EdPolicyWorks

Working Paper:

The Effects of Accountability Incentives in Early Childhood Education

Daphna Bassok[1,], Thomas S. Dee[2] & Scott Latham[2]

In an effort to enhance the quality of early childhood education (ECE) at scale, nearly all U.S. states have recently adopted Quality Rating and Improvement Systems (QRIS). These accountability systems give providers and parents information on program quality and create both reputational and financial incentives for program improvement. However, we know little about whether these accountability reforms operate as theorized. This study provides the first empirical evidence on this question using data from North Carolina, a state with a mature QRIS. Using a regression discontinuity design, we examine how quasi-random assignment to a lower quality rating influenced subsequent outcomes of ECE programs. We find that programs responded to a lower quality rating with comparative performance gains, including improvement on a multi-faceted measure of classroom quality. Programs quasi-randomly assigned to a lower star rating also experienced enrollment declines, which is consistent with the hypothesis that parents responded to information about program quality by selectively enrolling away from programs with lower ratings. These effects were concentrated among programs that faced higher levels of competition from nearby providers.

[1]University of Virginia
[2]Stanford University

Updated September 2017

EdPolicyWorks
University of Virginia
PO Box 400879
Charlottesville, VA 22904

Acknowledgements: The authors would like to thank North Carolina's Department of Health and Human Services for generously providing the data used in this research. We would also like to thank Vivian Wong, Jim Wyckoff, Robert Pianta, Chloe Gibbs, Kimberly Boller, and seminar participants at the Association for Education Finance and Policy (AEFP), Society for Research in Child Development (SRCD), Association for Public Policy Analysis & Management (APPAM), the Stanford Center for Education Policy Analysis (CEPA), and the University of Virginia's EdPolicyWorks for their helpful comments on earlier drafts of this paper. This research was partially supported by grant 20150907 from the Smith Richardson Foundation, by IES grant R305B090002 to the University of Virginia, and by IES grant R305B130017 to Stanford University.

EdPolicyWorks Working Paper Series No. 58. September 2017.
Available at http://curry.virginia.edu/edpolicyworks/wp
Curry School of Education | Frank Batten School of Leadership and Public Policy | University of Virginia

INTRODUCTION

High-quality early child education (ECE) programs have the potential to narrow achievement gaps and improve children's life trajectories (Heckman, 2006; Yoshikawa et al., 2013). Motivated by this potential, public investment in ECE programs has increased dramatically in recent years. For instance, state spending on preschool more than doubled between 2002 and 2016, from $3.3 to $7.4 billion (constant 2017 dollars) as did the number of 3 and 4 year olds enrolled in public preschool, from 700,000 to nearly 1.5 million (Barnett et al., 2017).

Although access to ECE programs has grown rapidly, many programs are of low quality, particularly in low-income communities (Burchinal et al., 2010; Bassok & Galdo, 2016). Further, two recent experiments tracking the impacts of scaled-up ECE programs found only short-term benefits that faded quickly (Lipsey, Farran, & Hofer, 2015; Puma et al., 2012). Variation in program quality is one of the most common candidate explanations for the quickly fading impacts of some scaled-up public preschool initiatives (Yoshikawa et al, 2013).

In light of these findings, policymakers have increasingly focused on improving the quality of ECE programs *at scale*. For instance, through two large federal programs (i.e., Race to the Top – Early Learning Challenge and Preschool Development Grants), the Federal government competitively allocated a combined $1.75 billion to states between 2011 and 2016, and tied those resources to explicit investments in quality-improvement infrastructures (Congressional Research Service, 2016). The recent federal reauthorization of the Child Care and Development Fund also included provisions aimed at increasing quality in the child care sector (U.S. Department of Health and Human Services, 2014).

As part of this wave of support for increased ECE quality, Quality Rating and Improvement Systems (QRIS) have emerged as a widespread and potentially powerful policy lever. QRIS are accountability systems that seek to drive, at scale, improvements in ECE quality. As of February 2017, 38 states have statewide QRIS, and nearly all others are in the planning or piloting phases (QRIS National Learning Network, 2017). Most of these state systems are quite recent; as of 2005, for instance, only 10 states had QRIS in place.

Similar to accountability reforms in a variety of other organizational contexts, QRIS aim to drive improvements through two broad channels. One is to establish quality standards for programs and to disseminate this information among program operators. A second QRIS mechanism is to create incentives and provide supports that encourage broad improvements in program quality. State QRIS typically provide financial rewards for meeting standards, and many also offer technical assistance or professional development to help programs improve. They seek to indirectly encourage program improvement by making information on program quality publicly available in an easily digestible format for parents and other stakeholders. In fact, arguably the most visible and defining trait of QRIS is that states rate programs on a single, summative, and discrete scale (e.g., 1 to 5 stars) meant to distinguish ECE programs of varying quality. In theory, this information allows parents to “vote with their feet,” and puts pressure on low-quality programs to improve or risk drops in enrollment.

Despite substantial investment in ECE accountability efforts, there is no evidence on whether these accountability systems have improved the quality of ECE programs or whether their primary mechanisms work as theorized. This project provides the first such evidence on this high-profile policy initiative by examining North Carolina’s Star Rated License (SRL) system, one of the oldest and most well established QRIS in the country. We provide causal evidence on the effects of the incentive contrasts created by the SRL system by evaluating the effect of receiving a lower “star” rating on several program outcomes. Specifically, we examine the effects of a lower rating on several subsequent measures including overall program quality scores, independent ratings of classroom quality as measured through observations, and the revealed preferences of parents as measured by program enrollments.[1] We also examine the effects of a lower rating on whether a program later closes or opts out of the opportunity for more comprehensive assessment and higher ratings.

We estimate the causal effects of a lower QRIS rating on these outcomes using a fuzzy regression discontinuity (RD) design based on a continuous measure of baseline program quality (i.e., classroom observation ratings). We demonstrate that the variation in this measure around a state-determined threshold value leads to large and discontinuous changes in the probability of earning a lower QRIS rating. We find that quasi-random assignment to a lower rating led programs to improve the quality of their services as measured by increases to their overall rating and by large gains in their score on a multifaceted measure of classroom quality (effect size = 0.34). We also find that a lower QRIS rating led to reductions in program enrollment. Our findings indicate that the causal effects of a lower rating are concentrated among programs that face higher levels of competition (i.e., those with more programs nearby). These three results provide evidence consistent with the basic QRIS theory of change in that QRIS incentives led to meaningful changes in program performance, particularly in contexts where there was greater competition.

However, our results also underscored the importance of policy design that mitigates the possibly unintended consequences of such accountability systems. For instance, our findings

[1] We note that reduced enrollment could instead reflect center efforts to improve quality through an intentional reduction in scale or their response to the lower state subsidy rate associated with a lower star rating. However, we also find evidence that that lower ratings reduced the capacity utilization centers reported, a finding more consistent with parents choosing not to enroll in centers with lower ratings than with centers lowering enrollment targets. Similarly, the lagged response of enrollment to a lower rating (i.e., several years) is more likely due to parents' enrollment decisions than the more immediate response we might expect from centers assigned a lower subsidy rate.

show that quasi-random assignment to a lower rating led programs to make improvements on one specific quality measure that contributed to their lower rating, but we found no effects on a wide range of other quality measures, suggesting the importance of ensuring that quality features that are incentivized in accountability systems are well aligned with strategies for improving quality. Further, we find weakly suggestive evidence that quasi-random assignment to a lower QRIS rating increased the probability that a program opted out of the opportunity for more exhaustive assessment (and, correspondingly, the opportunity for the highest ratings). This evidence indicates that the extent to which programs can choose not to participate in QRIS may be another salient design feature.

ACCOUNTABILITY IN EARLY CHILDHOOD EDUCATION

States regulate ECE quality by establishing minimum requirements that programs must meet. For example, all ECE programs face specific licensing requirements in terms of class size, ratios, or staff qualifications. Given concerns about the generally low levels of quality of ECE programs, recent federal initiatives have sought to create incentives to move beyond these “quality floors” for staffing and facilities (U.S. Department of Health and Human Services, 2014). For instance, the U.S. Department of Education has competitively allocated $1.75 billion to states from 2011-2016 through the Race to the Top – Early Learning Challenge and Preschool Development Grants. To be eligible for these grants, states were required to demonstrate their commitment to systematically assessing the quality of ECE programs, including through QRIS (Congressional Research Service, 2016).

