*Educational Evaluation and Policy Analysis*
*June 2019, Vol. 41, No. 2, pp. 145–163*
*DOI: 10.3102/0162373718820307*
*Article reuse guidelines: sagepub.com/journals-permissions*

# What Is in a Definition? The *How* and *When* of Special Education Subgroup Analysis in Preschool Evaluations

**Anna Shapiro**
**Christina Weiland**
*University of Michigan School of Education*

*There are unique challenges to estimating causal effects of preschool for students with special needs that have not received attention in the literature. We revisit the Head Start Impact Study (HSIS) to illustrate that when and how special needs is defined has implications for the internal validity of and interpretation of special needs subgroup impact estimates. We find that the treatment group in the HSIS was three percentage points more likely to be classified as special education (*SD *= 0.11,* p *< .001) at baseline, likely biasing the impact estimates for this subgroup. We also find that the estimated intent-to-treat effects of Head Start on cognitive and socioemotional measures are sensitive to subgroup definition.*

Keywords: *preschool evaluation, special education, measurement*

Six percent of the preschool-aged children enrolled in early childhood education programs in the 2013–2014 school year received special education services for a diagnosed disability, approximately 90% of whom were served in public programs (Brault, 2011; Chaudry & Datta, 2017; National Center for Education Statistics [NCES], 2016b). Some of these children enter school with a diagnosis while others, particularly those with disabilities tied to academic functioning, are referred for evaluation by their educators (Hebbler & Spiker, 2016). To date, more than 80 rigorous studies have examined the effect of preschool on typically developing children's kindergarten readiness, but only three have examined the effects of preschool on young children with special needs (Duncan & Magnuson, 2013). Due to this paucity of evidence and the policy goals of early intervention and diagnosis for children with disabilities, two different groups of early childhood experts recently emphasized the critical need for more studies that estimate the causal effects of preschool on children with special needs (Phillips, Johnson, & Weiland, 2017; Yoshikawa et al., 2013).

To inform the field's increasing focus on this group, we revisit data from one of the three studies to date that has estimated the impacts of public preschool on young children with special needs—the Head Start Impact Study (HSIS). We do so to draw attention to two previously overlooked issues in this area of research that should be taken into consideration in designing future studies—*when* special needs status should be collected for valid causal inference and *how* special needs should be defined. Like many early childhood program evaluations, the HSIS measured special needs after children were assigned to a study condition and used only one of several potential indicators of disability to determine special needs status.

The Head Start program, a federal early childhood health and education intervention for low-income children, is an ideal context in which to study these questions because of its strong emphasis on serving children with special needs and its well-established early screening and diagnosis procedures (Cooke, 1965; Zigler & Meunchow, 1992). At least 10% of center seats must be reserved for children with special needs, and Head Start centers are required to engage in recruitment activities to locate children with disabilities (Office of Child Development, 1975). Furthermore, Head Start evaluates all enrollees within 45 days of enrollment using health and developmental screenings, much like the majority of state-funded public preschool programs (Head Start Bureau, 2015; National Institute for Early Education Research, 2017). Conducting developmental screenings before children enter formal school settings generally is logistically challenging and early childhood programs such as Head Start play an important role in the early identification of disabilities. For these reasons, post-random assignment collection of special needs status is likely to be particularly problematic in this study context and also may characterize future public preschool program evaluations.

We find evidence that *when* the HSIS measured baseline special needs status—post-random assignment and several months into the beginning of the program—resulted in imbalanced treatment and control special needs subgroups that likely biased the estimated effects of the program for this group. We also find that *how* the special needs subgroup is defined matters. Whereas the original study used a single question from the parent survey to define special needs, we use multiple questions from the survey to construct three measures of special needs status that encompass a range of students, including those likely to be receiving services for a diagnosed disability and those who may not be officially diagnosed. We find suggestive evidence that the impact estimates on measures of literacy, language, numeracy, and externalizing behaviors are sensitive to which measure we use to define the subgroup.

Our findings have implications for the design of future research studies on the effects of preschool on young children with special needs and can be used to inform the analytic decisions required to generate internally valid, interpretable evidence for this understudied group of students. Our findings serve as an illustrative case study that may be more broadly useful to evaluations that measure the effects of educational interventions on other important student subgroups that are commonly defined by time-variant and/or schooling-dependent factors (such as English Language Learners). As more studies estimate heterogeneous program impacts for groups defined by malleable characteristics, careful consideration of when and how to measure subgroup membership is warranted.

## The "When" and "How" of Special Needs Measurement in the Early Childhood Literature

The question of *when* to measure special needs status presents a substantial challenge in studies of preschool, particularly as it relates to the internal validity of special education subgroup[1] estimates. To make causal inferences about program effects for a subgroup, the characteristics that could be influenced by treatment status must be measured prior to random assignment, ensuring that subgroup membership is exogenous to treatment status (Murnane & Willett, 2010). Unlike time-invariant characteristics, such as race or ethnicity, often used to examine heterogeneous treatment effects, special education status is both time-variant and dependent on schooling experiences. For this reason, it might be particularly important that special needs status is measured prior to treatment assignment.

Logistically, however, preassignment measurement of a school-related characteristic is a challenge for this age group. For one, preschool is often the first nonfamily care setting for young children, so administrative records that can be used in studies of older children to define preintervention special needs status may not be available. Furthermore, special education determination can take up to 120 days from referral and requires a team of professionals to evaluate a variety of data sources, many of which are collected in an academic setting (Individuals With Disabilities in Education Act [IDEA], 2004a). The relatively long timeline for special needs diagnosis and the likelihood that entry to preschool is the first opportunity for diagnosis for many students introduces two competing considerations. On one hand, an adequate preassignment measure of special needs status will be difficult to generate without ensuring that all participants have had time to be referred and evaluated for services. On the other hand, waiting for this period to pass before measuring special needs status may threaten the internal validity of the subgroup estimates if assignment to treatment (i.e., preschool) is likely to increase, in the short term, the likelihood of special education screening and diagnosis.

The challenge of *when* to measure special needs status also has implications for the external validity of produced subgroup estimates. Given the relatively long timeline for disability diagnosis, measures of special needs that use beginning of the year records of Individualized Education Plans (IEPs) are likely to underrepresent the population of students who will ultimately receive special education services during their preschool year. The earlier that special needs status is measured, the more likely it is that the special needs subgroup will represent students with earlier access to formal diagnoses prior to preschool who may be more advantaged (Bassok, Finch, Lee, Reardon, & Waldfogel, 2016) or students with physical or more severe disabilities that were identified prior to exposure to academic environments. Given that 75% of children receiving services in preschool are diagnosed with either speech/language impairment or developmental delay, which may not be detected until children spend some time in a preschool classroom, earlier measurement would generate estimates that may not be generalizable to the majority of the special education preschool population (National Center for Special Education Research [NCSER], 2006).

There are also external validity implications for *how* special needs are measured. How the group is defined will determine to whom a given study's special education subgroup estimates are likely to generalize, which is of particular importance when estimating effects for such a developmentally diverse group of students. Special education identification in a research study can be accomplished through the use of administrative records, parent surveys, and teacher reports, all of which have benefits and drawbacks. Most typically in recent evaluations, IEP status has been used to determine which students have special needs (Conyers, Reynolds, & Ou, 2003; Jenkins et al., 2006; Justice, Logan, Lin, & Kaderavek, 2014; Lipsey, Hofer, Dong, Farran, & Bilbrey, 2013; Muschkin, Ladd, & Dodge, 2015; Phillips, Gormely, & Anderson, 2016; Phillips & Meloy, 2012; C. Ramey, Campbell, et al., 2000; S. Ramey, Ramey, et al., 2000; Ricciuti, St. Pierre, Lee, Parsad, & Rimdzius, 2004; Weiland, 2016).

Although using IEP records has the benefit of indicating which students are legally guaranteed to receive specialized services, there are several reasons that the students receiving services under IDEA might not represent the true population of students with a disability. State-by-state variation in both the overall number of children receiving special education services and their disability classifications from prekindergarten through high school suggests that there are important differences in how schools, school districts, and states identify students with disabilities (Aron & Loprest, 2012; Hebbler & Spiker, 2016). In addition, there is a considerable body of literature that indicates that racial minority groups are either over- or underrepresented in special education, depending on the analytic strategy used and the study setting (Artiles, 2003; Harry & Klingner, 2006; Morgan et al., 2015; Morrier & Gallagher, 2012; Oswald, Coutinho, Best, & Singh, 1999; Sullivan & Bal, 2013). Therefore, if we are interested in understanding the impact of preschool attendance for children with disabilities, using IEP status will restrict us to the students with a diagnosed disability, which may in turn be influenced by differential selection into or access to special needs services.

