Will public pre-K really close achievement gaps? Gaps in prekindergarten quality between students and across states

Rachel Valentino
Stanford University

DRAFT: Do Not Cite or Distribute Without Permission

This research was supported by grant award #R305B090016 from the Institute for Education Sciences, U.S. Department of Education to Stanford University. The author thanks Sean Reardon, Deborah Stipek, Kenji Hakuta, and Claude Goldenberg for their invaluable help and feedback on earlier versions of this work. The author would also like to thank Steve Barnett and the National Institute for Early Education Research for access to these data.

Keywords: Early childhood education, public pre-K, classroom quality, gaps, segregation

Abstract

Publicly funded pre-K has often been touted as a means to narrow the achievement gap, but this goal is much less likely to be achieved if poor/minority children do not, at a minimum, attend equal quality pre-K as their non-poor/non-minority peers. In this paper I find large "quality gaps" in public pre-K between poor/minority students and non-poor/non-minority students, ranging from 0.3 to 0.7 SD on a range of classroom observational measures. I also find that even after adjusting for a series of classroom characteristics, significant and sizable quality gaps remain. Finally, I find much between-state variation in gap magnitudes, and that state-level quality gaps are related to state-level residential segregation. These findings are particularly troubling if a goal of public pre-K is to minimize inequality.

I. Introduction

By the time children first enter kindergarten, there are already large gaps in achievement between students of different racial, socioeconomic, and language backgrounds. More specifically, achievement gaps between black and Hispanic children and their white peers, and between children from lower socioeconomic backgrounds and their higher socioeconomic counterparts are about two thirds of a standard deviation at the start of kindergarten (Duncan & Magnuson, 2005; Reardon & Portilla, 2015; Loeb & Bassok, 2009; Reardon & Robinson, 2008) – the equivalent of about three years of learning in later grades. Even larger gaps exist between dual language learners[1] (DLL) and their native-English speaking peers. Such gaps are as large as a standard deviation in early elementary grades (NCES, 2011; Reardon & Galindo, 2009).

Publicly funded prekindergarten has long been touted as a means to bolster disadvantaged children's academic skills to help reduce gaps in achievement before children enter kindergarten, with the hope that such an approach will prevent "at-risk" children from falling further behind their peers in later grades (Pianta & Howes, 2009; Lee & Burkam, 2002; Camilli, Vargas, Ryan, & Barnett, 2010). There are three ways that public pre-K might close achievement gaps: (1) If pre-K quality experienced by poor/minority students is higher, on average, than that of non-poor/non-minority students; (2) If, despite equal levels of quality, poor/minority students benefit more from pre-K; and (3) If public pre-K is disproportionally attended or attended for more years by poor and minority students. This paper focuses primarily on investigating (1) whether pre-K is higher (or, at least not different) in quality for poor/minority children compared to that received by their non-poor/non-minority peers.

[1] Dual language learner (DLL) refers to students whose home language is not English. It is used instead of English language learner (ELL) because in pre-K children are not yet classified as ELLs. DLLs are still acquiring their first language simultaneously with English, so DLL is a more appropriate description of children in pre-K.

Overwhelmingly the research indicates that the highest quality programs yield the largest benefits for children (e.g. LoCasale-Crouch et al., 2007; Mashburn et al., 2008), and especially children of disadvantaged backgrounds (Garces, Thomas, & Currie 2002; Magnuson, Ruhm, & Waldfogel, 2007; Vandell, 2004). While this is good news, as it indicates that there are levers that can be manipulated to improve the outcomes of young "at-risk" children, it also indicates that if children do not have equal access to high-quality programs across different racial, socioeconomic, and language groups, the goal of narrowing achievement gaps may be difficult to achieve. Thus, understanding whether there are gaps in the quality of programs attended across groups and if so what factors may be driving such gaps is critical to increasing the likelihood that public pre-K will have its desired effect.

In this paper, I pursue four goals. First, I examine 42 different indicators of classroom quality to measure the average differences in quality in state-funded pre-K programs attended by students of different racial, ethnic, and socioeconomic groups and between DLL students and their native-English speaking peers. Second, I examine whether these gaps can be explained by: structural program characteristics, differences in average initial academic ability and social competence among students in the classroom upon entry to the program, classroom composition as measured by student demographics, and teacher race and use of a language other than English in the class. Third, I examine whether quality gaps vary among states. Finally, I consider whether the magnitude of state-level quality gaps can be explained by between-state differences in the rate of expansion of the state pre-K programs over time, in levels of public pre-K spending, and levels of residential segregation. This last piece could have implications for how to improve policies in states where quality gaps are currently large.

II. Background Literature

What is high quality pre-K?

The quality of children's early childhood learning experiences can be measured in a number of different ways, but the various measures of early childhood education (ECE) quality can generally be grouped into those that measure structural features of ECE programs and those that measure classroom processes and interactions among teachers and children.

Structural features of pre-K programs. Structural quality refers to factors that might indirectly affect child outcomes through classroom processes (NICHD ECCRN, 2002; Justice, Mashburn, Hamre, & Pianta 2008). Structural features of quality include things such as spending per child, operation schedule as full-day or half-day, teacher credentials, staff-child ratio, early learning standards, etc. Many of these indicators of quality are of interest to policy-makers because they are easily regulated. Structural indicators of quality also include factors such as classroom materials, commonly captured in some items of scales such as the Early Childhood Environmental Rating Scale (ECERS; Harms, Clifford, & Cryer, 2005), and time allocation across various classroom activities. The latter two can often fall under the category of both structural and process quality. For example, time allocation is comprised of two factors: *what* kinds of learning activities children are engaged in (structure) and *how* those learning activities are executed (process; discussed further below). There is some evidence that although structural dimensions of quality are not strong direct predictors of child outcomes, structural characteristics of settings are important prerequisites for fostering high quality classroom interactions or process quality, which in turn impact child outcomes (Nores & Barnett, 2014; NICHD, 2002).

Classroom process in pre-K programs. Measures of classroom process focus on what actually happens in classrooms, rather than on structural features of the program. Process generally captures teacher-child interactions, including teachers' sensitive and responsive

caregiving, attunement to children's cognitive and emotional needs, use of strategies that will scaffold children's learning, and use of open-ended questions and expansions to facilitate complex thought development among children. A large body of theory and research suggest that high-quality interactions between teachers and children are principle mechanisms that drive children's development (e.g. NICHD ECCRN, 2002; Morrison & Connor, 2002; Pianta, 2006; Pianta, Steinberg, & Rollins, 1995; Mashburn et al., 2008). Classroom quality measures such as the Classroom Assessment Scoring System (CLASS; Pianta, La Paro, & Hamre, 2008) are often used to measure classroom process. Such measures include indicators of both social and emotional classroom climate and instructional quality. Also, ECERS subscale measures teacher-child interactions and teachers' attempts to foster children's language and reasoning.

Is the definition of process quality universal?

Classroom processes are generally strong predictors of student outcomes (Mashburn et al., 2008), but there are many measures of classroom process, and while the measures generally provide information about what is happening in different classrooms (e.g. proportion of the day spent in free play), the interpretations of measures as indicating "quality" isn't always clear. Furthermore, what pre-K quality measurement tools reward as "high quality" may impact children of different backgrounds in different ways.

For instance, one indicator of pre-K classroom process is the amount of time spent in free-play versus direct instruction, with much free-play often seen as the gold standard. However, Chien and colleagues (2010) find that pubic pre-K attending children who spend more time in free-choice activities realize significantly *smaller* gains on a range of academic outcomes than their peers in classrooms with more direct instruction. Although these short term findings do not eliminate the possibility that free-play is beneficial for other child outcomes in the long-term,

they do suggest that pre-K programs engaging in much free-play may be less likely to increase students' school readiness through academic knowledge and skills.

Definitions of quality may further be culturally defined (Howes, 2010; Fuller & Clarke, 1994; Ladson-Billings, 1995; Kermani & Brenner, 2000; Baumrind, 1972; Chao, 2000). In some communities (African American ones in particular), didactic and directive instruction is often seen as more desirable than student-directed exploratory play, while in others (mostly white, middle class ones) scaffolded instruction is seen as the ideal (Slaughter, 1987; Pellegrini, Perlmutter, Galda, & Brody, 1990; Stipek, 2004). For this reason it is not clear that classrooms serving students of different backgrounds *should* look the same, as children who are used to certain styles of caregiving may respond most positively to the styles that they are most used to.

Because of these ambiguities in what defines "high quality" pre-K, I will first present gaps in the classroom environments of children of different groups. Later, in the discussion, I will elaborate on whether one should think of these gaps as "unjust" differences in quality, appropriate differences in instructional differentiation, or neutral differences in preferences.

The link between pre-K and child outcomes, by subgroup

When states fund pre-K, often the intention of such programs is to improve the future school performance of poor and minority children and ultimately narrow the achievement gap. It presumes to do this by (a) expanding access to early childhood educational experiences for disadvantaged children and (b) providing programs that are high in quality, particularly for the most disadvantages students.

We know that the largest effects of preschool tend to result from the highest quality programs (Barnett, 2011; Burchinal, Vandergrift, Pianta, & Mashburn, 2010; Camilli et al., 2010). Furthermore, the overwhelming majority of rigorous early childhood evaluations are of

programs serving children in poverty (Leak, Duncan, Magnuson, Schindler, & Yoshikawa, 2010; Yoshikawa et al., 2013), and these studies generally find that high intensity programs (both in terms of quantity and quality of instruction) yield large benefits for poor and/or minority children (see Schweinhart, Montie, Xiang, Barnett, Belfield, & Nores, 2005; Campbell, Ramey, Pungello, Sparling, & Miller-Johnson, 2002). Research is mixed on whether pre-K is *more* effective for disadvantaged children than their more advantaged peers, but at a minimum this research suggests that high quality pre-K programs (Phillips, Mekos, Scarr, McCartney, & Abbott-Shim, 2000; Barnett, Carolan, Squires, & Brown, 2013) are equally beneficial for poor and minority children as they are non-poor non-minority children (Gormley, Gayer, Phillips, & Dawson, 2005; Gormley, 2008; Weiland & Yoshikawa, 2013; Magnuson, Lahaie, & Waldfogel, 2006), and in a handful of cases may be more effective for the former group than the latter.

Given the above research, pre-K certainly has the *potential* to narrow achievement gaps, but mainly if (a) and/or (b) described above are true. Less is known about whether the quality of state pre-K programs that disadvantaged children attend are, on average, high, and/or as high as those attended by their more advantaged peers. Perhaps more importantly, if quality is unequal between these groups, it begs the question of what might be driving these inequalities.

Differential access to quality pre-K by student backgrounds

Across preschool settings (including child care and Head Start), research suggests that poor and minority children (and black children in particular) are less likely to be in the highest quality and most stimulating programs than their white non-poor peers (Barnett, Carolan, & Johns, 2013; Early et al., 2010; see also Bassok, Fitzpatrick, Greenberg, & Loeb, 2013). Presumably, the purpose of public state pre-K is, in part, to narrow this gap in access to quality.

A handful of studies have considered whether state-funded pre-K programs serving higher proportions of poor and minority children are lower in quality than those serving lower proportions of these groups of children. A 2010 case study of just under 4,000 Georgia pre-K classrooms found that while higher minority, lower-income communities had higher scores on structural measures of quality than did lower minority and higher income-communities, they had the lowest scores on the CLASS (Bassok & Galdo, 2015). Another study of 238 classrooms in six state pre-K programs (six of the 11 considered in the current study) examined whether classrooms with higher proportions of minority children and children living in poverty were linked to lower classroom quality. Using broadly defined high versus low quality "profiles", these researchers found that programs scoring the highest in quality served the lowest percentage of non-white and poor children (49%) compared to the lowest quality programs (73% and 65%, respectively). Finally, using data of the same approximately 700 classrooms in 11 state pre-K programs analyzed in the this paper, Chien and colleagues (2010) found that among four profiles of classroom instruction (free-play, scaffolded, individual, and group instruction), higher proportions of black, Hispanic, and poor children were enrolled in classrooms of the "individual instruction" profile than white non-poor children. The reverse was true of the other three profiles.

While the above three studies shed light on the disparities in state pre-K quality and classroom process, the current study extends the Pianta et al (2005) and Chien et al (2010) studies in several important ways. First, the prior literature focuses on broad profiles of quality, which is useful for thinking about equity to access overall, but makes is difficult to draw detailed policy implications. For this reason, the current study conducts analyses using 42 very specific indicators of quality. Second, prior research focuses on the proportion of children of various backgrounds enrolling in classrooms of different levels of quality. Instead, presenting quality

differences in the form of standardized gaps at the student level, as the current paper does, has three main advantages; (1) it allows for the comparison of gap magnitudes across measures on a uniform metric to understand whether quality gaps are a bigger issue for some dimensions of quality than others, (2) it allows one to compare the magnitude of quality gaps to the magnitude of achievement gaps, and (3) it describes average differences in quality experiences of children of different subgroups rather than describing the average demographic composition of classrooms meeting some quality criteria. Third, the prior literature considers quality differences by income and race, with little attention to an ever growing population of pre-K attendees – DLL students. This paper fills this gap. Fourth, while the prior literature speaks, in part, to the first question of this study (about whether there are differences in pre-K quality across student groups), it does not address any of the remaining questions; namely whether the size of "quality gaps" varies across states, which could have important policy implications if some states have large gaps while others do not. Finally, the prior literature does not investigate what state- and classroom-level factors most strongly predict quality gaps. This last piece is critical for understanding how to close quality gaps if they exist, and for crafting future policy to ensure that high quality pre-K is evenly and equitably distributed across groups.

Factors that predict classroom process

There are a number of different factors that could lead to measured differences in pre-K process. The most obvious and perhaps salient include: (1) structural differences, (2) differences in students' skills prior to program entry, (3) differences in teacher and/or student race/ethnicity, and (4) state differences in pre-K policies and neighborhood characteristics.

Insert Figure 1

(1) Differences in structural features of quality could lead to differences in process quality if, for example, programs serving higher proportions of poor and minority children are those employing teachers with less training and/or skills. If pre-K is anything like K-12, programs in the poorest neighborhoods may have the most difficulty attracting and retaining good teachers (Lankford, Loeb, & Wyckoff, 2002). (2) Differences in students' skills prior to program entry could also drive differences in classroom process if (a) classrooms serving higher portions of disadvantaged children are more difficult to manage (because they enroll more children with behavioral challenges), (b) teachers target instruction towards more basic academic skills to catch disadvantaged students up academically, or (c) teachers of lower credentials/skills are differentially sorted into these classrooms. (3) Differences in teacher race across programs could drive differences in classroom process if teachers have different cultural ideals about child rearing and thus engage in practices that are not typically rated high by conventional measures of pre-K process (Slaughter, 1987; Chao, 2000; Kermani & Brenner, 2000), whereas student race could if teachers adapt their teaching styles to be observant of community definitions of appropriate caregiving styles (see Fuller & Clarke, 1994; Howes, 2010; Slaughter, 1987; Stipek, 2004). Finally, (4) states could influence differences in classroom processes either through their influence on state policies (e.g. state-mandated pre-K funding per child), which in turn would influence structural features of quality and thus classroom policies, or through their lack of oversight of or ability to change characteristics of their states (see next section for further discussion). Figure 1 illustrates these relationships.

