AERA Open
April-June 2018, Vol. 4, No. 2, pp. 1–16
DOI: 10.1177/2332858418769287
© The Author(s) 2018. http://journals.sagepub.com/home/ero

New Findings on Impact Variation From the Head Start Impact Study: Informing the Scale-Up of Early Childhood Programs

Pamela A. Morris
New York University
Maia Connors
Ounce of Prevention Fund
Allison Friedman-Krauss
Rutgers University
Dana Charles McCoy
Harvard University
Christina Weiland
University of Michigan
Avi Feller
UC Berkeley
Lindsay Page
University of Pittsburgh
Howard Bloom
MDRC
Hirokazu Yoshikawa
New York University

This article synthesizes findings from a reanalysis of data from the Head Start Impact Study with a focus on impact variation. This study addressed whether the size of Head Start's impacts on children's access to center-based and high-quality care and their school readiness skills varied by child characteristics, geographic location, and the experiences of children in the control group. Across multiple sets of analyses based on new, innovative statistical methods, findings suggest that the topline Head Start Impact Study results of Head Start's average impacts mask substantial variation in its effectiveness and that one key source of that variation was in the counterfactual experiences and the context of Head Start sites (as well as the more typically examined child characteristics; e.g., children's dual language learner status). Implications are discussed for the future of Head Start and further research, as well as the scale-up of other early childhood programs, policies, and practices.

Keywords: *Head Start, impact variation, early childhood*

There is a strong evidence base that preschool programs for low-income children have the potential to improve their short- and longer-run outcomes and address income-related development gaps (Barnett, 1995, 2011; Dodge, Bai, Ladd, & Muschkin, 2016; Gormley, Phillips, & Gayer, 2008; Weiland & Yoshikawa, 2013; Yoshikawa et al., 2013). In addition, a number of studies highlighted the returns to society of such early investments (Barnett & Masse, 2007; Belfield, Nores, Barnett, & Schweinhart, 2006; Campbell et al., 2012; Elango, Garcia, Heckman, & Hojman, 2016; Garcia, Heckman, Lead, & Prados, 2016; Reynolds, Temple, Robertson, & Mann, 2002; Schweinhart et al., 2005). But whether those longer-run positive impacts can emerge in programs delivered at scale is still an open question. That is, our ability to extrapolate directly from smaller-scale efficacy studies—of either the value of preschool itself or enhancements to preschool quality among children already attending preschool—is complicated by the fact that testing programs under tightly controlled conditions may offer little guidance regarding the challenges of implementation within "real

Creative Commons Non Commercial CC BY-NC: This article is distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 License (http://www.creativecommons.org/licenses/by-nc/4.0/) which permits non-commercial use, reproduction and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access pages (https://us.sagepub.com/en-us/nam/open-access-at-sage).

world" city-, state-, or national-level contexts, characterized by more varied samples (Duncan & Magnuson, 2013; Pianta, Barnett, Burchinal, & Thornburg, 2009). Such a perspective highlights the value of research on larger "scaled up" preschool programs (see Durlak & Dupre, 2008; Granger, 2010; Shadish, Cook, & Campbell, 2002). Head Start is perhaps the oldest example of such a program, given its national scale, serving nearly 1 million children annually at a cost >$7 billion (Office of Head Start, 2015).

Head Start has been referred to as the nation's premier federally sponsored early childhood education program (Barnett, 1995; Lombardi, Harding, Connors, & Friedman-Krauss, 2016). Since 1965, it has provided comprehensive services to low-income preschool children and their families across the United States in an attempt to "narrow the gap" between disadvantaged children and their more affluent peers. If our goal is to meet the needs of large numbers of low-income children in preparing them for schooling, then Head Start is a pioneer as one of the largest, most comprehensive programs for 3- and 4-year-old children and a signature of the mid-1960s War on Poverty. Moreover, as a national program serving a diverse group of children across the country, Head Start offers a rare opportunity to understand the role of heterogeneity of program impacts—not only by child and family characteristics but also by neighborhood and state contexts, which until recently have received less attention and discussion in studies of preschool program effectiveness.

The largest and, arguably, most rigorous test of Head Start, a nationally representative randomized trial, was launched in the late 1990s. The Head Start Impact Study (HSIS; Puma, Bell, Cook, Heid, & Lopez, 2005; Puma et al., 2010a; Puma et al., 2012) was designed in response to a 1998 congressional mandate to provide a national estimate of Head Start's average impact on child outcomes and to explore for whom and under what circumstances its impacts are the greatest.

The HSIS confirmed and expanded our understanding of the impacts of Head Start (discussed later) but only scratched the surface on whether, how, and why impacts of Head Start might vary across the United States. In short, despite a long history of research on Head Start and some initial inquiry regarding the ways that program impacts might vary (particularly across groups of children and families), the field is only now beginning to explore what differentiates those programs, sites, and children for which Head Start is highly effective and those for which it is less so. In particular, information is still relatively limited on contextual characteristics and counterfactual experiences that may moderate the effects of a program such as Head Start. Findings reported here represent an important contribution to this body of literature.

Prior Research on Head Start's Effects

Head Start began with a strong commitment to research (Zigler & Styfco, 2010). Early descriptive studies showed Head Start–related gains on measures of cognitive achievement (Zigler & Muenchow, 1992). Subsequent quasi-experimental studies also generally demonstrated positive impacts on school achievement and attainment in the short and long term, suggesting that the program may be "working" and "cost-effective" (Currie & Thomas, 1995; Deming, 2009; Garces, Thomas, & Currie, 2002; Ludwig & Miller, 2007).

The design of the national HSIS has a number of important strengths relative to prior Head Start research, including its large sample (from 84 nationally representative delegate agencies), its rigorous random assignment design that tests the effects of Head Start versus alternate care arrangements, its inclusion of follow-up information on children through early elementary school, and its collection of outcomes from a range of developmental domains (social-emotional, cognitive, and health). As such, the HSIS addressed a critical question: What is the average effect of the offer of Head Start services on children's care experiences and developmental outcomes? The HSIS also aimed to address questions of variation in Head Start impacts (Puma et al., 2010a, p. xiii), but the study's lead story dominating policy discussions is largely about average impact.

First, the HSIS (Puma et al., 2010a) showed that children's care experiences were, perhaps not surprising, affected by the offer to enroll in Head Start: approximately 85% of children in the group assigned to Head Start (the treatment group) attended Head Start, as compared with <20% of children in the control group. What was perhaps more surprising is that half the control group members (and 90% of treatment group members) attended some form of center-based care setting (an early childhood care and education [ECCE] setting) rather than being cared for at home. Importantly, for children assigned to receive the offer of Head Start, most of their care was of high quality, with more than two-thirds of the treatment group in high-quality ECCE settings, as opposed to a quarter of children in the control group (based on one albeit-limited measure of quality).

Findings from the HSIS showed modest positive effects of Head Start on immediate cognitive and social-emotional outcomes (Puma et al., 2010a), which is consistent with prior Head Start research, including a recent meta-analysis of Head Start's short-term effects across earlier studies (Shager et al., 2013). HSIS results in the longer term (i.e., kindergarten through third grade) have been mixed: consistent with earlier work (Barnett, 1995; Deming, 2009), test scores for children attending Head Start and their control counterparts tend to converge over time in middle childhood (Puma et al., 2010a; Puma et al., 2012), although positive impacts on parenting practices persist through the early elementary grades (Gelber & Isen, 2013; Puma et al., 2012).

Prior research in the HSIS demonstrates some evidence for variation in impacts on child outcomes, most often exploring child and parent characteristics as moderators of that effect. For example, subgroup analyses in the original HSIS (Puma et al., 2010a) suggested that Head Start impacts may be larger for children who are dual language learners (DLLs),

have low baseline preacademic skills, have special needs, are from high-risk households, and live in nonurban settings. Subsequent analyses of HSIS data by other researchers (much of which was conducted simultaneous to the current set of studies) provide additional evidence of larger Head Start impacts among Spanish speakers and children with low baseline skills (Bitler, Hoynes, & Domina, 2014), children experiencing low and moderate levels of preacademic stimulation provided by parents at home (Miller, Farkas, Vandell, & Duncan, 2014), and children of mothers who were former Head Start participants themselves (Chor, 2016).

As we turn to questions about how impacts vary due to context, rather than individual characteristics, it is important to reemphasize that every impact is a *comparison*—between children in the treatment and those in the control group (see Figure 1). Because this difference is due to a *contrast* in the ECCE experiences of those two groups, the difference in outcomes that we observe as an "impact" is a function of the characteristics of the treatment group and the control group; any variation that we observe in these impacts may be due to variation in characteristics of the treatment group, the control group, or both.

Some studies have examined the implementation (or treatment) side of the impact comparison, while others focused on the control side of this comparison. For example, Head Start impacts on child outcomes were found to vary by characteristics of the Head Start center, with larger impacts in centers that offer full-day and home visiting services (Walters, 2015). Other analyses found that impacts vary by the ECCE experiences of children in the control group, with larger impacts among children who would not otherwise enroll in center-based ECCE (Kline & Walters, 2016; Zhai, Brooks-Gunn, & Waldfogel, 2014). All of this work was conducted at the same time of the current studies and similarly relied on methods for addressing these questions that were relatively new at the time. As such, the literature that focuses on the control side of the comparison is very much in its infancy, limiting the extent to which it has transformed the discourse regarding Head Start effects. This is unfortunate, given that one of the most common concerns about comparisons between the HSIS and older studies, such as Perry Preschool and the Abecedarian Program, is that the counterfactual has shifted (Duncan & Magnuson, 2013; Ludwig & Phillips, 2008). Without this empirical work, these discussions are relegated to well-reasoned hypotheses about study differences without the kind of empirical data that could speak to either the strength of those hypotheses or the magnitude of effects driven by differences in the counterfactual.