Other preschool evaluations have used parent and teacher reports instead of IEP status, asking parents and educators to identify students with "perceived need" or "with a disability" based on secondary reporting from health and other educational professionals (Lazar et al., 1982; Madden, O'Hara, & Levenstein, 1984; McCarton et al., 1997; Morgan et al., 2015; Puma et al., 2005, 2010, 2012; Seitz, Rosenbaum, & Apfel, 1985). Despite the potential for a more inclusive special needs definition using parent or teacher measures, this approach also introduces greater fuzziness to the measure and may not be comparable across contexts. For example, how a parent or teacher survey is worded, or how much secondary information from health care professionals or additional caretakers the parent or teacher has access to when responding to the survey, could influence responses. To this point, in an analysis of the Special Education Elementary Longitudinal Study (SEELS), teacher and parent reports of primary disability were incongruous, with parents more likely to report medical disabilities and teachers more likely to focus on academic disabilities (Marder, 2009).

Notably, *when* and *how* special needs status was measured varies considerably across the three studies to date that have estimated the effects of preschool on children with special needs. The findings from these studies were promising, with positive impacts on literacy (Phillips & Meloy, 2012; Weiland, 2016), language (Bloom & Weiland, 2015; Weiland, 2016), math, and executive function skills (Weiland, 2016). The evaluations of public programs in Tulsa, Oklahoma, and Boston, Massachusetts, used a regression discontinuity design that required special needs to be measured, using administrative IEP records, at the end of kindergarten to ensure equal time for diagnosis in the treatment and control groups (Phillips & Meloy, 2012; Weiland, 2016). A sensitivity analysis in the Boston study found that the students who entered preschool with a diagnosed disability were more likely to have more severe disability classifications, which could explain why the Boston program appeared to benefit these children more than children diagnosed by the end of kindergarten (Weiland, 2016). The HSIS, which was a randomized control trial, used questions from a parent survey about disability taken in the fall of the first year of preschool (Bloom & Weiland, 2015; Puma et al., 2010).

Prior research in other areas of education shows that *how* a subgroup is defined can be nontrivial, particularly when there are multiple measures with strong theoretical reasons for being valid indicators of the characteristic of interest. For example, a re-analysis of a New York City voucher experiment found that the impact estimates for African American students were not robust to how race was defined, with the estimated effects sensitive to which parent's race was used to construct the subgroup (Krueger & Zhu, 2004). Similarly, a recent paper examining how income is usually measured in the education literature finds that that the relationship between income and standardized test scores varies with how the free and reduced price lunch variable is used as a proxy for socioeconomic status (Michelmore & Dynarski, 2017).

## Current Study

Given the lack of consensus around special education measurement in the early childhood program evaluation literature, we use data from the HSIS as a case study to explicitly study the policy and empirical implications of the *when* and *how* of special needs measurement in the preschool period, with the goal of informing these decisions in future studies. Specifically, we address two research questions:

**Research Question 1:** Did the post-random assignment timing of the baseline survey in the HSIS likely bias the estimates of the effects of Head Start for children with special needs? (*when*)

**Research Question 2:** Are the estimated effects of Head Start sensitive to *how* special needs is defined at baseline? (*how*)

## Method

### *Sample*

Study children were part of the HSIS, the first nationally representative impact study of the Head Start program, conducted between 2002 and 2008. The HSIS was conducted with a nationally representative sample of oversubscribed centers for which there was greater demand for seats than were available (~85% of all Head Start centers), and did not include tribal centers, programs serving migrant and seasonal farm workers, or Early Head Start centers (Puma et al., 2010). The present study uses the restricted-use file, which excludes study centers in Puerto Rico, resulting in 4,440 children in 351 Head Start centers.

The HSIS followed two cohorts of Head Start applicants, 3-year-old applicants and 4-year-old applicants, from the start of their first program year through the end of third grade. Within each oversubscribed center, children who applied were randomly offered a seat in the program. Due to imperfect compliance to assigned condition, 81% of treatment children enrolled in Head Start, as did approximately 12% of the control group. In total, approximately 45% of control group reported enrollment in either Head Start or another center-based program by the spring of the first intervention year. Therefore, the HSIS estimates the impact of attending Head Start against a variety of counterfactual conditions, including home care, private early childhood centers, and public preschool programs (Puma et al., 2010). Given our aims in the present study, we estimate only intent-to-treat (ITT) effects, meaning our estimates should be interpreted as the effect of an offer to attend Head Start.

Our sample selection process is summarized in Figure 1. For the *when* analysis (RQ1), we restricted the analysis to participants with baseline parent interview data, resulting in an overall sample of 3,577 students in 345 centers, or 80% of the original HSIS sample (Box B-1, Figure 1). On baseline demographic characteristics and baseline measures of our outcomes of interests, the *when* analytic sample is nearly identical to the original HSIS sample (see column B, Appendix A in the online version of the journal). For the *how* analysis (RQ2), we limited our analytic sample to students with complete baseline and follow-up measures at the end of the first Head Start year, the spring of first grade, and the spring of third grade (Box C, Figure 1) to allow us to better compare the sensitivity of the estimates across time by using the same sample across periods. We relaxed this restriction by conducting analyses separately at each time point for all students with outcome data and found the results were not meaningfully different (results available upon request).

Finally, we limited the *how* analysis to Head Start centers in which there was at least one student with and without a baseline need in both the treatment and control conditions (Box D, Figure 1). While restricting the analysis to complete randomization blocks limits the external validity of our analysis, we cannot compare treatment and control students with and without special needs unless there was at least one of each at baseline in a given center. Overall, the subgroup effects in the *how* analysis are estimated using approximately 55% HSIS participants within slightly less than three fourths of the centers in the original study (Box 2, 3, 4, 5, Figure 1). Like the *when* sample, the *how* samples are nearly identical to the original sample on observable baseline characteristics (see columns C–F in Appendix A in the online version of the journal).

### *Procedures*

We use baseline data from direct child assessments and parent interviews conducted in the fall of the first program year, and outcome data from child assessments and parent interviews collected in the spring of the first program year, the spring of first grade, and the spring of third grade. Trained assessors administered the direct child assessments one-on-one either in the child's main care setting or in the child's home in the preschool years, and in the home for the first- and third-grade follow-ups. At baseline, children were assessed in the child's primary language when possible. All outcome data were collected using the English assessment battery.

Parent interviews were conducted in October (~35% of the sample), November (~35%), or between January and February (~20%) of the first program year with the parent or primary caregiver. Notably, when the interview was conducted was not randomly determined by the researchers; later respondents were those who required follow-up after first contact (Puma et al., 2005). The interview was conducted in the child's home in either English or the language the parent was most comfortable speaking if possible and included questions about child and parent physical and mental health, child and parent educational experiences, and home experiences. The questions about student disability and IEP status used in the current study to determine special education status were taken from these parent interviews (Puma et al., 2005).

### *Measures*

For the *how* analysis (RQ2), we estimated effects of Head Start assignment on four outcome measures of children's early literacy, language, numeracy, and socioemotional development. The first, the Peabody Picture Vocabulary Test–III (PPVT) is a measure of receptive vocabulary in which children identify which picture represents a given word (Dunn & Dunn, 1997). The PPVT is a nationally normed measure that has been used widely in early childhood program evaluations, and has strong split-half and test–retest reliability (Dunn & Dunn, 1997). The HSIS used a shortened item response theory (IRT)–scored