Notably, measuring the quality of ECE programs at scale (i.e., outside of small, carefully controlled studies with expensive longitudinal data collection) is difficult. In contrast to the K-12 context where accountability systems often define quality based on students' gains on test-based measures, quality measurement in ECE rarely focuses on direct measures of children's skills because these measures can be both expensive to administer and highly reliant on the timing of assessment, as children's skills change quickly at these early developmental stages (Snow & Van Hemel, 2008).

Instead, the measurement of quality in ECE programs is generally divided into measures of "structural" and "process" quality. Structural quality measures are program-level inputs that are straightforward to quantify and regulate (e.g., teacher education and experience levels, class size, and staff-child ratios) and are hypothesized to facilitate high-quality learning experiences for young children. In contrast, process measures aim to capture more directly, through classroom visits, the quality of a child's experience in a classroom (e.g., the extent to which the classroom is stimulating, engaging, and positive). A large body of research has demonstrated that, although they are costlier to collect, measures of process quality (e.g., the Classroom Assessment Scoring System [CLASS]) are generally stronger and more consistent predictors of children's learning than are structural measures (Araujo et al., 2014; Hamre & Pianta, 2005; Howes et al., 2008; Mashburn et al., 2008, Sabol et al., 2013).

QRIS typically include measures of both structural and process quality. QRIS establish multiple "tiers" of quality (e.g., 1 to 5 stars) with benchmarks for each. They then rate programs based on their adherence with these measures. Programs often receive direct financial incentives for meeting higher-quality benchmarks (e.g., subsidy reimbursement rates; merit awards), and states and/or local organizations may also provide support such as professional development and technical assistance (QRIS National Learning Network, 2015). The ratings are also publicly available to parents and other stakeholders, who often struggle to discern quality on their own (Bassok, Markowitz, Player & Zagardo, 2017; Mocan, 2007).

Like accountability reforms in the K-12 sector, the design of QRIS policies implicitly reflects two broad theoretical concerns. One involves how imperfect information may contribute to the prevalence of low-quality ECE. It may be that well-intentioned staff and leaders in ECE programs lack a full understanding of appropriate quality standards or the extent to which their program meets those standards. If so, the dissemination of information on standards and a program's performance on those standards may be an effective way to remediate an information problem. An empirical literature has examined the effects of such information efforts in K-12, and shows that simply providing information about the quality of schools did not lead to improvements in performance (Hanushek & Raymond, 2005). However, the ECE landscape is far more diverse and fragmented than the K-12 sector (Bassok et al., 2016), which may exacerbate the imperfect information problem. In this context, providing information about quality and performance to ECE programs may have a greater impact than in K-12 settings.

A second theoretical motivation for QRIS is that ECE programs may underperform, in part, because they lack high-powered incentives to focus their efforts on the desired dimensions of structural and process quality. There is a substantial body of evidence that K-12 accountability systems such as the federal No Child Left Behind (NCLB) can yield meaningful organizational improvements as evidenced by gains in student achievement (Dee & Jacob, 2011; Figlio & Loeb, 2011; Wong, Cook, & Steiner, 2015). For example, a 2011 report from the National Research Council concluded that school-level incentives like those in NCLB raised achievement by about 0.08 standard deviations (particularly in elementary-grade mathematics).

Providing information to parents can also add market-driven incentives to improve quality. A compelling research base suggests that parents are responsive to clear information about school quality in the K-12 context (Friesen et al., 2012; Koning & van der Wiel, 2013). For instance, Hastings & Weinstein (2008) provide experimental evidence that parents who received simplified information about school quality selected higher-quality schools for their children, and that these choices in turn led to improvements in children's test scores. In the ECE context, parents tend to overestimate the quality of ECE programs and their satisfaction with their child's program is *unrelated* to any observed quality characteristics (Cryer & Burchinal, 1997; Mocan, 2007; Bassok et al., 2017). The provision of simplified, reliable information about the quality of available ECE may thus allow parents to make informed decisions and selectively place their children with higher quality providers.

QRIS policies typically combine multi-faceted performance measurement with financial and reputational incentives, and thus resemble consequential accountability policies in K-12 education; reforms for which there is evidence of modest but meaningful efficacy. The K-12 literature and the broader literature on accountability do suggest that QRIS policies may be effective tools for driving improvements in ECE quality at scale. However, there is scant evidence as to whether QRIS, or accountability efforts more broadly defined, are effective in the ECE context. Most of the existing research on QRIS has focused on establishing the validity of QRIS ratings by comparing them to other measures of quality or to child outcomes (Sabol et al., 2013; Sabol & Pianta, 2014). Whether these new rating systems are sufficiently clear, well designed, and powerful to change the performance of ECE programs is an open, empirical question.

In the next sections, we describe the unusually mature QRIS policies in North Carolina and how we use longitudinal data on program performance to identify the causal effects of the incentive contrasts embedded in this system. We also consider the possibility of heterogenous impacts, depending on the extent to which programs face competition. The K-12 literature suggests that effects may be most pronounced among ECE programs that face higher levels of competition (Waslander, Pater, & van der Weide, 2010). For instance, Hoxby (2003) finds that metro areas with many school districts have significantly higher productivity than those with fewer districts, which she attributes to the higher level of choice, and, implicitly, the higher level of local competition.

QRIS IN NORTH CAROLINA

North Carolina provides a compelling context to study the effects of a large-scale ECE accountability effort for several reasons. First, North Carolina's Star Rated License (SRL) program is one of the oldest QRIS in the country. It was instituted in 1999, and has operated in its current form since 2005. The state spends more than $13 million yearly to administer its QRIS, more than any other state, and maintains nearly a decade of program-level data on star ratings as well as the underlying quality measures that go into calculating the ratings.

The program has all the key features of a mature QRIS including (1) well-defined quality standards linked to financial incentives; (2) support for program improvement through technical assistance and local partnerships; (3) regular quality monitoring and accountability and; (4) easily accessible quality information provided to parents (Tout et al., 2009; Zellman & Perlman, 2008; The Build Initiative and Child Trends, 2015).

Furthermore, while most state QRIS are voluntary, in North Carolina, all non-religious programs are automatically enrolled at the lowest (i.e., one star) level when they become licensed. Thus, the vast majority of licensed ECE programs participate in the SRL program, including all Head Start programs, all state pre-kindergarten programs, and most programs that operate in local public schools. Programs may apply for higher ratings after a temporary waiting period. In total, roughly 88% of licensed programs received star ratings in any given year. The 12% that do not receive star ratings consist primarily of religious sponsored facilities (10%), with a smaller number having temporary/provisional licenses (2%). This high rate of participation is crucial for understanding how QRIS function when implemented at scale, rather than targeted to a small and self-selected portion of the ECE market.

Another crucial feature of North Carolina's rating system relevant to the current study is that programs' star ratings are determined, in part, by a continuous measure of observed classroom quality. In contrast to other components of the QRIS, which are scored as discrete measures, this continuous measure of quality allows us to leverage a regression discontinuity (RD) design. Specifically, providers must exceed a set of thresholds on the observation metric to attain credit toward a higher star rating. This means that small differences in programs' observation scores can make the difference between earning a higher or lower star rating (e.g., 3 versus 4 stars). We leverage the idiosyncratic differences in these continuous scores to estimate the causal impact of receiving a higher vs. lower star rating on subsequent measures of program quality and on enrollment. Taken together, the North Carolina context and data provide a compelling setting to conduct the first study on the effects of a scaled-up ECE accountability system.

The Star Rated License (SRL) System

North Carolina's Division of Child Development and Early Education rates ECE programs on a scale of one to five stars.[2] The number of stars that a program receives is based on an underlying 15-point integer scale. The mapping of these points into star ratings is as follows: 1 star (0 to 3 points), 2 stars (4 to 6 points), 3 stars (7 to 9 points), 4 stars (10 to 12 points), and 5 stars (13 to 15 points). Programs' ratings on the underlying 15-point scale are primarily earned across two subscales, each worth up to 7 points.