Present Article

In this article, we summarize key findings from the Secondary Analysis of Variation in Impacts (SAVI) Center, a multiyear collaboration intended to extend the findings of the HSIS to rigorously address questions about impact variation. We use an ecological framework (Bronfenbrenner, 1979) to explore impact variation. This provides a theoretical model to guide our understanding of sources of variation in Head Start impacts among child outcomes, as well as two primary hypothesized mechanisms of assignment to Head Start on outcomes for children: type and quality of care experiences. In addition to determining the amount of variation that exists across sites in impacts on enrollment and outcomes for children, we examine characteristics of the macrosystem (state policy), the exosystem (neighborhoods), the microsystem (counterfactual care arrangements, or the type of setting in which children received care when not assigned to receive Head Start), and the individual child that may explain some of this variation (DLL status and baseline skill level). Because the design of the original HSIS was optimized to detect average program impacts, not impact variation, our studies leverage the experimental evaluation to answer nonexperimental questions about these sources of impact variation.

To address these questions, we draw on methodological advances in estimating variation in program impacts across sites (Raudenbush & Bloom, 2015), in the science of linking implementation to impact (Bloom, Hill, & Riccio, 2003), and in principal stratification within randomized trials (Frangakis & Rubin, 2002; Page, Feller, Grindal, Miratrix, & Somers, 2015). By using multinomial logit approaches, we are also the first to account for missing data in the estimate of early childhood program impacts on program type and quality (Friedman-Krauss, Connors, & Morris, 2017). These methods aim to disentangle predictors of variation that are likely conflated but cannot definitively point to the "cause" of impact variation; doing so would require different research designs that intentionally manipulate these sources of variation. Moreover, limitations of the current methods do not allow all these sources of variation to be tested simultaneously, although we do attempt to understand the extent to which findings stand alongside one another versus explain one another, wherever possible.

This work focuses primarily on impacts after 1 year of Head Start. Although "fadeout" of initial impacts has been observed in evaluations of preschool programs (including Head Start), it is not well understood why, how, and under what conditions the outcomes of treatment and control groups converge over the elementary years. To advance our understanding of the potential impacts of Head Start, our goal is to understand for whom and under what conditions initial impacts are largest. Nevertheless, we acknowledge the need for future work to consider impact variation over time.

Although the SAVI Center's findings have been published in disparate outlets, this article represents our first attempt to examine these findings side by side to tell the "story of variation" in the HSIS in a single place and to explore how such findings extend our understanding of Head Start's impacts. Our goal is not to pit sources against

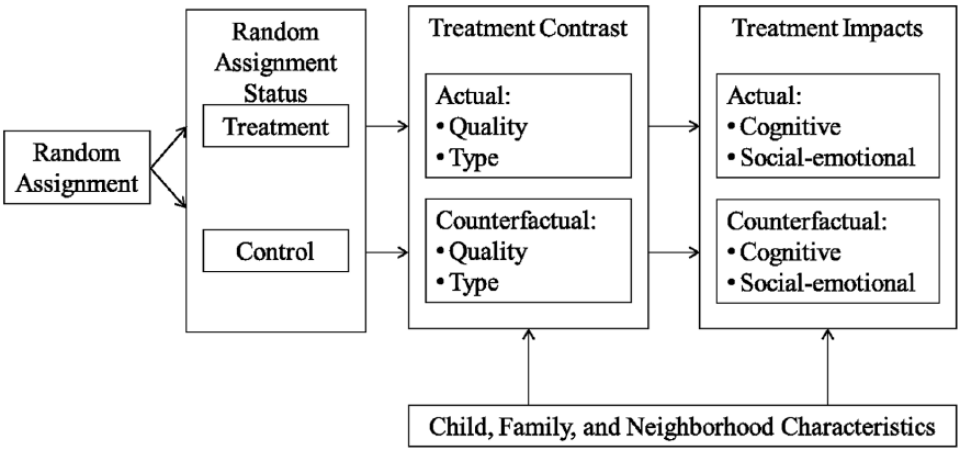

FIGURE 1. *Unpacking the treatment contrast and treatment impact due to random assignment, as well as the cross-site and cross-group variation in such contrasts and impacts.*

one another or to determine which source is most important. Instead we describe the variety of factors that together help to explain variation in the impacts of the Head Start program, and we consider new potential sources of variation that have received less attention in policy discussions. In short, this article extends the findings of prior work by (1) examining contextual influences on the counterfactual experience, (2) leveraging methodological advances, and (3) perhaps most important, putting this new work on impact variation into a single outlet to inform the field. Our primary aims are to contribute to the "story" of Head Start and to inform the next generation of Head Start and early childhood research and policy in the context of ongoing expansions to preschool education nationwide.

Method

Sample and Random Assignment

The original HSIS randomized 4,667 eligible 3- and 4-year-old first-time Head Start applicants from a national sample of oversubscribed Head Start centers (Puma et al., 2010a). The HSIS restricted use file, which is the basis for the present analyses, omits sample members from Puerto Rico, resulting in a sample of 4,440 children[1] from 351 Head Start centers across 81 Head Start grantees in 22 states (see Puma et al., 2010a). We often pool data for the HSIS 3- and 4-year-old cohorts to maximize statistical power, given that they were randomized together in a single block per Head Start site and many were in the same classrooms during the preschool year on which our analyses focus.

Within Head Start centers, children were randomly assigned to an offer of a Head Start slot (the study's treatment group, n = 2,644) or to the control group (n = 1,796). Members in the control group could not enroll in the Head Start center in which they were randomized. However, they could attend other ECCE settings, including Head Start programs not participating in the study, or be cared for at home.

Procedure

Recruitment and data collection. HSIS baseline and follow-up data were collected in the 2002–2003 Head Start year. Subsequent waves of data were collected in the spring of the kindergarten, first-grade, and third-grade years (Puma et al., 2012) and, for 3-year-olds only, the spring of the second preschool year.

Geocoding. As part of an agreement with the Administration for Children and Families, the SAVI Center utilized restricted geocoded location data (i.e., latitude and longitude) of each random assignment Head Start center. Using these data and ArcGIS (Version 10.1; ESRI, 2011), we identified the state, census tract, zip code, and county of each center.

Measures

Child outcomes. Our analyses focus on measures of cognitive development during Head Start that have strong psychometric evidence, are commonly used in early childhood research, and tap domains that were shown to predict later outcomes. These include an assessment of receptive language with the Peabody Picture Vocabulary Test (Dunn & Dunn, 1997) and three subscales of the Woodcock-Johnson III (Woodcock, McGrew, & Mather, 2001): Letter-Word, Oral Comprehension, and Applied Problems. See Bloom and Weiland (2015) and Puma et al. (2010b).

ECCE characteristics. Classroom quality was measured with three widely used observational tools: the Early Childhood Environment Rating Scale–Revised Edition (Harms, Clifford, & Cryer, 1998), the Family Day Care Rating Scale (Harms & Clifford, 1989), and the Arnett Caregiver Interaction Scale (Arnett, 1989). See Puma et al (2010b). Although some of our analyses include "overall" quality

scores from these tools, others leverage exploratory and confirmatory factor analyses conducted by Connors, Friedman-Krauss, Jones, Morris, and Yudron (2013) in the HSIS data set that identified three quality domains across items from the Early Childhood Environment Rating Scale–Revised Edition and Arnett Caregiver Interaction Scale: materials and space for learning, positive teacher-child interactions, and negative teacher-child interactions (for details, see Connors et al., 2013).

Child and family characteristics. Covariates included in each analysis vary somewhat; additional details are found in each article. Child-level covariates include gender, age, cohort (an indicator for whether the child was 2 years vs. 1 year away from kindergarten entry), and race (an indicator variable for children who were Hispanic and an indicator for children who were Black). Family covariates include maternal education, maternal age, home language, recent immigrant status, mother's marital status, whether the mother was previously married, a teenage mom, and whether the child lives with both biological parents. Mothers' depressive symptoms were also measured with the Center for Epidemiologic Studies–Depression Scale (Seligman, 1993). The date of the spring assessment, child's age at the spring assessment, and whether the child was assessed in English were also used as covariates in some analyses.

Neighborhood characteristics. Some analyses also include a set of neighborhood-level characteristics (e.g., urbanicity, poverty, racial/ethnic composition, crime rates, the number of alternative early education and care arrangements, and the availability of neighborhood resources) obtained via geocoded data from the 2000 census, the 2002 business census, the Department of Education, and the Federal Bureau of Investigation crime database at the census tract, zip code, or county level (for details, see McCoy et al., 2015).

