, and the potential control outcomes of those in the treatment group who experience it, are compared in order to understand the relative influence of the endogenous experience. Some settings are relevant for this approach. For example, Page (2012) examines how treatment and control exposure to the “world of work” experiences exist for both treatment and control cases, where the controls had access to such experiences but where the treatment group members had such access with higher intensity. Another example is that of Feller, Grindal, Miratrix, and Page (2016) who examine the HSIS data to consider how the type of alternative care setting is associated with variation in children’s vocabulary development. Both in the presence of a Head Start offer (treatment group) and in its absence (control group), parents make choices about what settings are best for their children, be it at home with a parent, with a friend or family member, or in a more formal center setting. What is distinctive about our endogenous subgroup analysis is that we are interested in the experiences of
treatment group members alone: that is we are interested in the \textit{treatment-induced} experiences of Head Start programs, and so this approach is not fully optimal. One can think of the approach we do take as a “one-sided” principal stratification analysis, where the treatment group experiences and their potential control outcomes help determine the principal effects of interest (Bein, 2013).

To analyze the role of teacher-child interactions and level of academic content in children’s development, we choose to use ASPES—analysis of symmetrically-predicted endogenous subgroups—to identify equivalent treatment and control group subsamples in each category. We believe it is fitting in this setting and overcomes some of the limits of other, related potential analyses (though of course coming at a cost of its own assumptions). ASPES is a technique introduced to the literature by the authors (e.g., Bell & Peck, 2013; Harvill, Bell & Peck; Peck, 2003, 2013, 2015a, 2015b) as a general tool for learning how groups defined by post-random assignment events in experimental evaluations differ in their program impacts.

ASPES is one of a few methods for exploring the effects of programs and policies on endogenous subgroups (e.g., Peck, 2015a; Solmeyer & Constance, 2015). As elaborated above, ASPES is akin to other related methods—such as principal stratification (e.g., Feller, et al., 2016) and principal score analysis (e.g., Zhai, Brooks-Gunn & Waldfogel, 2014)—that have been recently used to examine variation in Head Start impacts, and ASPES allows us to conduct our analyses within the context of the HSIS experimental design. The next section details ASPES steps along with specifics of the classroom traits and outcome measures we use and the originating data source, the HSIS.
METHODOLOGY

The Head Start Impact Study

The congressionally-mandated Head Start Impact Study randomly assigned newly entering three- and four-year-old children in a nationally representative set of Head Start sites to either (1) a Head Start group given access to Head Start services or (2) control group that could not participate in Head Start for a year (but whose children could receive any other non-Head Start services available in their communities chosen by their parents; see Puma et al., 2005). Under this randomized design, a simple comparison of outcomes for the two groups—treatment and control—yields an unbiased estimate of the impact of the initial year of access to Head Start 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, making observed differences in subsequent outcomes unambiguous evidence of the impact of program access.

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. 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 three-year-olds and 2,108 four-year-olds.

The study collected data from parents, children, teachers, and other care providers. While the data cover outcomes through third grade, we consider the Head Start years alone as the most proximate to the Head Start classroom factors of interest to our analysis—and as the years in which the program’s influence in general is known to have been strongest and therefore has the
greatest chance of varying meaningfully from one set of classrooms to another. The HSIS data include 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. Follow-up data 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.

**Measurement of Program Traits**

We examine two dimensions of Head Start programs in this study: the nature of the interactions between teachers and children within the Head Start classroom, and the degree to which the content of instruction was academically focused. These measures capture distinct dimensions of what the field considers to be “quality” and we choose to use each of them, independently, to analyze the ways in which children’s Head Start experiences influence their developmental outcomes. We discuss the specific operationalization of each of these constructs below.

**Teacher-Child Interactions**

We use an index computed from 31 variables to measure the level of teacher-child interactions in Head Start classrooms. Eight of these variables come from the Early Childhood Environment Rating Scale (ECERS) and 23 from the Arnett Caregiver Interaction Scale. The eight ECERS elements include the following: encouraging children to communicate, developing reasoning skills, and teacher-child interactions, for example. Each of these could range from 1 to 7 in value. The Arnett elements include the following characteristics of these 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 recode them to put them on a 1-to-7 range and make them comparable for averaging with the ECERS items. We define high levels of teacher-child interactions as an average score of 6 or higher and low as an average score below 6.

**Academic Content**

Academic content considers the frequency of academically-focused activities that children experience in the classroom. That 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 the week, counting, playing math games, working with rulers and measuring cups, for example. Each of the items within this scale can range from 1 to 7, and our aggregate measure is an average of all the items on this scale those. Those average scores of 6 or greater are identified as having high levels of academic content (recoded as 2) by this measure, and those with a lower average score are identified as having low levels of academic content (recoded as 1).

Exhibit 1 summarizes these two classroom measures. It shows that about 17 percent of the three-year-old cohort and 23 percent of the four-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 Head Start, for even one day. Among those who did attend Head Start, 72 percent of the three-year-old cohort and 79 percent of the four-year-old cohort experienced high levels of teacher-child interactions.
A smaller proportion of each cohort—27 percent and 25 percent, respectively—experienced high levels of academic content.