Virtually all of the existing literature on which factors predict quality have considered the degree to which structural indicators of quality predict classroom process. These studies tend to indicate that albeit weak, there is some relationship between factors such as teacher credentials

and instructional quality (Pianta et al., 2005; LoCasale-Crouch et al., 2007; NICHD, 2002; Nores & Barnett, 2014), but not higher order teacher skills like language modeling (Justice et al., 2008). Other factors such as child-teacher ratio and length of the program day (full vs. half) predict quality as measured by classroom time allocation (LoCasale-Crouch et al., 2007) but not overall classroom quality as measured by the CLASS and the ECERS (Pianta et al., 2005).

State pre-K policy and between-state variation in quality

There is much between-state variation in how state pre-K is regulated, which in turn likely has implications for the quality of pre-K classroom processes as described above. To date just over half of all state pre-K programs require lead teachers to have a B.A., 28% require that the assistant teacher have a CDA, and 60% require regular site visits (Barnettet al., 2013). The majority, but still not all (about 85% each) require teachers to have specialized training in ECE, class sizes to be 20 children or lower, and classrooms have a student-teacher ratio of 1:10 or better. One study has found that there is more between-state variation in pre-K quality than there is within-state variation (Pianta et al., 2005). If this is true, then considering which state-level features are the strongest predictors of high quality pre-K seems critical for helping to reform state pre-K policy to the end of improving quality for all kids.

There is also much between-state variation in state pre-K spending per-child, the rate at which states have increased the proportion of children served in pre-K over time, and the state's level of residential segregation (see Table iii-a and Figure iv-h of appendix; see also, Barnett et al., 2013). Any of these factors could explain the magnitude of quality gaps if they lead to self selection of the best pre-K teachers into the lowest need pre-K programs. One hypothesis is that states that pay more per child might be able to recruit pre-K teachers from a broader, more highly qualified labor force to fill positions in densely populated poor and minority neighborhoods.

Another is that states that expanded their pre-K programs the quickest are also those that will have the largest quality gaps because as more pre-K teaching positions become available within a state, the best pre-K teachers sort into the highest SES schools. Similarly, as research in K-12 has found, teachers have preferences to teach in schools with large numbers of white, high-ability students (Boyd, Lankford, Loeb, Ronfeldt, & Wyckoff, 2011; Jackson, 2009), so states with high levels of residential segregation may also be those with the largest gaps in the quality of pre-K experiences across groups simply because of limited access to high quality teachers.

While we know a lot about differences across states in pre-K regulations, we know less about differences across states in process quality experienced by children of different racial, language, and socioeconomic backgrounds.

III. The Current Study: Data and Methods

Sample

Data for this study came from the National Center for Early Development and Learning (NCEDL) Multi-State Study of Pre-Kindergarten, and the State-Wide Early Education Programs Study (SWEEPS). Data for the studies were collected in a total of 11 states -- 6 states (California, Georgia, Illinois, Kentucky, New York, and Ohio) in 2001-2002 and 5 states (Massachusetts, New Jersey, Texas, Washington, and Wisconsin) in 2003-2004, respectively. These two studies implemented the same measures with the intention of combining data sets for the purpose of analysis. The 11 states were selected to represent those that had committed significant resources to state pre-K initiatives and that had been in operation for several years. At the time of data collection, 80% of all children in the U.S. participating in state-funded pre-K were enrolled in one of these 11 states, and 83% of all state dollars spent in pre-K during the

time of study were spend in these states (Barnett, Robin, Hustedt, & Shulman, 2003)[2]. The Multi-state study involved a stratified random sampling of 40 state pre-K sites within each of the 6 selected states, while the SWEEP study involved a stratified random sample of 100 sites in each of the 5 study states. Finally, across both studies, one classroom was randomly selected within each selected pre-K site. A total of 647 classrooms with 12,334 pre-K students (aged 3 and 4) are included in the final sample, which represents 90% of all sampled classrooms that had complete data on all classroom quality measures.

Classroom Quality Measures[3]

I use measures of two main aspects of ECE classrooms in this study: structure and process. Structural measures were primarily collected through teacher and program director surveys, while process quality was collected through direct classroom observation. It should be noted that two direct classroom observation protocols (the ECERS and Snapshot) capture both classroom structure and process. For example, the ECERS measures *what* kind of activity (e.g. dramatic play) children are engaged in (structure), and *how* teachers facilitate learning (process). For this reason, the measures section is organized by survey or observation tool, rather than by category of quality. One additional measure is also included – teacher beliefs about childrearing. While this measure does not cleanly fall into either the structure or process quality buckets, and is not typically thought of as a measure of preK quality, it is included because teacher beliefs may play a critical role in determining how teachers structure their classrooms and facilitate classroom process. Means and standard deviations of all measures are presented in Table 1.

[2] Although new states have created pre-K programs, and existing states have expanded their pre-K programs since the time of data collection, these 11 states still represent the majority (approximately 60%) of all children enrolled and funds allocated to pre-K in the present day (Barnett, Carolan, Squires, & Brown, 2013)

[3] Additional measure details, including extended descriptions and psychometric properties, can be found in Appendix A.

Survey measures of structural quality. In addition to aspects of the ECERS and Snapshot, four main measures of structural quality are used in this paper. These include staff-child ratio, class size, teachers' years of experience with preschool-aged children, and teachers' years of education. In this sample, the average child-staff ratio was 8.7 (*sd* = 3.6), class size was 17.6 (*sd* = 4.4), teacher years of education was 16 (*sd* = 1.8), and teacher years of experience with preschool children was 8.9 (*sd* = 7.0). These measures were assessed through director surveys for the former two questions and teacher surveys for the latter two.

Classroom Assessment Scoring System (CLASS). The pre-K version of the CLASS, a measure based solely on classroom process, was used in this paper (Pianta et al., 2008). The most current version of the CLASS measures three dimensions of quality: instructional climate, emotional support, and behavior management. A factor analysis of the version of the CLASS used in these studies was conducted by the original study researchers (see Pianta et al., 2005). This analysis yielded two factors of process quality, which are those used in this paper: Emotional Climate (*m* = 5.56, *sd* = 0.68) and Instructional Climate (*m* = 2.07, *sd* = 0.83).

Individual items that were used to derive the Emotional Climate Dimension include: Positive Climate (enthusiasm, enjoyment, and respect during interactions among children or between teachers and children), Negative Climate (level of negative emotional tone of the classroom as expressed through emotions like anger), Teacher Sensitivity (degree of teacher comfort, reassurance, and encouragement), Over Control (rigidity and structure of the classroom environment), and Behavior Management (use of effective methods to anticipate and redirect children's misbehaviors). Items used to derive the Instructional Climate Dimension include: Productivity (effective management of instructional time and routines to ensure learning occurs), Concept Development (teacher strategies used to promoted higher order thinking), Learning

Formats (availability and arrangement of activities to maximize student engagement), and Quality of Feedback (quality of verbal feedback provided in response to children's ideas and work). Each of the 9 items is rated on a 7-point scale, with a score of a 1 or 2 for low quality, 3, 4, or 5 for mid-range quality, and 6 or 7 for high quality.

Early Childhood Environmental Rating Scale-Revise (ECERS-R). The ECERS-R, a 43-item measure of structural and process classroom quality was also used. Each item is scored on a 7-point scale, with odd-anchor scores 1 "inadequate", 3 "minimal", 5 "good", and 7 "excellent. The instrument evaluates programs across seven domains: Space and Furnishings; Personal Care Routines; Language and Reasoning; Activities; Interactions; Program Structure; and Provisions for Parents and Staff. Although there is some overlap (e.g. many of the activities items measure aspects of both structure and process), by and large the Language and Reasoning, Activities, and Interactions subscales measure process, while the remaining subscales measure structure.

Consistent with prior research (e.g. Peiser-Feinberg et al., 2001), the original study researchers found two main factors of the ECERS-R (see Pianta et al., 2005) that are also used in the current analysis: Language and Interactions ($m = 4.71$, $sd = 1.18$) and Provisions for Learning ($m = 3.75$, $sd = 0.99$). Language and Interactions is a composite of indicators measuring factors like staff-child interactions, discipline, supervision, encouraging children to communicate, and using language to develop reasoning skills. Provisions for learning measures factors that are more structural in nature (i.e. the materials and environmental structure in place to facilitate learning) such as furnishings, room arrangement, blocks, dramatic play, art, and nature or science materials.

Emerging Academics Snapshot (Snapshot). The Snapshot (Ritchie, Howes, Kraft-Sayre, & Weiser, 2001) is a pre-K observation measure that captures the moment-by-moment

experiences of children in the classroom. In the Multi-State Study, the Snapshot was conducted on two separate days during the spring, but due to funding constraints in the SWEEP study it was conducted on a single day in the spring.

The Snapshot is conducted by observing 4 randomly selected children within each classroom for 20 second observation periods followed by 40 second coding periods. Each of the 4 children are observed in succession before returning to observe the first child again. The cycle continued for 5 iterations, at which point the observer stopped coding the Snapshot for at least five minutes to code other measures of overall classroom quality before returning to the Snapshot to complete additional cycles. On average, children in the current subsample were observed and coded across 50 one minute cycles (*sd* = 22.4). The snapshot consists of a list of codes that fall in three domains to be coded as present or absent during each 20 second observational period. In classes where a language other than English was spoken, the measures were still coded regardless of which language the indicator took place in. Codes for each domain and item were averaged across kids within classrooms to obtain classroom-level measures of the proportion of time children spent, on average, engaged in various activities and settings

The snapshot is comprised of three main domains, indicators of which were coded in each domain if present during each 20 second observation interval: (1) *Setting*, (2) *Activity*, and (3) *Teacher-Child Interactions*. The (1) *Setting* domain is captured, through three, mutually exclusive categories, whether children's activities could be categorized as *free choice*, *teacher-assigned*, or *meals/routines*. Also captured in this domain was whether the activity during the 20 second period was a *whole-group* or *small-group* activity. The (2) *Activity* domain captures any of 11 learning activity codes, which were not mutually exclusive (e.g. a child could have been read to about science). *Read To* was coded whenever a child was being read to by an adult, *Pre-*

Read/Reading was coded whenever a child was reading (or pretending to read) without an adult. Additional activity codes include: *Letter/Sound*, *Oral Language Development*, *Writing, Math*, S*cience*, *Social Studies, Art*, *Gross Motor,* and *Fine Motor* activities. Capturing all literacy related activities, I also consider one combined *Engaged in a Literate Activity* measure. Finally, (3) *Teacher-Child Interactions* were coded in two ways. First, each observation interval coded whether teacher-child interactions (if present) were *Scaffolded* or *Didactic*. Interactions were also coded on a more continuous scale to measure whether interactions were present at all (*None*), were *Minimal*, *Routine*, *Simple*, or *Elaborated*. These were mutually exclusively coded, whereas Scaffolded and Didactic could have each been present and coded during a given observational cycle. Finally, whether teachers were *Distracted* was also measured.

Teacher Beliefs about Child Rearing. The Modernity Scale (Schaefer & Edgerton, 1985) was used to measure teachers' beliefs about child rearing to differentiate teachers with much more traditional adult-centered (or more authoritarian) beliefs from those with more child-centered (or authoritative) beliefs. The latter are generally associated with higher quality. The measure is a 15-item Likert questionnaire ranging from 1 (do not agree) to 5 (strongly agree). Child-centered belief items were reverse coded, and scores on the measure were derived by taking the average of all items. Teachers holding a more adult-centered view would agree with items like, "Children must be carefully trained early in life or their natural impulses make them unmanageable" and "Children should always obey the teacher." Teachers with more child-centered beliefs would more strongly agree with statements such as "Children should be able to disagree with their parents if they feel their ideas are better." Chronbach's alpha for this scale is 0.84. Prior research has shown that in child-care homes, caregiver attitudes about children and

childrearing significantly predict both classroom quality and the presence of behavior problems in young children (Clarke-Stewart, Lowe Vandell, Burchinal, O'Brien, & McCartney, 2002).

Principal Components Analysis (PCA). PCA was used to compress the large number of quality gap indicators into a more concise subset of indicators. The factor loadings (greater than the absolute value of 0.2) are presented in Table i of Appendix B and a Scree plot from Horn's parallel analysis (described in further detail in Appendix B) is presented to visually display the elbow/leveling of components in Figure ib. The process component is largely derived from the CLASS, ECERS, and a series of Snapshot indicators (free choice, elaborated instruction, and proportion of instructional time allocated to various academic activities) while the structural component is comprised of teacher-child ratio, class-size, teacher years of education, and teacher years of pre-K experience.

Additional Variables

Family Questionnaire. Questionnaires were sent home with all children in each randomly selected classroom regardless of whether children were one of the 4 randomly selected for participation on measures of child outcomes. This questionnaire collected information about the ethnicity, gender, income-to-needs ratio, maternal education level, and language status of children. Across all 647 classrooms, this lead to a sample of 12,334 children with racial/ethnic and socioeconomic information, which totals 19 children per class on average.

Classroom-level explanatory variables. In addition to the above described variables, additional variables were used to explain gaps in classroom quality to answer question two. First, the average student performance in classrooms (computed as the average score among the four randomly sampled children per classroom) was used in a subset of analyses. A small number of children (n=176) were excluded due to missing income or family size data. Teacher reports were

used to supplement data when questionnaires were missing. Both reports are believed to be accurate estimates of the ethnic composition of classrooms because the questionnaires correlate highly ($r > 0.95$) with teacher reports of the demographic composition of classrooms.

Three variables of interest were used: (1) *The Peabody Picture Vocabulary Test- 3rd edition* (PPVT-III) to measure children's receptive vocabulary skills and scholastic aptitude (Dunn & Dunn, 1997). Test-retest reliability of the PPVT is 0.93 and split-half reliability of internal consistency is 0.94. (2) The *Woodcock-Johnson Psycho-Educational Battery* (WJ; Woodcock, McGrew, & Mather, 2001) Applied Problems (AP) subtest, which was administered in both the Multi-State study and SWEEP, was used to assess academic achievement. WJ-AP assesses emergent math reasoning and problem-solving skills. Children assessed on the WJ-AP subtest analyze and solve math problems while performing simple calculations. Spanish versions of both of these tests were administered to DLLs whose first language was Spanish. This score was used in place of the English score when applicable. (3) Finally, the *Teacher-Child Rating Scale* (TCRS; Hightower et al., 1986), a scale of behavioral skills was completed by teachers. The social competence scale of the TCRS was used in this study. The social competence scale consists of 20 items rated on a 5-point scale with odd anchor points 1 "not at all," 3 "moderately well," 5 "very well." The scale was computed as the average of the 20 items. Examples of social competence items include: "participates in class discussions" and "well-liked by classmates." Chronbach's alpha for the social competence scale is 0.95.

Additional measures of structural quality were used to explain the size of process quality gaps. These measures include whether the class was full-day (53.1 percent), hours/week students spent in class (m=24.5, sd=12.7), and the teacher's wage on a per hour basis in 2014 dollars

(*m*=$26.42, *sd*=12.6), in addition to the structural measures described earlier that were used to compute quality gaps (child-staff ratio, class size, teacher years of school and experience).