Results

We present findings from our collective effort to address questions of moderation of the Head Start program's impacts. First, we address questions about the amount of variation in impacts on outcomes of enrollment and quality, and we identify contextual sources of that variation in state policy. Next, we consider questions regarding the variation of impact on outcomes for children across ecological levels as outlined in the introduction. We briefly highlight the methodological approach utilized and summarize our findings. Table 1 summarizes key information for each study including research questions, sources of variation, methodology, and sample. We conclude the results by interpreting the findings together.

Quantifying Variation in Impacts on Outcomes of Enrollment and Quality

In an earlier work (Bloom & Weiland, 2015), we asked, are impacts on enrollment and quality the same across all sites? We used standard subgroup analysis (Bloom & Michalopoulos, 2013) and a newer, innovative method for predicting impact variation across centers (Bloom, Raudenbush, Weiss, & Porter, 2017; Raudenbush & Bloom, 2015). Our impact variation work capitalized on the original HSIS design as a multisite trial in which children were randomized within centers. We treated each site as a "mini randomized trial" to examined whether impacts on outcomes of care enrollment and care quality varied across sites.

We found that impacts of assignment to Head Start differed substantially across sites on measures of enrollment and exposure to high-quality care (Table 2). Cross-site standard deviations of impacts across measures of Head Start enrollment, center-based care enrollment, and enrollment in high-quality care ranged from 21.4 to 28.4 percentage points. This means that there is significant variation by site in the size of Head Start's impacts on children's access to Head Start, center-based ECCE, and high-quality ECCE.

Sources of Variation in Impacts on Care Enrollment and Quality

Given vast differences in ECCE regulations from state to state, we expect that the impacts of Head Start on children's access to high-quality ECCE may vary according to the state policy context. To address this, we used multinomial logistic regression to estimate the average impact of random assignment to Head Start on various aspects of enrollment in ECCE (Friedman-Krauss et al., 2017) and then explored how these impacts vary across state contexts (Connors & Friedman-Krauss, 2017).

We first sought to understand whether children in the HSIS treatment group enrolled in higher-quality ECCE than would otherwise have been available to them (Friedman-Krauss et al., 2017). The HSIS final report findings suggest that the answer is yes. However, we reexamined this question given differential rates of missing data in the control and treatment groups (73% vs. 27%) and our interest in concurrently estimating the impact of random assignment on enrollment in any formal ECCE and that in high- and low-quality formal ECCE.

In this work, we used multinomial logistic regression (Thiel, 1969) to estimate impacts of random assignment on enrollment in formal ECCE and that in formal ECCE that is high or low quality, while accounting for missing data. Our models leveraged the random assignment design of the HSIS to predict a child's membership in five mutually exclusive and collectively exhaustive categories: (1) cared for exclusively in his or her own

TABLE 1
Secondary Analysis of Variation in Impacts Center Studies

Study: Sources of variation	Research question	Method	Sample
Bloom and Weiland (2015): (1) Head Start random assignment center and (2) children	(1) Do some Head Start centers appear to be more effective than others, relative to their alternatives in (a) access to Head Start or (b) children's developmental outcomes? (2) Do Head Start's effects differ by children's pretest performance, dual language learner status, home language, special-needs status, age cohort, gender, or race?	Impact variation (Bloom, Raudenbush, Weiss, & Porter, 2017; Raudenbush & Bloom, 2015) with subgroup analysis (Bloom & Michalopoulos, 2013)	Sample includes children in complete randomized blocks with nonzero compliance and nonmissing child outcome data (n = 3,465–3,529, depending on outcome)
Friedman-Krauss, Connors, and Morris (2017): Head Start random assignment center	To what extent does the opportunity to enroll in Head Start affect children's access to high-quality ECCE?	Multinomial logistic regression (Thiel, 1969)	Sample includes children in complete randomized blocks (N = 4,385)
Connors and Friedman-Krauss (2017): State policy	Do Head Start's impacts on children's access to high-quality ECCE vary by state policy context?	Multinomial logistic regression (Thiel, 1969), subgroup analysis	Sample includes children in complete randomized blocks (N = 4,385)
McCoy, Morris, Connors, Gomez, and Yoshikawa (2016): Neighborhood	To what degree does the effectiveness of Head Start for children's developmental outcomes vary by Head Start centers' locations in urban vs. rural communities?	Impact variation (Bloom et al., 2017; Raudenbush & Bloom, 2015) with moderation	Sample includes children in complete randomized blocks with nonzero compliance and nonmissing child outcome data (N = 3,503)
Feller, Grindal, Miratrix, and Page (2016): Family choice of care arrangement	How do impacts of Head Start differ according to the settings in which children would have received care if not enrolled in Head Start (i.e., the counterfactual care type)?	Principal stratification (Page, Feller, Grindal, Miratrix, & Somers, 2015)	Sample includes children in complete randomized blocks (N = 4,385)

Note. All studies, except that by Feller et al. (2016), use data from only the first year of the Head Start Impact Study. Bloom and Weiland (2015) and Feller et al. use data from Year 1 through children's first-grade year. ECCE = early childhood care and education.

TABLE 2
Treatment Contrasts on Outcomes of Enrollment and Quality From Bloom and Weiland (2015)

	Grand mean				Cross-site *SD*	
Percentage in . . .	Treatment group	Control group	Difference	p	Difference	p
Head Start	86.6	16.6	70.0***	<.0001	22.3***	<.0001
Any center care	90.6	49.3	41.3***	<.0001	21.4***	<.0001
Nonrelative care with an ECERS-R score	69.8	27.0	42.8***	<.0001	28.4***	<.0001

Note. Samples include children in complete randomized blocks with nonzero compliance and nonmissing Woodcock-Johnson III–Letter-Word outcome data. Estimation models used as covariates: nonresidualized pretest scores, standard Head Start Impact Study covariates, a binary indicator for age cohorts, and fixed intercepts for Head Start centers. For all percentage outcomes, the cross-site *SD* is expressed in percentage points. ECERS-R = Early Childhood Environment Rating Scale–Revised Edition.
***p < .01.

home, (2) enrolled in high-quality ECCE (defined as ≥5 on the Early Childhood Environment Rating Scale–Revised Edition), (3) enrolled in lower-quality ECCE (<5), (4) enrolled in an ECCE program but quality not observed, and (5) type of care setting and ECCE quality both missing. The overall multinomial logit model was statistically significant (Wald's chi-square = 929.15, df = 4, p < .001), indicating that the distribution of the five quality and missing data categories differed significantly across random assignment groups.

We found that children randomized to the treatment group were more likely to enroll in formal ECCE: they were 45 percentage points more likely than children in the control

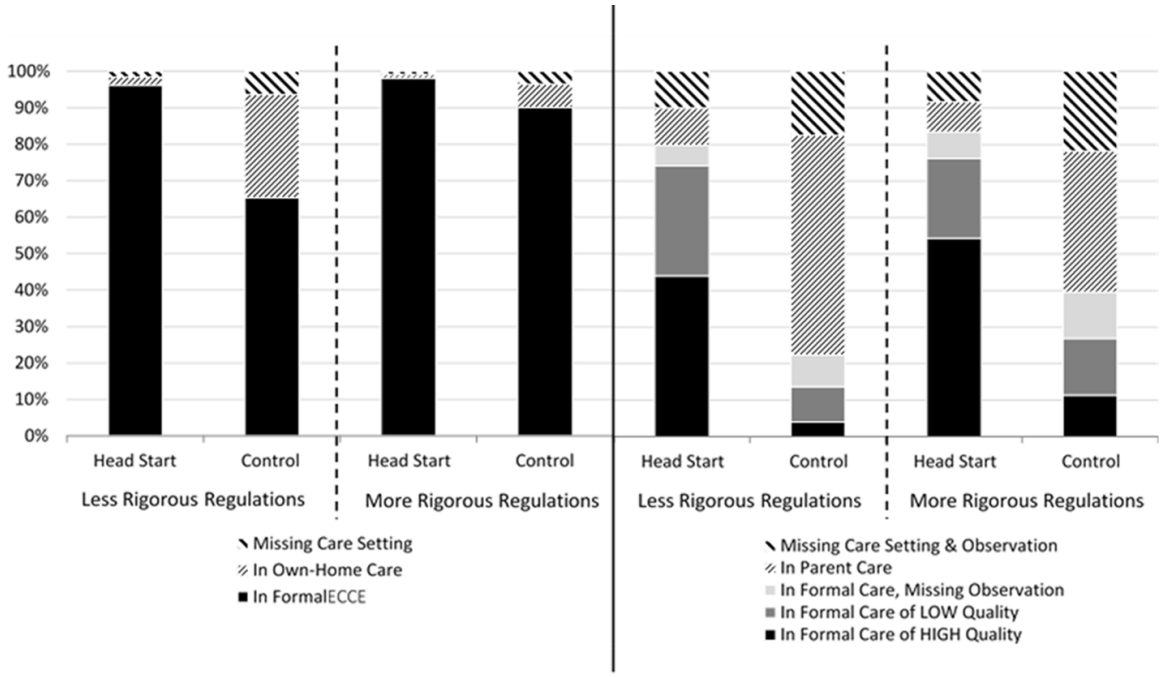

FIGURE 2. *Predicted probabilities of the treatment (Head Start) and control groups' ECCE enrollment by the overall rigor of states' child care licensing regulations. The left side compares predicted probabilities of enrolling in any formal ECCE and the right side compares predicted probabilities of enrolling in ECCE with high quality materials and space for learning. Adapted from Connors and Friedman-Krauss (2017).*

group to enroll in high-quality ECCE and about 10 percentage points more likely to enroll in low-quality ECCE. Additional analyses suggested that treatment impacts were the largest at the high end of the quality distribution and were driven by increased enrollment in Head Start rather than other types of ECCE (see Friedman-Krauss et al., 2017).