**Exhibit 1. Descriptive Statistics of Three Head Start Traits among Treatment Group Members who Participated in Head Start, by Age Cohort**

<table>
<thead>
<tr>
<th>Head Start Treatment Group</th>
<th>Three-Year Old Cohort</th>
<th>Four-Year Old Cohort</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Number</td>
<td>Percent</td>
</tr>
<tr>
<td>Never participated in HS</td>
<td>243</td>
<td>16.6</td>
</tr>
<tr>
<td>Participated in HS</td>
<td>1,223</td>
<td>83.4</td>
</tr>
</tbody>
</table>

Among those who participated in HS…

**Teacher-Child Interactions** (range = 1-7)

- High (6+)
  - Three-Year Old Cohort: 764 (71.7%)
  - Four-Year Old Cohort: 617 (79.0%)
- Lower (<6)
  - Three-Year Old Cohort: 302 (28.3%)
  - Four-Year Old Cohort: 164 (21.0%)

**Academic Content** (range = 1-7)

- High (6+)
  - Three-Year Old Cohort: 278 (27.4%)
  - Four-Year Old Cohort: 188 (24.7%)
- Lower (<6)
  - Three-Year Old Cohort: 735 (72.6%)
  - Four-Year Old Cohort: 574 (75.3%)

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

With no field-accepted designated threshold for these measures, we chose the cutoff of 6. The teacher-child 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” levels of teacher-child interactions seems appropriate because of the distribution of resulting values on this measure: as Exhibit 1 shows, about 72 percent of the three-year-old cohort in the treatment group and 79 percent of the four-year-old cohort in the treatment group had high levels of teacher-child interactions. Because 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. The academic content measure draws from teacher reports and—like the teacher-child interaction measure—does not have a field-accepted designated threshold for what one might consider to be “high” levels of academic content. As a result, we chose the cut-point of 6 to create a relatively high bar and to maintain consistency with the teacher-child interaction classification scale.
Outcome Measures

Although the Head Start Impact Study explores many outcomes, for this analysis we examine five specific outcomes across two domains—pre-academic skills and social-emotional development. In the domain of pre-academic skills we include the PPVT, 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 and Total Child Behavior Problems. We choose these specific measures because they represent important early skills that are predictive of children’s success in school. These measures are also useful because they are consistently measured across all points of HSIS follow-up and represent outcome measures commonly used in the literature on the impact of preschool programs. The current analysis considers the Head Start years, which is when we know Head Start’s overall influence to be strongest and we therefore have the greatest chance of detecting impacts from varying levels of these traits of Head Start programs. Each of the selected outcome variables is detailed below.

Pre-academic Skills Domain

Within the pre-academic skills domain, we use the Peabody Picture Vocabulary Test (PPVT-III, adapted), a standardized measure children’s receptive vocabulary. The PPVT in an individually administered test in which the assessor says a word and the child is then asked to identify which of four pictures best represents the word. We use to Woodcock-Johnson III (WJ3) Letter-Word Identification as a measure of children’s “pre-reading skills”. This individually administered oral test of children’s reading skills asks children to identify letters and pronounce words from an increasingly difficult vocabulary list. Woodcock-Johnson III Applied Problems subset provides our measure of children’s “early math skills.” This measure asks children to
analyze and solve simple mathematical problems: listen to the problem, recognize the procedure to be followed, and perform relatively simple calculations.

**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; 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 (abbreviated simply as “social skills”), 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 program traits specifically on this subset of outcomes in the
pre-academic and socio-emotional domains because prior theory and evidence indicate these are
most proximally related to the programs levels of teacher-child interactions and academic
content.

**Analytic Approach**

As discussed earlier, comparing children in the Head Start treatment group with high
levels of teacher-child interactions or academic content with those in the treatment group with
low levels of those factors would involve children with different outcomes levels not caused by
their Head Start experiences and thus generate biased estimates of impact. So too would
comparison of any one of these treatment group subsamples to the full control group, which is
comprised of many different types of children. Unfortunately, the control group children who
comprise each of the three conceptually distinct subpopulations of interest here (who *would*
constitute valid counterfactuals for various treatment group subsamples of interest) cannot be
identified in the data and used separately in the analysis.

To avoid these problems and capitalize on the experimental design of the HSIS, we use
an approach established in Peck (2003) to create equivalent predicted 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. Because some misclassification of children is inevitable—e.g., predicting a child as likely to receive low academic content who in fact receives high academic content—we convert results for predicted subgroups to results for actual subgroups by making certain assumptions described below. We thereby—subject to the validity of the assumptions—translate internally valid impact estimates for predicted subgroups into externally valid—and more policy relevant—impact estimates for the actual subgroups of interest.

ASPES identifies sample members from the treatment and control groups predicted to participate in Head Start classrooms with high levels of teacher-child interactions or academic content in identical fashion, then estimates impacts on those matched subsamples as one would in any experimental subgroup analysis. The symmetry of the identification procedure ensures that equivalent subgroups are compared and guarantees that the resulting impact estimates are free from differential selection bias or other sources of internal bias. However, the subgroup for which the methodology produces unbiased impact estimates—children with the highest predicted probabilities of being in Head Start classrooms with high levels of teacher-child interactions or academic content—is not necessarily the subgroup of policy interest—children who actually experience Head Start with high levels teacher-child interactions or academic content. 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 convert results from impacts on predicted subgroups to impacts on actual subgroups subject to certain assumptions.