Raw scores on all continuous measures (both classroom quality and student measures) were standardized relative to the full sample in all 11 states, to reflect classroom's or children's performance relative others in the study population.

Finally, teacher race and whether a teacher spoke a language other than English in the classroom were used to predict quality gaps. Fourteen percent of teachers were African American, 17% were Latino/a, 4% were Asian, and the remaining were White. Further, 37.5% of teachers spoke a language other than English in the classroom, 91% of whom spoke Spanish.

State-level explanatory variables. To answer the final question about which of three state-level factors are correlated with state-level quality gaps, three sets of variables were constructed. First, state pre-K spending per child in the year of study was collected from the State of Preschool Yearbooks (Barnett et al., 2003). Figures were adjusted for inflation to 2014 dollars. Second, rate of pre-K expansion was calculated using two data sources. The Current Population Survey (CPS) was used to construct a variable indicating the proportion of age-eligible children enrolled in public preschool by year and state. This variable was the combination of two variables, whether the child was in preschool, and whether it was public or private. Because responses to this variable could include Head Start or state pre-K, additional figures on Head Start enrollment rates by state and year were collected from the Kids Count data center and subtracted off of the CPS figures. Analyses were run both ways (subtracting and not subtracting Head Start figures) and results were consistent – likely because Head Start enrolment rates have remained fairly stable over time across states. Finally, the Information Theory Index (Theil's H) was used to compute a measure of state-level residential segregation. The Theil index

measures entropic distance the population is away from the optimal fully integrated state. Census tract-level data from the American Community Survey (ACS) was used to compute white-black, white-Hispanic, and poor-non-poor residential segregation within states. The index increases from 0 to 1, with 0 representing complete integration and 1 representing complete segregation.

Sample of children. Of the children enrolled in the classrooms used to compute quality gaps, all were between the ages of 3 and 4. Furthermore, 19.2% were black, 27% were Hispanic, 20.5% were DLL, and 54.1% were poor (defined as 150% of the poverty line). Given that the sample of students in the current paper are those enrolled in state pre-K programs, which are in many cases targeted towards low-income children, it is important to describe the "non-poor" children in the sample. The average family income of the non-poor children in the sample is $58,610 per year, compared to $29,790 for the poor children. The income distribution in the sample is presented in Figure 2, with a dashed line indicating the poor-non-poor cut-off used for an average-size family in the current paper. The median family income in the U.S. is captured by the solid line. Generally, this distribution looks similar to the income distribution of the overall population in the U.S. with two exceptions – (1) there is a slightly higher proportion of poor individuals in the current sample, and (2) there is an income ceiling of just over $110,000/year in the current non-poor sample. While not entirely representative of the population, this income ceiling likely indicates that any poor-non-poor quality gaps are *underestimates* of the potential gap in quality of pre-K experiences between these two groups. Income distributions by state are presented in Figure ia of the appendix.

Analytic Strategy

Data were analyzed using linear regression models that adjusted for the survey design structure of the data, including the primary sampling units (PSUs) or clusters within which

classrooms were randomly selected. Probability weights were also used to account for the number of classrooms in the population that each observation represents. The sampling weights took the form of the inverse probability of selection to ensure unbiased parameter estimation. In order to calculate quality gaps at the student-level, two variables were used to first expand[4] the classroom-level data set: proportion of students in each class that were of each ethnic or socioeconomic background and class-size. The product of these two variables allowed me to calculate the number of students of each background in each class, and in turn create a long data set with dummy variables for whether a given student *i* (for all students enrolled) in classroom *c* was black, Hispanic, poor, and/or DLL or not. This approach enabled me to capture all students in each classroom for the purpose of data analysis (totaling 12,334 students), rather than relying on the four randomly sampled children within each classroom (which would have totaled approximately 2,500 students) to compute quality gaps. This method is also the most straightforward approach to calculating gaps, as it allows one to incorporate analytic weights through the expansion process, and then adjust for probability weights during regression models. Regression models (described in further detail below) were used to take into account the complex sampling design and sample weights. Jackknife standard errors[5] were also estimated in all regression models (see Kolenikov, 2010).

To answer the first question of interest, as to whether there are standardized black-white, Hispanic-white, poor-non-poor, and DLL- non-DLL gaps in pre-K classroom quality, a

[4] Analyses were also conducted on a non-expanded data set as a specification check to ensure that the expansion method did not produce artificially small standard errors due to increased sample size. Here, Q is a classroom quality measure. To examine the average difference in Q between the classrooms of students of group A and B, I first standardize Q, and then compute $\bar{Q}_A - \bar{Q}_B$, where $\bar{Q}_A = \frac{\sum_c w_c t_c \pi_{Ac} Q_c}{\sum_c w_c t_c \pi_{Ac}}$. c indexes classrooms, w is a classroom weight, t is the total enrollment in the classroom c, and π_{AC} is the proportion of students in classroom c who are in group A. The same is computed for group B and the two values are differenced. Both methods produced identical gap and standard error estimates. It can be shown mathematically that the variance of the analytic weighting approach (described in this footnote) is identical to that produced through expansion (Gould, 1999).

[5] Jackknife standard errors should produce results that are asymptotically the same as bootstrap standard errors, and are the best suited to estimate standard errors with complex survey designs.

standardized (mean 0) version of each quality measure (Q_{ic}) for each student *i* in each classroom *c* was regressed on a dummy variable, G_{ic}, to represent student *i*'s group (i.e. race, SES, or language status). Quality gaps are standardized relative to the entire sample of students. Four separate regressions were run for each quality outcome (to capture the black-white, Hispanic-white, poor-non-poor, and DLL-non-DLL gaps, each separately), using a subpopulation[6] option within the survey regression models to yield a constant sample across models estimated. Models were equivalent in form to those described in reference to question two below, except without the covariates X_c and state fixed effects Γ_s.

To answer question two, about which factors might explain the size of quality gaps between groups, a model of the following form is estimated:

$$Q_{ics} = \beta_0 + \beta_1 G_{ics} + X_c + \Gamma_s + e_{ics}$$

Where Q_{ics} is the standardized quality experienced by student *i* in classroom *c* in state *s*. β_0 captures the standardized classroom quality experienced by the reference group (white, non-poor, or non-DLL students). The coefficient on G_{ics} captures the standardized difference between the group of interest and the reference group (e.g. black-white) in the classroom quality experienced, conditional on controls. For most measures (i.e. those where higher scores represent higher quality), a negative coefficient indicates that the poor and minority students receive lower average quality pre-K than their respective counterparts on that measure. A handful of measures are coded in the opposite way by design. For example, higher scores on the Modernity Scale represent more traditional child-rearing views, so a positive coefficient on student dummy variables represents more traditional or authoritarian (rather than more progressive or

[6] Subpopulation estimation involves computing point and variance estimates for part of the population (e.g. just between black and white students), but is different from just restricting the sample to the observations within the subpopulation prior to running the model because variance estimation for survey data measures sample-to-sample variability (West, Berglund, & Heeringa, 2008).

authoritative) child-rearing views. Similarly, higher class sizes not the ideal for students, so a positive gap (indicating larger classes for the minority group) still favors the majority group.

X_c captures a vector of classroom-level control variables used to explain classroom quality. The vector includes four categories of variables: (1) structural characteristics of classrooms to examine whether, for instance, variation across classrooms and groups in teacher training and experience explains the size of gaps, (2) average academic and social skills of students in the classroom upon fall entry to the program to explore the possibility that if the initial skills of the students are low, teachers may target teaching to the ability of their students, (3) demographic classroom composition to consider the possibility that teachers teach differently in classrooms that are densely populated by one group over another, and (4) teacher race and a dummy for whether teachers speak a language other than English in the class to examine whether teacher race and/or language spoken may be tied to different kinds of practices, perhaps because beliefs about child rearing vary by race/ethnicity. Each set of controls are added to the model separately in four iterative regressions, followed by two additional models: a model with state fixed effects, Γ_s, and a model with all controls plus Γ_s to estimate within-state effects and control for all unobservable idiosyncratic state factors that influence classroom quality such as state pre-K policy. Including these controls and observing the degree to which the coefficient, β_1, changes in magnitude and significance indicates how much of the relationship between a student race and classroom quality is explained by other classroom and/or state factors.

To answer the third question, whether there are differences in the size of classroom quality gaps across the 11 states in the study sample, I estimate a model of the following form:

$$Q_{ics} = \Gamma_s + \Gamma_s * G_{ic} + e_{ics}$$

Where $\Gamma_s * G_{ic}$ represents a vector of state-by-subgroup dummy variables to estimate black-white, Hispanic-white, poor-non-poor, and DLL-non-DLL quality gaps within each state.

Finally, because of the limited number of states, question four, regarding which state-level factors are correlated with quality gaps is considered qualitatively through figures. Due to data-use agreements, state names have been redacted from figures answering questions three and four. States are instead assigned a state number that is constant across figures.

IV. Results

Question 1: Magnitude of standardized quality gaps

Results for the first research question, regarding the magnitude of standardized white-black, white-Hispanic, poor-non-poor, and DLL-non-DLL gaps in pre-K quality are presented in Table 2 and graphically in Figures 3 and 4. These tables and figures capture most, but not all gaps tested in the current paper. A full list of gaps (including individual items of the CLASS and extended items of the Snapshot) are presented in Tables i-a and i-b of the appendix.

*** Insert Table 2 ***

Results for this question indicate that quality gaps are large and significant on most measures, generally ranging from about 0.3 to 0.7 standard deviations (*sd*)[7], with the largest gaps on measures of emotional and instructional climate, as measured by the CLASS and ECERS. Notably, the magnitudes of these quality gaps mirror the magnitudes of achievement gaps in the sample (see Figure ib of the appendix). It is also noteworthy that on the CLASS, black-white gaps are particularly large (*gap* = -0.659 *sd* for the overall CLASS), and about double the size of Hispanic-white and poor-non-poor gaps on the same measure.

[7] Signs on gaps presented in tables and graphs represent the direction indicated by the order of subgroup names. For example, in the case of black-white, gaps are calculated as black relative to white. For this reason most negative gaps indicate that black students experience lower quality pre-K, on average, than their white peers.

Further, the Snapshot reveals some interesting patterns. Black students are significantly less likely to experience elaborated teacher-child interactions (e.g. reciprocal conversation that validates a child's feelings or conversation that expands play or ideas; *gap* = -0.371 *sd*) than their white peers, and black, Hispanic, and poor students are significantly more likely to be engaged in activities focused on basics (e.g. toileting or clean-up, *gap* = 0.281 to 0.493 *sd*). In practical terms[8], the elaborated gap equates to about 2.4% less of the instructional day consisting of elaborated interactions for black students compared to white students (for comparison's sake, consider that elaborated interactions make up 10.9% of the day, on average). The basics gap in portion of the day terms is approximately 2.4% to 4.1% more of the day (the average in the sample is 21.5% of the day). All groups are significantly less likely than their white, non-poor, and non-DLL peers, respectively, to be engaged in free-choice activities (*gap* = -0.396 to -0.582 *sd*, or 6.2% to 9.1% less of the day) and are significantly more likely to be engaged in individual time (assigned to work individually on worksheets, independent projects, computer work, etc; *gap* = 0.348 to 0.423 *sd* or 2.3% to 2.6% more of the overall day). The same pattern is apparent for didactic (*gap* = 0.212 to 0.527 *sd* or 2.6% to 6.5% more of the day) versus scaffolded instruction (*gap* = -0.199 to -0.301 *sd* or 1.4% to 1.9% less of the day), though the latter was only significant in the case of black-white and poor-non-poor gaps. Finally, the potential academic tradeoff between time allocated to letters/sounds and away from science is apparent, such that black, Hispanic, poor, and DLL students spend significantly more time engaged in activities related to learning letters/sounds (*gap* = 0.182 to 0.424 *sd* or 0.7% to 1.6% more of the day than their non-minority peers) and significantly less time engaged in science activities (*gap* = -0.259 to -0.469 *sd*, or 1.4% to 2.6% less of the instructional day), though the effect was not significant for DLL students. One possibility for this finding is simply that poor/minority

[8] Using information from Table 1, a crude estimate of the practical significance of these gaps can be calculated.

children enter programs substantially behind their non-poor/non-minority peers in literacy skills, so teachers allocate more instructional time toward learning letters/sounds to help students catch up. While sometimes gap magnitudes, in terms of proportion of the overall day seem small (i.e. a few percentage points), when on average pre-K classrooms spend only 4.1% of the day on letters/sounds and 7.2% of the day on science, for example[9], these gaps are quite sizable.

*** Insert Figures 3 & 4 ***

Gaps on structural measures of quality are less salient than those on the prior measures discussed, which may in part be a reflection of state-mandates for things like minimum teacher-child ratios and teacher degrees. Still, a few significant gaps stand out. First, the class sizes of Hispanic students are significantly larger than those of their white peers (*gap* = 0.40) [10]. This translates to Hispanic children having approximately 1.6 additional students in their classrooms, on average, than their white peers. Second, Hispanic, poor, and DLL students all have teachers with fewer years of experience than their white, non-poor, and non-DLL peers (*gap* = -0.138 to -0.396). In practical terms, these estimates equate to about 1.6, just over 1, and 2.7 fewer years of experience in pre-K, respectively, among teachers teaching these students versus their non-minority peers. Finally, it is noteworthy that DLLs, however, have teachers with *more* years of education than do their non-DLL peers (*gap* = 0.219, or approximately 0.4 years of education more). This may be because some states require additional credentials or a Master's degree to teach in classrooms serving DLLs – bilingual classrooms in particular.

In addition, there are significant gaps for all groups on the measure of traditional child-rearing views, such that minority and poor students are more likely to have teachers who endorse more adult-centered authoritarian beliefs than their non-minority and non-poor peers.

[9] Reference Table 1 for a full list of the proportion of the day spent on the discussed activities/settings.
[10] Note that this gap is positive because class-size is an increasing number. But larger class sizes are generally viewed unfavorably, so still, this gap favors white students.

Finally, gaps the process and structural components from the PCA, which generally summarize all of these results, can be found at the top of Table 2 and in the left most panel of Figure 3. There are very large gaps for all groups, on the process component, ranging from -0.98 to -1.44 *sd*. The gap on the Structural component is only significant for the Hispanic-white (-0.609 *sd*) and DLL-non-DLL (-0.508 *sd*) gaps.

Question 2: Classroom factors that explain gap magnitudes

Results for the second research question as to what classroom factors explain the size of gaps are presented in Tables 3a and 3b. These tables present results for a subset of the significant quality gaps, but all quality gaps can be found in Figures 5-8, which visually depict the findings. The remaining significant quality gaps from Table 2 are presented in the appendix in Tables ii-a and ii-b. Four separate models (presented in the columns of Tables 3a and 3b) demonstrate how the inclusion of each set of variables changes gap magnitudes. P-values of the Joint F-test that the covariates in each model jointly equal zero are presented beneath each adjusted gap estimate. Each row represents a different race gap. Each cell is estimated from a separate model[11].