To unpack our findings of Head Start's overall positive impact on children's access to formal and high-quality ECCE (Friedman-Krauss et al., 2017) and the variation that we found in those impacts (Bloom & Weiland, 2015),[2] we leveraged variation in state child care licensing regulations and the random assignment design of the HSIS (Connors & Friedman-Krauss, 2017). In the absence of federal policy governing accessibility and quality standards for non–Head Start ECCE programs (which serve the majority of children in out-of-home care), individual states have developed their own definitions and regulations. We hypothesized that characteristics of states' child care licensing regulations—the policies that set minimum health, safety, and quality standards for the majority of legally operating ECCE programs—may be particularly important in this regard (Pianta et al., 2005).

In this work, we relied on similar multinomial logistic regression models (as in Friedman-Krauss et al., 2017) to predict unordered categorical outcome variables within two subsamples of children: those in states with more rigorous licensing regulations and those in states with less rigorous licensing regulations. Regulation rigor was measured with primary documentation of states' child care licensing regulations pertaining to center-based programs that serve preschool-age children, which was coded to create an overall child care licensing score (see Connors & Friedman-Krauss, 2017).

We found that impacts of the offer of Head Start on enrollment do indeed vary by the rigor of state regulations: impacts on likelihood to enroll in formal ECCE are smaller in states with more rigorous child care licensing regulations (and the overall multinomial logit model was statistically significant in states with less, but not more, rigorous licensing regulations). Moreover, descriptive post hoc analyses revealed that variation in impacts on enrollment in formal ECCE was due primarily to variation in the enrollment behavior of children in the control group (there is very little variation in the enrollment behavior of children in the treatment group across licensing characteristics; Figure 2). Children in the control group were 25 percentage points more likely to enroll in formal ECCE in states with more rigorous licensing regulations than in states with less rigorous licensing regulations, while this difference was only 2 percentage points among children in the treatment group. The result is a much stronger impact on enrollment in states with less rigorous regulations (an effect size of 31 vs. 8 percentage points).

Turning to the question of quality, we found statistically significant Head Start impacts on enrollment in ECCE with high-quality teacher-child interactions in states with less, but not more, rigorous licensing regulations. Conversely, Head Start impacts on enrolling in ECCE with high-quality materials and space for learning were similar across states. Post hoc

analyses suggest that the reason is that all children were more likely to enroll in ECCE with high-quality materials and space in states with more rigorous licensing regulations than they were in states with less rigorous regulations (see Figure 2).

Quantifying Variation in Impacts on Outcomes of Children

We then asked whether some Head Start centers appear to be more effective than others in supporting developmental outcomes for children, relative to their alternatives (Bloom & Weiland, 2015), and about the magnitude of any impact variation observed. Utilizing the same approach that we used for estimating cross-site variation in enrollment and quality, we found substantial variation across centers for three of four cognitive outcomes in the effects of assignment to Head Start (intent to treat; ITT) and in the effects of enrolling in Head Start (treatment on the treated; i.e., local average treatment effect [LATE]); this was true even for outcomes for which average treatment effects were not statistically significant. The amount of significant variation in the cognitive outcomes across centers was substantial (0.12–0.25 *SD* for assignment, 0.15–0.26 *SD* for enrollment; see Table 3). Effects did not vary significantly across centers for early numeracy, perhaps because preschool teachers at this time tended (universally) to spend little time on math (Clements & Sarama, 2007; Early et al., 2010).

Given that compliance with treatment assignment was imperfect (and could have varied across sites), it was important to determine if variation in impacts on outcomes for children was solely due to these compliance differences. We directly account for these differences across sites by using the compliance rate for each site to estimate its LATE estimate (Option A in Raudenbush, Reardon, & Nomi, 2012). Similar findings across the ITT and LATE analyses bolster the conclusion that ITT effects are not due to compliance differences: for example, the cross-site *SD* for receptive vocabulary was 0.12 for ITT and 0.15 for LATE. We also examine differences in compliance across key subgroups (presented later).

Sources of variation in impacts on outcomes for children. Given substantial variation across site in Head Start's impacts on nearly all child outcomes, we examine sources of impact variation across the multiple bioecological levels outlined in the introduction.

Variation due to child characteristics. Prior work showed that ECCE programs can have differential impacts on child outcomes depending on several child characteristics (Cooper & Lanza, 2014; Gormley et al., 2008; Lipsey, Hofer, Dong, Farran, & Bilbrey, 2013; Phillips & Meloy, 2012; Weiland, 2016; Weiland & Yoshikawa, 2013). In earlier work (Bloom & Weiland, 2015), we examined whether Head Start effects differed by children's pretest performance (before random assignment), DLL status, home language, special needs status, age cohort, gender, or race. To do so, we conducted subgroup moderation analyses (with fixed effects for centers).

Our analyses revealed substantial variation across centers due to many of the child characteristics examined. We found much larger treatment impacts on language and numeracy for low-pretest performers versus nonlow performers and for DLLs relative to English-only children (Table 4). For example, the effect size of Head Start random assignment on vocabulary was 0.26 *SD* ($p < .01$) for DLLs and 0.10 *SD* ($p < .01$) for their non-DLL peers, a 0.16-*SD* difference in impacts across groups. Similar results were found for children who spoke Spanish as their home language versus English.

Because of the overlap among these subgroups, we also examined whether effects were larger for low- versus higher-pretest performers within language subgroups. Low-pretest DLL children showed substantially larger positive effects as compared with their nonlow-pretest DLL counterparts for receptive vocabulary and early numeracy. There was no such pattern by pretest score within the non-DLL sample. Thus, Head Start appears to have substantially compensated for the limited prior English of some DLL children.

In this work, it was critical to determine if there were compliance differences across these subgroups of children for whom we found differing impacts. We found that compliance differences across subgroups were quite small, with compliance rates across these four groups ranging from 74.1% to 84.7% (Bloom & Weiland, 2015). These differences are likely too small to explain the observed subgroup differences in impacts on assignment on child outcomes.

Finally, to what extent does the concentration of children with these characteristics matter to site-level impacts? Findings show that it does matter for receptive vocabulary—the one outcome for which there is substantial cross-site variability and a sizable differential effect for low-pretest DLLs as compared with their counterparts with greater language skills at program entry (Bloom & Weiland, 2015). For sites with the mean percentage of low-pretest DLL students (just under one in five students), the grand mean ITT effect size was 0.15 *SD* for receptive vocabulary. For sites in which all students were low-pretest DLLs, the grand mean ITT effect size was 0.40 *SD*, a 0.25-*SD* difference. Comparing across models with and without this predictor of variation, Bloom and Weiland (2015) estimated that about 15% of the cross-site effect size variance was explained by cross-site variation in the representation of low-pretest DLLs.

Variation due to counterfactual care arrangements. Another potential source of impact heterogeneity is the type of care that children would have received had they not been offered access to Head Start. Exploring these alternative care arrangements is particularly important given that nearly half the children in the control group enrolled in center-based ECCE—an important distinction between the HSIS and earlier ECCE

TABLE 3
Cross-Site Grand Means and SD *for Head Start Effect Sizes From Bloom and Weiland (2015)*

	Cognitive outcomes			
	Receptive vocabulary (PPVT)	Early reading (WJ-LW)	Oral comprehension (WJ-OC)	Early numeracy (WJ-AP)
Effects of assignment to Head Start (ITT)				
Grand mean	0.14***	0.17***	0.01	0.12***
	(<.001)	(<.001)	(.625)	(<.001)
SD	0.12**	0.25***	0.12*	0.07
	(.030)	(<.001)	(.097)	(.230)
Effects of participation in Head Start (LATE)				
Grand mean	0.17***	0.25***	0.03	0.15***
	(<.001)	(<.001)	(.354)	(<.001)
SD	0.15***	0.26***	0.20*	0.00
	(.004)	(.002)	(.057)	(.560)
Children, *N*	3,523	3,529	3,465	3,491
Centers, *N*	297	297	296	296

Note. Within each relevant subgroup, models were fit by using children with available outcome data in nonzero compliance and complete randomized blocks, including the standard Head Start Impact Study covariates, using fixed intercepts for centers, using the appropriate nonresidualized pretest, using data from both cohorts, and including a binary indicator for age cohort. Effect sizes were calculated by dividing the estimated Head Start effect on each outcome in its original units by the control group *SD* for that outcome; *p* values are in parentheses below each parameter estimate. PPVT = Peabody Picture Vocabulary Test; WJ-LW = Woodcock-Johnson III–Letter-Word; WJ-OC = Woodcock-Johnson III–Oral Comprehension; WJ-AP = Woodcock-Johnson III–Applied Problems; ITT = intent to treat; LATE = local average treatment effect (treatment on the treated).
*$p < .10$. **$p < .05$. ***$p < .01$.