The first subscale, "education standards" (i.e., ≤ 7 points), is determined by the education and experience levels of administrators, lead teachers, and the overall teaching staff. For instance, programs receive more points for a staff with more years of ECE teaching experience or more advanced training in the field. Each component of the staff education subscale is scored on a discrete scale.

The second subscale, "program standards" (also, ≤ 7 points), includes measures of quality such as staff-child ratios and square footage requirements. Each of these measures is scored on a discrete scale. As described in detail below, the program standards subscale also includes an observational component, the Environment Rating Scale (ERS), scored on a continuous scale.

Finally, each program can also receive an additional "quality point" by meeting at least one of a variety of other education or programmatic criteria (e.g., using a developmentally appropriate curriculum, combined staff turnover of ≤ 20%, 75% of teachers/lead teachers with at least 10 years of ECE experience).

A feature of the SRL system that is centrally relevant for this study is that providers are eligible for more points on the program-standards subscale (and, in turn, higher star ratings) if

[2] We focus here on the specific features of North Carolina's QRIS that are crucial for understanding and interpreting this research. For a more comprehensive description of this program, see the website for North Carolina's Division of Child Development and Early Education website (ncchildcare.nc.gov).

they exceed specified thresholds on the ERS. ERS is a widely used observation tool, currently included in 30 QRIS throughout the country. It is a broad measure of classroom quality, and incorporates both structural features of the classroom (e.g., space and layout, daily schedules) as well as measures of "process" quality like student-teacher interactions and classroom activities.

In North Carolina, the Division of Child Development contracts with the North Carolina Rated License Assessment Project (NCRLAP) to conduct these assessments. Programs must submit a request to be rated, and they receive a four-week scheduling window during which assessors may visit at any time. NCRLAP stresses the importance of evaluations occurring on a "typical day," and, to this end, programs may designate up to five days as non-typical days during which assessments will not occur. Each rating is valid for three years and the state provides one free assessment every three years. Programs wishing to re-rated sooner must wait a minimum of six months after their previous rating, and must cover the cost of assessment on their own (North Carolina Rated License Assessment Project, n.d.).

During the rating process, assessors conduct site visits where they randomly select a third of classrooms to be rated, including at least one classroom for every age group served (i.e., infants/toddlers, 3-4 year olds, school-aged children). Assessors spend a minimum of 3 hours in each classroom, recording notes on a wide variety of interactions, activities, and materials. They also spend 30-45 minutes interviewing the lead classroom teacher. This information is used to rate providers across 38 or more distinct items, depending on the version of the assessment used.[3] Each item is scored either a 1, 3, 5, or 7, indicating "inadequate," "minimal," "good," or

[3] Four different versions of the ERS are available depending on the age of children and the type of care setting. Specifically, care settings may be rated using the Early Childhood Environment Rating Scale - revised (ECERS-R, 47 items; Harms, Clifford & Cryer, 1998), the Infant/Toddler Environment Rating Scale - revised (ITERS-R, 39 items; Harms, Cryer, & Clifford, 2003), the School-Aged Care Environment Rating Scale (SACERS, 49 items; Harms, Jacobs, & White, 1996), or the Family Child Care Environment Rating Scale - revised (FCCERS-R, 38 items; Harms, Cryer, & Clifford, 2007). Although the scales are tailored to specific age groups, each is scored on the

"excellent" quality, respectively. The scores are then averaged across items to determine each program's overall ERS rating (The Build Initiative & Child Trends, 2015). In our data, these ratings are defined out to 2 decimal places.

Programs are not required to receive ERS ratings, but those that elect to be rated are eligible for higher overall star ratings. For example, programs who opt to forego an ERS rating can earn a maximum of 2 out of the 7 possible program score points, and just 10 of the 15 total points possible This means a program choosing not to receive an ERS rating cannot receive a 5-star rating (which requires 13 points), and must earn every other point possible to receive a 4-star rating (which requires 10 points). In practice, most programs opt to receive ERS ratings, and the percentage has been increasing over time, from 52% in 2008 to 66% by 2014. The decision to opt out of receiving an ERS rating is one of the policy-relevant outcomes we study.

Among programs that elect to receive an ERS rating, both the *average* ERS score that a program receives across classrooms and the *lowest* ERS score received can influence the total number of points earned. Programs earn additional points by exceeding a series of thresholds along each of these. For instance, a program whose lowest classroom ERS is above 4.0 can earn a maximum of 6 points on program standards, while a program with a lowest classroom rating below 4.0 can only earn a maximum of 2 points. Similarly, a program with an *average* ERS rating of 4.5 is eligible for up to 4 points on program standards, whereas a program that receives just below a 4.5 is only eligible for 3 points (see the Appendix for full details of how program standards scores are calculated). This means that small, and arguably random, differences in ERS ratings can be the difference between a program earning a higher or lower point total on the program standards scale. Because each point constitutes roughly a third of a star, these same

same 1-7 scale, and contains measures of basic care provision, physical environment, curriculum, interactions, schedule/program structure, and parent/staff education.

small differences can lead to meaningful differences in the probability of earning a higher versus lower star rating.

The Treatment Contrast

In the regression-discontinuity design we describe below, each program's baseline ERS rating serves as an assignment variable that influences the program's star rating. We focus on whether a program's average ERS rating was at or above 4.5, a necessary condition for receiving 4 or more points on the program standards subscale. We show that programs' baseline scores relative to this threshold generate a discontinuous "jump" in the likelihood a program earns more stars.

The character of the treatment contrast defined by this "intent to treat" (ITT) merits careful scrutiny. The star ratings received by ECE programs are critical components in the QRIS theory of action, creating incentives for program improvement through direct financial rewards and, indirectly, through the effects of information and market pressure. First, in North Carolina, ECE programs receive higher per-student reimbursements for subsidy-eligible children for every additional star they earn. These increases vary by county and by the age of children served but, in most cases, they are substantial. For instance, in 2007, a 5-star program averaged an 11% higher reimbursement per subsidy-eligible student than a 4-star program. A 4-star program averaged a 5% higher reimbursement than a 3-star program, and, strikingly, a 3-star program averaged a 35% higher per-student reimbursement than a 2-star program (NC Division of Child Development and Early Education, 2007). These performance-defined differences in subsidy rates may encourage lower-rated programs, particularly those who enroll many subsidy-eligible children, to improve their quality to qualify for higher reimbursement rates.

Second, star ratings are publicly available, and may create market pressure through their effect on parents' choices about where to enroll their children. North Carolina has implemented multiple strategies to increase awareness of the Star Rated License program. These include requiring star rated licenses to be displayed prominently within each program, publishing star ratings through a searchable tool on North Carolina's Department of Health and Human Services website, distributing posters, business cards, and postcards with the web address for this tool, and arranging for media coverage of highly rated programs (National Center of Child Care Quality Improvement, 2015; see Figure A1 in the Appendix for an example of a star-rated license).

Because North Carolina's QRIS simultaneously embeds non-trivial financial incentives and the market incentives created by publicizing ratings, it provides a compelling setting for evaluating the theorized mechanisms that motivate these ECE accountability reforms. Our RD approach examines the effects of credibly random incentive contrasts that exist within North Carolina's QRIS. We hypothesize that programs who receive lower ratings will likely focus on making improvements in their ERS ratings, because small improvements along this dimension are likely to lead to higher star ratings. We first expect to see improvements along this measure three years after the initial ratings occurred, because ERS ratings are technically valid for three years. However, in practice, about 12% of programs did not receive new ratings until at least 4 years after the initial rating, so any improvements may not be apparent until even later.[4] We also hypothesize that lower rated programs will face a decrease in enrollment as a result of lower demand, though this will depend both on whether parents are aware of star ratings and whether they use them to make ECE decisions. We expect that the effects of QRIS incentives will vary based on the context of the local ECE market. In local markets where providers face high levels

[4] Programs can also opt to obtain an earlier ERS assessment but at their own cost. We examine such early ERS assessments as another behavioral response to a star rating.

of competition, QRIS incentives are likely to be particularly salient and powerful. In markets with low levels of competition, these incentives may be relatively weak.