Insert Figures 5-8

The first noteworthy finding is that with the exception of the process quality component and the DLL gap on free choice and individual time, the set of structural classroom indicators (including class size, teacher-child ratio, teacher degree, years of experience, whether class is full-day, hours/week spent in class, and teacher wage) explain virtually none of the quality gap magnitudes. For example, the overall CLASS score – comparing across the unconditional (left-most) column and the structural column – gaps remain equally large in magnitude and equally significant before and after controlling for these factors. This suggests that differences across groups on structural measures of quality do not account for the size of process quality gaps.

[11] Estimates on all covariates included in each model made available upon request.

However, these structural indicators are often jointly significant predictors of quality, indicating that structural factors covary with process quality, but do not vary much between groups.

Second, the set of average student ability variables (as measured by average student vocabulary, math, and competence in the classroom; column 3) and classroom racial/socioeconomic composition variables (column 4), each separately explain about half of the size of the quality gaps and approximately 50% of gaps remain significant in both cases. Most notable, average class ability explains virtually all of gap magnitudes on the ECERS and classroom composition explains virtually all of the gap on the scaffolds measure of classroom process. Still, despite the finding that these factors explain a substantial amount of the magnitude of quality gaps, in many cases, as in the CLASS, free choice, and other indicators for some groups, statistically significant gaps of approximately -0.15 to -0.30 *sd* remain even after adjusting for these factors.

Thirdly, teacher race and whether a teacher speaks a language other than English in class explains little to none of the magnitude of quality gaps on the CLASS, ECERS, individual time, and overall process, but explains all of the size of black-white and Hispanic-white gaps on teachers' beliefs about child-rearing, and in most cases at least half of the size of the free-choice and scaffolding gaps. Implications discussed further in the discussion.

After adjusting for all controls, about a third of all estimated gaps remain significant, though mostly only showing coefficients of half to a third as large as the original magnitude. Controlling for state fixed effects explains about 50% of gaps, which suggests that half of quality gap magnitudes are due to the fact that disadvantaged students are concentrated in states with lower pre-K quality than the average. This may suggest that it is just as important to improve overall quality as it is to close quality gaps in some states. Best seen in the last panel of Figure 8,

controlling for all factors and state fixed effects explains virtually all of the magnitude and significance of gaps (only 23% of gaps remain significant, and those that do are quite small).

*** Insert Table 3a & 3b ***

Question 3: Variation across states

Results for question three, whether there is between-state variation in the magnitude of quality gaps are presented in Figure 9 for the overall CLASS and the ECERS. Figures iii-a through iii-g of the appendix present results on additional measures. Figures are sorted by the average size of the quality gap across groups (black-white, Hispanic-white, poor-non-poor, and DLL-non-DLL) and by state. It is noteworthy that some state gap estimates are highly imprecisely estimated, while others are quite precise. The imprecision could in part be because of a small population for the relevant gap group within the state. In a handful of cases, confidence intervals were truncated because they would extend beyond the range of the x-axis.

*** Insert Figure 9***

There are three main takeaways from these figures. First, there is sizable variation in gaps across states on some measures, but not others. The most precise state estimates and also the most variation across states is evident on the ECERS and the CLASS (Figure 9) but especially the provisions for learning factor of the ECERS and the instructional climate factor of the CLASS (Figure iii-a of the appendix). There is also much variation across states in the proportion of the instructional day spent as free choice and that consists of scaffolded interactions (Figure iii-c of the appendix). Less variation is apparent on measures such as proportion of the instructional day spent as individual time, teachers' belief about child-rearing (Figure iii-d of the appendix), and measures of proportion of the instructional day spent on academic content such as science and letters/sounds (Figure iii-e of the appendix).

Second, these figures can be viewed in clusters of quality gaps by state within a figure. It is particularly noteworthy that when a state has a large gap on the measure for one of the gap groups, it tends to have a large gap on that measure for all gap groups. In other words, the sorting by state does not vary dramatically by gap group within a given quality measure. This is in part because gaps are necessarily correlated. Black-white and Hispanic-white gaps, for example, both rely on the average classroom quality of white students in the state to be calculated. Similarly, language status, race, and income are correlated, which would result in correlated quality gaps.

Third, there are some consistent patterns of state rankings across measures. Note that the sorting of states may not be consistent across all quality measures in part because some states are legitimately better performers on some measures and not others, but also in part because quality gaps for any two states ranked next to each other are rarely statistically distinguishable from each other, so some of what determines where they rank is due to random noise. Still, it is noteworthy that across gap type, states six and 11 are among those that tend to display quality gaps favoring the poor and minority students, or at the very least gaps that are not distinguishable from zero. On the other hand, states eight and 2 are among those that tend to consistently display large quality gaps that favor non-poor and white students. These findings may have future implications for learning from states that don't display large quality gaps or display "favorable" quality gaps. Finally, Figure iii in the appendix displays between-state variation in overall quality on 8 of the quality measures. This figure shows that there is much less between-state variation in overall quality than there is in gaps, but that overall quality rankings tend to mimic those of quality gaps.

Question 4: State-level factors that explain between-state differences

Finally, this section reports on the fourth research question, as to whether any of three state level factors – (1) the rate at which states expanded access to pre-K since 1995, (2) state-

level spending per child in pre-K, and (3) state-level residential segregation – correlate with the size of state-level quality gaps. Simple correlational illustrations demonstrate that both state-level rate of pre-K expansion and spending per child are not clearly predictive of state-level quality gaps. A table with the rates of pre-K expansion by state and state pre-K spending by state is presented in table iii-a of the appendix.

However, there are relationships between quality gaps and residential segregation on a number of quality measures, particularly for black-white and Hispanic-white gaps. Figure 10 and Figures iv-a through iv-g in the appendix display results of these relationships. First, it is noteworthy that there is a clear inverse relationship between segregation and black-white and Hispanic-white gaps on the ECERS and CLASS (overall and both sets of factors); states with the largest gaps (i.e. favoring white students) tend to be those with the highest levels of residential segregation for that gap group, and vice versa. The same relationship is evident for proportion of the day engaged in free choice activities and the process quality component from the PCA. For proportion of the day engaged in individual time, the pattern is similar for black-white gaps, but visually looks the reverse (because more individual time engaged in teacher-assigned activities is generally a less desirable characteristic of pre-K). States with less residential segregation tend to be those with non-significant individual time gaps, or gaps favoring black students, while those with higher levels of segregation tend to have quality gaps favoring white students. Interestingly, there seems to be a clear positive relationship between Hispanic-white gaps in teachers' traditional beliefs about child rearing and residential segregation, such that as segregation increases, so do quality gaps, indicating more traditional beliefs among teachers serving disproportionate numbers of Hispanic students compared to their white peers. Finally, the structural quality component from the PCA shows a positive relationship between a state's

gap in structural quality (particularly the black-white gap) and the state's level of residential segregation. There is little/no relationship for gap measures of science, letters/sounds, and scaffolds/didactic. Further, there is not enough variation in poor-non-poor residential segregation to strongly detect a relationship, but of all of the measures, if any there seems to be a similar relationship between poor-non-poor gaps and segregation on the ECERS and CLASS as there is for black-white and Hispanic-white gaps.

*** Insert Figure 10***

V. Discussion

This paper estimates the magnitude of gaps in pre-K quality and classroom process between black/Hispanic and white, poor and non-poor, and DLL and non-DLL students. It is motivated by the need to understand whether there are quality gaps and if so how large they are to consider whether quality pre-K is equitably distributed across students and whether it has the potential to close the achievement gap. We know that pre-K can close as much as 50% of the achievement gap if children attend the highest quality programs (Camilli et al., 2010). Findings suggest that while pre-K has potential to do so, this end may not be achieved unless quality gaps are closed. Further, the paper explores what factors may be driving gaps to consider what levers might be able to narrow gaps through policy changes. There are four key sets of findings.

First, results highlight that pre-K quality gaps are large, ranging from about 0.3 to 0.7 standard deviations in magnitude, and mirror the size of achievement gaps (see appendix Figure ia) of children at school entry. There are three ways that public pre-K might close achievement gaps (take for example, black-white gaps): (1) If pre-K quality experienced by black students is higher, on average, than that experienced by white students; (2) If, given equal levels of quality, black students benefit more from pre-K; and (3) If public pre-K is disproportionally attended by

black students or attended for more years and pre-K is better than the counterfactual care they would have experienced. Findings from this paper conflict with the first of these conditions, and therefore suggest that state pre-K is unlikely to narrow achievement gaps as it currently exists. It is possible that achievement gaps could still be reduced through (2) or (3), but this is not likely; first, because pre-K is generally found to be equally beneficial for kids of different subgroups, but not necessarily *more* beneficial for poor or minority children than their non-poor non-minority peers (Gormley et al., 2005; Gormley, 2008; Weiland & Yoshikawa, 2013; Magnuson et al., 2006); and second, because the current push for universal over targeted pre-K may make it less likely over time that poor and minority children will disproportionately attend pre-K.

There is, however, a caveat to these findings. It is unclear that *all* of these gaps in classroom process are problematic. The magnitude of these gaps may rather be indicative of differences in racial and cultural norms of how children learn. For instance, parenting literature suggests that white parents are more likely to rely on scaffolding approaches to instruction, while black parents are more likely to engage in didactic and directive interactions (Slaughter, 1987; Pellegrini, Perlmutter, Galda, & Brody, 1990; Stipek, 2004). If these are the norms that some children are most used to, they may respond more positively than expected, academically and socially, to instructional techniques that are not conventionally rated highly by standard measures of pre-K quality. Take didactic and directive versus scaffolded and free-play instruction for example. On the one hand, poor and minority children (and black children in particular) may fair well under didactic/directive instructional conditions because it is the communication style their parents and the parents of others in their community most frequently use. On the other hand, such instructional approaches could be more likely to foster feelings of insecurity and thus lower academic performance in white non-poor children, in part because such

children *interpret* such an instructional approach more negatively (e.g. as reprimanding) because they are not used to authority figures facilitating learning in such a way. Teacher race and/or student race should be significant drivers of some gaps if this is true (discussed more below). Still, this paper finds that gaps are also present for the most objective measures of quality such as proportion of instructional time spent in science, indicating that on the whole, these quality differences are troubling.

Second, results for question two find that structural characteristics of classrooms explain little to none of the magnitude of most process quality gaps, while peers in the classroom (both in terms of average academic ability/social competence and racial/socioeconomic composition) explain 50-65% of the magnitude of most gaps. Teacher race explains the magnitude of some gaps, but not others, and is not as salient of a predictor as the average ability of students in the class. There are several key takeaways from these findings. First, the limited explanatory power of structural characteristics is perhaps in part due to the fact that there is little variation across groups on these factors, as states tend to regulate structural factors such as teacher degree and wage. Further, as this paper finds, gaps on structural measures are rarely statistically significant. Taken together, this suggests that the increasing use of Quality Rating Improvement Systems (QRIS) among states may not be very effective at ensuring that children have equal access to high quality programs across groups. This is because QRIS's overly rely on structural components of quality (Tout, Starr, Soli, Moodie, Kirby, & Boller, 2011). As this paper shows, structural features may be necessary but not sufficient conditions for ensuring equity of pre-K process quality across student groups. A handful of states, like California, Arizona, Oklahoma, Virginia, and Georgia incorporate measures such as the CLASS or other metrics of environmental quality into their QRIS, but even in many of these states, participation is often

voluntary and observational measures are often administered by in-house personnel who have incentives to rate classrooms higher than they are in reality. Perhaps a model QRIS would, among other things, include valid and reliable process quality measures, be conducted on a random sample of classrooms annually by an external observer, and be mandatory that all ECE programs participate. Many states are beginning to tie monetary incentives to scoring higher on their QRIS. Incorporating process measures into the overall QRIS scale could help to incentivize the lowest scoring process quality programs to improve in this dimension.

In addition, the finding that average student ability in the class explains a substantial portion of gap magnitudes could in part be explained by the fact that poor and minority students enter programs with lower average skills than their non-poor and non-minority peers, and thus teachers adapt how and what they teach to target instruction in a way that catches students up academically. For example, some teachers may spend more time teaching letters/sounds because their students are trailing in literacy skills, and may also spend less time in free play for fear that too much unstructured free time could slow students' progress on academic skills (e.g. Chein et al., 2010). Another possibility is that process quality measures are sensitive to much more than teachers' behaviors and practices, and rather are also a function of the kinds of kids in the classroom. Poor and minority students are more likely to have behavioral challenges than their non-poor non-minority peers (see Figure ii of the appendix; see also Reardon & Portilla, 2015), so classrooms with higher proportions of the former group of students may experience more disruptions. This could (a) lead to less instructional time spent on academic or other important content because teachers are spending more time on classroom management, and (b) decrease teacher patience and sensitivity, both of which would be reflected in process quality scores.

Teacher race and whether a teacher speaks a language other than English in class explains little to none of the magnitude of quality gaps on the CLASS, ECERS, individual time, and overall process, but explains all of the size of black-white and Hispanic-white gaps on teachers' beliefs about child-rearing, and in most cases at least half of the size of the free-choice and scaffolding gaps. There are two potential explanations for this. First, the teacher-beliefs about child-rearing finding is consistent with the notion that black and Hispanic caregivers are more likely to endorse authoritarian than authoritative beliefs about child-rearing (see Howes, 2010; Fuller & Clarke, 1994; Ladson-Billings, 1995; Kermani & Brenner, 2000; Baumrind, 1972; Chao, 2000). Second, black and Hispanic teachers may be more likely to teach in classrooms serving higher portions of poor and minority children, as is true in this data set (e.g. on average 78% of students in a black teacher's class are black/Hispanic and 70% are poor, compared to 5% and 42%, respectively in the case of white teachers). Anecdotal evidence in Head Start suggests that minority teachers are highly concerned with helping their poor and minority students to catch up to their non-poor non-minority peers, particularly academically and in the domain of literacy. If the same is true in pre-K, this could be why minority teachers are more likely to shy away from time spent in free-play. They may fear that time spent in free-play is at the expense of time spent on academics. Finally, it should be noted that many of these significant explanatory factors from question two are interrelated and may be interacting simultaneously.

Third, the paper finds that there is much between-state variation in gap magnitudes, such that some states consistently have large gaps in quality across groups, while others have null gaps, or gaps that favor poor/minority preschoolers. This finding of gaps favoring the disadvantaged students in somewhat surprising. It may be explained by particularly progressive pre-K policies in some states. This could include differential pre-K funding *within* states for slots

serving more disadvantaged children or higher salaries for teachers teaching in higher need areas. It is also possible that in these states, stronger attention is paid to monitoring process quality in programs serving the most disadvantaged children. On the other hand, it is possible that states with favorable quality gaps just have fewer disadvantaged students overall, making it more difficult to experience extreme differential sorting of disadvantaged students into lower quality programs that are more densely populated by other disadvantaged students. This latter explanation is, however, less likely, as marker weighting on figures for question four do not show a strong relationship between minority population size and segregation within states.