TABLE 4
Differential Effects of Head Start on Children's Cognitive Outcomes by Child Subgroups (ITT) From Bloom and Weiland (2015)

	Receptive vocabulary (PPVT)	Early reading (WJ-LW)	Oral comprehension (WJ-OC)	Early numeracy (WJ-AP)
Pretest performance				
Low performers effect size	0.02***	0.16**	0.03	0.20**
Other performers effect size	0.09***	0.18***	−0.02	0.06*
Difference	0.11*	−0.02	0.05	0.14*
Dual language learner status				
Dual language learner effect size	0.26***	0.23***	−0.01	0.30***
English only effect size	0.01***	0.15***	0.02	0.06**
Difference	0.16*	0.08	0.01	0.24*

Note. Within each subgroup, models were fit by using children with available outcome data in nonzero compliance and complete randomized blocks, including the standard Head Start Impact Study covariates, using fixed intercepts for centers, using the appropriate nonresidualized pretest, using data from both cohorts, and including a binary indicator for age cohort. Effect sizes were calculated by dividing the estimated Head Start effect on each outcome in its original units by the control group standard deviation for that outcome. Statistical significance indicating differences in subgroup impacts was determined by a *t* test of the interaction between the subgroup characteristic and the treatment variable. ITT = intent to treat; PPVT = Peabody Picture Vocabulary Test; WJ-LW = Woodcock-Johnson III–Letter-Word; WJ-OC = Woodcock-Johnson III–Oral Comprehension; WJ-AP = Woodcock-Johnson III–Applied Problems.
*$p < .10$. **$p < .05$. ***$p < .01$.

evaluations in which control children were primarily cared for at home (Duncan & Magnuson, 2013). In previous work (Feller, Grindal, Miratrix, & Page, 2016), we investigated whether the impact of enrolling in Head Start varies depending on the setting in which children would otherwise receive care (i.e., the counterfactual care type). These analyses focus on two groups of Head Start enrollees: those who would otherwise be cared for at home or in a home-based setting and those who would otherwise be cared for in a non–Head Start center.

The main challenge in estimating these effects is that we cannot observe counterfactual care type directly: for children assigned to the treatment condition, we observe their care setting under treatment but not their care setting if they had instead been assigned to control. To handle this challenge, we

utilize the principal stratification framework (Page et al., 2015) to define groups of children based on their observed care setting and their counterfactual care setting. The terminology used to define the groups is similar to that used in instrumental variables analysis for handling noncompliance in treatment participation. Therefore, our classifications incorporate classical noncompliance with treatment assignment (LATE vs. ITT) as well as additional information on the specific care setting. As in classical noncompliance, we assume that random assignment has no impact on outcomes for the always and never takers (i.e., the exclusion restriction).

There are five groups that we consider: *always Head Start*—children who would always enroll in Head Start, irrespective of condition; *always other center-based care*—children who would never take up Head Start and would always enroll in another non–Head Start center-based setting; *always home-based care*—children who would never take up Head Start and would always receive care in a home-based setting; *center care compliers*—children who would participate in Head Start under assignment to treatment but who would receive care in a non–Head Start center-based setting under assignment to control; and *home care compliers*—children who would participate in Head Start under assignment to treatment but who would receive care in a home-based setting under assignment to control. Children in the last 2 groups cannot be observed directly.

Our primary goal is to estimate the effect of assignment to Head Start for children who are center care compliers and home care compliers. We use a Bayesian model to combine information from covariates, outcomes, centers of random assignment, and observed care settings to predict the group to which each child belongs. We then estimate subgroup ITT effects within each predicted group (known as "principal strata"), allowing us to explore whether the impact of Head Start varies according to the opportunities that children and families would take up otherwise.

In Table 5, we present the impacts of Head Start on receptive vocabulary for all children, all compliers, and separately for center compliers and home compliers. For the home-based group, we estimate that, after 1 year, enrollment in Head Start improved scores by roughly 0.20 *SD*—comparable to the difference in effects that we find for DLL versus non-DLL children. This impact is >50% larger than the corresponding HSIS ITT estimates (Puma et al., 2010a). We find no evidence that other center-based alternatives are more effective than Head Start, on average.

Additional analyses found that the magnitude of the impact of Head Start participation for the home-based group declines gradually after the first year. This finding is consistent with the original HSIS results (Puma et al., 2010a) but stands in contrast to the rapid attenuation identified by prior work (Gibbs, Ludwig, & Miller, 2011). Most important, for children who would otherwise have no exposure to center-based care prior to kindergarten, the impact of Head Start on receptive vocabulary remains positive for several years, suggesting an important nuance to the "long-term" effects question of Head Start (Feller et al., 2016).

TABLE 5

Impact of Head Start Offer, Head Start Participation, and Head Start Participation by Alternative Care Type on PPVT in Year of Randomization From Feller et al. (2016)

	Point estimates[a]
Panel A: ITT model—ITT	**0.14 (0.11, 0.16)**
Panel B: IV model—overall LATE	**0.18 (0.14, 0.23)**
Panel C: Principal stratification model	
LATE for center compliers	0.00 (−0.13, 0.14)
LATE for home compliers	**0.23 (0.15, 0.30)**
p ($LATE_{hc} > LATE_{cc}$)	.99

Note. PPVT = Peabody Picture Vocabulary Test; ITT = intent to treat; IV = instrumental variable; LATE = local average treatment effect (treatment on the treated).

[a]Point estimates are posterior medians with 2.5 and 97.5 quantiles of posterior distribution in parentheses; 95% posterior intervals that exclude zero are printed in bold.

Variation due to the neighborhood location of the Head Start site. Research has highlighted the role of neighborhoods in children's early development and learning (e.g., Brooks-Gunn, Duncan, & Aber, 1997; Leventhal & Brooks-Gunn, 2000), but we know less about the ways that neighborhoods may enhance or restrict the effectiveness of the social services available within them. As a national program serving rural and urban communities, Head Start provides an opportunity to explore whether such dynamics may explain the cross-site variation in treatment impacts that we identified previously (Bloom & Weiland, 2015).

In an earlier study (McCoy, Morris, Connors, Gomez, & Yoshikawa, 2016), we examined the degree to which the effectiveness of Head Start for improving children's receptive vocabulary and early literacy outcomes differed according to centers' locations in urban versus rural communities. To do so, we built on the basic multilevel impact variation models from Bloom and Weiland (2015) to include the urbanicity of the Head Start center as a predictor of child-level impacts across sites. We also examined whether urban-rural differences could be explained (attenuated) by the inclusion of other neighborhood, center, family, and child characteristics.

Results of this work suggest that Head Start was more than twice as effective in improving children's receptive vocabulary scores in urban versus rural communities: the effect of random assignment to Head Start versus a control condition was 0.14 *SD* in urban communities versus 0.07 *SD* in rural communities, a 0.07-*SD* difference in impacts (much smaller than the differences observed earlier but still

TABLE 6

Grand Mean ITT Effects on Features of the HSIS Treatment Contrast for Dual Language Learners and English-Only Learners, by Pretest Performance Subgroup From Bloom and Weiland (2015)

	Estimated ITT effect			
	Percentage in Head Start	Percentage in any center care	Percentage in nonrelative care with an ECERS-R ≥5	Percentage in parent care
Dual language learners				
Low-pretest performers	84.7***	53.5***	44.9***	−40.8***
Other sample members	79.1***	44.0***	50.9***	−35.3***
English-only sample members				
Low-pretest performers	80.7***	49.2***	45.5***	−30.6***
Other sample members	74.1***	45.6***	47.2***	−32.3***

Note. Within each subgroup, samples include children in complete randomized blocks with nonzero compliance and nonmissing Woodcock-Johnson III–Letter-Word outcome data. Estimation models used as covariates: nonresidualized pretest scores, standard HSIS covariates, a binary indicator for age cohorts, and fixed intercepts for Head Start centers. ITT = intent to treat; HSIS = Head Start Impact Study; ECERS-R = Early Childhood Environment Rating Scale–Revised Edition.

***$p < .01$.

statistically significant). Conversely, Head Start's impact on oral comprehension skills was significantly smaller in urban relative to rural communities: the effect of Head Start was 0.10 *SD* in rural communities versus −0.02 *SD* in urban communities. This work did not address differences in compliance between urban and rural sites; as such, some of these differences in impacts could be due to differences in compliance rates across context.

Putting the Findings Together

A number of questions arise from this body of work when we consider these findings side by side. Here we discuss the subset of questions that fall within the bounds of what existing methodology enables us to address. See Supplementary Table S1 in the online materials for a summary of the results.

First, to what extent are the findings for DLL and low-pretest performers due to differences in the HSIS treatment contrast? For example, are the differences that we found between impacts for DLL and non-DLL children due to differences in the counterfactual care arrangements of these two groups (where differences in impact were also found)? This question can be addressed from two angles: from the perspective of the DLL findings (based on the methods and approaches of Bloom & Weiland, 2015) and from the perspective of the findings on counterfactual care (per the methods and approaches of Feller et al., 2016).

Recall the findings presented earlier that Head Start impacts on receptive vocabulary were largest for DLL children who were also low-pretest performers—to what extent is this due to differences in care experiences? There were very few differences in impacts between these groups in care type or quality (Table 6). In effect, the differences between the groups on counterfactual care arrangements are just too small to explain the much larger differences between these groups of children in their impacts on receptive vocabulary.