DATA

Our analysis leverages program-by-year data for all licensed ECE programs in the state of North Carolina in the years 2007-2014 (N=6,929 unique programs across the entire panel). These data, generously provided by the North Carolina Department of Health and Human Services, span nearly the entire period since the last major revision to North Carolina's rating system in 2005. For each program-year observation, these data include street addresses as well as information about the type of program (e.g., independent program, Head Start), enrollment, and capacity. We also have unusually detailed information about program quality as measured through the QRIS, including overall star ratings, program standards and staff education scores, ERS ratings, and indicators for whether each program earned a quality point.

North Carolina revised its QRIS in 2005, which changed the relationship between ERS ratings and star ratings. For this reason, we define our ITT sample using a program's first rating under the revised regime. This is somewhat complicated by the fact the rollout of the updated system was staggered across multiple years. In particular, ratings that took place on or after January 1, 2006 were scored under the new regime, but pre-existing programs had until January 1, 2008 to transition to the new system (NC Division of Child Development and Early Education, n.d.). Our data begin in 2007, and, because ratings are valid for multiple years, some of the ERS ratings we observe reflect ratings from the previous regime. To determine each provider's first rating under the new regime, we rely on recorded ERS visit dates where possible (about 47% of observations), and classify all recorded visits that occurred in 2007 or later as belonging to the new regime. In cases where ERS visit dates are not recorded, we use several decision rules to determine which ERS ratings were scored under the new regime.[5]

We limit our ITT sample to programs observed in the three-year window 2007-2009, which allows us to track program outcomes for each of five years after the baseline observation. Our data include 5,866 unique programs that were observed at baseline. However, we exclude 844 programs that never had a star rating (i.e., those operating under a religious-sponsored, temporary, or provisional license), as well as 1,865 programs that had a star-rated license but chose not to receive an ERS rating during our baseline window. These sample exclusions are necessary as the baseline assignment variable is not defined for these programs. The programs observed over our baseline period but excluded from our analysis differ from those in our analytical sample in several ways (Table A1 in the Appendix). For example, in 2007, the excluded programs were more likely to have religious sponsorship (e.g., 21% versus 8% in our study sample) and to be independently operated (53% versus 44%). Excluded programs were less likely to be located in local public schools (17% versus 27%). Furthermore, only 1% of excluded programs were Head Start programs, compared with 10% of programs in the sample. The programs included in our analysis also have higher average enrollment, both overall and relative to capacity. Finally, and unsurprisingly, programs that are in the sample have higher star ratings at baseline than those that are excluded. Though these restrictions imply a caveat regarding generalizability, we note that, given the broad coverage of North Carolina's system, our sample

[5] Because ERS ratings are valid for three years, we assume that ratings were initially conducted in 2007 if we observe the same rating throughout the years 2007-2009. In cases where we observe a rating in 2008 or 2009 that differs from the 2007 rating, we include the first *changed* rating in our ITT sample.

includes a larger portion of the state's programs than the portion included in most state's QRIS (The Build Initiative & Child Trends, 2015).[6]

Our final ITT sample includes 3,157 unique ECE programs. Table 1 presents descriptive statistics for this sample in the baseline year (T) and for subsequent years through T+5. At baseline, the vast majority of programs (97%) had earned at least a 3-star rating, 81% has at least a 4-star rating, and 44% had earned a 5-star rating. The average enrollment was about 53 children, and programs were operating, on average, at 71% of their total capacity. The average ERS rating was 5.21, indicating relatively high quality across the sample.

REGRESSION DISCONTINUITY DESIGN

Our RD analysis compares outcomes among programs whose average ERS rating at baseline is above or below an ERS threshold that influences star ratings. This contrast implies a fuzzy regression discontinuity design, as programs that are just below this cutoff – those with an intent to treat (ITT) equal to one – are significantly less likely to receive a higher star rating compared to programs just above the cutoff (i.e., ITT=0). In this design, treated programs (i.e., ITT=1) are more likely to receive lower star ratings and face incentives to improve quality both directly through reduced subsidy rates and indirectly through reputational effects and parents' enrollment decisions. As is common practice (e.g., Lee & Lemieux, 2010; Schochet et al., 2010), we employ a combination of graphical and statistical evidence in our analysis. We estimate the magnitude and statistical significance of receiving a higher vs. lower star rating using reduced-form specifications that take the following form for outcome Y_i associated with program *i*:

$$Y_i = \gamma I(S_i < 0) + k(S_i) + \alpha_i + \varepsilon_i \qquad (1)$$

[6] A related external-validity caveat is that the privately run ECE programs in our sample are disproportionately likely to be "compliers" with the intent to treat (ITT) in our RD design. This is because Head Start and public pre-K programs are required to maintain 4+ star ratings.

The variable S_i is the assignment variable (i.e., the program's average ERS rating at baseline) centered at 4.5, the focal RD threshold in the current analysis, and k is a flexible function of the centered assignment variable.[7] We condition on a fixed effect, $\alpha_{i,}$ for the specific year in which a program's ERS rating occurred (2007-2009), and ε_i is a mean-zero random error term. We report heteroscedastic-consistent standard errors throughout. The parameter of interest, γ, identifies the effect of having an ERS rating just below the 4.5 threshold (and, by implication, an increased likelihood of a lower star rating), relative to a rating at or above 4.5 (i.e., the estimated effect of the ITT).

To examine effects on program quality, our outcome measures include future star ratings, ERS ratings, and other indicators of quality measured as part of North Carolina's QRIS such as staff-child ratios and teacher qualifications. We also consider enrollment (both total and as a proportion of program capacity), as potential proxies for program demand. Finally, we examine the heterogeneity of these effects by the extent to which programs faced local competition. Specifically, we calculate the number of other ECE programs located within 5 miles of each program in the baseline year. We divide our sample into "low competition" and "high competition" at the median number of nearby programs (22), and estimate RD results separately for these low- and high-competition subsamples.

Assignment to Treatment

A regression discontinuity design relies on institutional circumstances in which small changes in

[7] The SRL system also implies other candidate thresholds that may be leveraged using a regression discontinuity. Specifically, centers are eligible for more QRIS points when their *lowest* ERS rating across classrooms exceeds either 4.0 or 5.0, or when their *average* ERS rating exceeds 4.75 or 5.0. We ultimately focus on the average ERS rating as a forcing variable to address the potential manipulation concerns discussed below. We focus on the 4.5 cut-off primarily because it offers the strongest "first stage" relationship (i.e., this cutoff is most strongly related to star ratings).

an assignment variable lead to large and discontinuous changes in treatment status. In the North Carolina context, the scoring procedures for star ratings implies that small differences in ERS ratings may lead to discontinuous probabilities of earning a higher star rating. For this project, we leverage the fact that earning an *average* ERS rating just below 4.5 makes a program less likely to earn a higher star rating. In Figure 1, we illustrate two "first-stage" relationships implied by the 4.5 threshold. Here, we organize programs into bins of size .1 on either side of the threshold, and show the proportion of programs who earned a 3+ or 4+ star rating in each bin. We restrict these figures to a bandwidth of 1 around the focal RD threshold and superimpose regression lines from parametric estimates with quadratic splines.

Figure 1 shows that in North Carolina, programs whose average ERS rating was < 4.5 were significantly less likely to receive a 3+ star rating than those just at or above 4.5. These programs were also significantly less likely to receive a 4+ star rating. In Table 2, we present analogous regression estimates. These estimates show that, for the full sample, programs just below the RD threshold were 13 percentage points less likely to earn 3+ stars and 29 percentage points less likely to earn 4+ stars than programs just above the threshold. Table 2 also presents "local linear" first-stage estimates, including linear splines for the full sample and for increasingly narrow bandwidths down to the recommended Imbens & Kalyanaraman (2012) bandwidth of 1. These estimates are quite similar to the quadratic specification, which we ultimately prefer based on the Akaike information criterion (Akaike, 1974; Schochet et al., 2010).