Fourth this paper finds that state-level residential segregation and quality gaps are correlated, particularly in the case of black-white and Hispanic-white segregation and process quality gaps, but not structural quality gaps. These findings suggest that increasing spending per child in pre-K alone is unlikely to do much to close quality gaps, and that continued pre-K expansion is unlikely to exacerbate the quality gap problem. But, where one lives is tied to the quality of pre-K he/she receives. When states have more residential segregation, black and Hispanic children in particular tend to experience worse pre-K environments than their white peers. In some ways, this finding is very predictable. Without segregation there necessarily cannot be a gap in quality, because students of different backgrounds would attend, in expectation, programs of equal quality. Still, this pattern is consistent with what is seen in K-12, where neighborhood is tied to school quality, in large part because many teachers prefer to teach in schools serving lower proportions of African American, poor, or low-achieving students (Boyd et al., 2011; Jackson, 2009). These findings suggest that one of two policy approaches might be effective at reducing gaps: incentives to recruit higher quality teachers to high needs

neighborhoods, or interventions such as professional development (Phillips et al., 1992; Pianta et al., 2005) to increase quality among existing teachers.

Limitations and Future Research

This study provides evidence of large gaps in the quality of public pre-K experiences between disadvantaged children and their non-disadvantaged peers. There are, however, study limitations. The findings about classroom factors that explain gap magnitudes should be interpreted with caution, as they certainly are not causal in nature. It is possible that other unobservable school- or classroom-level factors explain the size of gap magnitudes. Further, while compelling, the between-state variation findings and relationship between segregation and gaps warrant further research. Although this study represents the states enrolling the majority of children enrolled in public pre-K, data is only available for 11 states, and therefore cannot be extrapolated to other states or the U.S. in general. The small sample of states made it difficult to quantitatively investigate the relationship between state-level factors and quality gaps. For this reason, future efforts to explore such relationships with larger samples is warranted. Still, this evidence does provide cause to consider the possibility that segregation is a driver of disparities in quality of pre-K experiences across groups. Finally, although speculated, this paper does not reconcile the possibility that not all of the gaps are indicative of "unjust" experiences for poor- and minority children. More research is necessary to parse out the benefits and drawbacks of indicators of quality such as free play and scaffolded instruction versus individual time and didactic instruction across different student groups. It is possible that although didactic instructional styles are related to larger gains in academic outcomes in the short term (Chein et al., 2010), in the longer term free play and scaffolded approaches are be more optimal for students of all subgroups.

References

Barnett, W.S. (2011). Effectiveness of early educational intervention. *Science, 333*, 975-978.

Barnett, W.S., Carolan, M., & Johns, D. (2013). *Equity and excellence: African-American children's access to quality preschool.* Center for Enhancing Early learning Outcomes & National Institute for Early Education Research. New Brunswick, NJ.

Barnett, W.S., Carolan, M.E., Squires, J.H., & Brown, K.C. (2013). The state of preschool 2013: State preschool yearbook. New Brunswick, NJ: National Institute for Early Education Research.

Barnett, S.W., Robin, K.B., Hustedt, J.T., & Karen, L.S. (2003). The state of preschool 2003: State preschool yearbook. New Brunswick, NJ: National Institute for Early Education Research.

Boyd, D., Lankford, H., Loeb, S., Ronfeldt, M., & Wyckoff, J. (2011). The effect of school neighborhoods on teachers' career decisions. In G.J. Duncan & R.J. Murnane (Eds.), Whither Opportunity (pp. 377-395). New York, NY: Russell Sage.

Bassok, D., Fitzpatrick, M., Greenberg, E., & Loeb, S. (2013). The extent of within- and between-sector quality differences in early childhood education and care. *Unpublished Manuscript*.

Bassok, D., & Galdo, E. (2015). Inequality in preschool quality? Community-level disparities in access to high quality learning environments. *Unpublished Manuscript.*

Baumrind, D. (1972). An exploratory study of socialization effects on black children: Some black-white comparisons. *Child Development, 43*, 261-267.

Burchinal, M.R., & Cryer, D. (2003). Diversity, child care quality, and developmental outcomes. *Early Childhood Research Quarterly, 18*, 401-426.

Burchinal, M., Vandergrift, N., Pianta, R., & Mashburn, A. (2010). Threshold analysis of association between child care quality and child outcomes for low-income children in pre-kindergarten programs. Early Childhood Research Quarterly, 25, 166-176.

Camilli, G., Vargas, S., Ryan, S., & Barnett, W.S. (2010). Meta-analysis of the effects of early education interventions on cognitive and social development. *Teachers College Record, 112*, 579-620.

Campbell, F.A., Ramey, C.T., Pungello, E., Sparling, J., & Miller-Johnson, S. (2002). Early childhood education: Young adult outcomes from the Abecedarian Project. *Applied Developmental Science, 6*, 42-57.

Chao, R.K. (2000). The parenting of immigrant Chinese and European American mothers: Relations between parenting styles, socialization goals, and parenting practices. *Journal of Applied Developmental Psychology, 21*, 233-248.

Chien, N.C., Howes, C., Pianta, R.C., Burchinal, M., Ritchie, S., Bryant, D.M., Clifford, R.M., Early, D.M., & Barbarin, O.A. (2010). Children's classroom engagement and school readiness gains in prekindergarten. *Child Development, 81*, 1534-1549.

Clarke-Stewart, K.A., Lowe Vandell, D., Burchinal, M., O'Brien, M., & McCartney, K. (2002). Do regulable features of child-care homes affect children's development? *Early Childhood Research Quarterly, 17*, 52-86.

Dinno, A. (2009). Implementing Horn's parallel analysis for principal component analysis and factor analysis. *The Stata Journal, 9*, 291-298.

Duncan, G.J., & Magnuson, K.A. (2005). Can family socioeconomic resources account for racial and ethnic test score gaps? *Future of Children, 15*, 35-54.

Dunn, L.M., & Dunn, L.M. (1997). *Peabody picture vocabulary test* (3rd edition). Circle Pines, MN: American Guidance Service, Inc.

Early, D.M., Iruka, I.U., Ritchie, S., Barbarin, O.A., Winn, D.C., Crawford, G.M., Frome, P.M., Clifford, R.M., Burchinal, M., Howes, C., Bryant, D.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.

Fuligni, A.S., Howes, C., Huang, Y., Hong, S.S., & Lara-Cinisomo, S. (2012). Activity settings and daily routines in preschool classrooms: Diverse experiences in early learning settings for low-income children. *Early Childhood Research Quarterly, 27*, 198-209.

Fuller, B. & Clarke, P. (1994). Raising School Effects While Ignoring Culture? Local Conditions and the Influence of Classroom Tools, Rules, and Pedagogy. *Review of Educational Research, 64,* 119-157.

Garces, E., Thomas, D., & Currie, J. (2002). Longer term effects of Head Start. *The American Economic Review, 92*, 69-123.

Gormley, W.T., Gayer, T., Phillips, D., Dawson, B. (2005). The effects of universal pre-k on cognitive development. *Developmental Psychology, 41*, 872-884.

Gould, W. (1999). What is the effect of specifying aweights with regress? Stata Technical Bulletin, 20. Available at: http://www.stata.com/support/faqs/statistics/analytical-weights-with-linear-regression/.

Harms, T., Clifford, R.M., & Cryer, D. (2005). *Early Childhood Environmental Rating Scale Revised Edition*. New York: Teachers College Press.

Hightower, A.D., Work, W.C., Cowen, E.L., Lotyczewski, B.S., Spinell, A.P., Guare, J.C., et al. (1986). The Teacher-Child Rating Scale: A brief objective measure of elementary children's school problem behaviors and competencies. *School Psychology Review, 15*, 393-409.

Howes, C. (2010). *Culture and Child Development in Early Childhood Programs: Practices for Quality Education and Care*. New York, NY: Teachers College Press.

Jackson, C.K. (2009). Student demographics, teacher sorting, and teacher quality: Evidence from the end of school desegregation. *Journal of Labor Economics, 27*, 213-256.

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

Kermani, H. & Brenner, M.E. (2000). Maternal scaffolding in the child's zone of proximal development across tasks: Cross-cultural perspectives. *Journal of Research in Childhood Education, 15*, 30-52.

Kolenikov, S. (2010). Resampling variance estimation for complex survey data. *The Stata Journal, 10*, 165-199.

Ladson-Billings, G. (1995). But that's just good teaching! The case for culturally relevant pedagogy. *Theory into Practice, 34*, 159-165.

Lankford, H. Loeb, S., & Wyckoff, J. (2002). Teacher sorting and the plight of urban schools: A descriptive analysis. *Educational Evaluation and Policy Analysis, 24*, 37-62.

Leak, J., Duncan, G., Li, W., Magnuson, K., Schindler, H., & Yoshikawa, H. (2012). Is timing everything? How early childhood education program cognitive and achievement impacts vary by starting age, program duration and time since the end of the program. *Presented at the Association for Policy Analysis and Management annual Meeting.*

Lee, V.E., & Burkam, D.T. (2002). *Inequality at the starting gate.* Washington, DC: Economic Policy Institute.

LoCasale-Crouch, J., Konold, T., Pianta, R., Howes, C., Burchinal, M. Bryant, D., Clifford, R., Early, D., & Barbarin, O. (2006). Observed classroom quality profiles in state-funded pre-kindergarten programs and associations with teacher, program, and classroom characteristics. *Early Childhood Research Quarterly, 22*, 3-17.

Loeb, S., & Bassok, D. (2008). Early childhood and the achievement gap. In H.F. Ladd, & E.B. Fiske (Eds.), *Handbook of research in education finance and policy.* (pp. 497-516). New York, NY: Routledge.

Magnuson, K.A., Ruhm, C., & Waldfogel, J. (2007). The persistence of preschool effects: Do subsequent classroom experiences matter? *Early Childhood Research Quarterly, 22*, 18-38.

Mashburn, A.J., Pianta, R.C., Hare, B.K., Downer, J.T., Barbarin, O.A., Bryant, D., Burchinal, M. & Early, D.M. (2008). Measures of Classroom Quality in Prekindergarten and Children's Development of Academic, Language, and Social Skills. *Child Development, 79*, 732-749.

Morrison, F.J., & Connor, C.M. (2002). Understanding schooling effects on early literacy: A working research strategy. *Journal of School Psychology, 40*, 493-500.

NICHD Early Child Care Research Network (2002). Child-care structure → process → outcome: Direct and indirect effects of child-care quality on young children's development. *Psychological Science, 13*, 1999-206.

Nores, M. & Barnett, W.S. (2014). Preschool quality in structure and process. Why the same story can have different endings. Unpublished Manuscript

Peisner-Feinberg, E.S., Burchinal, M.R., Clifford, R.M., Culkin, M.L., Howes, C., Kagan, S., & Yazejian, N. (2001). The Relation of Preschool Child-Care Quality to Children's Cognitive and Social Developmental Trajectories through Second Grade. *Child Development, 72*, 1534-1553.

Pellegrini, A., Perlmutter, J., Galda, L., & Brody, G. (1990). Joint reading between Black Head Start children and their mothers. *Child Development, 61*, 443-453.

Phillips, D.A., Gormley, W.T., & Lowenstein, A.E. (2009). Inside the pre-kindergarten door: Classroom climate and instructional time allocation in Tulsa's pre-K programs. *Early Childhood Research Quarterly, 24*, 213-228.

Phillips, D., Mekos, D., Scarr, S., McCartney, K., & Abbott-Shim, M. (2000). Within and beyond the classroom door: assessing quality in child care centers. *Early Childhood Research Quarterly, 15*, 475-496.

Pianta, R.C., 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*, 144-159.

Pianta, R.C. (2006). Teacher-child relationships and early literacy. In D. Dickinson & S.Newman (Eds.), *Handbook of early literacy research* (Vol. 2, pp.149-162). New York: The Guilford Press.

Pianta, R.C., & Howes, C. (2009). *The Promise of Pre-K*. (Eds). Baltimore, MD: Brookes Publishing.

Pianta, R.C., La Paro, K.M., & Hamre, B.K. (2008). *Classroom Assessment Scoring System Pre-K*. Baltimore, MD: Brookes Publishing.

Pianta, R.C., Steinberg, M.S., & Rollins, K.B. (1995). The first two years of school: Teacher-child relationships and deflections in children's classroom adjustment. *Development and Psychopathology, 7*, 295-312.

Reardon, S.F., & Galindo, C. (2009). The Hispanic-white achievement gap in math and reading in the elementary grades. *American Educational Research Journal, 46*, 853-891.

Reardon, S.F., & Portilla, X.A. (2011). *Recent trends in socioeconomic and racial school readiness gaps at kindergarten entry*. Unpublished manuscript.

Reardon, S.F., & Robinson, J.P. (2008). Patterns and trends in racial/ethnic and socioeconomic academic achievement gaps. In H.F. Ladd, & E.B. Fiske (Eds.), *Handbook of research in education finance and policy.* (pp. 497-516). New York, NY: Routledge.

Ritchie, S., Howes, C., Kraft-Sayre, M., & Weiser, B. (2001). *Emerging Academic Snapshot*. Unpublished measure. University of California at Los Angeles.

Schaefer, E., & Edgerton, M. (1985). Parental and child correlates parental modernity. In I.E. Sigel (Ed.), *Parental belief systems* (pp. 121-147). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

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: High/Scope Press.

Slaughter, D. (1987). The home environment and academic achievement of Black American children and youth: An overview. *Journal of Negro Education, 56*, 3-20.

Stipek, D. (2004). Teaching practices in kindergarten and first grade: different strokes for different folks. *Early Childhood Research Quarterly, 19*, 548-568.

Tout, K., Starr, R., Soli, M., Moodie, S., Kirby, G., & Boller, K. (2011). Compendium of quality rating systems and evaluations. Prepared for: Office of Planning, Research, & Evaluation, Washington, D.C. Available at: http://www.acf.hhs.gov/sites/default/files/opre/qrs_compendium_final.pdf .

Weiland, C., & Yoshikawa, H. (2013). Impacts of a prekindergarten program on children's mathematics, language, literacy, executive function, and emotional skills. *Child Development, 84*, 2112-2130.

West, B.T., Berglund, P., & Heeringa, S.G. (2008). A closer examination of subpopulation analysis of complex-sample survey data. *The Stata Journal, 8*, 520-531.

Woodcock, R.W., McGrew, K.S., & Mather, N. (2001). *Woodcock-Johnson III: Tests of achievement*. Itasca, IL: Riverside Publishing.

Yoshikawa, H., Weiland, C., Brooks-Gunn, J., Burchinal, M.R., Espinosa, L.M., Gormley, W.T., Ludwig, J., Magnuson, K.A., Phillips, D., & Zaslow, M.J. (2013). Investing in our future: The evidence base on preschool education. *Foundation for Child Development*.