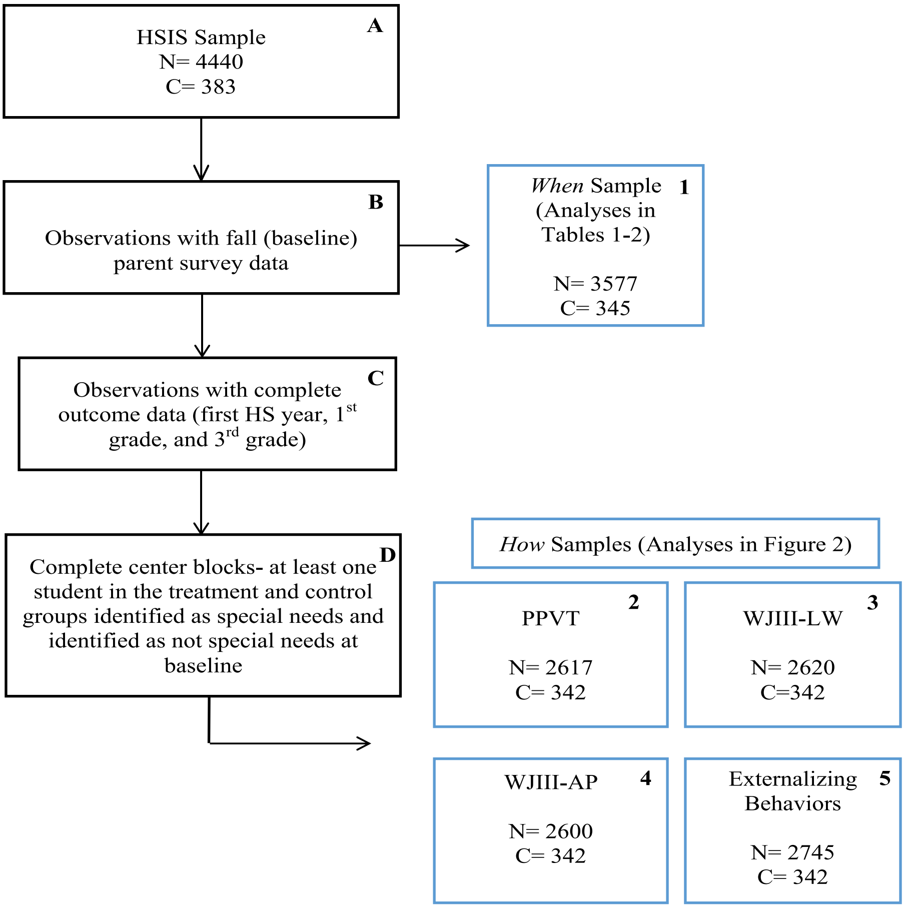

HSIS Sample
N= 4440
C= 383
A
B
Observations with fall (baseline) parent survey data
*When* Sample (Analyses in Tables 1-2)
N= 3577
C= 345
1
Observations with complete outcome data (first HS year, 1st grade, and 3rd grade)
C
Complete center blocks- at least one student in the treatment and control groups identified as special needs and identified as not special needs at baseline
D
*How* Samples (Analyses in Figure 2)
PPVT
N= 2617
C= 342
2
WJIII-LW
N= 2620
C=342
3
WJIII-AP
N= 2600
C= 342
4
Externalizing Behaviors
N= 2745
C= 342
5

FIGURE 1. *Sample selection.*

*Note.* HSIS = Head Start Impact Study; HS = Head Start; PPVT = Peabody Picture Vocabulary Test–III; WJIII-AP = Woodcock–Johnson III applied problems; WJIII-LW = Woodcock–Johnson III letter–word; N = participants; C = centers.

version of the PPVT, which we likewise use in our analyses.

The second and third measures are taken from subscales of the Woodcock–Johnson III assessment. The Woodcock–Johnson III is also widely used and nationally normed. The Letter–Word Identification subscale ($\alpha = .96$) asks children to identify and read individual letters and words fluently. The Applied Problems subscale ($\alpha = .90$) is a measure of early numeracy that requires children to solve arithmetic problems using simple calculations (Woodcock, McGrew, & Mather, 2001). We use the *W* scores (i.e., IRT-scored) from both subscales in our analyses.

The fourth measure is a measure of externalizing behavior using seven parent survey questions on the Child Behavior Checklist, which is commonly used to assess early childhood socioemotional development (Duncan et al., 2007; Raver et al., 2009). Parents were asked whether it was "very true," "true," or "not true" that their child displayed each of seven behaviors: temper tantrums, difficulty paying attention, restlessness, difficulty getting along with other children, fighting with other children, nervousness, and disobedience at home. Following the method used by Bloom and Weiland (2015), we generated a composite externalizing measure ($\alpha = .71$) for all students with answers to at least five of the seven questions. Scores on the measure range from 0 to 2. Children with composite scores closer to 2 had more "very true" or "true" answers to the questions than did students with scores closer to 0. For this reason, negative differences over time on this measure indicate a "positive" outcome.

*Covariates.* Using baseline data collected through the parent interview, we used a set of child and parent demographic characteristics following the original study team (Puma et al., 2010) and Bloom and Weiland (2015). The child characteristics are binary indicators of gender, race/ethnicity (Black, Hispanic, and White), whether the child speaks English at home, and whether the child is living with both biological parents. The parent characteristics are mother's age in years, binary indicators of educational attainment (less than high school, high school), mother's race/ethnicity, marital status, and an indicator of whether the mother was a teenager when she gave birth to the participant. We also include a measure of the child's age at the time the outcome measure was taken, baseline assessment scores for the outcome of interest, and a measure of the number of weeks elapsed between September 1, 2002, and the baseline assessment. We use multiple imputation (with 100 imputations) to address missingness in baseline covariates, which ranges from 0.2% and 5% depending on the covariate and analytic sample (Graham, 2009).

### *Special Needs Subgroup Construction*

Using the full HSIS sample, we constructed three measures of baseline special education status using questions from the parent survey conducted in the fall of the first year of the HSIS. All three measures were parent reported, which is consistent with similar measures taken in the national Early Childhood Longitudinal Study–Birth Cohort (ECLS-B) and Pre-Elementary Education Longitudinal Study (PEELS) surveys (National Center for Education Statistics [NCES], 2018; National Center for Special Education Research [NCSER], 2006). In the first question, parents were asked to report whether a doctor, health professional, or education professional had ever informed them that their child had any special needs or disabilities. This is the question that was used by the original HSIS team to define special needs. In a second follow-up question, parents were asked whether said professional had told them their child had any of 13 specific disabilities. Finally, parents were asked whether their child had an IEP or Individual Family Service Plan (IFSP) in place (see Appendix B in the online version of the journal for the text of the survey questions from the original study). Importantly, the IEP measure is a measure of parents reporting that a child has an IEP in place, not a measure of which students had an IEP in place as determined by administrative records.

The affirmative respondents to each of these three questions are neither the same children, nor subsets of the same group of children. While almost all those who responded "yes" to the doctor report also reported a disability classification, only 25% of doctor report "yes" respondents also reported having an IEP. And only 61% of the respondents who answered "yes" to having an IEP also answered "yes" to the doctor report. Although it is not clear whether these discrepancies are a function of reporting error or representative of true differences in incidence, the imperfect overlap across the questions suggests that affirmative answers to the three questions might capture different populations of students.

For this reason, we used the parent interview questions described above to construct three binary measures of special needs status at baseline. The first indicator, "doctor report," is the same definition used in the original HSIS and is set to one if the parent answered "yes" to the doctor or health professional report (Puma et al., 2010). The second measure, "IEP," is set to one if the parent answered "yes" to their child having an IEP or IFSP. The third measure, "any of the above," is set to one if the parent answered affirmatively to the doctor report, the specific disability question, or the IEP question to account for the broadest possible population of students with a disability. Given that the disability-specific report was intended to be a follow-up to the doctor report, we did not use the disability report independently as a special needs definition in our analyses. If respondents either refused to answer or answered "I don't know," we marked them as missing in subsequent analyses. As a result, four respondents who otherwise answered the parent interview questions were set to missing for the baseline IEP report. Across all analytic "how" samples, the "all of the above" measure includes the largest number of students ($N$ = 370–536) and the IEP measure includes the smallest number of students ($N$ = 87–115; see Appendix C in the online version of the journal).

### *Data Analytic Plan*

To answer our first research question assessing the implications of *when* the baseline survey in the HSIS was conducted, we compare within-center baseline means for each of our three special education measures in the treatment and control groups using the following specification:

$$Y_{ij} = \theta_0 + \theta_1 \left( \mathrm{HS}_{ij} \right) + \delta_j + \varepsilon_{ij}, \qquad (1)$$

where $Y_{ij}$ is an indicator of special needs at baseline for student $i$ in center $j$ for the relevant special needs definition; $\mathrm{HS}_{ij}$ is the treatment indicator, set to 1 if student $i$ was assigned an offer to attend Head Start center $j$ and 0 otherwise; $\delta_j$ is a vector of center fixed effects; and $\varepsilon_{ij}$ is the student-level error term.

To address our second question—whether the estimated effects of Head Start are sensitive to *how* special needs status is defined—we estimated an ITT effect of being assigned to attend Head Start for each of our three special education subgroups on measures of early language, numeracy, and socioemotional outcomes taken at the end of the first Head Start year, the end of first grade, and the end of third grade. To do so, we use the following linear regression model:

$$Y_{ij} = \beta_0 + \beta_1 \left( \mathrm{HS}_{ij} \right) + \gamma_{ij} + \delta_j + \varepsilon_{ij}, \qquad (2)$$

where $Y_{ij}$ is the relevant outcome for student $i$ in center $j$; $\mathrm{HS}_{ij}$ is the treatment indicator, set to 1 if student $i$ was assigned an offer to attend Head Start center $j$ and 0 otherwise; $\gamma_{ij}$ is vector of student and family baseline characteristics; $\delta_j$ is a vector of center fixed effects; and $\varepsilon_{ij}$ is the student-level error term. The coefficient $\beta_1$ represents the ITT effect of Head Start assignment on outcome $Y$.