Internal Validity

A key identifying assumption of regression discontinuity designs is that no one is able to manipulate the value of their baseline ERS rating relative to the RD threshold. In this context, either ECE programs or raters could be a source of such manipulation. Although programs are able and encouraged to conduct self-assessments on the ERS, these self-assessments do not provide precise information about the ERS ratings programs will ultimately receive. Raters, who likely know the implications of receiving scores above or below particular thresholds, could manipulate scores by “bumping up” ERS ratings for programs that fall just below an ERS threshold. However, because we rely on each program’s *average* ERS (and more than half of the programs in our sample have two or more classrooms), a single classroom’s rating cannot as easily determine where a program’s score falls relative to the RD threshold.

These features imply that precise manipulation of the assignment variable is unlikely in this context. To corroborate this empirically, we examine a standard battery of tests for manipulation. First, we perform a visual inspection of the density of the assignment variable. Here we construct binned density plots, organizing the assignment variable into 0.05 and 0.025 point bins on either side of the 4.5 threshold (Figure 2a). These plots suggest no discontinuity in density at the 4.5 threshold. We test for a discontinuity formally using the commonly employed McCrary density test (McCrary 2008, Figure 2b) as well as a newly developed alternate procedure proposed by Cattaneo, Jansson, & Ma, 2017.[8] With both tests, we fail to reject the null hypothesis of no discontinuity. Finally, we conduct auxiliary RD regressions to test for differences in the observed baseline traits of programs above and below the 4.5 threshold (Table 3). We find no evidence of differences in these programs across the threshold. Both the smoothness of the assignment variable’s distribution and the covariate balance are consistent with the causal warrant of the RD design.

[8] The Cattaneo et al. (2017) procedure (“rddensity” in Stata), in contrast to McCrary (2008), does not “pre-bin” the data into a histogram, which requires researchers to specify the width and position of bins. Instead, this procedure requires only the choice of bandwidth associated with the local polynomial fit.

Another potential threat to internal validity involves program closure. Five years after our baseline observation, 24 percent of programs have closed. Our findings might be biased if programs with lower ratings were differentially likely to close and thus have no defined outcomes. We examine this possibility and report our findings in the Appendix (Table A2 and Figure A2). Specifically, we estimate versions of equation (1) in which indicators for program closure are the dependent variables. We find no evidence that programs on either side of the RD threshold were differentially likely to be closed at any point in the five years after ERS ratings were assigned (i.e., both in the full sample and in the samples defined above and below-median competition). This finding strongly suggests that program closure does not constitute an empirically relevant internal-validity threat.

A somewhat related issue is that, five years after our baseline observation, roughly 8 percent of the programs that remained open also chose to opt out of ERS ratings. Although ERS ratings are provided for free, and cannot lower a program's overall star rating, these programs may have decided that they prefer no public ERS rating rather than a low rating. Using our RD specification, we examined whether programs with average ERS less than 4.5 were more likely to opt out of future ERS assessments (Table A3 in the Appendix). We found weakly significant evidence that such programs were indeed more likely to opt out. This pattern does not complicate our analysis using future star ratings and program enrollment as outcomes. Those outcomes are defined for *all* the open programs in our ITT sample (i.e., *including* those that opted out of ERS assessments). This finding suggests that the ERS assessment gains we observe among programs assigned to lower ratings could reflect a separating equilibrium created by the treatment contrast (i.e., some lower-rated programs improving and others opting out). However, we also find that, five years after our baseline observation, there is *not* a statistically significant opt-out effect in the high-competition sample (i.e., where and when the ERS gains are concentrated). Nonetheless, we return to this finding when discussing the normative and policy-design implications of our results.

RESULTS

We begin illustrating our main findings graphically. Figure 3 illustrates the relationship between initial ERS ratings and star ratings at baseline (T) and in each of five subsequent years, using binned scatter plots analogous to the first stage plots presented above. Panel (a) focuses on the likelihood a program has 3 or more stars. For programs to the left of the 4.5 threshold (which is centered on zero), the ITT value was one. For those to the right, it was zero. The gap in the probability of having 3 or more stars narrowed rapidly in the first few years following the initial rating. This gap appears to have closed completely by T+4. This may partially reflect a ceiling effect, in that nearly all programs in our sample were rated at least 3 stars in T+5. By contrast, panel (b) of Figure 3 considers the probability that a program earned 4 or more stars, and shows no evidence of a ceiling effect. In this panel, we observe similar patterns with respect to the effect of the ITT: three years after the initial ERS rating, the gap at the threshold in the likelihood of being rated 4 or 5 stars had closed almost completely.

At the top of Table 4, we report RD estimates and standard errors that correspond to these figures. As Figure 3 suggests, these RD results indicate that the baseline ratings gap created by a program’s position relative to the 4.5 threshold shrunk and was no longer statistically significant within 3 years of the initial ratings assignment. These results suggest that quasi-random assignment to a lower star rating and the incentives that implies (i.e., lower financial subsidies, market pressures) led programs to improve their measured performance over the subsequent years.

Another useful outcome measure is the ERS rating received by each program if and when they are re-rated. These measures provide a more direct assessment of the developmental experiences of children within each program. Furthermore, we might expect programs close to the 4.5 threshold to be uniquely responsive with regard to this particular outcome. RD estimates for average ERS ratings are also shown in Table 4. Because ERS ratings are renewed every 3 years, we are most interested in estimates from periods T+3, T+4, and T+5. We find that in T+3 programs below the 4.5 threshold had somewhat higher ERS ratings (i.e., an increase of 0.13) but that this difference was not statistically significant.[9] However, in T+4 and T+5, we find that average ERS ratings jumped by 0.23 and 0.20, respectively, among programs to the left of the threshold. Figure 4a illustrates this relationship graphically in T+5. An ERS gain of 0.20 constitutes a 0.34 effect size with respect to the standard deviation observed at baseline (i.e., 0.20/0.58).[10] Given our first-stage estimates (Table 2), this ITT estimate implies that the estimated effect of receiving a 3-star rating instead of a 4-star rating is nearly 1.2 *program-level* standard deviations (i.e., 0.34/0.29). Such large "treatment on the treated" (TOT) estimates may reflect the unique salience of gains in ERS performance for ECE programs just below the 4.5 threshold. However, these large estimated effects may also reflect the stigma of receiving fewer than four stars. Such comparatively low star ratings would place a program in the lowest quintile of our baseline sample and, five years later, in the lowest decile (Table 1).

We also found additional supporting evidence that programs responded to the incentive contrasts created by their QRIS rating by examining their more immediate behavior. Specifically,

[9] As mentioned above, about 12% of the programs in our sample did not receive a new ERS rating until 4 or more years after the initial rating. When we limit the sample to centers that had received a new rating 3 years after the initial rating, we observe a statistically significant effect on average ERS ratings in T+3.

[10] As noted earlier, in the full sample, we find weakly significant evidence that centers below the 4.5 threshold at baseline were more likely to opt out of these ERS assignments. This suggests that the ERS gains we observe here could reflect both improvements among some poorly rated centers and the differential attrition of others. However, as we discuss below, there is no statistically significant opt-out effect in the high-competition sample where the ERS gains are concentrated.

if a program does not want to wait three years for its next free ERS assessment, it can choose to pay for an earlier re-rating. Using our RD specification, we find weakly significant evidence that programs below the 4.5 threshold were more likely to be re-rated in period T+1 (see Table A4 in the Appendix). However, by period T+2, this differential has shrunk considerably and become statistically insignificant. Nonetheless, the evidence of this early response is consistent with the hypothesis that ECE programs were both aware of their ERS and star ratings and seeking to improve them.