The following steps are involved in carrying out the ASPES analytic approach:

---

1 Further discussion of ASPES appears as a Method Note in Three Parts in Peck (2013), Bell and Peck (2013), and Harvill, Peck and Bell (2013).
1. Select random subsamples of the treatment group from which to predict the level of Head Start teacher-child interactions or academic content experienced by individual children in those samples.

2. Using baseline characteristics, predict levels of teacher-child interactions or academic content (and participation in Head Start in the first place).

3. Use the fitted predictive model to generate probabilities of receiving different types of Head Start experiences that symmetrically identify subsets of the treatment and control groups for matched analyses.

4. Analyze the impact of Head Start in each of the predicted subgroups by comparing mean outcomes between the symmetrically-derived treatment and control group subsamples.

5. Convert results for predicted subgroups to represent impacts on actual subgroups under certain assumptions.

   **Step 1.** Select random subsamples of the treatment group to predict Head Start levels of teacher-child interactions or academic content. 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 teacher-child interactions or academic are identified in both treatment and control groups.\(^2\,^3\) In order to ensure this subgroup symmetry, in this application, we select ten random

---

\(^2\) 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
90-percent subsets of the treatment group from the combined three-year-old and four-year-old cohorts for predictive modeling, as elaborated below.

**Step 2. Using baseline characteristics, predict program characteristics.** In this application, we create three distinct indicators for the traits of interest for all members of the three-year-old and four-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” level, among those who participated in the program; and a value of 2 represents a high” level Head Start, also among those who participated in the program. The specific threshold for dividing the trait into high and low is measure-specific, as defined in our measurement sub-section above. With this categorical measure as our dependent variable, we used a generalized logit procedure to predict no-show, low and high status on each of the traits, 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 Head Start program’s traits; we further allow this to proxy other community characteristics that 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 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.
might be associated with higher levels of teacher-child interactions or academic content.\(^5\) Other family- and child-level characteristics might also be associated with the traits 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 experiences 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 program trait-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 traits’ level variable to identify subgroups.** Within the sample, each individual is designated to a subgroup (no-shows, low level and high level of the trait) based on which category (0, 1 or 2) he or she has the highest probability of belonging to, given baseline characteristics.

\(^5\) To gauge the extent to which our assertion that “the center of random assignment would be the best predictor” of Head Start traits 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 4. Analyze the impact of levels of teacher-child interactions or academic content 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).

Step 5. Convert impacts for predicted trait subgroups to impacts on actual 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 to Peck and Bell (2014) elaborates on two alternative sets of assumptions in the Head Start application and the 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 level participants, and high level 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.

\[
I_N = s_NN_N + w_NL_N + g_NH_N
\]

\[
I_L = s_LL_L + w_LL_L + g_LH_L
\]

\[
I_H = s_HH_H + w_HL_H + g_HH_H
\]

where the following notation applies:

- \(I_N\) is the impact on predicted no-shows
- \(I_L\) is the impact on predicted low level participants
- \(I_H\) is the impact on predicted high level participants
- \(N_N\) is the impact on predicted no-shows who are actual no-shows
- \(N_L\) is the impact on predicted low level participants who are actual no-shows
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 level participants who are actual no-shows is zero
3. $N_H = 0$ – the impact on predicted high level participants who are actual no-shows is zero
(4) \(L_H = L_L\) – the impacts on low level participants are the same for children predicted to be high quality participants and children predicted to be low level participants.

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

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

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:

\[
L = \left( \frac{1 - r}{w_N + g_N} \right) I_N - \left[ \frac{(1 - r)g_N(w_H + g_H) + rg_H(w_N + g_N)}{(w_Hg_L - w_Lg_H)(w_N + g_N)} \right] I_L \]

\[
+ \left[ \frac{(1 - r)g_N(w_L + g_L) + rg_L(w_N + g_N)}{(w_Hg_L - w_Lg_H)(w_N + g_N)} \right] I_H
\]

\[
H = \left( \frac{1 - p}{w_N + g_N} \right) I_N - \left[ \frac{(1 - p)w_N(w_H + g_H) + pw_H(w_N + g_N)}{(w_Hg_L - w_Lg_H)(w_N + g_N)} \right] I_L \]

\[
+ \left[ \frac{(1 - p)w_N(w_L + g_L) + pw_L(w_N + g_N)}{(w_Hg_L - w_Lg_H)(w_N + g_N)} \right] I_H
\]

where

\(I - r\) is the proportion of low level participants who are predicted as no-shows; and

\(I - p\) is the proportion of high level 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 assumption (Bloom, 1984; Puma et al., 2005). Although these equations

---

6 Or whether one looks at high and low level cases predicted to be high level participants, once one combines this final assumption with the previous two assumptions to derive \(H_N - L_N = H_L - L_L\).

---
appear complex, they simply permit reallocating of the results pertaining to predicted subgroups to be interpreted for actual subgroups by using the information that is known to infer the information that is not.