### *Differences Between Our Approach and the Original Study*

The current study differs from the original HSIS team's work in several ways. First, we pool the two cohorts (e.g., 3-year-olds and 4-year-olds) of students, rather than analyzing the cohorts separately. We do this because participants were randomized at the center level without regard to age cohort and because 43% of the centers served both groups of children, with most children enrolled in mixed-aged classrooms (Bloom & Weiland, 2015). To account for the 3-year-old cohort experiencing up to 2 years of Head Start compared with up to 1 year for the 4-year-old cohort, we use outcomes measured at the end of the *first* Head Start year, the end of first grade, and the end of third grade. As the sample size for the special education subgroup is already very small, analyzing the two cohorts together improves our power to detect statistically significant differences between treatment and control students. Furthermore, the two cohorts did not differ in composition by special needs measures or by disability profile (see Appendix D in the online version of the journal).

Second, we do not use the HSIS sampling weights in our analyses. In the original HSIS study, the weights were used to make the study sample representative of the national population of newly entering Head Start children in 2002 (Puma et al., 2010). The weights were constructed separately for the 3- and 4-year-old cohorts at each collection point. Because we combine cohorts and restrict our sample to children with outcome data at all follow-up collections, the original weights are not clearly transferrable to our analytic samples and are not easily recreated with a limited re-analysis sample. Although we cannot directly examine whether our estimates would change with the inclusion of similarly constructed child weights, we find that removing the weights from the original study models does not greatly impact the magnitude of the impact estimates (results available upon request). However, the weights do increase the standard errors of the estimates, reducing the likelihood of detecting statistically significant results (Bloom & Weiland, 2015). Therefore, we interpret the statistical significance testing of our results with caution. Finally, we use multiple imputation to address missing baseline data on covariates other than special needs, whereas the HSIS team used hot deck imputation.

## Findings

### When *Analysis*

As the means in Table 1 demonstrate, the treatment group is three to four percentage points more likely to have a special need at baseline across our three definitions ($p < .05$), which indicates that there may be a substantial threat to the internal validity of special education subgroup analyses that arises from measuring special needs *after* treatment has begun. Notably, the treatment and control group are largely balanced on other baseline characteristics (see Appendix E in the online version of the journal); accordingly, the special education imbalance is likely not a function of random assignment but of post-assignment timing on defining a time-varying characteristic.

More concerning than the special education imbalance, there is considerable differential missingness in the special education variables as a result of who responded to the baseline parent survey in the fall of 2002, with treatment group members 11 percentage points less likely (15%) to be missing baseline parent interview data than the control group (26%). Taken with the overall missingness rate on the parent interview measures (~20% missing in the full HSIS sample), the magnitude of missingness between treatment and control ($SD = 0.27$) falls in the "unacceptable" category of differential attrition levels set by the What Works Clearinghouse, indicating potential for bias (2017).

Differential missingness in parent survey responses could have important implications for our estimates, particularly if families in the control group with less access to formal care were also less likely to respond to the parent interview. If this were the case, students in the control group would be underidentified for special education at baseline. Should this underidentification be correlated with other student or parent characteristics, we would be particularly concerned that the differential missingness coupled with the timing-induced imbalance might bias any special education subgroup estimates.

*Extending the "when" findings.* To better understand our *when* findings, we explored whether the fall imbalance between treatment and control groups reflected treated students having greater access to diagnosis than control students in the months between random assignment and the baseline survey. Accordingly, we reanalyzed the treatment and control balance in three ways. First, we compared the fall rates of special education using the "any of the above" definition, comparing treatment and control students who were enrolled in either a Head Start center or another center-based program at the fall collection period, which we refer to as "formal care," and then comparing treatment and control students in "informal care," which includes both family day care and children who stayed at home with a family member.

If the imbalance was generated through an access-to-diagnosis mechanism, we would expect that children in the treatment and control conditions enrolled in formal care would have similar baseline special education rates. From column 3 of the fall panel in Table 2, the overall difference in baseline special education status between treatment and control groups is approximately 4.07 percentage points ($p < .01$). However, when these differences are estimated separately by fall care setting in columns 4 and 5, the estimates are no longer statistically significant but the imbalance appears to be concentrated in the informal care setting (3.82 compared with 1.62 percentage points), lending support to an access-to-diagnosis effect in our *when* analysis.

Alternatively, we might expect that access to diagnosis would lead to differing rates of special education status only among children with disabilities most commonly diagnosed in educational settings. Some disabilities, including those disabilities most common in preschool, are more likely to be diagnosed when children begin school because these disabilities are closely tied to academic functioning (Markowitz et al., 2006). Students with these types of disabilities, which we refer to as "higher incidence" disabilities, are those whose parents reported "speech and/or language impairment," "developmental delay," "emotional disturbance," and "other disability/impairment." Other disabilities, such as physical impairments and severe cognitive disabilities, may be more likely to be diagnosed before entering the academic environment. Students included in this category, which we refer to as "physical/severe" disabilities, are those students who reported "orthopedic impairment," "visual impairment," "hearing impairment," "mental retardation (now intellectual disability)," "traumatic brain injury," and "autism." If access to diagnosis explains the fall imbalance, we would expect higher rates of differential special education diagnosis in the higher incidence category than in the physical/severe category.

To explore this hypothesis, in our second extension approach, we compared baseline equivalence in special needs status after splitting

TABLE 1

*Difference in Baseline Special Education Rates in the Treatment and Control Groups for the When Sample*

| | Treatment | Control | Difference | ES |
|---|---|---|---|---|
| Doctor report | 14.26 | 11.16 | 3.10*<br>(1.17) | 0.10 |
| IEP report | 6.37 | 3.70 | 2.67*<br>(0.77) | 0.14 |
| All of the above | 16.82 | 12.75 | 4.07*<br>(1.24) | 0.12 |

*Note.* All models included center fixed effects. Standard errors are in parentheses. The treatment mean is the unadjusted mean and the control mean was calculated by subtracting the estimated difference in means from the treatment mean. Fifteen percent of the treatment group and 26% of the control group were missing baseline parent survey data, including the special education measures. All participants with baseline survey data (3,577, the "when" sample as denoted in Figure 1) answered the question about doctor report. Of this sample of 3,577, four were missing an answer on the IEP report and three were missing answers on at least two of the questions and were not included in the "all of the above" measure. IEP = Individualized Education Plan; ES = effect size.
$^{\dagger}p < .10$. $^{*}p < .05$. $^{**}p < .01$. $^{***}p < .001$.

the group with reported disabilities in the fall into higher incidence and physical/severe disability categories. Looking at the physical/severe and higher incidence rows of the fall column of Table 2, several patterns support our hypothesis. First, the imbalance in all care settings appears to be concentrated in the higher incidence disability group (2.44 percentage points, $p < .05$, compared with 0.71 percentage points in physical/severe disability). Also, for the higher incidence disability group, the imbalance between treatment and control is larger in magnitude for the informal care setting group than for the formal care group, although not substantially so.

If likelihood of diagnosis is impacted by a child's access to formal early learning environments but diagnosis may be delayed for several months, then we might expect that any differences between treatment and control students found in the fall would be larger in magnitude in the spring. Following this logic, in our third extension approach, we examined the rate of special education definition in the treatment and control centers by care type and by disability type using the spring 2003 follow-up parent survey. Interestingly, in the spring column of Table 2, we see that while the overall imbalance between treatment and control decreases slightly in magnitude at the spring collection period (3.56 percentage points compared with 4.07 in the fall), the difference in higher incidence disability classification rates in the spring is highly concentrated in the informal care setting and is now marginally significant (4.73 percentage points, $p < .06$). Given that reported disabilities changed between the fall and spring collection period, and the difficulty in determining why these changes occurred, we see the spring analysis as only suggestively supporting the notion that greater time for diagnosis in formal care environments is a component of the treatment effect.