Next, we examined the effects of the intent to treat with a lower rating on future enrollment. Like star ratings, enrollment is also defined for all programs (i.e., regardless of whether they opted out of a future ERS rating). In panel B of Table 4, we report RD estimates from specifications in which enrollment and the proportion of capacity filled are the dependent variables. We see that, in T+3, programs with initial average ERS ratings below 4.5 enrolled nearly 5 fewer students. This estimate became smaller and statistically insignificant in T+4. However, the results for T+5 indicate that the intent to treat lowered enrollment by slightly more than 7 children. We also find that, by T+5, programs that were initially to the left of the 4.5 threshold had a reduction in their capacity utilization of 7 percentage points. We illustrate these findings graphically in Figures 4b and 4c. These findings suggest that parents were less willing to enroll children in programs assigned to a lower rating.[11] Interestingly, this enrollment reduction occurs despite the eventual recovery in star ratings among programs that received a lower baseline rating. There are at least two explanations for why the enrollment decisions made by parents may respond to a program's rating with a lag. First, parents may be somewhat

[11] It may also be that program operators intentionally reduced their scale to improve their quality (or did so in response to the lower state subsidy rate). However, the corresponding reduction in capacity utilization is inconsistent with this hypothesis as a reduction in enrollment targets would, ceteris paribus, *increase* utilization. Also, the lagged effect on enrollment is more consistent with the effects of parent demand given that we might expect a more immediate response by early-childhood centers to a lower subsidy rate.

unwilling to transfer already enrolled children. Second, the information set used by parents making enrollment decisions may depend largely on sources (e.g., the input from other parents) that respond sluggishly to changes in a program's official rating.

As a complement to our main outcomes (i.e., future star ratings, program enrollment, and ERS assessments), we also examined the effect of lower quality ratings on other program quality traits collected by North Carolina as part of its SRL program. These include staff education and experience, space requirements, and staff-child ratios. We find no evidence that the intent to treat with a lower star rating significantly influenced any of these measures. These null findings are likely to reflect in part the comparative relevance of the ERS rating for programs close to the threshold.

As noted above, our preferred full-sample specification conditions on both linear and quadratic splines of the assignment variable. However, to examine the robustness of our findings, we report the results of models predicting T+5 outcomes based on alternative functional forms and additional covariate controls (Table A5 in the Appendix). These specifications include local linear regressions that condition on a linear spline of the assignment variable using the data from increasingly tight bandwidths around the threshold. This includes the bandwidth of 1, a value chosen by the Imbens & Kalyanaraman (2012) procedure. We also show the results from RD specifications weighted by a triangular kernel. We also note that our findings are similar when we also condition on other baseline covariates like those in Table 3. The consistency of the findings across these specification choices suggests that our findings are not an artifact of functional form or omitted variable biases.

In Table 5, we examine how the effects of the intent to treat with a lower star rating differ by the level of competition that programs face from nearby programs. We present results separately for programs that faced “below median competition” and “above median competition,” where competition was defined as the number of other ECE programs within a five-mile radius. Treated programs in the high-competition sample had larger gains in ERS ratings. In T+4 and T+5, these programs improved relative to untreated programs by 0.23 and 0.27 points, respectively. This effect in T+5 is shown in Figure 5a. Treated programs in the low-competition sample improved by 0.08 and 0.07 points relative to untreated programs, gains that are not significantly different from zero in either year.

Five years after ERS ratings were issued, treated programs in the high-competition sample also enrolled almost 12 fewer students on average than untreated programs. By contrast, there was no detectable effect on enrollment among programs in the low-competition sample. The same pattern holds true when considering the proportion of capacity enrolled. These results are depicted for the high-competition sample in Figures 5b and 5c. The findings in Table 5 suggest that the presence of competition (i.e., nearby alternatives for ECE) is a substantively important moderator of whether incentives are effective in influence program performance. However, this heterogeneity might reflect the influence of other unobserved community traits that correlate with the presence of competition. To examine this issue, we also estimated these RD specifications controlling for zip code level characteristics (i.e., percent black, percent Hispanic, percent below poverty line, median income) and county fixed effects (results not shown). These results were quite similar to those presented in Table 5, suggesting that these differences are not likely to be due to other local characteristics related to the presence of ECE alternatives.

DISCUSSION

This paper examines the causal effects of the incentive contrasts created by a widely adopted policy innovation: state-level Quality Rating and Improvement Systems (QRIS) for ECE programs. Specifically, we examined the effects of receiving a lower versus higher star rating under North Carolina's Star Rated License program on subsequent program quality and enrollment. Understanding the effects of such QRIS incentives is critical as these accountability systems are among the most important policy efforts seeking to drive at-scale improvements in ECE. Using a regression-discontinuity (RD) design, we find that the lower star ratings caused ECE programs to substantially improve their performance as measured both by their summative star ratings and by the state's observations of their classrooms. Our RD results also indicate that a lower star rating eventually led to reduced enrollments suggesting the revealed preferences of parents.[12] Taken together, our results provide the first causally credible evidence on the key incentive mechanisms by which QRIS are intended to operate. They show that program rating cause significant changes in both program and parental behaviors.

Notably, we *did not* find that receiving a lower versus higher star rating under North Carolina's Star Rated License program led to improvements along a large set of other measured dimensions of quality. For instance, we did not find that missing the cut-off for a star rating led to improvements in child-staff ratios or teacher/administrator credentials. In part, the lack of improvement along these other dimensions is an artifact of our research design. Specifically, we leverage a treatment contrast in which treated programs stood to improve their overall star

[12] This parallels findings by Hastings & Weinstein (2008), who found that parents responded to information about quality by selectively enrolling their children into higher-quality care. One possibility for distinguishing between changes in enrollment driven by parents and by providers would be to compare effects across centers that face different enrollment incentives. For instance, Head Start providers, which are fully funded by the federal government, are not likely to be responsive to potential increases in state subsidies for child care. However, we are unable to examine the differential effect of this RD threshold on Head Start centers in North Carolina because these centers are required to maintain at least a 4 star rating, which means almost no Head Start centers fall below the 4.5 ERS threshold.

ratings by improving their ERS ratings by only a small amount. Programs could not necessarily improve their star ratings by improving a similar amount along other dimensions. This suggests that programs responded narrowly to the particular incentives that they faced.

Although our key findings suggest that both programs and families respond to QRIS ratings and the associated incentives, in some cases programs responded in ways counter to the intentions of the policy. For instance, we document suggestive (but weakly significant) evidence that a lower rating led some programs to opt out of participating in classroom observations (and the opportunity for higher ratings) in the future.[13] This effect was not sufficiently large or common enough to nullify the performance gains among programs assigned to a lower rating. However, it suggests that the ability to opt out of QRIS assessments is a policy design feature that merits careful attention as these accountability systems evolve. In North Carolina, QRIS incentives drove performance gains, on average, even when programs could opt out of an ERS assessment. However, this finding may reflect the fact that programs could not easily opt out of receiving an *overall* star rating. Many state QRIS systems are voluntary, and in those contexts QRIS may not lead to similar performance gains. Another related and open empirical question is whether a further narrowing of opt-out options (e.g., not allowing ECE programs in North Carolina to opt out of ERS assessments as easily) would amplify the incentive effects we found.

Another critical finding is that the effects of QRIS incentives appear concentrated in communities with higher levels of competition from other ECE providers. In fact, we do not find statistically significant effects of receiving a lower quality rating among those programs located in communities with few other ECE options, even when controlling for a host of community characteristics or including county fixed effects. This finding is consistent with research from K-

[13] This is consistent with experimental evidence that the effects of incentives can turn on whether the targeted behavior is perceived as responsive to effort (e.g., Camerer et al. 1999). Studies in education (e.g., Dee and Jacob 2006, Dee and Wyckoff 2015) similarly find that incentives can encourage attrition as well as performance gains.

12 that shows the effects of market-based reforms are larger when schools face greater competition (e.g., Belfield & Levin, 2002; Hoxby, 2003). This context-dependent evidence of moderation is important given that a fundamental motivation for state QRIS is the imperative to improve ECE at scale. Our evidence indicates that the performance effects of QRIS incentives may be limited to those communities with more extensive options. As other state QRIS mature, this will be another important area of inquiry.