**FINDINGS**

This section reports the estimated impacts of low and high levels of teacher-child interactions and academic content in Head Start classrooms on the skill development of children who actually experienced those inputs, subject to the assumptions described earlier. Like previous Head Start Impact Study reports involving subgroups, we discuss 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. In addition—given the limited potential for multi-year patterns to emerge in just one or two preschool years of follow-up (an issue not faced by the writers of the final published report of the HSIS, who also examined child outcomes in kindergarten, first grade, and third grade)—we also discuss impacts that we are confident (3) occur for more than one outcome measure in a particular domain of child development (pre-academic, social-behavioral) in a given year. 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 the above informal adjustments in selectively using the results.

Exhibit 2 reports impacts for the three-year-old cohort in the first and second Head Start years. Exhibit 3 reports results for the four-year-old cohort in their single Head Start year. By assumption, for children in the experimental treatment group who never participated in Head Start no impacts occurred ($N = 0$)—a finding provided at the top of every panel in both of the
exhibits. We focus here on significant findings for individual high and low subgroups and for differences in impact between subgroups.

Turning first to Exhibit 2, we see that for the children in the three-year-old cohort, the extent of teacher-child interactions in Head Start classrooms during their initial year in the program matter to the size of the program’s impact on the development of pre-academic skills but not on social-emotional skills. Greater levels of teacher-child interactions (the “High” subgroup category in the exhibit) produce favorable impacts on PPVT, WJ3 Letter-Word, and WJ3 Applied Problems in the first year of Head Start participation, with the WJ3 Letter-Word effect continuing into the second year. No statistically significant impacts are found for children attending Head Start classrooms with low teacher-child interactions. Most strikingly, in the second follow-up year the impact of Head Start is statistically significantly larger for children who attended high teacher-child interaction classrooms the prior year than for children who did not. These suggest that the level teacher-child interactions provided during the initial year facilitates additional skills development in the following year. The effect sizes that correspond to the statistically significant impacts in the pre-academic skills domains for high teacher-child interaction children (i.e., estimated impact divided by standard deviation of the outcome in the control group) range from 0.20 to 0.35. The differential impact of high rather than low teacher-child interactions for the WJ3 Letter-Word outcome in the second year—fully 16.2 points on that assessment measure’s scale—represents an effect size of 0.59, quite large relative to the existing literature on how much different developmental inputs affect cognitive progress.
Exhibit 2. Estimated Impacts on Pre-academic and Social-Behavioral Outcomes for the Three-Year-Old Cohort, by Teacher-child Interaction Level, at the End of the First and Second Head Start Years (2003, 2004)

<table>
<thead>
<tr>
<th>Control Group Average</th>
<th>PPVT</th>
<th>WJ3 Letter-Word</th>
<th>WJ3 Applied Problems</th>
<th>Social Skills</th>
<th>Problem Behaviors</th>
</tr>
</thead>
<tbody>
<tr>
<td>(standard deviation)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No-shows</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>(34.3)</td>
<td>(27.4)</td>
<td>(27.4)</td>
<td>(1.8)</td>
<td>(3.6)</td>
</tr>
</tbody>
</table>

End of First Head Start Year

<table>
<thead>
<tr>
<th></th>
<th>PPVT</th>
<th>WJ3 Letter-Word</th>
<th>WJ3 Applied Problems</th>
<th>Social Skills</th>
<th>Problem Behaviors</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>7.6</td>
<td>9.7</td>
<td>5.4</td>
<td>0.0</td>
<td>-0.6</td>
</tr>
<tr>
<td>Low</td>
<td>11.2</td>
<td>5.0</td>
<td>4.8</td>
<td>-0.1</td>
<td>-0.4</td>
</tr>
<tr>
<td>Difference</td>
<td>-3.5</td>
<td>4.7</td>
<td>0.6</td>
<td>0.1</td>
<td>-0.1</td>
</tr>
</tbody>
</table>

End of Second Head Start Year

<table>
<thead>
<tr>
<th></th>
<th>PPVT</th>
<th>WJ3 Letter-Word</th>
<th>WJ3 Applied Problems</th>
<th>Social Skills</th>
<th>Problem Behaviors</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>4.8</td>
<td>7.5</td>
<td>5.1</td>
<td>0.4</td>
<td>-0.9</td>
</tr>
<tr>
<td>Low</td>
<td>-4.3</td>
<td>-8.8</td>
<td>-8.3</td>
<td>-0.4</td>
<td>1.0</td>
</tr>
<tr>
<td>Difference</td>
<td>9.0</td>
<td>16.2</td>
<td>13.4</td>
<td>0.8</td>
<td>-2.0</td>
</tr>
</tbody>
</table>

Notes: Impact is estimated as the regression-adjusted difference between the treatment and control group mean outcomes for the number of points on a given child assessment scale. In some instances reported “Difference” in impacts does not equal reported “High” impact minus reported “Low” impact due to rounding.