Similarly, analyses presented in Feller and colleagues (2016) show that variation of Head Start impacts by counterfactual conditions stands alongside, rather than being explained by, impact variation by child characteristics (i.e., DLL status and pretest score). Across all four subgroups, Head Start effects for home-based compliers are positive, while effects for center-based compliers are negligible. Home-based complier effects are even larger for DLL children (0.35 *SD*, more than double that for non-DLL children) and for children who are low-pretest performers. In sum, DLL and low-pretest children cannot "explain" impact variation by families' choice of counterfactual—the home complier effect is apparent even among DLL and low-pretest performers.

Could the variation in impacts by urbanicity be due to the different concentrations of DLL children in those settings? Analyses by McCoy and colleagues (2016), who examined the urbanicity coefficient in the context of alternative models with a large set of controls, indicated that the most likely explanation for differences in impacts on receptive language across context is differences in the concentration of Spanish-speaking children, rather than a broad set of alternative neighborhood characteristics, center characteristics, or classroom composition characteristics, with few exceptions (Table 7).

Discussion

The findings from this collective set of work by the SAVI Center lead to a number of important additions to (and perhaps a retelling of) the "story" of Head Start's effects—most notably, that Head Start is not a monolithic program that functions in the same way for all children and in all locations. Indeed, our findings (and those of our colleagues who were

TABLE 7
Impact of Head Start on Peabody Picture Vocabulary Test Scores From McCoy et al. (2016)

	Basic model	With neighborhood characteristics	With center characteristics	With composition characteristics	With home language
Urban	0.160***	0.162***	0.157***	0.149***	0.147***
Rural	0.070	0.062	0.073	0.090**	0.099**
Difference	0.090*	0.100	0.084	0.059	0.048

*$p < .10$. **$p < .05$. ***$p < .01$.

conducting complementary work about moderated effects of Head Start) show that the topline HSIS results of its average impacts—on the type and quality of care that children experienced and on developmental outcomes for children—actually mask a great deal of variation by alternative care type, child characteristics, and geographic location. Notably, this variation was found in the program itself and in the counterfactual due, at least in part, to the context in which Head Start was delivered. As such, our findings suggest that sweeping claims of Head Start's ineffectiveness (e.g., Burke & Muhlhausen, 2013; Whitehurst, 2013) or even characterizations of the HSIS impacts as small but meaningful (Ludwig & Phillips, 2007) are misleading, at least in terms of impact on the key enrollment and child skills that we examined.

Our work tells a critical story regarding for whom and under what circumstances Head Start is effective at improving children's access to high-quality ECCE and their school readiness. The national scope of the HSIS (including 351 Head Start sites in 22 states) allowed us to uncover the ways in which characteristics of a Head Start site—including the state policy context, the urbanicity of the neighborhood, or the proportion of DLL children served—matter to the impacts of assignment to and attendance in Head Start. With regard to impacts on children's ECCE experiences, we found that, overall, the offer of Head Start successfully moved children from home-based settings into high-quality center-based care (primarily Head Start) but that it altered children's care experience the most (in terms of enrollment in formal care and the quality of that care) in states with less rigorous child care licensing regulations. The quality of Head Start, too, appears to vary more in states with less-rigorous regulations, suggesting that the combination of federal standards and state policy context may be important to ensuring a system of uniformly high-quality ECCE. This makes sense—Head Start has the potential to make a greater difference when the counterfactual landscape is weaker (as critics have long pointed out when studies of prior cohorts are compared with those from ECCE studies conducted today).

Moreover, we find differential effectiveness of Head Start based on location in urban versus rural environments, with greater benefits for receptive vocabulary in urban environments and for oral comprehension in rural settings that are not due to many of the expected differences in poverty and ethnicity across these settings (although the results for receptive vocabulary may be at least partly due to differences in children served). Prior research suggested, for example, that Head Start and other ECCE programs face different constraints regarding the availability of wrap-around services, alternative care options, transportation, and other resources depending on their urbanicity (Chertow, 1968; National Advisory Committee on Rural Health and Human Services, 2012; Rural Poverty Research Institute, 2008). Such differences in the broader context may also lead to important differences in the composition, quality, or nature of the interactions that take place in the Head Start classroom, possibly explaining differential impacts on developmental outcomes.

With regard to variation by families and children, our analyses identified key features of children's counterfactual experiences and characteristics that affected the impact of Head Start on children's development and learning. Impacts on child outcomes are larger for children from families for whom the offer of Head Start led them to enroll in Head Start rather than a home-based setting, highlighting the important role that Head Start plays for some families. Head Start also provides the strongest benefits for DLLs and children with low levels of baseline English proficiency, which is in line with its mission of cultural and linguistic sensitivity and may reflect Head Start's strengths in helping DLL children to build their English language skills over the Head Start year. All of this heterogeneity appears to play a distinct role and to have meaningful implications for Head Start's ability to provide uniformly positive impacts on children's outcomes.

Future research is needed to fully understand all the reasons behind these moderating effects, but perhaps the most interesting story is about the counterfactual, not the treatment itself. Although some findings might lead us to look more closely at what Head Start in particular is doing well to support children with weak English language skills, the findings on policy context, neighborhoods, and counterfactual arrangements require us to also consider the experiences of children and families who do not enroll in Head Start. In part, the reason might be that measures of program and classroom practice available within the HSIS may not have the sensitivity needed to fully address variations in teacher practice and classroom quality. Future research should use finer-grained measures focused on instructional quality and program

practices related to professional development, teaching, and learning and that information should be collected within Head Start and the comparison ECCE settings. Given the current diversity of ECCE policy contexts across the United States—in which ECCE programs vary greatly in number, type, quality, and funding source—and families' preferences regarding early care settings, doing so seems critical.

How do these findings align with a recently growing body of other work on impact variation? Supplementary Table S2 in the online materials summarizes the findings from several other articles that emerged over the last 4 years, as we were conducting the work of the SAVI Center, with additional detail regarding the questions, source of variation, method, and findings emerging from each. Five articles examined effects by child and family characteristics in the HSIS, also finding that impacts are largest for children who have low baseline skills and who speak Spanish (according to instrumental variable quantile treatment effects; Bitler, Hoynes & Domina, 2014). The articles also showed impact variation by family characteristics not examined in the current study (by Head Start generational status, Chor, 2016; by parental stimulation, Miller et al., 2014), although differences in the impacts in terms of effect size are not large. Analyses based on latent classes further show large impact differences among families defined by marital status, education, employment, and English language learner status considered together (Cooper & Lanza, 2014). Two studies (Kline & Walters, 2016; Zhai et al., 2014) examined the role of counterfactual care as we did, using methods different from those of Feller and colleagues (2016). Both found that Head Start's effects are more positive when the counterfactual is parent care, although the differences in impacts found by Kline and Walters (2016) are much larger than what we found (Head Start impact of 0.37 *SD* compared with home care vs. no impact compared with other center-based care), and those found by Zhai and colleagues (2014) are much smaller (Head Start impact of 0.30 compared with parent care, but impacts for Head Start compared with center care are 0.18 for 3-year-olds and 0.07 for 4-year-olds). Still, the consistency in conclusions between our own study and these two other independent efforts increases our confidence in the results presented here.

Limitations

This work is not without its limitations. Three were noted at the outset of this article—the nonexperimental nature of this work, our inability to disentangle sources of variation, and our focus on initial rather than sustained impact. To those, we add two more: our inability to examine all possible sources of variation and the historical context of the HSIS that predated recent expansions of state-funded preschool and dramatic changes in the ECCE policy landscape as well as the changing demographic makeup of the United States (Johnson & Lichter, 2010). These changes are likely to alter the "impact" of Head Start into the future.

What Does This Mean for Head Start?

Our results suggest that Head Start is effective in meeting the most basic benchmark of ECCE programming: moving children (particularly low-income children) into higher-quality care. This finding underscores Head Start's importance in the American ECCE landscape, its potential for continued impact, as well as the importance of federal and state policies that guarantee access to high-quality care for young children. At the same time, our results highlight the ways in which policies that support children's development and learning must be "customized" for local contexts. Specifically, to optimize effectiveness, policy and program designers need to consider many forms of diversity (children and place) when planning scale-up, resource allocation, and differentiated programming. For example, our findings may suggest that, to be most effective, Head Start would be wise to consider geographic targeting of services based on the local policy context, availability of alternative high-quality center-based ECCE, or neighborhood concentration of DLLs (without ignoring key child characteristics; e.g., DLL status). Such an approach stands alongside more established approaches of targeting Head Start services to individuals based on family characteristics. Of course, this discussion does not address the costs and relative costs of opening a Head Start center across these communities, which likely vary dramatically across places. Additionally, if we ignore costs and targeting, a new Head Start program is likely "valuable" wherever it opens in that it is an important source of high-quality ECCE for low-income families, even in places that have good ECCE alternatives (although Head Start may or may not be more effective or cost-effective than these alternatives). However, a geographic approach to concentrating expansion of slots could extend Head Start's impacts in two ways: by "filling in" the most acute gaps in access and quality left by alternative ECCE systems and by positioning Head Start sites to serve the subpopulations for whom it has proven most effective.