There are two notable caveats to our findings. One is that our study tests a key theorized QRIS mechanism (i.e., the effects of incentives) but does *not* identify the average treatment effect (ATE) of introducing a QRIS. Stated differently, our RD design studies the effects of the incentive contrasts created by North Carolina's QRIS among ECE programs, all of whom are QRIS participants. However, the overall effects of introducing QRIS may differ from those of the incentives we study. For instance, between 2007-2014 North Carolina's licensed ECE programs made significant improvements on many of the quality indicators included in North Carolina's QRIS, and these improvements may have been driven by aspects of the QRIS. Our RD design cannot test that. Future studies may be able to leverage differences across states or across regimes to estimate the average treatment effect (ATE) of a state QRIS on program quality more directly.

A second caveat is that we are limited in our ability to make conclusions about *how* these improvements occurred and whether programs improved in ways that were meaningful for student learning. For example, although we see improvement in ERS ratings overall, these ratings encompass a diverse set of classroom measures, and we do not observe the specific dimensions on which these programs improved. A higher ERS rating could equate to added classroom materials, better personal care routines, more enriching interactions between children and staff, or a number of other possibilities. Some areas are likely to be easier to improve than others, and some may be more salient for student learning. This raises the possibility that program responses in North Carolina may have been concentrated along easily improved, but less important, dimensions of quality.

Relatedly, although ERS ratings are among the most widely used measures of quality in ECE programs, some studies have raised concerns that these summative ratings are not strongly related to student outcomes (e.g. Perlman, Zellman, & Le, 2004; Gordon et al., 2013). Similarly, Cannon, Zellman, Karoly & Schwartz (2017) raise concerns about the inconsistent and sometimes weak associations between QRIS ratings and children's learning. Further research on the validity and reliability of ECE quality measures will provide essential guidance to the designers of state QRIS. Despite these important design concerns, our findings from North Carolina provide seminal evidence consistent with the fundamental motivation for state QRIS; namely, that the incentives created by these accountability reforms influence the behaviors of both ECE programs and the parents of the children they serve.

REFERENCES

Akaike, H. (1974). A new look at the statistical model identification. IEEE transactions on automatic control, 19, 716-723.

Araujo, M., Carnerio, P., Cruz-Aguayo, Y., & Schady, N. (2014). A helping hand? Teacher quality and learning outcomes in kindergarten. Inter-American Development Bank.

Barnett, W. S., Friedman-Krauss, A., Gomez, R., Horowitz, M., Weisenfeld, G. G., & Squires, J. (2017). The state of preschool 2016: State preschool yearbook. National Institute for Early Education Research. Retrieved from http://nieer.org/sites/nieer/files/2015%20Yearbook.pdf

Bassok, D., Fitzpatrick, M., Greenberg, E., & Loeb, S. (2016). Within- and between-sector quality differences in early childhood education and care. Child Development, n/a-n/a. http://doi.org/10.1111/cdev.12551

Bassok, D., & Galdo, E. (2016). Inequality in preschool quality? Community-level disparities in access to high-quality learning environments. Early Education and Development, 27(1), 128–144. http://doi.org/10.1080/10409289.2015.1057463

Bassok, D., Markowitz, A., Player, D., & Zagardo, M. (2017). Do parents know "high quality" preschool when they see it? EdPolicyWorks working paper.

Belfield, C. & Levin, H. (2002). The effects of competition between schools on educational outcomes: A review for the United States. Review of Educational Research. 72(2), 279-341.

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. Volume 25 pp. 166_176.

Camerer, Colin F., et al. "The effects of financial incentives in experiments: A review and capital-labor-production framework." *Elicitation of Preferences*. Springer Netherlands, 1999. 7-48.

Cannon, J., Zellman, G. L., Karoly, L. A., & Schwartz, H. L. (2017). *Quality Rating and Improvement Systems for Early Care and Education Programs: Making the Second Generation Better*. Santa Monica, CA: RAND Corporation.

Cattaneo, M. D., Jansson, M., & Ma, X. (2017). Simple local polynomial density estimators. Working Paper. Retrieved July 22, 2017 from http://www-personal.umich.edu/~cattaneo/papers/Cattaneo-Jansson-Ma_2017_LocPolDensity.pdf

Congressional Research Service (2016). Preschool Development Grants (FY2014-FY2016) and Race to the Top – Early Learning Challenge Grants (FY2011-FY2013). Retrieved July 22, 2017 from https://www.everycrsreport.com/reports/R44008.html

Cryer, D., & Burchinal, M. (1997). Parents as child care consumers. Early Childhood Research Quarterly, 12, 35–58.

Dee, T. S., & Jacob, B. (2011). The impact of No Child Left Behind on student achievement. Journal of Policy Analysis and Management, 30, 418–446. http://doi.org/10.1002/pam.20586

Dee, T. S., & Jacob, B. A. (2006). Do high school exit exams influence educational attainment or labor market performance? National Bureau of Economic Research. Retrieved Jun 1, 2017 from http://www.nber.org/papers/w12199

Dee, T. S., & Wyckoff, J. (2015). Incentives, selection, and teacher performance: Evidence from IMPACT. Journal of Policy Analysis and Management. 34, 267-297.

Figlio, D., & Loeb, S. (2011). School Accountability. In E. A. Hanushek, S. Machin, & L. Woessmann (Eds.), Handbook of the economics of education. Vol. 3: [...] (1. ed). Amsterdam: North-Holland.

Friesen, J., Javdani, M., Smith, J., & Woodcock, S. (2012). How do school "report cards" affect school choice decisions? Canadian Journal of Economics/Revue Canadienne D'économique, 45, 784–807.

Gordon, R. A., Fujimoto, K., Kaestner, R., Korenman, S., & Abner, K. (2013). An assessment of the validity of the ECERS-R with implications for assessments of child care quality and its relation to child development. Developmental Psychology, 49, 146–160. http://doi.org/10.1037/a0027899

Hamre, B. K., & Pianta, R. C. (2005). Can instructional and emotional support in the first-grade classroom make a difference for children at risk of school failure? Child Development, 76, 949–967.

Hanushek, E. A., & Raymond, M. E. (2005). Does school accountability lead to improved student performance? Journal of Policy Analysis and Management, 24, 297–327.

Harms, T., Clifford, R., & Cryer, D. (1998). Early Childhood Environment Scale - Revised Edition.

Harms, T., Cryer, D., & Clifford, R. (2003). Infant/Toddler Environment Rating Scale Revised Edition.

Harms, T., Cryer, D., & Clifford, R. (2007). Family Child Care Environment Rating Scale - Revised Edition.

Harms, T., Jacobs, E., & White, D. (1996). School-Age Care Environment Rating Scale.

Hastings, J. S., & Weinstein, J. M. (2008). Information, school choice, and academic achievement: Evidence from two experiments. The Quarterly Journal of Economics, 123, 1373–1414. http://doi.org/10.1162/qjec.2008.123.4.1373

Heckman, J. J. (2006). Skill formation and the economics of investing in disadvantaged children. Science, 312, 1900–1902.

Howes, C., Burchinal, M., Pianta, R., Bryant, D., Early, D., Clifford, R., & Barbarin, O. (2008). Ready to learn? Children's pre-academic achievement in pre-Kindergarten programs. Early Childhood Research Quarterly, 23, 27–50. http://doi.org/10.1016/j.ecresq.2007.05.002

Hoxby, C. M. (2003). School choice and school productivity. Could school choice be a tide that lifts all boats? In The Economics of School Choice (pp. 287-342). University of Chicago Press.

Imbens, G., & Kalyanaraman, K. (2012). Optimal bandwidth choice for the regression discontinuity estimator. The Review of Economic Studies, 79, 933-959.