- No statistical significance noted because no-show impact estimates are derived by assumption (to be zero).
- **statistically significant: p<0.05; * statistically significant: p<0.10**

A second interesting story emerges from these results (see Exhibit 3), which examines the role of academic content in the first Head Start year on concurrent and subsequent year impacts from the program. As can be seen, lower levels, not higher levels, of this input at age three prove beneficial to children. For example, WJ3 Letter-Word scores for the three-year-old cohort show favorable impacts from low academic content in both pre-school years—the latter year’s effect implying that scaling back academic content when one has three-year-olds in Head Start classrooms produces favorable impacts on language development in the first year (including also vocabulary skills that year, as measured by the PPVT) in that year and equips children to continue to acquire letter-word skills more rapidly in the subsequent year as a further benefit of Head Start participation. Neither of these results are produced when high academic content is provided to three-year-olds. The effect sizes that correspond to these statistically significant impacts range from 0.21 to 0.33. These effects are somewhat higher than those reported in a recent meta-analysis of early childhood education programs (Duncan & Magnuson, 2013)
Exhibit 3. Estimated Impacts on Pre-academic and Social-Behavioral Outcomes for the Three-Year-Old Cohort, by Academic Content, Level at the End of the First and Second Head Start Years (2003, 2004)

<table>
<thead>
<tr>
<th></th>
<th>PPVT</th>
<th>WJ3 Letter-Word</th>
<th>WJ3 Applied Problems</th>
<th>Social Skills</th>
<th>Problem Behaviors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control Group Average</td>
<td>251.4</td>
<td>307.6</td>
<td>373.6</td>
<td>12.4</td>
<td>6.2</td>
</tr>
<tr>
<td>(standard deviation)</td>
<td>(34.3)</td>
<td>(27.4)</td>
<td>(27.4)</td>
<td>(1.8)</td>
<td>(3.6)</td>
</tr>
<tr>
<td>No-shows</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
</tbody>
</table>

End of First Head Start Year

<table>
<thead>
<tr>
<th>Level</th>
<th>PPVT</th>
<th>WJ3 Letter-Word</th>
<th>WJ3 Applied Problems</th>
<th>Social Skills</th>
<th>Problem Behaviors</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>12.5</td>
<td>7.2</td>
<td>2.5</td>
<td>0.0</td>
<td>1.2 *</td>
</tr>
<tr>
<td>Low</td>
<td>7.1</td>
<td>9.1 **</td>
<td>6.2</td>
<td>-0.1</td>
<td>-1.1 ***</td>
</tr>
<tr>
<td>Difference</td>
<td>5.4</td>
<td>-1.9</td>
<td>-3.7</td>
<td>0.1</td>
<td>2.4 **</td>
</tr>
</tbody>
</table>

End of Second Head Start Year

<table>
<thead>
<tr>
<th>Level</th>
<th>PPVT</th>
<th>WJ3 Letter-Word</th>
<th>WJ3 Applied Problems</th>
<th>Social Skills</th>
<th>Problem Behaviors</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>2.4</td>
<td>-3.7</td>
<td>6.6</td>
<td>0.5</td>
<td>0.8</td>
</tr>
<tr>
<td>Low</td>
<td>2.6</td>
<td>5.7 *</td>
<td>-0.4</td>
<td>0.1</td>
<td>-0.9</td>
</tr>
<tr>
<td>Difference</td>
<td>-0.1</td>
<td>-9.5</td>
<td>7.0</td>
<td>0.4</td>
<td>1.7</td>
</tr>
</tbody>
</table>

Notes: Impact is estimated as the regression-adjusted difference between the treatment and control group mean outcomes for the number of points on a given child assessment scale. In some instances reported “Difference” in impacts does not equal reported “High” impact minus reported “Low” impact due to rounding.

* 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

In addition, the low academic content approach to Head Start services also produces a more desirable impact on one of two social-emotional outcome indicators. We note that reducing children’s Problem Behaviors is a good thing; in the first Head Start year low academic content achieves this whereas high academic content in fact increases the number of behavior problems that year. Neither of these effects carries over to the subsequent year, however. But while they exist the opposite consequences of low versus high academic content produce a differential impact of 2.4 problem behaviors per child, compared to a control group mean of 6.2 problem behaviors. This represents an effect size of 0.67 and nearly a 40 percent reduction in the number of problem behaviors achieved on average.

Next, Exhibits 4 and 5 report impacts for both high and low teacher-child interactions and high and low academic content in the one preschool year that children from the four-year-old cohort participated in Head Start. Effects broken out by the level of teacher-child interaction provided in children’s Head Start classrooms that year look much like those seen previously for the three-year-old cohort. Specifically, children receiving high levels of teacher-child interaction
experience positive impacts on PPVT and WJ3 Letter-Word outcomes while those receiving low levels of teacher-child interactions do not. As with the three-year-old cohort, no notable findings emerge for the two social-behavioral outcomes considered.

The picture regarding academic content *reverses* that seen earlier for the younger cohort. For children entering Head Start at age four, *high* academic content is shown to be beneficial in the pre-academic skills domain while low academic content is not. In particular, both PPVT and WJ3 Letter-Word scores are improved by Head Start participation when academic content is high but by Head Start classrooms with low content. Remarkably, the *difference* in impact magnitude for WJ3 Letter-Word is fully 27.8 points on a scale where control group children average 325.5 points—a 9 percent swing and a differential effect size of 0.98. So it seems that if Head Start participation is to be delayed to an older age—4 rather than 3—letter-word and vocabulary acquisition will be best advanced by providing *high* academic content during the first year of children’s participation rather than *low* academic content as works best for the younger Head Start entrants.

Effect sizes corresponding to the statistically significant impacts on pre-academic skill development in Exhibits 4 and 5 range from 0.21 to 1.03. No statistically significant findings by subgroup emerge there in the social-behavioral realm.