Conclusion

The information that the SAVI Center has produced may inform policy makers and practitioners in making concrete improvements in their programs through better understanding about where and for whom those programs might be most effective. Moreover, our findings demonstrate the value in designing studies that allow for the examination of variation in program impacts and counterfactual conditions. In short, the findings from this article—emphasizing the salience of variation, context, and the counterfactual—contribute to a growing body of literature that can inform the future of Head Start policy, programming, and research, as well as intentional preschool improvement and expansion strategies to support children's school readiness at scale.

Notes

1. Note that we do not use the HSIS sampling weights that were developed to extrapolate the study's findings to the 2002–2003 national population of oversubscribed Head Start centers (Puma et al., 2010b, chap. 2). Doing so made it possible to avoid the added complexity that would result from their use when computing statistical tests for analyses of variation in Head Start effects. Fortunately, these weights have little effect on HSIS point estimates of average program effects and only increase their standard errors (Bloom & Weiland, 2015).

2. Notably, limitations in modeling cross-site variation do not allow us to estimate it in the same models in which we conduct multinomial logit regressions—hence, our need to rely on Bloom and Weiland (2015) to motivate the analyses on cross-site variation, even if our analytic approach draws from Friedman-Krauss and colleagues (2017).

References

Arnett, J. (1989). Caregivers in day-care centers: Does training matter? *Journal of Applied Developmental Psychology*, *10*, 541–552. doi:10.1016/0193-3973(89)90026-9

Barnett, W. S. (1995). Long-term effects of early childhood programs on cognitive and school outcomes. *Future of Children*, *5*(3), 25–50.

Barnett, W. S. (2011). Effectiveness of early educational intervention. *Science*, *333*, 975–978. doi:10.1126/science.1204534

Barnett, W. S., & Masse, L. N. (2007). Early childhood program design and economic returns: Comparative benefit-cost analyses of the Abecedarian program and policy implications. *Economics of Education Review*, *26*, 113–125.

Belfield, C., Nores, M., Barnett, W. S., & Schweinhart, L. (2006). The High/Scope Perry Preschool Program: Cost-benefit analysis using data from the age 40. *Journal of Human Resources*, *16*(1), 162–190. doi:10.3368/jhr.XLI.1.162

Bitler, M. P., Hoynes, H. W., & Domina, T. (2014). *Experimental evidence on distributional effects of Head Start* (NBER Working Paper No. w20434). Cambridge, MA: National Bureau of Economic Research.

Bloom, H. S., Hill, C. J., & Riccio, J. A. (2003). Linking program implementation and effectiveness: Lessons from a pooled sample of welfare-to-work experiments. *Journal of Policy Analysis and Management*, *22*, 551–575.

Bloom, H. S., & Michalopoulos, C. (2013). When is the story in the subgroups? *Prevention Science*, *14*(2), 179–188.

Bloom, H. S., Raudenbush, S. W., Weiss, M. J., & Porter, K. (2017). Using multisite experiments to study cross-site variation in treatment effects: A hybrid approach with fixed intercepts and a random treatment coefficient. *Journal of Research on Educational Effectiveness*, *10*(4), 817–842.

Bloom, H. S., & 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.

Bronfenbrenner, U. (1979). *The ecology of human development: Experiments by nature and design*. Cambridge, MA: Harvard University Press.

Brooks-Gunn, J., Duncan, G., & Aber, J. L. (Eds.). (1997). *Neighborhood poverty, Volume 1: Context and consequences for children*. New York, NY: Russell Sage Foundation.

Burke, L. M., & Muhlhausen, D. B. (2013). *Head Start impact evaluation report finally released* (Issue Brief No. 3823). Washington, DC: Heritage Foundation.

Campbell, F. A., Pungello, E. P., Burchinal, M., Kainz, K., Pan, Y., Wasik, B., . . . Ramey, C. T. (2012). Adult outcomes as a function of an early childhood education program: An Abecedarian Project follow-up. *Developmental Psychology*, *48*, 1033–1043. doi.org/10.1037/a0026644

Chertow, D. S. (1968). *Project Head Start, the urban and rural challenge: Final report submitted to Office of Economic Opportunity* (Doctoral dissertation). Syracuse University, Syracuse, NY.

Chor, E. (2016). Multigenerational Head Start participation: An unexpected marker of progress. *Child Development*. Advance online publication. doi:10.1111/cdev.12673

Clements, D. H., & Sarama, J. (2007). Effects of a preschool mathematics curriculum: Summative research on the Building Blocks project. *Journal for Research in Mathematics Education*, *38*, 136–163.

Connors, M. C., & Friedman-Krauss, A. H. (2017). Varying states of Head Start: Impacts of a federal program across state policy contexts. *Journal of Research on Educational Effectiveness*. Advance online publication. doi:10.1080/19345747.2017.1320736

Connors, M. C., Friedman-Krauss, A. H., Jones, S. M., Morris, P. A., & Yudron, M. (2013, November). *Refining early measures of early childhood classroom quality*. Paper presented at the Association for Public Policy Analysis and Management Fall Research Conference, Washington, DC.

Cooper, B. R., & Lanza, S. T. (2014). Who benefits most from Head Start? Using latent class moderation to examine differential treatment effects. *Child Development*, *85*, 2317–2338.

Currie, J., & Thomas, D. (1995). Does Head Start make a difference? *American Economic Review*, *85*, 341–364.

Deming, D. (2009). Early childhood intervention and life-cycle skill development: Evidence from Head Start. *American Economic Journal: Applied Economics*, *1*(3), 111–134.

Dodge, K. A., Bai, Y., Ladd, H. F., & Muschkin, C. G. (2016). Impact of North Carolina's early childhood programs and policies on educational outcomes in elementary school. *Child Development*. Advance online publication. doi:10.1111/cdev.12645

Duncan, G. J., & Magnuson, K. (2013). Investing in preschool programs. *Journal of Economic Perspectives*, *27*(2), 109–132. doi:10.1257/jep.27.2.109

Dunn, L. M., & Dunn, L. M. (1997). *Peabody Picture and Vocabulary Test, Third Edition (PPVT)*. Circle Pines, MN: American Guidance Service.

Durlak, J. A., & Dupre, E. P. (2008). Implementation matters: A review of research on the influence of implementation on program outcomes and the factors affecting implementation. *American Journal of Community Psychology*, *41*, 327–350. doi:10.1007/s10464-008-9165-0

Early, D. M., Iruka, I. U., Ritchie, S., Barbarin, O. A., Winn, D. M. C., Crawford, G. M., . . . Pianta, R. C. (2010). How do pre-kindergarteners spend their time? Gender, ethnicity, and income as predictors of experiences in pre-kindergarten classrooms. *Early Childhood Research Quarterly*, *25*, 177–193.

Elango, S., García, J. L., Heckman, J. J., & Hojman, A. (2016). *Early childhood education*. Retrieved from https://heckman.uchicago.edu/sites/heckman2013.uchicago.edu/files/uploads/Papers/Moffitt-ECE-Paper_2016-08-29a_jld.pdf

ESRI. (2011). *ArcGIS desktop: Release 10*. Redlands, CA: Environmental Systems Research Institute

Feller, A., Grindal, T., Miratrix, L., & Page, L. C. (2016). Compared to what? Variation in the impacts of early childhood education by alternative care type. *Annals of Applied Statistics*, *10*, 1245–1285.

Frangakis, C. E., & Rubin, D. B. (2002). Principal stratification in causal inference. *Biometrics*, *58*, 21–29.

Friedman-Krauss, A. H., Connors, M. C., & Morris, P. A. (2017). Unpacking the treatment contrast in the Head Start Impact Study: To what extent does assignment to treatment affect quality of care? *Journal of Research on Educational Effectiveness*, *10*, 68–95. doi:10.1080/19345747.2016.1147627

Garces, E., Thomas, D., & Currie, J. (2002). Longer-term effects of Head Start. *American Economic Review*, *92*, 999–1012.

García, J. L., Heckman, J. J., Lead, D. E., & Prado, M. J. (2016). *The life-cycle benefits of an influential early childhood program*. Cambridge, MA: National Bureau of Economic Research.

Gelber, A., & Isen, A. (2013). Children's schooling and parents' behavior: Evidence from the Head Start Impact Study. *Journal of Public Economics*, *101*, 25–38.

Gibbs, C., Ludwig, J., & Miller, D. L. (2011). *Does Head Start do any lasting good?* (NBER Working Paper No. 17452). Cambridge, MA: National Bureau of Economic Research.

Gormley, W. T., Phillips, D., & Gayer, T. (2008). Preschool programs can boost school readiness. *Science*, *320*, 1723–1724.

Granger, R. C. (2010). Understanding and improving the effectiveness of after-school practice. *American Journal of Community Psychology*, *45*(3–4), 441–446.

Harms, T., & Clifford, R. M. (1989). *Family Day Care Rating Scale (FDCRS)*. New York, NY: Teachers College Press.

Harms, T., Clifford, R. M., & Cryer, D. (1998). *Early Childhood Environment Rating Scale (Revised Edition)*. New York, NY: Teachers College Press.

Johnson, K. M., & Lichter, D. T. (2010). Growing diversity among America's children and youth: Spatial and temporal dimensions. *Population and Development Review*, *36*(1), 151–176.