Tables

Table 1. Descriptive statistics: Means and standard deviations of unstandardized quality measures, child, and classroom characteristics

Teacher Beliefs ECERS Structural		CLASS: Overall, Factors, and Items		Snapshot: Teacher-Child Interactions & Setting		Snapshot: Activity		Additional Child & Classroom Characteristics	
	mean/sd		mean/sd		mean/sd		mean/sd		mean/sd
Traditional Child-Rearing Views	38.281	Total Score	4.464	Didactic	0.291	Aesthetics	0.099	Black	0.192
	(9.356)		(0.622)		(0.123)		(0.058)		(0.300)
ECERS (Total Score)	3.828	Emotional Climate Factor	5.512	Scaffolded	0.085	Letters/Sounds	0.039	White	0.402
	(0.793)		(0.678)		(0.064)		(0.041)		(0.367)
ECERS (Language/Interactions)	4.659	Instructional Climate Factor	2.036	Elaborated	0.108	Gross Motor	0.059	Latino	0.27
	(1.161)		(0.736)		(0.065)		(0.045)		(0.350)
ECERS (Provisions for Learning)	3.709	Positive Climate	5.21	Simple	0.121	Math	0.07	Poor	0.541
	(0.964)		(0.860)		(0.081)		(0.045)		(0.316)
Ratio (Children enrolled to staff)	8.846	Negative Climate	1.592	Routine	0.011	Oral Language Development	0.061	DLL	0.205
	(3.218)		(0.643)		(0.012)		(0.052)		(0.318)
Class Size	17.477	Teacher Sensitivity	4.72	Minimal	0.025	Science	0.071	Maternal Education	12.855
	(3.913)		(0.948)		(0.019)		(0.056)		(1.409)
Teacher Years of Education	16.054	Over Control	1.85	None	0.418	Social Studies	0.113	PPVT/TVIP (Vocab)	91.936
	(1.780)		(0.913)		(0.108)		(0.067)		(11.033)
Teacher Yrs Experience in pre-K	8.552	Behavior Management	5.071	Distracted	0.017	Writing	0.009	WJ (Applied Problems)	95.554
	(6.782)		(0.975)		(0.021)		(0.017)		(10.728)
Teacher Degree ECE specific	0.576	Productivity	4.53	Basics	0.212	Fine Motor	0.1	Competence (TCRS -- 5-pt)	3.439
	(0.495)		(0.884)		(0.083)		(0.053)		(0.518)
Full-Day (vs. Half-Day)	0.531	Concept Development	2.178	Free Choice	0.293	Gross Motor	0.073	Teacher is Black	0.138
	(0.499)		(0.840)		(0.156)		(0.052)		(0.345)
Hours Per Week in Class	24.496	Learning Formats	4.013	Individual Time	0.044	Engaged in literate Activity	0.161	Teacher is Asian	0.042
	(12.680)		(1.073)		(0.065)		(0.080)		(0.200)
Teacher Hourly Wage (2014 $)	26.423	Quality of Feedback	1.893	Meals/Snacks	0.117			Teacher is Latino/a	0.166
	(12.600)		(0.779)		(0.064)				(0.372)
				Small Group	0.058			Teacher is White	0.696
					(0.081)				(0.461)
				Whole Group	0.276			Instruction in Other Lang	0.345
					(0.125)				(0.476)
N (Classrooms)	647		647		647		647		647

Table 2. Classroom quality gaps between groups

Quality Indicator		Black-White	Hispanic-White	Poor-Non-poor	DLL-Non-DLL
		Principal Components Analysis			
PCA	Process Component	-1.44***	-1.35***	-.980***	-1.08***
		[.262]	[.311]	[.181]	[.272]
	Structural Component	-0.323	-.609**	-.226+	-.508*
		[.221]	[.254]	[.13]	[.239]
		CLASS			
CLASS	Total Score	-.659***	-.374***	-.377***	0.02
		[.164]	[.118]	[.096]	[.094]
	Emotional Climate Factor	-.662***	-.238*	-.296***	.144+
		[.193]	[.118]	[.108]	[.077]
	Instructional Climate Factor	-.334***	-.313***	-.298***	-0.092
		[.117]	[.099]	[.065]	[.078]
		ECERS			
ECERS	Total Score	-.567***	-.523***	-.43***	-.353***
		[.157]	[.143]	[.095]	[.102]
	Language/Interactions Factor	-.647***	-.471***	-.428***	-.17*
		[.149]	[.142]	[.1]	[.085]
	Provisions for Learning Factor	-.421**	-.547***	-.395***	-.572***
		[.17]	[.162]	[.102]	[.101]
		Snapshot			
Snapshot	Elaborated	-.371***	-0.182	-.188+	0.013
		[.124]	[.119]	[.099]	[.108]
	Free Choice	-.501***	-.582***	-.396***	-.522***
		[.142]	[.147]	[.092]	[.13]
	Individual Time	.394***	.348*	.359***	.423+
		[.136]	[.162]	[.067]	[.22]
	Didactic	.25+	.495***	.212***	.527***
		[.132]	[.129]	[.081]	[.175]
	Scaffold	-.301*	-0.111	-.199*	0.05
		[.13]	[.101]	[.099]	[.099]
	Letters/Sounds	.306***	.394***	.182*	.424+
		[.104]	[.12]	[.082]	[.227]
	Math	-0.093	0.037	-0.059	.252**
		[.09]	[.082]	[.054]	[.101]
	Science	-.469***	-.302**	-.259***	-0.218
		[.075]	[.118]	[.066]	[.152]
		Structural			
Structural	Class Size (Reverse Coded)	.229+	.4***	0.093	0.144
		[.118]	[.122]	[.065]	[.114]
	Years Experience in Pre-K	0.024	-.243*	-.138*	-.396***
		[.166]	[.114]	[.069]	[.089]
	Ratio (Children enrolled to staff)	.259+	0.312	.157+	0.248
		[.154]	[.193]	[.089]	[.209]
	Teacher Years of Education	-0.003	0.093	-0.02	.219*
		[.162]	[.146]	[.086]	[.105]
		Teachers' Child Rearing Beliefs			
Schaefer	Traditional Child-Rearing Views	.481***	.367***	.186+	.288***
		[.152]	[.104]	[.097]	[.081]
	N (Students)	12,334	12,334	12,334	12,334

Table 3a. Explaining Quality Gaps Using Stable Classroom Characteristics

	Uncon'l	Structural	Average Student Ability	Classroom Comp	Teacher Race/Language	State FE	All
	b/se	b/se	b/se	b/se	b/se	b/se	b/se
Process Component							
Black	-1.375***	-1.396***	-.689***	-.933***	-1.124***	-.974***	-.458**
	[.256]	[.194]	[.22]	[.255]	[.249]	[.152]	[.186]
Joint F		p=.000	p=.000	p=.000	p=.106	p=.000	p=.000
Hispanic	-1.334***	-.996***	-.736***	-.669***	-.92***	-.667***	-0.146
	[.322]	[.186]	[.276]	[.255]	[.284]	[.128]	[.132]
Joint F		p=.000	p=.001	p=.000	p=.055	p=.000	p=.000
Poor	-.94***	-.791***	-.448***	-.357***	-.699***	-.435***	-.102+
	[.161]	[.099]	[.096]	[.073]	[.116]	[.08]	[.054]
Joint F		p=.000	p=.000	p=.000	p=.000	p=.000	p=.000
DLL	-1.046***	-.619***	-0.442	-.478+	-.874***	-.425***	-0.179
	[.3]	[.184]	[.272]	[.25]	[.324]	[.142]	[.161]
Joint F		p=.000	p=.000	p=.000	p=.003	p=.000	p=.000
CLASS (Total Score)							
Black	-.642***	-.604***	-.301***	-.378***	-.49***	-.516***	-0.076
	[.163]	[.175]	[.115]	[.121]	[.166]	[.144]	[.138]
Joint F		p=.542	p=.000	p=.000	p=.311	p=.000	p=.000
Hispanic	-.361***	-.35***	-0.098	-0.128	-.311***	-.251**	-0.054
	[.121]	[.093]	[.107]	[.098]	[.106]	[.108]	[.092]
Joint F		p=.000	p=.003	p=.000	p=.923	p=.000	p=.000
Poor	-.366***	-.338***	-.172***	-.2***	-.29***	-.241***	-0.074
	[.094]	[.078]	[.065]	[.071]	[.079]	[.059]	[.052]
Joint F		p=.008	p=.000	p=.001	p=.028	p=.000	p=.000
DLL	0.025	0.07	.314***	.289***	0.125	.135**	.24*
	[.099]	[.068]	[.068]	[.07]	[.112]	[.057]	[.106]
Joint F		p=.029	p=.000	p=.000	p=.021	p=.000	p=.000
ECERS (Total Score)							
Black	-.533***	-.633***	-0.149	-.312*	-.517***	-.253**	-0.153
	[.155]	[.117]	[.15]	[.152]	[.161]	[.101]	[.109]
Joint F		p=.000	p=.000	p=.000	p=.264	p=.000	p=.000
Hispanic	-.503***	-.423***	-0.198	-.227+	-.497***	0.005	0.05
	[.144]	[.1]	[.127]	[.135]	[.117]	[.077]	[.081]
Joint F		p=.000	p=.013	p=.000	p=.983	p=.000	p=.000
Poor	-.408***	-.372***	-.173***	-.202***	-.355***	-.127**	-.079+
	[.087]	[.066]	[.058]	[.072]	[.077]	[.05]	[.045]
Joint F		p=.000	p=.000	p=.002	p=.396	p=.000	p=.000
DLL	-.353***	-0.129	-0.045	-0.075	-.293*	0.005	0.053
	[.115]	[.083]	[.109]	[.103]	[.132]	[.064]	[.076]
Joint F		p=.000	p=.000	p=.000	p=.235	p=.000	p=.000
Traditional Child-rearing Roles							
Black	.462***	.504***	.256+	.465***	0.098	.334+	0.026
	[.155]	[.146]	[.138]	[.135]	[.125]	[.171]	[.103]
Joint F		p=.03	p=.019	p=.000	p=.009	p=.001	p=.000
Hispanic	.358***	.368***	0.248	.283**	0.154	0.146	0.024
	[.104]	[.098]	[.154]	[.111]	[.106]	[.114]	[.11]
Joint F		p=.001	p=.012	p=.002	p=.001	p=.000	p=.000
Poor	.168+	0.152	0.025	-0.024	0.017	0.102	-0.05
	[.094]	[.097]	[.086]	[.077]	[.078]	[.082]	[.067]
Joint F		p=.039	p=.003	p=.000	p=.000	p=.000	p=.000
DLL	.279***	.306***	0.105	0.159	.206***	.246***	.199**
	[.084]	[.087]	[.101]	[.104]	[.071]	[.082]	[.081]
Joint F		p=.098	p=.007	p=.015	p=.000	p=.000	p=.000
Structural Characteristics		X					X
Average Student Ability			X				X
Demographic Composition				X			X
Teacher Race					X		X
State FE						X	X
N	12,334	12,334	12,334	12,334	12,334	12,334	12,334

Table 3b. Explaining Quality Gaps Using Stable Classroom Characteristics

	Uncon'l	Structural	Average Student Ability	Classroom Comp	Teacher Race/Language	State FE	All
	b/se	b/se	b/se	b/se	b/se	b/se	b/se
Snapshot (Free Choice)							
Black	-.47***	-.514***	-.25+	-.328**	-.436***	-.291***	-.238*
	[.137]	[.098]	[.129]	[.131]	[.144]	[.108]	[.105]
Joint F		p=.000	p=.09	p=.005	p=.052	p=.000	p=.000
Hispanic	-.577***	-.428***	-.317**	-.289*	-.343**	-.314***	-0.006
	[.152]	[.11]	[.127]	[.132]	[.142]	[.088]	[.082]
Joint F		p=.000	p=.003	p=.000	p=.014	p=.000	p=.000
Poor	-.375***	-.333***	-.172***	-.143***	-.273***	-.15***	-0.038
	[.08]	[.052]	[.041]	[.041]	[.065]	[.055]	[.034]
Joint F		p=.000	p=.002	p=.000	p=.006	p=.000	p=.000
DLL	-.493***	-.274***	-.229*	-.264*	-.347***	-0.171	-0.001
	[.138]	[.089]	[.117]	[.12]	[.132]	[.13]	[.092]
Joint F		p=.000	p=.01	p=.000	p=.016	p=.000	p=.000
Snapshot (Scaffolds)							
Black	-.299*	-.281*	-0.028	-0.138	-0.11	-.262+	0.092
	[.131]	[.128]	[.126]	[.106]	[.172]	[.137]	[.119]
Joint F		p=.016	p=.018	p=.008	p=.003	p=.000	p=.000
Hispanic	-0.083	0.013	.219*	0.096	0.178	-0.173	0.115
	[.102]	[.098]	[.112]	[.098]	[.109]	[.15]	[.072]
Joint F		p=.000	p=.045	p=.003	p=.001	p=.000	p=.000
Poor	-.196*	-.151+	-0.052	-0.126	-0.11	-0.151	-0.034
	[.099]	[.085]	[.071]	[.078]	[.093]	[.101]	[.061]
Joint F		p=.000	p=.028	p=.011	p=.000	p=.000	p=.000
DLL	0.071	0.073	.239**	.161***	0.12	-0.018	0.047
	[.094]	[.081]	[.097]	[.062]	[.075]	[.062]	[.077]
Joint F		p=.003	p=.014	p=.000	p=.005	p=.000	p=.000
Snapshot (Individual Time)							
Black	.374***	.354***	0.103	0.147	.324***	.242*	-0.011
	[.137]	[.111]	[.14]	[.138]	[.098]	[.123]	[.083]
Joint F		p=.000	p=.000	p=.000	p=.836	p=.000	p=.000
Hispanic	.373*	.215+	0.143	0.018	0.237	0.11	-0.114
	[.161]	[.11]	[.195]	[.122]	[.153]	[.111]	[.133]
Joint F		p=.000	p=.13	p=.000	p=.003	p=.000	p=.000
Poor	.375***	.32***	.257***	.234***	.318***	.209***	.161***
	[.068]	[.051]	[.067]	[.041]	[.042]	[.056]	[.049]
Joint F		p=.000	p=.002	p=.045	p=.071	p=.000	p=.000
DLL	.432*	.284+	0.283	0.205	.447*	0.189	0.177
	[.215]	[.16]	[.234]	[.196]	[.218]	[.19]	[.174]
Joint F		p=.000	p=.012	p=.000	p=.112	p=.000	p=.000
Structural Characteristics		X					X
Average Student Ability			X				X
Demographic Composition				X			X
Teacher Race					X		X
State FE						X	X
N	12,334	12,334	12,334	12,334	12,334	12,334	12,334