Finally, as an auxiliary to the three extension approaches, we separated our first two extension analyses by survey month, given that the parent interview was given over the course of several months after the program began (see Appendix F in the online version of the journal). We restricted the extension analysis sample to only those participants who answered the fall survey in October and November, as these months would have fallen within the 120-day timeline for referral, evaluation, and diagnosis. We find even larger differences in the fall and spring for all care (5.70 percentage points, $p < .001$; 5.56, $p < .001$) that are particularly concentrated in the informal care setting. We also controlled for survey month using the original extension sample and found slightly attenuated differences, although the patterns remain (see row 3 in Appendix F in the online version of the journal). As sample sizes were too small with these further restrictions, we did not conduct these survey month analyses within disability groups. The results of these analyses further suggest that the timing of the baseline survey impacted baseline special needs rates; the treatment control imbalance was concentrated in

TABLE 2

*Extending the When Findings—Rates of Special Needs Identification by Disability Type and Focal Care Arrangement in the Fall and Spring of the First Head Start Year*

| | Fall | | | Spring | | |
|---|---|---|---|---|---|---|
| | All care | Formal | Informal | All care | Formal | Informal |
| All disabilities | | | | | | |
| Treat | 16.32 | 16.62 | 12.74 | 15.44 | 15.68 | 13.30 |
| Control | 12.25 | 15.00 | 8.92 | 11.88 | 15.56 | 9.53 |
| Diff | 4.07** | 1.62 | 3.82 | 3.56** | 0.12 | 3.78 |
| | (1.23) | (1.87) | (1.46) | (1.19) | (1.74) | (2.86) |
| Physical/severe | | | | | | |
| Treat | 2.35 | 2.35 | 2.55 | 1.77 | 1.72 | <0.01 |
| Control | 1.64 | 2.76 | 0.90 | 1.63 | 1.76 | 1.53 |
| Diff | 0.71 | −0.41 | 1.65 | 0.14 | 0.04 | −1.62 |
| | (0.51) | (0.78) | (1.36) | (0.45) | (0.65) | (1.14) |
| Higher incidence | | | | | | |
| Treat | 11.39 | 11.58 | 8.92 | 10.57 | 10.55 | 10.73 |
| Control | 8.95 | 10.11 | 6.88 | 8.91 | 11.78 | 6.00 |
| Diff | 2.44* | 1.47 | 2.04 | 1.66 | −1.23 | 4.73$^{\dagger}$ |
| | (1.07) | (1.61) | (3.05) | (1.02) | (1.48) | (2.53) |

*Note.* Standard errors are in the parentheses. All models include center fixed effects. Treatments means are unadjusted and control means are calculated by subtracting the center-adjusted difference from the unadjusted treatment mean. We were underpowered to test whether the differences between care setting and disability type estimates were statistically significantly different from each other.

$^{\dagger}p < .10$. $^{*}p < .05$. $^{**}p < .01$. $^{***}p < .001$.

the earlier survey respondents, when Head Start attenders were likely all screened.

Taken together, this descriptive, exploratory extension of our *when* findings supports the notion that baseline imbalance between treatment and control groups may have been a result of the timing of the baseline survey. By waiting to take pretreatment measures until up to three months into first program year, the treatment group had greater access to diagnosis, which appears to have been translated into greater reporting of special needs. This interpretation is supported by our extension analyses. This pretreatment imbalance threatens the internal validity of any subgroup estimates for children with special needs, particularly if we think that early diagnosis of a special need is part of the treatment effect of early childhood programs.

### How *Analysis*

To answer our second research question—whether estimates are sensitive to *how* the special needs subgroup was defined—we first examined the balance in baseline demographic characteristics within the three special education subgroups in each analytic sample. As shown in Table 3, while the detected differences are not statistically significant, we find that regardless of definition used, the treatment groups were more likely to be male (6–16 percentage points), more likely to have mothers with less than a high school degree (8–16 percentage points), and more likely to have teen mothers (7–9 percentage points). In addition to these common differences, within the IEP definition, the treatment group members were also less likely to live with both biological parents (–35 percentage points), more likely to have married mothers (13 percentage points), and more likely to be Hispanic (six percentage points) than were the control group members.

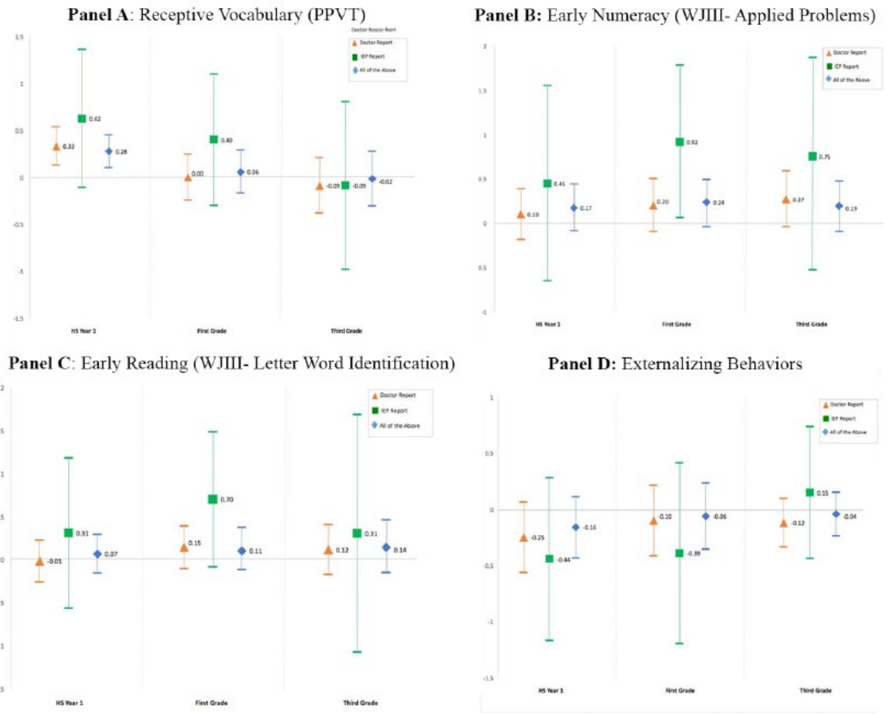
Digging deeper into Table 3, we also assessed how the treatment and control groups compare with one another across definitions to better understand how the three measures might be capturing different groups of students within the treatment and control groups (see Appendix G in the online version of the journal also for differences in

TABLE 3

*Differences Between Treatment and Control Groups on Baseline Demographic Characteristics and Baseline Assessments Within Special Education Subgroups for PPVT Outcome*

| | Doctor report | | | IEP report | | | All of the above | | |
|---|---|---|---|---|---|---|---|---|---|
| | T | C | Diff (*SE*) | T | C | Diff (*SE*) | T | C | Diff (*SE*) |
| Child characteristics | | | | | | | | | |
| % Male | 64.05 | 58.36 | 5.69 (7.58) | 67.02 | 50.92 | 16.10 (18.27) | 63.14 | 54.05 | 9.09 (6.89) |
| % Black | 21.07 | 21.40 | −0.32 (3.20) | 23.40 | 29.45 | −6.45 (11.68) | 21.17 | 22.81 | −1.64 (3.01) |
| % Hispanic | 32.23 | 33.23 | −1.00 (4.18) | 29.79 | 23.34 | 6.45 (14.73) | 31.75 | 33.01 | −1.26 (3.97) |
| % English at home | 77.27 | 81.38 | −4.11 (4.71) | 77.66 | 87.34 | −9.68 (12.59) | 75.55 | 79.80 | −4.25 (4.23) |
| % Living with both biological parents | 44.44 | 49.49 | −5.05 (8.11) | 50.00 | 14.50 | 35.50 (20.04) | 46.04 | 50.01 | −3.97 (7.41) |
| Receptive vocabulary (PPVT) | 252.06 | 257.58 | −5.52 (6.60) | 254.26 | 252.62 | 1.64 (17.26) | 251.06 | 255.45 | −4.39 (6.20) |
| Early numeracy (WJIII-AP) | 374.54 | 376.36 | −1.83 (5.03) | 372.17 | 376.63 | −4.46 (13.69) | 373.78 | 374.74 | −0.96 (4.59) |
| Early reading (WJIII-LW) | 298.92 | 298.38 | 0.54 (3.83) | 294.75 | 304.62 | −9.87 (12.16) | 297.85 | 296.38 | 1.47 (3.59) |
| Externalizing behaviors | 0.88 | 0.94 | −0.05 (0.08) | 0.88 | 1.09 | −0.22 (0.17) | 0.87 | 0.90 | −0.03 (0.07) |
| Mother characteristics | | | | | | | | | |
| Age | 30.11 | 29.68 | 0.43 (1.25) | 31.65 | 28.49 | 3.16 (3.15) | 30.34 | 30.05 | 0.29 (1.12) |
| Education—% less than HS | 38.84 | 29.99 | 8.85 (7.67) | 38.30 | 22.20 | 16.10 (16.50) | 38.69 | 29.72 | 8.97 (6.95) |
| Education—% HS | 31.82 | 42.32 | −10.50 (7.56) | 31.91 | 54.51 | −22.60 (20.18) | 33.21 | 42.41 | −9.20 (6.91) |
| % Black | 19.50 | 20.39 | −0.88 (3.34) | 22.34 | 28.79 | −6.45 (11.68) | 19.41 | 21.53 | −2.12 (3.11) |
| % Hispanic | 32.37 | 28.43 | 3.94 (4.29) | 29.79 | 7.19 | 22.60 (12.10) | 32.23 | 29.56 | 2.67 (3.89) |
| % Married | 39.58 | 47.71 | −8.13 (7.62) | 40.43 | 27.53 | 12.90 (18.22) | 40.07 | 49.11 | −9.04 (6.91) |
| % Previously married | 19.58 | 22.03 | −2.45 (6.87) | 19.15 | 44.95 | −25.80 (14.43) | 19.85 | 18.82 | 1.03 (6.12) |
| % Teen mother | 16.94 | 9.39 | 7.55 (6.16) | 15.96 | 6.28 | 9.68 (17.35) | 16.79 | 10.73 | 6.06 (5.58) |

*Note.* All models included center fixed effects. Standard errors are in parentheses. PPVT = Peabody Picture Vocabulary Test–III; IEP = individualized education plan; WJIII-AP = Woodcock–Johnson III applied problems; WJIII-LW = Woodcock–Johnson III letter–word; HS = high school.