<table>
<thead>
<tr>
<th></th>
<th>PPVT</th>
<th>WJ3 Letter-Word</th>
<th>WJ3 Applied Problems</th>
<th>Social Skills</th>
<th>Problem Behaviors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control Group Average</td>
<td>290.3</td>
<td>325.5</td>
<td>397.5</td>
<td>12.5</td>
<td>5.6</td>
</tr>
<tr>
<td>(standard deviation)</td>
<td>(35.9)</td>
<td>(28.5)</td>
<td>(24.0)</td>
<td>(1.8)</td>
<td>(3.8)</td>
</tr>
<tr>
<td>No-shows</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
</tbody>
</table>

*Teacher-Child Interaction Level*

<table>
<thead>
<tr>
<th>Level</th>
<th>PPVT</th>
<th>WJ3 Letter-Word</th>
<th>WJ3 Applied Problems</th>
<th>Social Skills</th>
<th>Problem Behaviors</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>7.4</td>
<td>8.8</td>
<td>5.3</td>
<td>0.0</td>
<td>-0.3</td>
</tr>
<tr>
<td>Low</td>
<td>3.8</td>
<td>7.0</td>
<td>-0.3</td>
<td>-0.4</td>
<td>-0.3</td>
</tr>
<tr>
<td>Difference</td>
<td>3.5</td>
<td>1.8</td>
<td>5.6</td>
<td>0.5</td>
<td>0.1</td>
</tr>
</tbody>
</table>

*Notes:* Impact is estimated as the regression-adjusted difference between the treatment and control group mean outcomes for the number of points on given child assessment scale. In some instances reported “Difference” in impacts does not equal reported “High” impact minus reported “Low” impact due to 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. Estimated Impacts on Pre-academic and Social-Behavioral Outcomes for the Four-Year-Old Cohort, by Academic Content Level, at the End of the Head Start Year (2003)

<table>
<thead>
<tr>
<th></th>
<th>PPVT</th>
<th>WJ3 Letter-Word</th>
<th>WJ3 Applied Problems</th>
<th>Social Skills</th>
<th>Problem Behaviors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control Group Average</td>
<td>290.3</td>
<td>325.5</td>
<td>397.5</td>
<td>12.5</td>
<td>5.6</td>
</tr>
<tr>
<td>(standard deviation)</td>
<td>(35.9)</td>
<td>(28.5)</td>
<td>(24.0)</td>
<td>(1.8)</td>
<td>(3.8)</td>
</tr>
<tr>
<td>No-shows</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
</tbody>
</table>

*Academic Content Level*

<table>
<thead>
<tr>
<th>Level</th>
<th>PPVT</th>
<th>WJ3 Letter-Word</th>
<th>WJ3 Applied Problems</th>
<th>Social Skills</th>
<th>Problem Behaviors</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>14.5</td>
<td>29.4</td>
<td>-5.0</td>
<td>-0.3</td>
<td>0.0</td>
</tr>
<tr>
<td>Low</td>
<td>4.2</td>
<td>1.6</td>
<td>7.2</td>
<td>0.0</td>
<td>-0.4</td>
</tr>
<tr>
<td>Difference</td>
<td>10.3</td>
<td>27.8</td>
<td>-12.1</td>
<td>-0.2</td>
<td>0.4</td>
</tr>
</tbody>
</table>

*Notes:* Impact is estimated as the regression-adjusted difference between the treatment and control group mean outcomes for the number of points on given child assessment scale. In some instances reported “Difference” in impacts does not equal reported “High” impact minus reported “Low” impact due to 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

**DISCUSSION & CONCLUSION**

This research contributes to a small but active and growing literature examining variation in Head Start impacts (see, Bitler, Hoynes & Domina, 2013; Feller et. al., 2016) noting that average program impact can mask important variation that exists among subgroups. This paper examines the impact of enrolling in Head Start classrooms with high levels of 1) teacher-child interactions and 2) academic content on the development of children’s pre-academic and social-emotional skills. Despite the importance to of understanding the role of these factors in
influencing children’s developmental progress, researchers examining the Head Start Impact Study have not previously sought to address these areas of program effectiveness. This is primarily because of the analytic challenges in (1) conceptually defining and selecting among the numerous dimensions of Head Start program experiences and inputs that might be thought of as “quality” with the potential to affect the program’s ability to support children’s development; (2) making the selected program facets to be investigated measurable with validity and reliability from the study’s data; and (3) determining impacts for children experiencing varying types of Head Start experiences along the dimensions of interest—observable in the experiment’s treatment group but undefined in its control group.

The methodology used in this paper effectively addresses all of these challenges. In addition to disaggregating “quality” into the policy-relevant program “traits” of teacher-child interaction and level of academic content, we use ASPES to minimize the influence of selection bias on impact estimates. Relative to simple high-low group comparisons, ASPES leverages the experimental design to identify symmetrically balanced treatment and control subgroups that are defined by exogenous characteristics but align with the endogenous program experiences of interest. Because we are interested in treatment-induced experiences, we do not use the control group’s experiences of these program traits (even though some children were enrolled in programs where teacher-child interaction and level of academic content could be measured). Perhaps future research might extend this work fully into the principal stratification framework in order to learn whether those treatment-excluded (control) experiences are useful in understanding the relationship between children’s program experiences and their outcomes.