Figures

Figure 1. Conceptual Framework

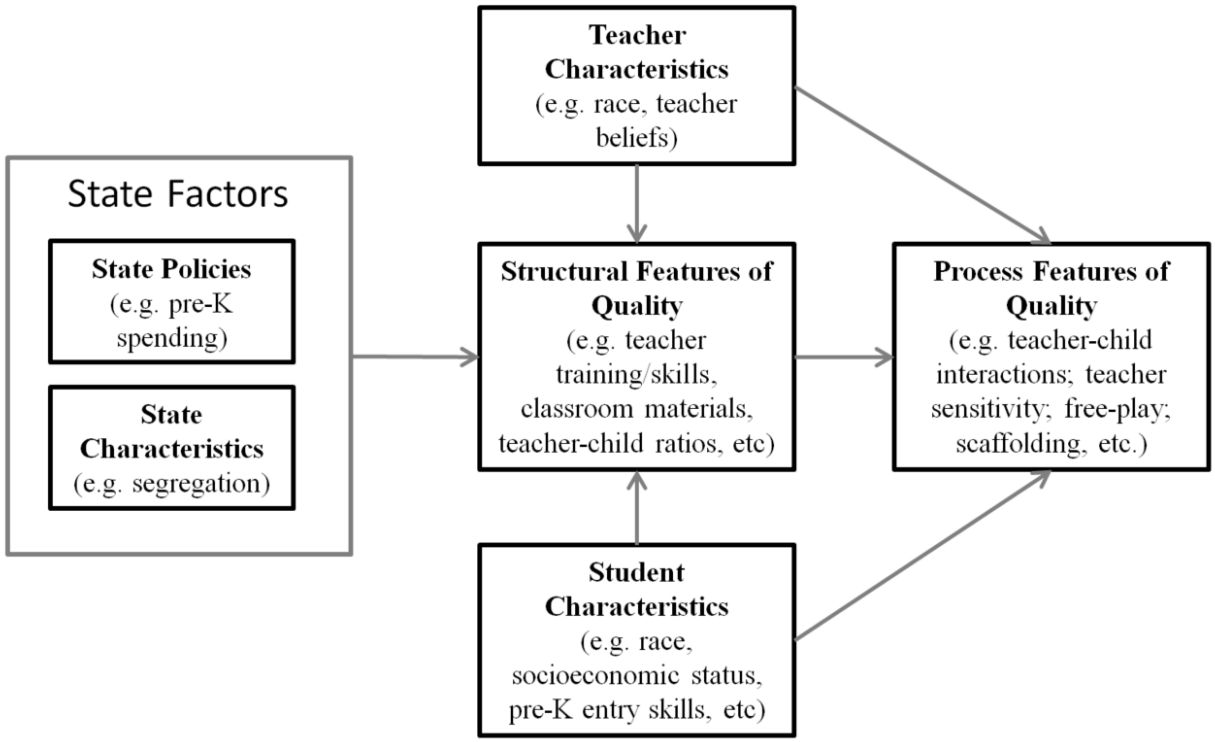


Figure 2. Income Distribution of Sample

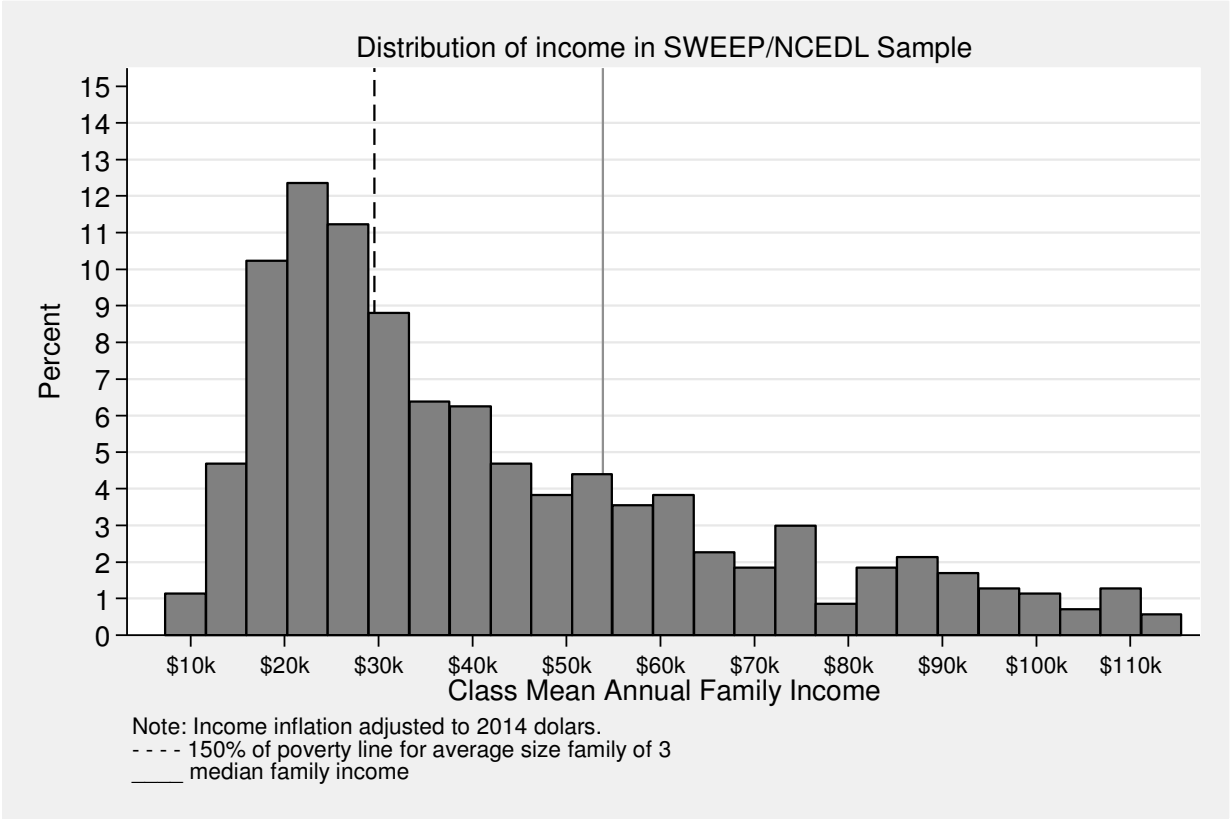


Figure 3.

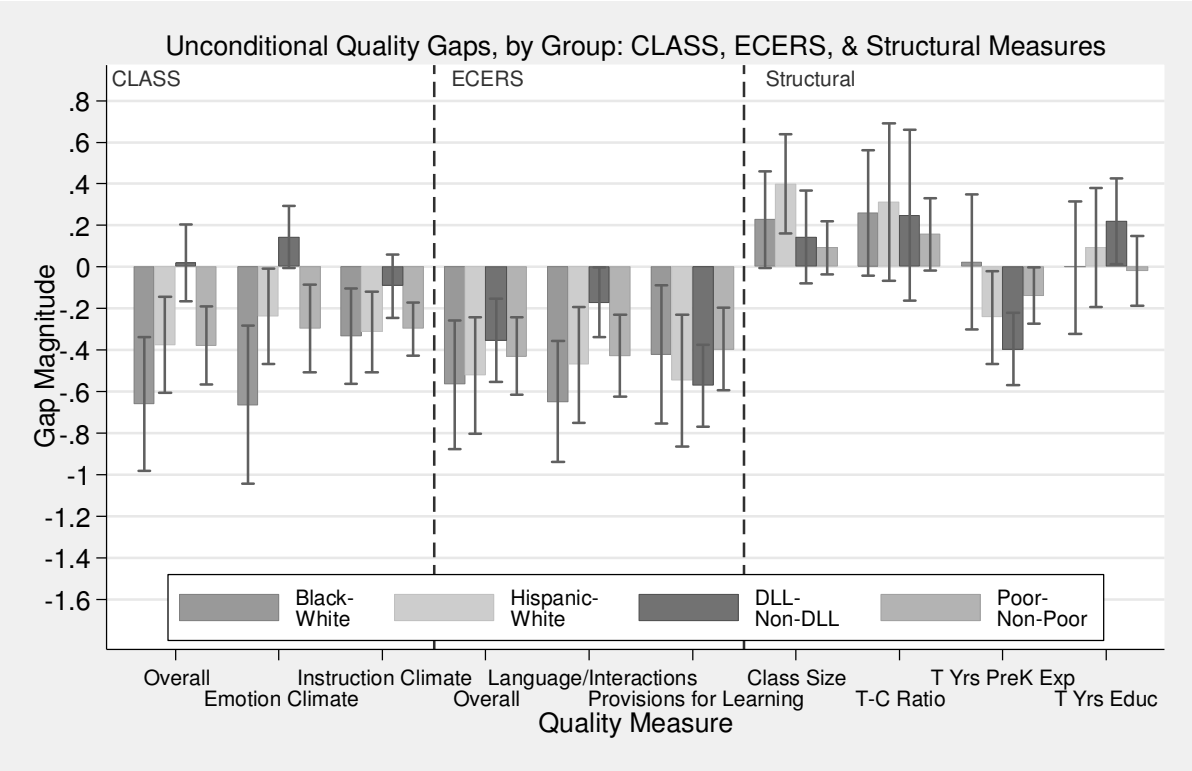


Figure 4.

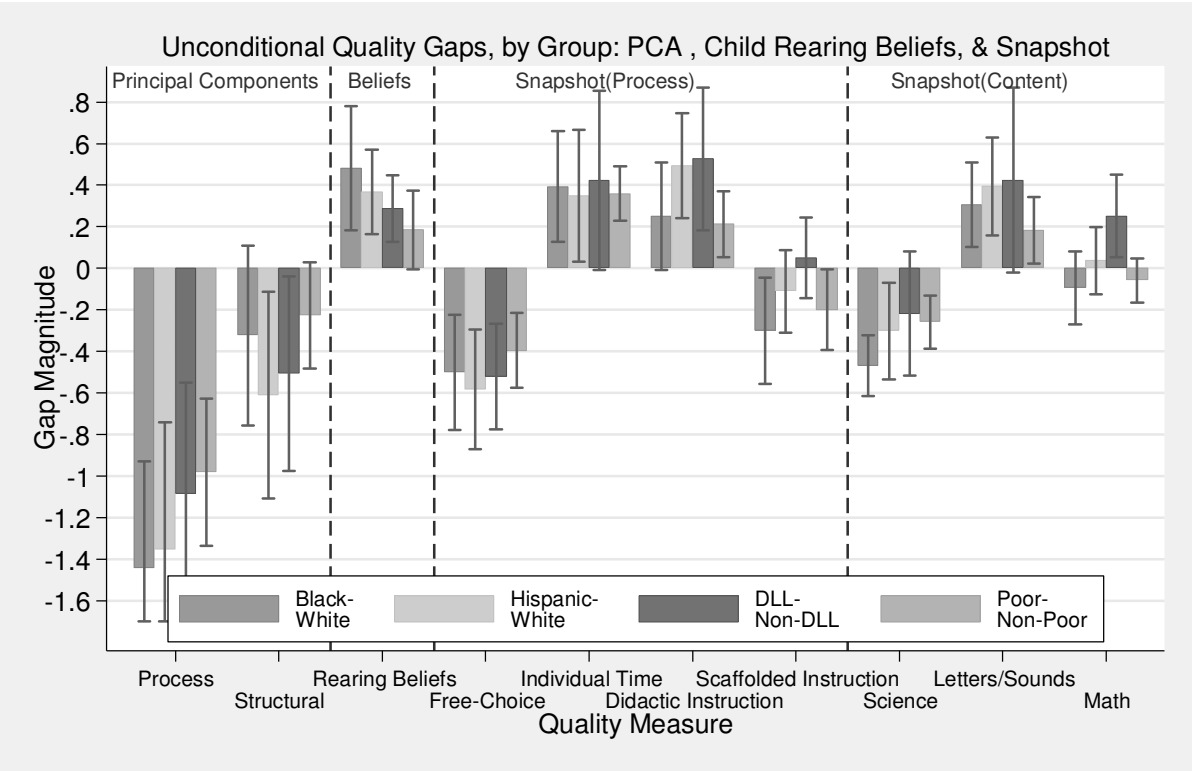


Figure 5

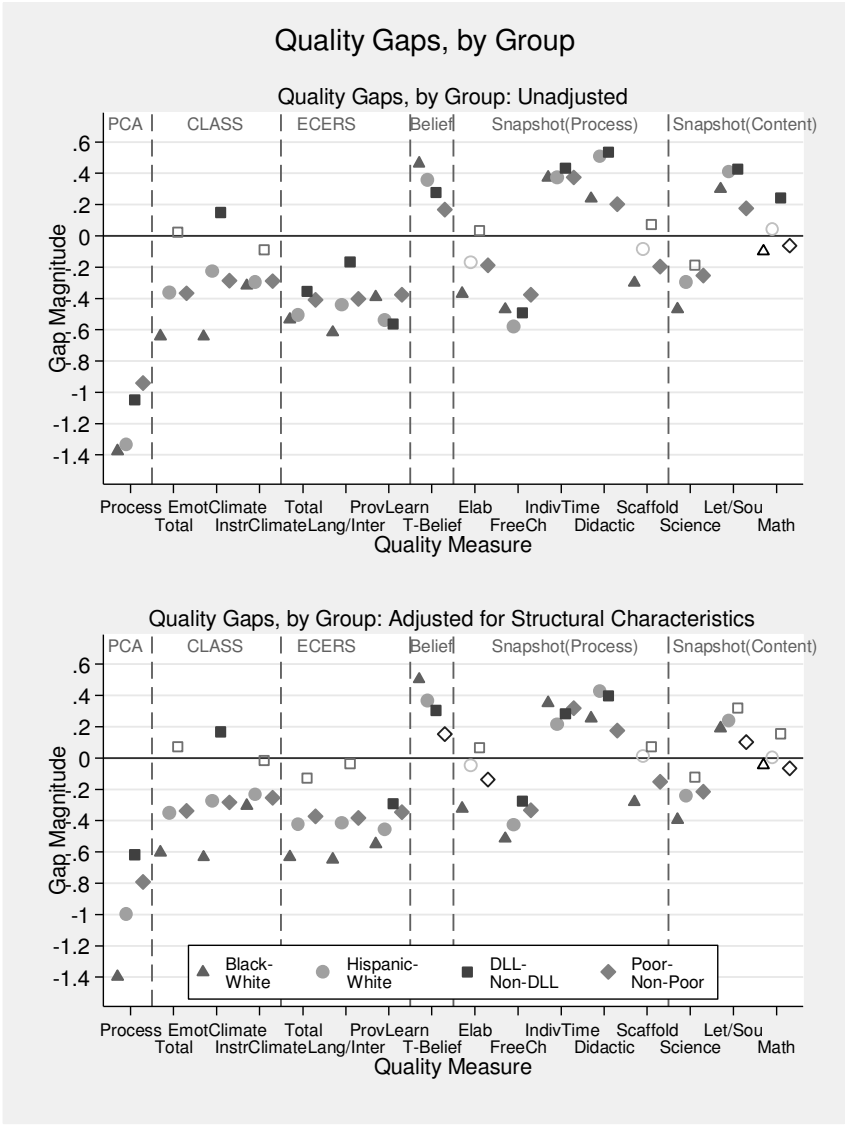
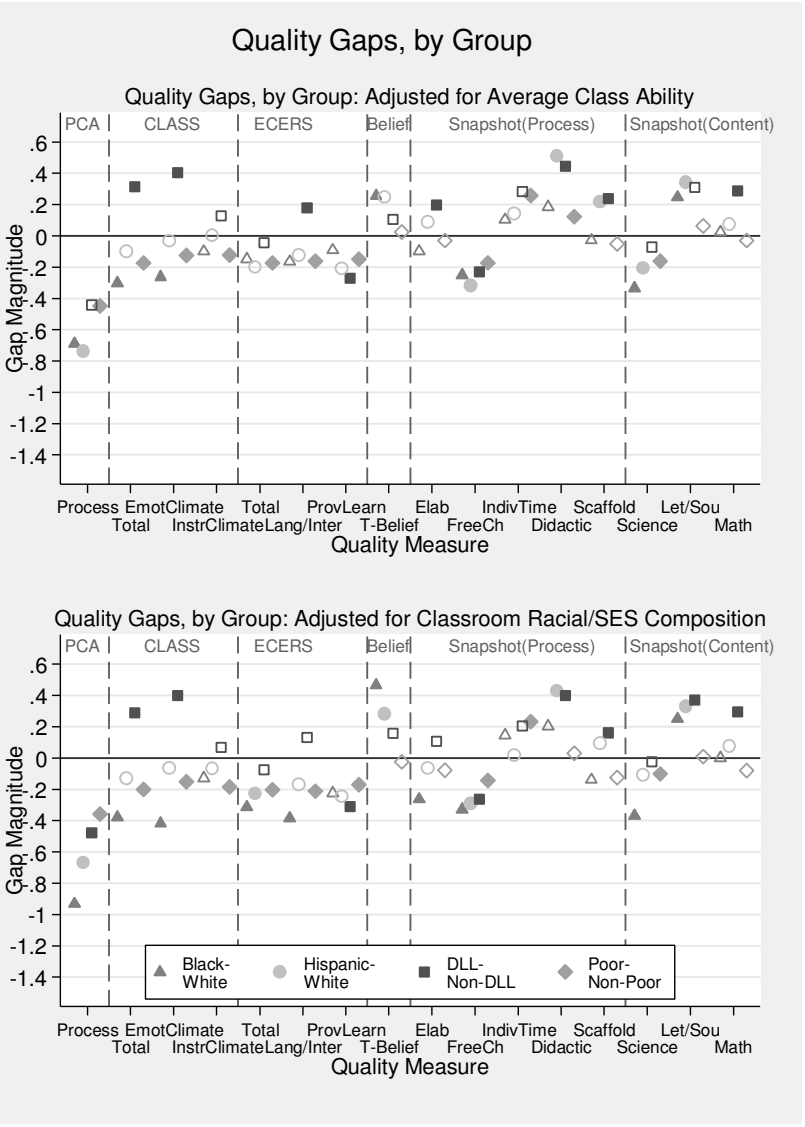


Figure 6.