Panel A: Receptive Vocabulary (PPVT)

Panel B: Early Numeracy (WJIII- Applied Problems)

Panel C: Early Reading (WJIII- Letter Word Identification)

Panel D: Externalizing Behaviors

FIGURE 2. *Estimated subgroup effects of Head Start on receptive vocabulary, early reading, early numeracy, and externalizing behaviors at the end of the first year of Head Start, end of first grade, and end of third grade for three special education subgroups.*

*Note.* Standard errors in parentheses. Missing covariates are imputed using multiple imputation. Models include center fixed effects and covariates (child's race, child's gender, mother's race, mother's marital status, mother's education, an indicator for whether the child lives with both biological parents, an indicator for whether the child's mother was a teen mother, age of child at time of outcome measurement, corresponding baseline scores for the outcome of interest, and the number of weeks between September 1, 2002, and the fall assessment). Effect sizes are calculated by dividing the point estimate by the standard deviation of the control mean for the full sample. PPVT = Peabody Picture Vocabulary Test–III; WJIII-AP = Woodcock–Johnson III applied problems; WJIII-LW = Woodcock–Johnson III letter–word.

baseline covariate means for the treatment and control groups, first comparing the doctor and all of the above groups [column A], then the doctor and IEP groups [column B], and finally the all of the above and IEP groups [column C]). Overall, we find that the IEP sample treatment and control groups differ from the doctor and all of the above treatment and control groups in a number of ways. Notably, these differences are particularly pronounced in the control group comparisons, suggesting that the IEP analytic sample captures a different treatment and control contrast compared with the other two measures.

Demographically, both IEP treatment and control groups were more likely to be Hispanic and less likely to be Black when compared with their doctor and all of the above counterparts. Interestingly, the treatment group was less likely to be male and the control group was more likely to be male using the IEP definition. The IEP treatment and control groups were also less likely to live with both biological parents and had mothers with lower educational attainment. Both of these descriptive analyses indicate that how the subgroup is defined not only changes which participants are considered to have a disability, but also changes the composition of the subgroup on other characteristics.

Keeping these underlying subgroup differences across definitions in mind, Figure 2 summarizes the estimated effects of being assigned the opportunity to attend Head Start across outcomes and over time for each special education subgroup (see Appendix H in the online version of the journal for the point estimates for each subgroup reflected in the figure). Although only a few of the estimated effects are statistically significant—for example, end of Head Start year effects on language for the "doctor report" (*SD* = 0.33, $p < .01$) and "all of the above" (*SD* = 0.28, $p < .01$) subgroups; first grade effects on numeracy (*SD* = 0.92, $p < .05$) for the "IEP" subgroup—we find a consistent pattern across our results suggesting that how the subgroup is defined changes the interpretation of the findings. First, there are consistently larger and positive (*SD* = 0.31–0.92) ITT effects for the IEP definition across the cognitive outcomes with the exception of third-grade receptive vocabulary (*SD* = −0.09), and a pattern of diminishing effects that become negative by third grade on externalizing behaviors. In contrast, using the doctor report and all of the above definitions, there are smaller effects that fade over time on language (*SD* = 0.33 to −0.09), smaller but consistently positive effects on literacy and numeracy (*SD* = 0.01–0.15 to 0.10–0.27, respectively), and smaller but consistently positive effects on externalizing behaviors (*SD* = −0.25 to −0.04) that diminish in magnitude over time.

Overall, our results suggest that the IEP definition may capture students who may have benefited more from being assigned to Head Start, as demonstrated by the consistently larger estimate treatment effects across outcome and time point, while the doctor report and all of the above definitions produce subgroup estimates of relatively smaller magnitudes. However, the IEP estimates are also the least precisely estimated, and are not statistically significantly different from those estimates when using the other two definitions in most cases. Coupled with the findings in Table 3, however, we see the divergent pattern of results as indicative that the IEP definition captures a group of students who may have benefited from the program differently.

The pattern of findings within subgroup definition leads to substantively different conclusions about the long-term effects of the program. Were the special needs subgroup to be defined using IEP report, one might conclude that there is suggestive evidence of large positive effects that persist over time on cognitive measures but not on socioemotional outcomes. On the contrary, were the special needs subgroup to be defined using one of the two broader definitions, one might conclude the benefit of being assigned to Head Start, while still positive, is smaller overall, fades out over time for language, and becomes stronger over time for prosocial behaviors. We see this analysis, therefore, as suggestive evidence that how the estimated effect of the program is sensitive to how the special needs subgroup is defined.

## Discussion

More than 2.5 million preschool-aged children are diagnosed with a disability and receive special education services in public preschool programs, but few studies have examined the causal effects of early childhood education for these students (National Center for Special Education Research, 2006). Of the studies that have estimated effects for students with special needs, the findings are promising, although the approaches to measuring special needs status have varied (Phillips & Meloy, 2012; Puma et al., 2010; Weiland, 2016). In the present study, we used data from the HSIS to explore the internal and external validity implications of *when* baseline special needs are measured and *how* the subgroup is defined. By examining how the timing of the measurement impacts the characteristics of the special needs subgroup, and how program impact estimates compare when using different special education definitions, we are able to make better conclusions about what preschool impact estimates mean and for whom they are relevant. Given the important role that early childhood programs play in diagnosing disabilities and providing early intervention services to ameliorate the effects of learning and developmental delays, a better understanding of how to estimate program effects for this group of students is not only important from a methodological perspective but also critical for informing early childhood policy.

Our findings suggest that the composition of the special needs subgroup can change substantially depending on when special needs is measured and what variables are used to construct the subgroup. First, we find evidence that post-assignment measurement (*when*) generated imbalanced treatment and control special education subgroups, threatening the internal validity of these estimates. Across special education subgroup definitions, treatment students were three to four percentage points more likely to be identified as having special needs at baseline. We found that this imbalance may have been generated by an access to diagnosis effect of treatment status, as evidenced by our exploratory analyses of baseline balance in which the treatment and control differences appear to be more concentrated among students who were diagnosed with higher incidence disabilities and who experienced informal child care during the first program year. Second, we also found that *how* matters; the estimated ITT effects of being assigned to Head Start are larger for the IEP-defined special education subgroup than are the estimates using two broader definitions, including the definition used by the original study. The consistent pattern across the outcomes and over time suggests that how the special needs group was constructed could lead to substantively different interpretations of the impact of Head Start for children with disabilities.

Taken together, our findings demonstrate that decisions around when and how to measure special needs for the purpose of subgroup analyses in evaluations of preschool programs have implications for the internal validity and external validity of subgroup estimates. With regard to *when* to measure, identifying special needs prior to preschool entry is necessary for making causal inferences about a program's effects, particularly if treatment status impacts the likelihood that a student will be identified as having special needs during the preschool year. Alternatively, prioritizing the internal validity of the estimates by measuring disability status prior to assignment introduces an external validity trade-off because the earlier the special needs are measured, the more likely the subgroup will be a narrower cross-section of the students for whom we are interested in estimating program effects. Given that the majority of preschool-aged children have academic-related disabilities, which are more likely to be diagnosed after exposure to formal education, this limitation to the external validity of the estimates is not a minor concern.

Researchers conducting experimental and quasi-experimental studies of future public preschool programs accordingly face a balancing act in estimating impacts for preschool children with special needs. Ultimately, we recommend that researchers be clear on which possible special needs subgroup they are interested in studying and that analytic decisions-related special education measurement should be guided by the researcher's hypotheses about why students identified under a particular definition at a given time point might experience preschool programs differentially. If, for example, we hypothesize heterogeneous effects for students who receive special education services for a disability in preschool, a definition that allows for adequate time for diagnosis and relies on IEP records might be more appropriate. On the contrary, if we are more interested in impacts for students at risk for later diagnosis, or students who are perceived to be developing differently than their peers, a more inclusive measure might be warranted. In addition, future studies should formally assess the impact of any measurement decisions, both in terms of describing the characteristics of students included in the subgroup with each definition and the sensitivity of subsequent impact estimates. In so doing, researchers will be able to make more informed interpretation of their findings and to whom they may apply.