Applying ASPES to the experimental HSIS evaluation data, we find some evidence that Head Start’s impact varies by levels of both teacher-child interaction and academic content. Overall, exposure to programs with high levels of teacher child interactions leads to generally
higher impacts on measures of pre-academic skills. Among children who begin Head Start at age three-years old, whether or not their Head Start classroom offered high or low levels of teacher-child interactions does not appear to lead to significant differences in the impacts during the child’s first year of enrollment. By contrast, enrolling in a classroom with high levels of academic content during the child's first year of enrollment leads to less favorable impacts on children’s problem behaviors and vocabulary development. During their first year of enrollment, three-year-old children in classrooms with high levels of academic content are rated as having substantially more behavior problems than those children in classrooms with low levels of academic content. The magnitude of this effect is sizable. Enrolling in a classroom with low levels of academic content was associated with a 40 percent reduction in problem compared to the control group mean of 6.2 problems and represents an effect size of 0.67.

As these three-year-old children progress into their second year of Head Start, enrolling in a classroom with high levels of teacher child interactions does appear to be beneficial for children's pre-academic skill development. Among children in their second year Head Start who entered at age 3, the impact difference between high and low academic content experiences is quite large: children’s early reading increased by 0.59 of a standard deviation more when teacher-child interactions are high than when they are low. The negative impact of being in a high academic content classroom for three-year-old children in their first year of Head Start is no longer apparent when the children are in their second year of Head Start. This suggests that although participating in a high academic content classroom may lead to challenges children in their three-year-old year, there do not appear to be continued negative impacts of this early exposure to a high academic content classroom.

Among children who enter Head Start at four years old, we do not observe any statistically significant differences between those who enroll in classrooms that offer high and
low levels of teacher child interactions. For these children, high levels of academic content does appear to produce stronger positive impacts on children’s pre-academic skills but not differential impacts of measures of children’s social behavior.

These have findings offer some potentially important implications for early childhood educators. Although high levels of teacher-child interactions appear to be beneficial for both ages of children, high levels of academic content appear to be beneficial to older children but detrimental to younger children. As expected and documented in prior work, high levels of teacher child interactions appear to be beneficial across a range of domains. This is consistent with prior research noting that the development of young children is supported through regular reciprocal interactions with warm and responsive caregivers (Shonkoff & Phillips, 2000). The current efforts in Head Start and other large-scale early care and education programs to monitor and support preschool teachers’ capacities to provide children with these sorts of teacher-child interactions is supported by the findings in this paper.

The findings related to academic content also provide much-needed guidance to the developers of preschool curricula, preschool teachers and program leaders. The findings suggest that providing three-year-old children with high levels of academic content may not be appropriate for supporting their early development. To the contrary, three-year-old children appeared to respond to these more academically oriented programs by exhibiting aggressive behaviors and challenges in concentration. By contrast, for four-year-old children enrolling in a more academically oriented program appears to be beneficial. This suggests that teachers should seek to differentiate the curriculum and activities for the younger and older students within the broader three to five year old continuum typically served by preschool programs. Based on the findings in this study it would appear that younger children would benefit from less academically oriented activities while these types of activities are more appropriate for older children.
References


Pianta, R. C., Barnett, W. S., Burchinal, M., & Thornburg, K. R. (2009). The effects of preschool education what we know, how public policy is or is not aligned with the evidence base, and what we need to know. Psychological Science in the Public Interest, 10(2), 49-88.


APPENDIX A: RESULTS OF PREDICTION PROCESS

As noted in the text, ASPES starts by predicting which individuals would not participate in Head Start or would experience low or high levels of either teacher-child interactions or focus on academic content in their Head Start programs. If there were perfect prediction, then the ultimate ASPES conversion step would be unnecessary. Our prediction is not perfect, but it is better than random, and so it is fitting to use the approach.

As explained in the Analytic Approach subsection’s Step 1, 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 level, and high level subgroups. Because we observe both the predicted and actual subgroup 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 two program measures that we use, following with an exhibit that presents the notation that identifies each of these elements for its use in the subsequent conversion process.

Exhibit A-1 cross-tabulates predicted teacher-child interactions in its rows by the actual subgroup measurement in its columns. The following percentages appear:

- Row percentages allocate members of a given predicted quality subgroup across actual quality categories: the top left entry in the exhibit indicates that 37.1 percent of predicted non-participants are actual non-participants; and

- Column percentages allocate members of a given actual subgroup across predicted categories: the top left bracketed entry indicates that 21.6 percent of actual non-participants are predicted as non-participants.