*Note: Solid markers represent a significant quality gap. Hollow markers represent a non-significant gap.

Figure 7

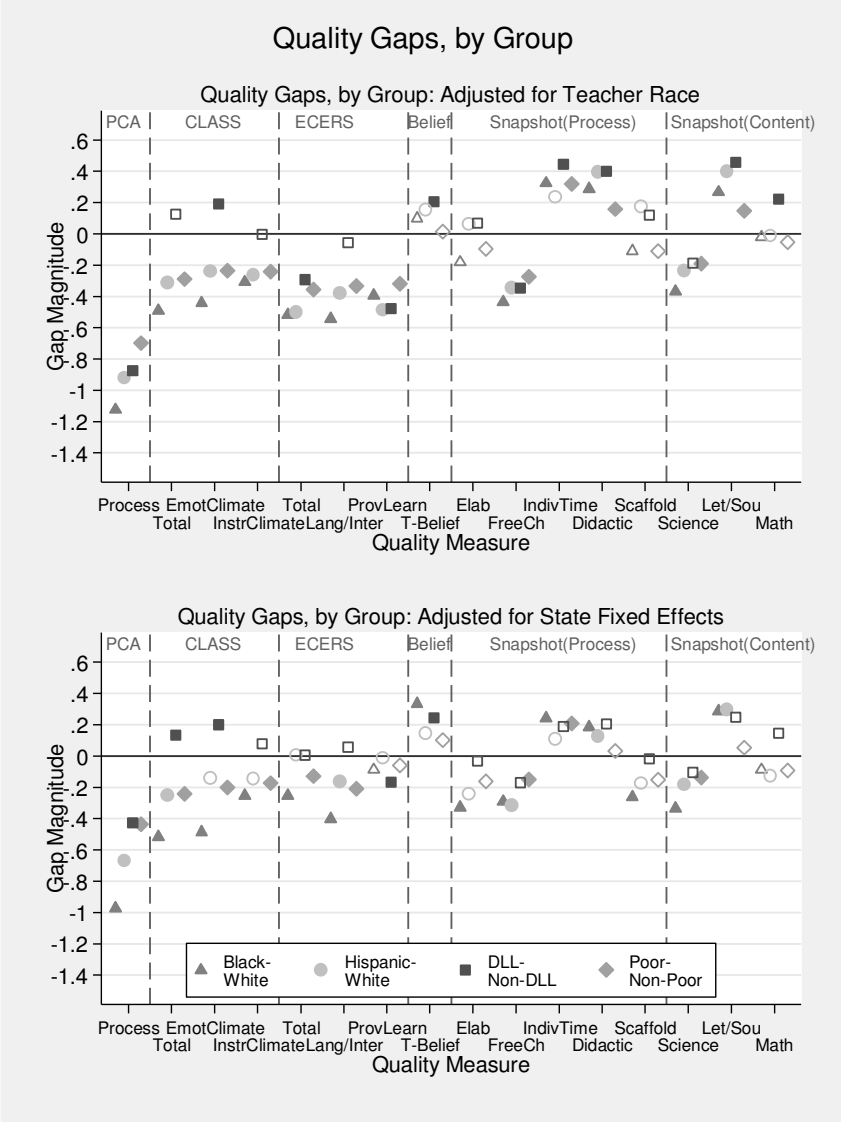


Figure 8

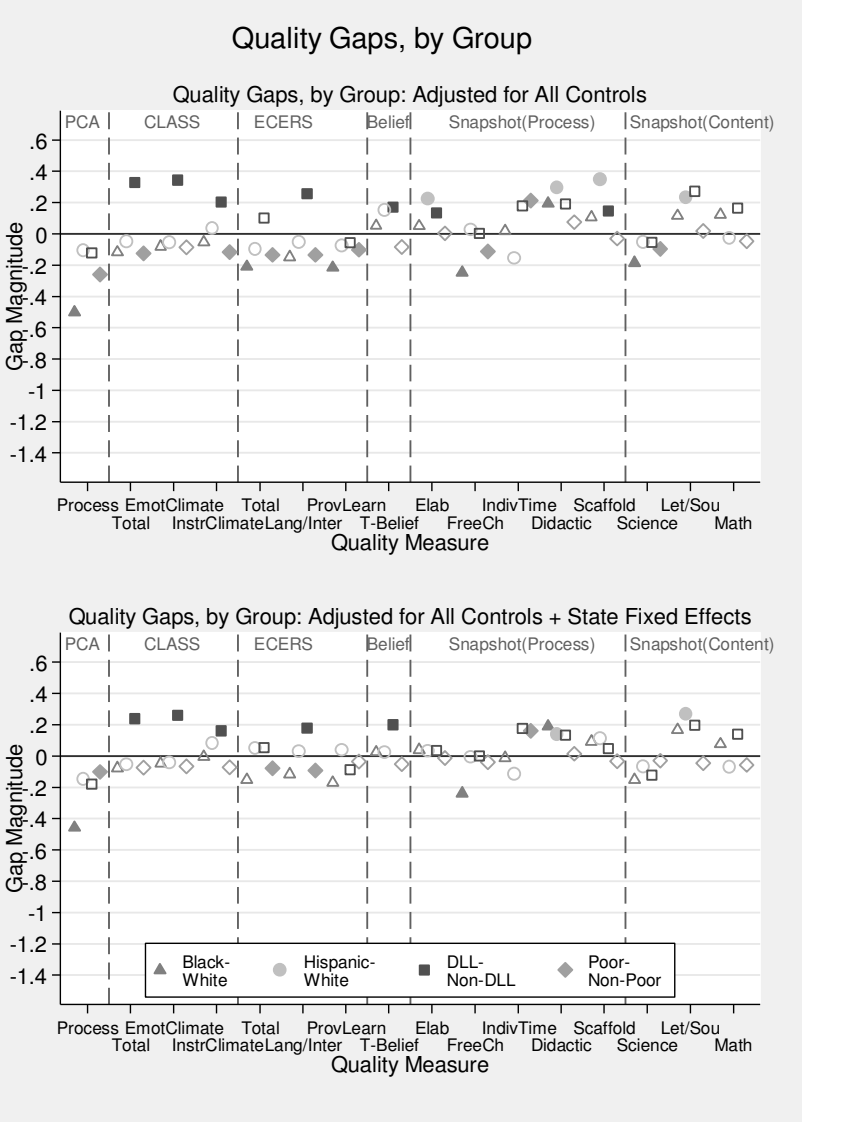


Figure 9. Between state variation in overall ECERS & CLASS scores

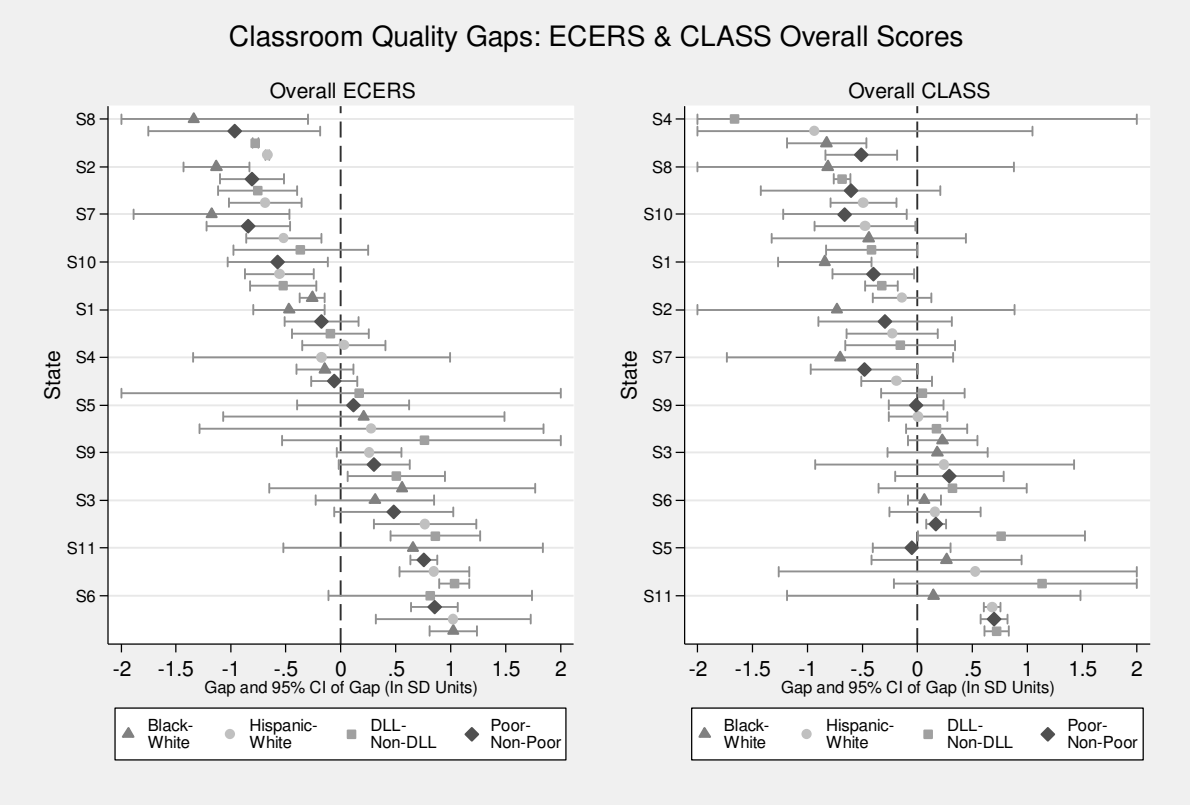
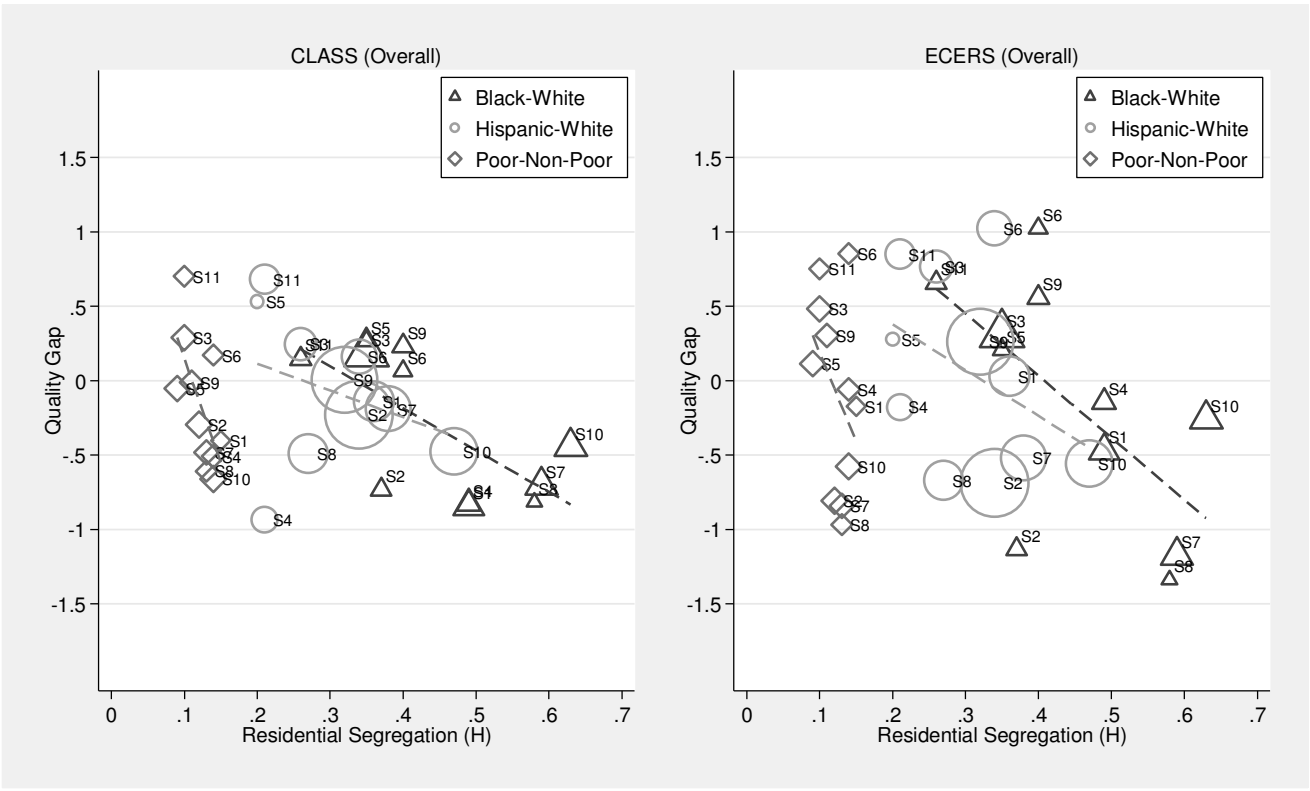


Figure 10. Quality gaps by level of residential segregation and state



*Size of shapes represent the size of the black, Hispanic, or poor population in that state