Future studies might also consider approaches not yet used in the literature but that are likely feasible in field-based trials. For example, rather than relying on administrative records that reflect who has had access to diagnosis at baseline, researchers might consider administering a consistent, research-validated screener for hearing, vision, and developmental delay to all program participants prior to random assignment. Researchers should also consider collecting and using referral and evaluation data from sources such as Child Find, an early identification program mandated by Part B of IDEA and carried out by local school districts, which would include disability status determined prior to preschool entry (IDEA, 2004b).

There are several limitations to the current study related to sample size, analytic sample selection, available data, and differential missingness in the baseline special education variables. First, although the HSIS provides a larger-than-typical sample of students with disabilities due to Head Start's long-standing emphasis on serving young children with disabilities, the special education subgroups are still small, particularly when using the IEP definition. The imprecision of the estimates for the special education subgroups limits our ability to detect statistically significantly different effects, even when the magnitudes of the estimated effects are large. For this reason, in the measurement sensitivity analysis, we focus on the pattern of the results for the special needs subgroups but do not make any substantive interpretations of the estimates. We also do not interpret the differences between the special education and nonspecial education groups. Second, limiting our *how* analysis to Head Start centers to complete randomization blocks also limits the external validity of our analysis, and the comparability of the *when* and *how* analyses. However, baseline equivalence tests from the full sample and our trimmed analytic samples indicate that this selection process did not change the average observable characteristics of the treatment and control groups.

We are also limited by having to rely on parent reports to determine IEP status rather than administrative records. Although this approach is common in national data collections such as ECLS-B and PEELS, many current studies use administrative records. Future work comparing parent reports of IEPs with administrative records would shed light on the reliability of parent reports to collect this information. Finally, our analysis is limited by the differential missingness between treatment and control groups on the parent interview survey used to construct the baseline special education measures, which suggest a potential for biased treatment effects generated by underreported diagnosis in the control group.

Despite these limitations, the present study adds to the field by highlighting largely overlooked internal and external validity implications of special education measurement decisions in early childhood program evaluations. Rather than attempt to identify an ideal measure of baseline special needs, this article presents evidence that these decisions matter both for which students the special education subgroup includes and the magnitude of the estimates associated with the special education subgroup. In addition, the measurement decisions will impact how the estimates should be interpreted and to whom they will generalize. As future studies include subgroup estimates for children with special needs, considerable attention should be paid to aligning measurement decisions with hypotheses on how a program might be expected to produce different impacts for children with disabilities, and efforts should be made to test the sensitivity of special education subgroup estimates to these measurement decisions.

## Acknowledgments

We thank the Inter-university Consortium for Political and Social Research (ICPSR, 2018) for providing the data used in this study. Thanks to Matthew Diemer, Brian Jacob, Donald Peurach, Matthew Ronfeldt, Kevin Stange, Howard Bloom, participants in the Causal Inference in Education Research Seminar, and members of the Equity in Early Learning Lab at the University of Michigan for their feedback.

## Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

## Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This study is funded by the Institute of Education Sciences R305B150012.

## Note

1. In this article, we use the term *subgroup* to refer to all students with disabilities, not to distinguish between disability classifications.

## ORCID iD

A. Shapiro https://orcid.org/0000-0001-9312-1300

## References

Aron, L., & Loprest, P. (2012). Disability and the education system. *The Future of Children*, *22*, 97–122. doi:10.1353/foc.2012.0007

Artiles, A. J. (2003). Special education's changing identity: Paradoxes and dilemmas in views of culture and space. *Harvard Educational Review*, *73*, 164–202. doi:10.17763/haer.73.2.j78t573x377j7106

Bassok, D., Finch, J., Lee, R., Reardon, S., & Waldfogel, J. (2016). Socioeconomic gaps in early childhood experiences: 1998 to 2010. *AERA Open*, *2*(3), 1–22. doi:10.1177/2332858416653924

Bloom, H., & Weiland, C. (2015). *Quantifying variation in Head Start effects on young children's cognitive and socio-emotional skills using data from the National Head Start Impact Study* New York, NY: MDRC. Retrieved from https://www.mdrc.org/sites/default/files/quantifying_variation_in_head_start.pdf.

Brault, M. (2011). *School-aged children with disabilities in U.S. metropolitan statistical areas: 2010* (American Community Survey briefs). U.S. Census Bureau. Retrieved from https://www.census.gov/library/publications/2011/acs/acsbr10-12.html

Chaudry, A., & Datta, A. R. (2017). *The current landscape for public pre-kindergarten: The current state of scientific knowledge on pre-kindergarten effects*. Washington, DC: Brookings Institution. Retrieved from https://www.brookings.edu/wp-content/uploads/2017/04/duke_prekstudy_final_4-4-17_hires.pdf

Conyers, L., Reynolds, A., & Ou, S.-R. (2003). Effect of early childhood intervention and subsequent special education services: Findings from the Chicago Child-Parent Centers. *Educational Evaluation and Policy Analysis*, *25*, 75–95. Retrieved from http://www.jstor.org/stable/3699518

Cooke, R. (1965, February 19). *Recommendations for a Head Start program by panel of experts chaired by Dr. Robert Cooke Johns-Hopkins University*. Washington, DC. Retrieved from https://eclkc.ohs.acf.hhs.gov/sites/default/files/pdf/cooke-report.pdf

Duncan, G. J., Dowsett, C. J., Claessens, A., Magnuson, K., Huston, A. C., Klebanov, P., . . . Japel, C. (2007). School readiness and later achievement. *Developmental Psychology*, *43*, 1428–1446. doi:10.1037/0012-1649.43.6.1428

Duncan, G. J., & Magnuson, K. (2013). Investing in preschool programs. *Journal of Economic Perspectives*, *27*, 109–132. doi:10.1257/jep.27.2.109

Dunn, L. M., & Dunn, L. M. (1997). *Peabody Picture Vocabulary Test* (3rd ed.). Bloomington, MN: Pearson Assessments.

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology*, *60*, 549–576. doi:10.1146/annurev.psych.58.110405.085530

Harry, B., & Klingner, J. K. (2006). *Why are so many minority students in special education? Understanding race and disability in schools*. New York, NY: Teachers College Press.

Head Start Bureau. (2015). *Head Start program facts fiscal year 2015*. Washington, DC: U.S. Department of Health and Human Services. Retrieved from https://eclkc.ohs.acf.hhs.gov/sites/default/files/pdf/head-start-fact-sheet-fy-2015.pdf

Hebbler, K., & Spiker, D. (2016). Supporting young children with disabilities. *The Future of Children*, *25*, 185–205.

Individuals With Disabilities in Education Act of 2004, Pub. L. No. 108-446 Sec. 303.300-346 (2004a).

Individuals With Disabilities in Education Act of 2004, Pub. L. No. 108-446 Sec. 655(a)(5) (2004b).

Jenkins, J., Dale, P., Mills, P., Cole, K., Pious, C., & Ronk, J. (2006). How special education preschool graduates finish: Status at 19 years of age. *American Educational Research Journal*, *43*, 737–781. doi:10.3102/00028312043004737

Justice, L. M., Logan, J. A. R., Lin, T.-J., & Kaderavek, J. N. (2014). Peer effects in early childhood education: Testing the assumptions of special-education inclusion. *Psychological Science*, *25*, 1722–1729. doi:10.1177/0956797614538978

Krueger, A., & Zhu, P. (2004). Another look at the New York City school voucher experiment. *American Behavioral Scientist*, *47*, 658–698. doi:10.1177/0002764203260152

Lazar, I., Darlington, R., Murray, H., Royce, J., Snipper, A., & Ramey, C. (1982). Lasting effects of early education: A report from the consortium for longitudinal studies. *Monographs of the Society for Research in Child Development*, *47*(2/3), 1–151. Retrieved from http://www.jstor.org/stable/1165938

Lipsey, M., Hofer, K., Dong, N., Farran, D., & Bilbrey, C. (2013). *Evaluation of the Tennessee voluntary prekindergarten program: Kindergarten and first grade follow-up results from randomized control design*. Nashville, TN: Peabody Research Institute. Retrieved from https://my.vanderbilt.edu/tnprekevaluation/files/2013/10/August2013_PRI_Kand1stFollowup_TN-VPK_RCT_ProjectResults_FullReport1.pdf

Madden, J., O’Hara, J., & Levenstein, P. (1984). Home again: Effects of the mother–child home program on mother and child. *Child Development*, *55*, 636–647. doi:10.2307/1129975

Marder, C. (2009). *Facts from SEELS: Perspectives on students’ disability classifications*. Office of Special Education Programs, U.S. Department of Education. Retrieved from https://www.seels.net/info_reports/DisabilityClassif1.9.09.pdf