<table>
<thead>
<tr>
<th></th>
<th>Actual Non-Participant</th>
<th>Actual Low Level</th>
<th>Actual High Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predicted Non-Participant</td>
<td>37.1%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Predicted Low Level</td>
<td></td>
<td>21.6%</td>
<td></td>
</tr>
<tr>
<td>Predicted High Level</td>
<td></td>
<td></td>
<td>100%</td>
</tr>
</tbody>
</table>
Exhibit A-1. Predicted by Actual Teacher-Child Interactions

<table>
<thead>
<tr>
<th>Predicted Teacher-Child Interactions</th>
<th>Actual Teacher-Child Interactions</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No-show</td>
<td>Low</td>
</tr>
<tr>
<td>No-show</td>
<td>37.1</td>
<td>15.7</td>
</tr>
<tr>
<td></td>
<td>[21.6]</td>
<td>[8.5]</td>
</tr>
<tr>
<td>Low</td>
<td>13.5</td>
<td>57.4</td>
</tr>
<tr>
<td></td>
<td>[12.7]</td>
<td>[49.7]</td>
</tr>
<tr>
<td>High</td>
<td>16.6</td>
<td>12.1</td>
</tr>
<tr>
<td></td>
<td>[65.1]</td>
<td>[41.9]</td>
</tr>
<tr>
<td>Total</td>
<td>19.0</td>
<td>20.5</td>
</tr>
</tbody>
</table>

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 the model correctly predicted 21.6 percent of no-shows, 49.7 percent of those with a low level of teacher-child interactions, and 82.8 percent of those with a high level of teacher-child interactions. The “Overall” rows and “Total” columns indicate that the predicted distribution of cases among the three groups (11, 18 and 71 percent for each of the non-participant, low level and high level groups, respectively), is not wildly different from the actual distribution (of 19, 21 and 61 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 Exhibit A-2 shows, the correct prediction rate for academic focus measure is 51.4 percent for the high level subgroup, 79.7 percent for the low level subgroup, and 22.7 percent for no-shows. As with the teacher-child interaction measure, the overall distribution of subgroups is quite similar for the level of academic content: 12, 68 and 21 percent, respectively, across the actual no-show, low-, and high-level groups, and 20, 59 and 21 percent for the predicted subgroups. Overall, this presentation of the correct placement rates that the analysis achieves leads us to conclude that the rates are acceptable for advancing this method of analyzing the effects of Head Start program traits.
Exhibit A. Predicted by Actual Level of Academic Content

<table>
<thead>
<tr>
<th>Predicted Level of Academic Content</th>
<th>Actual Level of Academic Content</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>No-show</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td>38.3 [22.7]</td>
<td>47.0 [9.2]</td>
<td>14.6 [8.0]</td>
</tr>
<tr>
<td>Low</td>
<td>69.7 [79.7]</td>
<td>12.8 [40.6]</td>
</tr>
<tr>
<td>High</td>
<td>15.9 [16.9]</td>
<td>52.7 [51.4]</td>
</tr>
<tr>
<td>Total</td>
<td>19.6</td>
<td>59.1</td>
</tr>
</tbody>
</table>

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 the 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-3 to identify the elements from Exhibits A-1 and A-2 that are needed as inputs into the conversion formulae to compute the conversion factors themselves.

Exhibit A-3. Predicted by Actual Levels, Notational Information for Conversion

<table>
<thead>
<tr>
<th>Predicted Level of Program Trait</th>
<th>Actual Level of Program Trait</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>No-show</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td>s_N</td>
<td>(W_N)</td>
<td>(g_N)</td>
</tr>
<tr>
<td>(1-r)</td>
<td>((1-p))</td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td></td>
<td></td>
</tr>
<tr>
<td>s_L</td>
<td>(W_L)</td>
<td>(g_L)</td>
</tr>
<tr>
<td>(1-r)</td>
<td>q</td>
<td></td>
</tr>
<tr>
<td>High</td>
<td></td>
<td></td>
</tr>
<tr>
<td>s_H</td>
<td>(W_H)</td>
<td>(g_H)</td>
</tr>
</tbody>
</table>

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).
### Exhibit A-3. Predicted by Actual Level of Academic Content

<table>
<thead>
<tr>
<th>Predicted Level of Academic Content</th>
<th>Actual Level of Academic Content</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>No-show</td>
<td>No-show</td>
<td>11.6</td>
</tr>
<tr>
<td></td>
<td>[22.7]</td>
<td>14.6</td>
</tr>
<tr>
<td></td>
<td>[9.2]</td>
<td>8.0</td>
</tr>
<tr>
<td>Low</td>
<td>17.5</td>
<td>69.7</td>
</tr>
<tr>
<td></td>
<td>[60.4]</td>
<td>12.8</td>
</tr>
<tr>
<td></td>
<td>[79.7]</td>
<td>40.6</td>
</tr>
<tr>
<td>High</td>
<td>15.9</td>
<td>31.4</td>
</tr>
<tr>
<td></td>
<td>[16.9]</td>
<td>52.7</td>
</tr>
<tr>
<td></td>
<td>[11.0]</td>
<td>51.4</td>
</tr>
<tr>
<td>Total</td>
<td>19.6</td>
<td>59.1</td>
</tr>
<tr>
<td></td>
<td>[16.9]</td>
<td>21.3</td>
</tr>
<tr>
<td></td>
<td>100.0</td>
<td></td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Predicted Quality</th>
<th>Actual Quality</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No-show</td>
<td>Low</td>
</tr>
<tr>
<td>No-show</td>
<td>$s_N$</td>
<td>$W_N$</td>
</tr>
<tr>
<td></td>
<td>(1-r)</td>
<td>(1-p)</td>
</tr>
<tr>
<td>Low</td>
<td>$s_L$</td>
<td>$W_L$</td>
</tr>
<tr>
<td></td>
<td>q</td>
<td></td>
</tr>
<tr>
<td>High</td>
<td>$s_H$</td>
<td>$W_H$</td>
</tr>
</tbody>
</table>

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).