Kline, P., & Walters, C. R. (2016). Evaluating public programs with close substitutes: The case of Head Start. *The Quarterly Journal of Economics*, *131*, 1795–1848. doi:10.1093/qje/qjw027

Leventhal, T., & Brooks-Gunn, J. (2000). The neighborhoods they live in: The effects of neighborhood residence on child and adolescent outcomes. *Psychological Bulletin*, *126*(2), 309.

Lipsey, M. W., Hofer, K. G., Dong, N., Farran, D. C., & Bilbrey, C. (2013). *Evaluation of the Tennessee Voluntary Prekindergarten Program: End of pre-K results from the randomized control design*. Nashville, TN: Peabody Research Institute.

Lombardi, J., Harding, J. F., Connors, M. C., & Friedman-Krauss, A. H. (2016). Prologue: Coming of age: A review of federal early childhood policy 2000–2015. In H. Dichter (Ed.), *Rising to the challenge: Building effective systems for young children and families*. Retrieved from http://www.buildinitiative.org/OurWork/StateandLocal/EarlyLearningChallenge.aspx

Ludwig, J., & Miller, D. L. (2007). Does Head Start improve children's life chances? Evidence from a regression discontinuity design. *Quarterly Journal of Economics*, *122*, 159–208. doi:10.1162/qjec.122.1.159

Ludwig, J., & Phillips, D. (2007). *The benefits and costs of Head Start*. Washington, DC: Society for Research on Child Development.

Ludwig, J., & Phillips, D. (2008). Long-term effects of Head Start on low-income children. *Annals of the New York Academy of Sciences*, *1136*, 257–268.

McCoy, D. C., Connors, M. C., Morris, P. A., Yoshikawa, H., & Friedman-Krauss, A. H. (2015). Neighborhood economic disadvantage and children's cognitive and social-emotional development: Exploring Head Start classroom quality as a mediating mechanism. *Early Childhood Research Quarterly*, *32*, 150–159.

McCoy, D. C., Morris, P. A., Connors, M. C., Gomez, C. J., & Yoshikawa, H. (2016). Differential effectiveness of Head Start in urban and rural communities. *Journal of Applied Developmental Psychology*, *43*, 29–42.

Miller, E. B., Farkas, G., Vandell, D. L., & Duncan, G. J. (2014). Do the effects of Head Start vary by parental preacademic stimulation? *Child Development*, *85*(4), 1385–1400.

National Advisory Committee on Rural Health and Human Services. (2012). *Challenges to Head Start and early childhood development programs in rural communities*. Rockville, MD: Author.

Office of Head Start. (2015). *Head Start Program facts fiscal year 2015*. Retrieved from https://eclkc.ohs.acf.hhs.gov/hslc/data/factsheets/docs/head-start-fact-sheet-fy-2015.pdf

Page, L. C., Feller, A., Grindal, T., Miratrix, L., & Somers, M. A. (2015). Principal stratification: A tool for understanding variation in program effects across endogenous subgroups. *American Journal of Evaluation*, *36*, 514–531.

Phillips, D. A., & Meloy, M. E. (2012). High-quality school-based pre-K can boost early learning for children with special needs. *Exceptional Children*, *78*(4), 471–490.

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*, 49–88. doi:10.1177/1529100610381908

Pianta, R., Howes, C., Burchinal, M., Bryant, D., Clifford, R., Early, D., & Barbarin, O. (2005). Features of pre-kindergarten programs, classrooms, and teachers: Do they predict observed classroom quality and child-teacher interactions? *Applied Developmental Science*, *9*(3), 144–159.

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*. Washington, DC: Administration for Children and Families.

Puma, M., Bell, S., Cook, R., Heid, C., & Lopez, M. (2005). *Head Start Impact Study: First year findings*. Washington, DC: Administration for Children and Families.

Puma, M., Bell, S., Cook, R., Heid, C., Shapiro, G., Broene, P., . . . Spier, E. (2010a). *Head Start Impact Study: Final report*. Washington, DC: Administration for Children and Families.

Puma, M., Bell, S., Cook, R., Heid, C., Shapiro, G., Broene, P., . . . Spier, E. (2010b). *Head Start Impact Study: Technical report*. Washington, DC: Administration for Children and Families.

Raudenbush, S., & Bloom, H. S. (2015). *Learning about and from variation in program impacts using multisite trials*. New York, NY: MDRC.

Raudenbush, S. W., Reardon, S., & Nomi, T. (2012). Statistical analysis for multi-site trials using instrumental variables. *Journal of Research and Educational Effectiveness*, *5*, 303–332.

Reynolds, A. J., Temple, J. A., Robertson, D. L., & Mann, E. A. (2002). Age 21 cost-benefit analysis of the Title I Chicago child-parent centers. *Educational Evaluation and Policy Analysis*, *24*(4), 267–303.

Rural Poverty Research Institute. (2008). *Human service programs serving rural communities: Head Start*. Corvallis, OR: Rural Poverty Research Institute

Schweinhart, L. J., Montie, J., Xiang, Z., Barnett, W. S., Belfield, C. R., & Nores, M. (2005). *Lifetime effects: The High/Scope Perry Preschool study through age 40*. Ypsilanti, MI: High/Scope Educational Research Foundation.

Seligman, M. E. P. (1993). *What you can change . . . and what you can't*. New York, NY: Ballantine Books.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Belmont, CA: Wadsworth.

Shager, H. M., Schindler, H. S., Magnuson, K. A., Duncan, G. J., Yoshikawa, H., & Hart, C. M. D. (2013). Can research design explain variation in Head Start research results? A meta-analysis of cognitive and achievement outcomes. *Educational Evaluation and Policy Analysis*, *35*(1), 76–95. doi:10.3102/0162373712462453

Thiel, H. (1969). A multinomial extension of linear logit model. *International Economic Review*, *10*, 251–259.

Walters, C. R. (2015). Inputs in the production of early childhood human capital: Evidence from Head Start. *American Economic Journal: Applied Economics*, *7*(4), 76–102.

Weiland, C. (2016). Impacts of the Boston prekindergarten program on the school readiness of young children with special needs. *Developmental Psychology*, *52*, 1763–1776. doi:10.1037/dev0000168

Weiland, C., & Yoshikawa, H. (2013). Impacts of a prekindergarten program on children's mathematics, language, literacy, executive function, and emotional skills. *Child Development*, *84*(6), 2112–2130. doi:10.1111/cdev.12099

Whitehurst, G. J. (2013). *New evidence raises doubts on Obama's Preschool for All*. Retrieved from https://www.brookings.edu/research/new-evidence-raises-doubts-on-obamas-preschool-for-all/

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., Gormley, W., Ludwig, J., . . . Zaslow, M. J. (2013). *Investing in our future: The evidence base on preschool education*. Ann Arbor, MI: Society for Research in Child Development.

Zhai, F., Brooks-Gunn, J., & Waldfogel, J. (2014). Head Start's impact is contingent on alternative type of care in comparison group. *Developmental Psychology*, *50*, 2572–2586.

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

Zigler, E., & Styfco, S. (2010). *The hidden history of Head Start*. New York, NY: Oxford University Press.

Authors

PAMELA A. MORRIS is a professor of applied psychology and vice dean at the Steinhardt School of Culture, Education and Human Development, New York University. Her works lies at the intersection of social policy, practice, and developmental psychology, testing promising interventions for low-income families and children.

MAIA CONNORS is a senior research associate at Ounce of Prevention Fund. Her research focuses on early childhood care and education policy and systems and she aims to translate findings into policy and practice. Her work focuses on understanding and improving early childhood care and education policy, systems' support of high quality early education and professional learning, and adults' support of young children's development.

ALLISON FRIEDMAN-KRAUSS is an assistant research professor at the National Institute for Early Education Research, Rutgers University. Her research focuses on unpacking the impacts of early education interventions and quality and the cognitive and social-emotional development of low-income children.

DANA CHARLES MCCOY is an assistant professor at the Graduate School of Education, Harvard University. Her work focuses on understanding the ways that poverty-related risk factors in children's home, school, and neighborhood environments affect the development of their cognitive and social-emotional skills in early childhood.

CHRISTINA WEILAND is an assistant professor at the School of Education, University of Michigan. Her research focuses on the effects of early childhood interventions and public policies on children's development, especially on children from low-income families, and she is particularly interested in the active elements that drive children's gains in successful at-scale public preschool programs.

AVI FELLER is an assistant professor at the Goldman School of Public Policy, UC Berkeley. His work lies at the intersection of public policy, data science, and statistics, and his applied research focuses on working with governments to use data to design, implement, and evaluate policies.

LINDSAY PAGE is an assistant professor at the School of Education, University of Pittsburg. Her work focuses on quantitative methods and their application to questions regarding the effectiveness of educational policies and programs across the preschool to postsecondary spectrum.

HOWARD BLOOM is a chief social scientist at MDRC. He leads the development of experimental and quasi-experimental methods for estimating program impacts across education policy and welfare employment policy.

HIROKAZU YOSHIKAWA is a professor of applied psychology at the Steinhardt School of Culture, Education and Human Development, New York University. He is a community and developmental psychologist who studies the effects of public policies and programs related to immigration, early childhood, and poverty reduction on children's development.