Markowitz, J., Carlson, E., Frey, W., Riley, J., Shimshak, A., Heinzen, H., . . . Lee, H. (2006). *Preschoolers’ characteristics, services, and results: Wave 1 overview report from the Pre-Elementary Education Longitudinal Study (PEELS)*. Rockville, MD: Westat. Retrieved from http://www.peels.org/

McCarton, C., Brooks-Gunn, J., Wallace, I., Bauer, C., Bennett, F., Berbaum, J., . . . Meinert, C. (1997). Results at age 8 years of early intervention for low-birth-weight premature infants: The Infant Health Development Program. *Journal of the American Medical Association*, *277*, 126–132. doi:10.1001/jama.1997.03540260040033

Michelmore, K., & Dynarski, S. (2017). The gap within the gap: Using longitudinal data to understand income differences in educational outcomes. *AERA Open*, *3*, 1–18. doi:10.1177/2332858417692958

Morgan, P., Farkas, G., Hillemeier, M., Mattison, R., Maczuga, S., Li, H., & Cook, M. (2015). Minorities are disproportionately underrepresented in special education: Longitudinal evidence across five disability conditions. *Educational Researcher*, *44*, 278–292. doi:10.3102/0013189X15591157

Morrier, M., & Gallagher, P. (2012). Racial disparities in preschool special education eligibility for five southern states. *The Journal of Special Education*, *44*, 152–169. doi:10.1177/0022466910380465

Murnane, R., & Willett, J. (2010). *Methods matter: Improving causal inference in educational and social science research*. Oxford, UK: Oxford University Press.

Muschkin, C., Ladd, H., & Dodge, K. (2015). Impact of North Carolina's early childhood initiatives on special education placements in third grade. *Educational Evaluation and Policy Analysis*, *37*, 478–500. doi:10.3102/0162373714559096

National Center for Education Statistics. (2016a). *ECLS-B Preschool National Study: Parent interview*. Retrieved from https://nces.ed.gov/ecls/pdf/birth/preschool_parent_interview.pdf

National Center for Education Statistics. (2016b). *Table 204.70: Number and percentage of children served under Individuals with Disabilities Education Act (IDEA), part B, by age group and state or jurisdiction: Selected years, 1990-91 through 2013-14*. Digest of Education Statistics. Retrieved from https://nces.ed.gov/programs/digest/d15/tables/dt15_204.70.asp?current=yes

National Center for Education Statistics. (2016c). *Table 202.20: Percentage of 3-, 4-, and 5-year-old children enrolled in preprimary programs, by level of program, attendance status, and selected child and family characteristics: 2014*. Retrieved from https://nces.ed.gov/programs/digest/d15/tables/dt15_202.20.asp

National Center for Special Education Research. (2006). *Preschoolers with disabilities: Characteristics, services, and results: Wave 1 overview report from the Pre-Elementary Education Longitudinal Study (PEELS)*. Retrieved from https://ies.ed.gov/pubsearch/pubsinfo.asp?pubid=NCSER20063003

National Institute for Early Education Research. (2017). *The state of preschool 2016*. Retrieved from http://nieer.org/wp-content/uploads/2017/09/Full_State_of_Preschool_2016_9.15.17_compressed.pdf

Office of Child Development. (1975, July). *Head Start program performance standards* (OCD-N-30-364-4). Washington, DC: U.S. Department of Health, Education, and Welfare. Retrieved from https://files.eric.ed.gov/fulltext/ED122936.pdf

Oswald, D. P., Coutinho, M. J., Best, A. M., & Singh, N. (1999). Ethnic representation in special education: The influence of school-related economic and demographic variables. *The Journal of Special Education*, *32*, 194–206. doi:10.1177/002246699903200401

Phillips, D., Gormely, W., & Anderson, S. (2016). The effects of Tulsa's CAP Head Start program on middle-school academic outcomes and progress. *Developmental Psychology*, *52*, 1247–1261. doi:10.1037/dev0000151

Phillips, D., Johnson, A., & Weiland, C. (2017, August). *Public preschool in a more diverse America: Implications for next-generation evaluation research* (Poverty Solutions at the University of Michigan Working Paper Series #2-17). Retrieved from https://poverty.umich.edu/research-publications/working-papers/preschool-diverse-america/

Phillips, D., & Meloy, M. (2012). High-quality school-based pre-k can boost early learning for children with special needs. *Exceptional Children*, *78*, 471–490. doi:10.1177/001440291207800405

Puma, M., Bell, S., Cook, R., Heid, C., & Lopez, M. (2005). *Head Start Impact Study: First year findings*. Washington, DC: U.S. Department of Health and Human Services, Administration for Children and Families. Retrieved from https://www.acf.hhs.gov/sites/default/files/opre/first_yr_finds.pdf

Puma, M., Bell, S., Cook, R., & Heid, C. (2010). *Head Start Impact Study: Final report*. Washington, DC: U.S. Department of Health and Human Services, Administration for Children and Families. Retrieved from https://www.acf.hhs.gov/sites/default/files/opre/hs_impact_study_final.pdf

Puma, M., Bell, S., Cook, R., Heid, C., Broene, P., Jenkins, F., . . . Downer, J. (2012). *Third grade follow-up to the Head Start Impact Study: Final report* (OPRE Report No. 2012-45). Washington, DC: Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services. Retrieved from https://www.acf.hhs.gov/sites/default/files/opre/head_start_report.pdf

Ramey, C., Campbell, F., Burchinal, M., Skinner, M., Gardner, D., & Ramey, S. (2000). Persistent effects of early childhood education on high-risk children and their mothers. *Applied Developmental Science*, *4*, 2–14. doi:10.1207/S1532480XADS0401_1

Ramey, S., Ramey, C., Phillips, M., Lanzi, R., Brezausek, C., Katholi, C., & Snyder, S. (2000). *Head Start children's entry into public school: A report on the National Head Start/Public School Early Childhood Transition Demonstration Study*. Washington, DC: Administration for Children and Families, U.S. Department of Health and Human Service. Retrieved from https://www.acf.hhs.gov/sites/default/files/opre/transition_study.pdf

Raver, C. C., Jones, S. M., Li-Grining, C. P., Zhai, F., Metzger, M., & Solomon, B. (2009). Targeting children's behavior problems in preschool classrooms: A cluster-randomized controlled trial. *Journal of Consulting and Clinical Psychology*, *77*, 302–316. doi:10.1037/a0015302

Ricciuti, A. E., St. Pierre, R. G., Lee, W., Parsad, A., & Rimdzius, T. (2004). *Third national even start evaluation: Follow-up findings from the experimental design study*. Washington, DC: Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, U.S. Department of Education. Retrieved from https://ies.ed.gov/ncee/pdf/20053002.pdf

Seitz, V., Rosenbaum, L., & Apfel, N. (1985). Effects of family support intervention: A ten-year follow-up. *Child Development*, *56*, 376–391. doi:10.2307/1129727

Sullivan, A. L., & Bal, A. (2013). Disproportionality in special education: Effects of individual and school variables on disability risk. *Exceptional Children*, *79*, 475–494. doi:10.1177/001440291307900406

Weiland, C. (2016). Impacts of the Boston prekindergarten program on the school readiness of young children with special needs. *Developmental Psychology*, *52*, 1763–1776.

United States Department of Health and Human Services, Administration for Children and Families, Office of Planning, & Research and Evaluation. (2018, February 8). Head Start Impact Study (HSIS), 2002-2006 [United States] (ICPSR 29462). Ann Arbor, MI: Inter-university Consortium for Political and Social Research. Retrieved from https://doi.org/10.3886/ICPSR29462.v7

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock–Johnson® III Tests of Achievement*. Itasca, IL: Riverside.

Yoshikawa, H., Weiland, C., Brooks-Gunn, J., Burchinal, M. R., Espinosa, L. M., Gormely, W. T., . . . Zaslow, M. J. (2013). *Investing in our future: The evidence base on preschool education*. Society for Research in Child Development. Retrieved from https://www.fcd-us.org/assets/2016/04/Evidence-Base-on-Preschool-Education-FINAL.pdf

Zigler, E., & Meunchow, S. (1992). *Head Start: The inside story of America's most successful educational experiment*. New York, NY: Basic Books.

## Authors

ANNA SHAPIRO is a PhD candidate at the University of Michigan School of Education. Her research areas include early childhood education program evaluation, special education policy evaluation, and the effects of early intervention programs for children with disabilities.

CHRISTINA WEILAND is an assistant professor at the University of Michigan School of Education. Her research focuses on the effects of early childhood interventions and public policies on children's development, especially for children from low-income families.

Manuscript received May 23, 2018
First revision received September 13, 2018
Second revision received November 20, 2018
Accepted November 21, 2018