Koning, P., & van der Wiel, K. (2013). Ranking the schools: How school-quality information affects school choice in the Netherlands. Journal of the European Economic Association, 11, 466–493. http://doi.org/10.1111/jeea.12005

Lee, D. S., & Lemieux, T. (2009). Regression discontinuity designs in economics (Working Paper No. 14723). National Bureau of Economic Research. Retrieved June 1, 2017 from http://www.nber.org/papers/w14723

Lipsey, M., Farran, D., & Hofer, K. (2015). Evaluation of the Tennessee voluntary prekindergarten program: Kindergarten and first grade follow-up results from the randomized control design. Nashville, TN: Peabody Research Institute. Retrieved June 3, 2017 from https://my.vanderbilt.edu/tnprekevaluation/files/2013/10/August2013_PRI_Kand1stFollowup_TN-VPK_RCT_ProjectResults_FullReport1.pdf

Mashburn, A. J., Pianta, R. C., Hamre, B. K., Downer, J. T., Barbarin, O. A., Bryant, D., … Howes, C. (2008). Measures of classroom quality in prekindergarten and children's development of academic, language, and social skills. Child Development, 79, 732–749.

McCrary, J. (2008). Manipulation of the running variable in the regression discontinuity design: A density test. Journal of Econometrics, 142, 698–714. http://doi.org/10.1016/j.jeconom.2007.05.005

Mocan, N. (2007). Can consumers detect lemons? An empirical analysis of information asymmetry in the market for child care. Journal of Population Economics, 20(4), 743–780.

National Center on Child Care Quality Improvement. (2015). QRIS Resource Guide. QRIS National Learning Network. Retrieved July 1, 2017 from https://qrisguide.acf.hhs.gov/files/QRIS_Resource_Guide_2015.pdf

National Research Council. (2011). Incentives and test-based accountability in education. Washington, D.C.: National Academies Press. Retrieved April 1, 2017 from http://www.nap.edu/catalog/12521

North Carolina Division of Child Development and Early Education. (n.d.). Retrieved July 1, 2017 from http://ncchildcare.nc.gov/general/home.aspf

North Carolina Division of Child Development and Early Education. (2007). Subsidized child care rates for child care centers. Retrieved May 1, 2017 from http://ncchildcare.nc.gov/providers/pv_marketrates.asp

North Carolina Rated License Assessment Project. (n.d.). Retrieved May 6, 2017 from www.ncrlap.org

Perlman, M., Zellman, G. L., & Le, V.-N. (2004). Examining the psychometric properties of the Early Childhood Environment Rating Scale-Revised (ECERS-R). Early Childhood Research Quarterly, 19, 398–412. http://doi.org/10.1016/j.ecresq.2004.07.006

Puma, M., Bell, S., Cook, R., Heid, C., Broene, P., Jenkins, F., … Downer, J. (2012). Third grade follow-up to the Head Start Impact Study: Final report. OPRE report 2012-45. Administration for Children & Families. Retrieved July 3, 2017 from http://eric.ed.gov/?id=ED539264

QRIS National Learning Network. (2017). QRIS state contacts & map. Retrieved May 20, 2017, from http://qrisnetwork.org/sites/all/files/maps/QRISMap_0.pdf

Sabol, T. J., Hong, S. S., Pianta, R. C., & Burchinal, M. (2013). Can rating pre-k programs predict children's learning? Science, 341, 845–846.

Sabol, T. J., & Pianta, R. C. (2014). Do standard measures of preschool quality used in statewide policy predict school readiness? Education Finance and Policy, 9, 116–164.

Schochet, P., Cook, T., Deke, J., Imbens, G., Lockwood, J. R., Porter, J., & Smith, J. (2010). Standards for regression discontinuity designs. What Works Clearinghouse. Retrieved May 9, 2017 from http://eric.ed.gov/?id=ED510742

Snow, C., & Van Hemel, S. (2008). Early childhood assessment: Why, what, and how? National Research Council of the National Academies.

The Build Initiative & Child Trends. (2015). A catalog and comparison of Quality Rating and Improvement Systems (QRIS) [Data system]. Retrieved June 3, 2017 from www.qriscompendium.org

Tout, K., Zaslow, M., Halle, T., & Forry, N. (2009). Issues for the next decade of quality rating and improvement systems. Washington, DC: Child Trends. Retrieved June 3, 2017, from http://www.acf.hhs.gov/sites/default/files/opre/next_decade.pdf

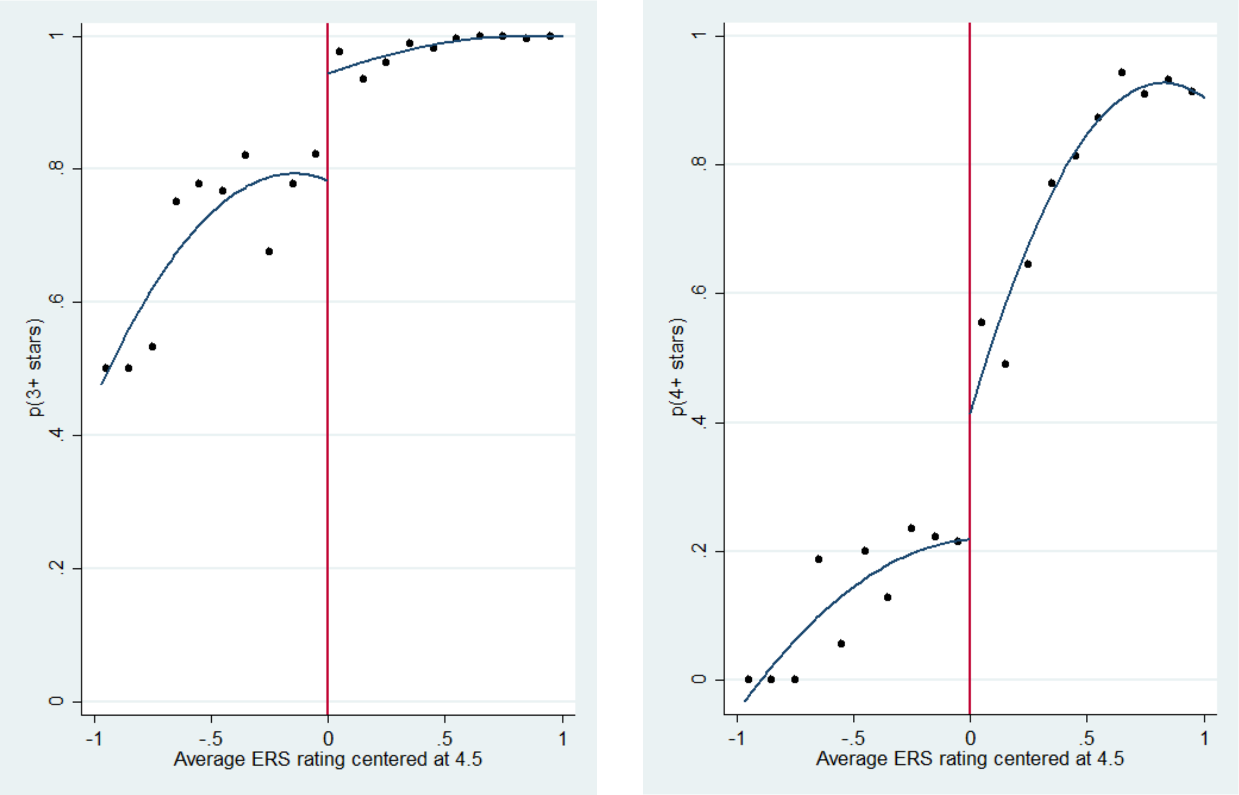
U.S. Department of Health and Human Services (2014). Child Care and Development Block Grant Act of 2014: Plain language summary of statutory changes. Retrieved July 1, 2017 from https://www.acf.hhs.gov/occ/resource/ccdbg-of-2014-plain-language-summary-of-statutory-changes

Waslander, S, Pater, C. & van der Weide, M. (2010). Markets in education: An analytical review of empirical research on market mechanisms in education. OECD Education Working Papers, No 52. OECD Publishing.

Wong, M., Cook, T. D., & Steiner, P. M. (2015). Adding design elements to improve time series designs: No Child Left Behind as an example of causal pattern-matching. Journal of Research on Educational Effectiveness, 8, 245–279. http://doi.org/10.1080/19345747.2013.878011

Yoshikawa, H., Weiland, C., Brooks-Gunn, J., Burchinal, M., Espinosa, L. M., Gormley, W. T., … Zaslow, M. J. (2013). Investing in our future: The evidence base on preschool education.

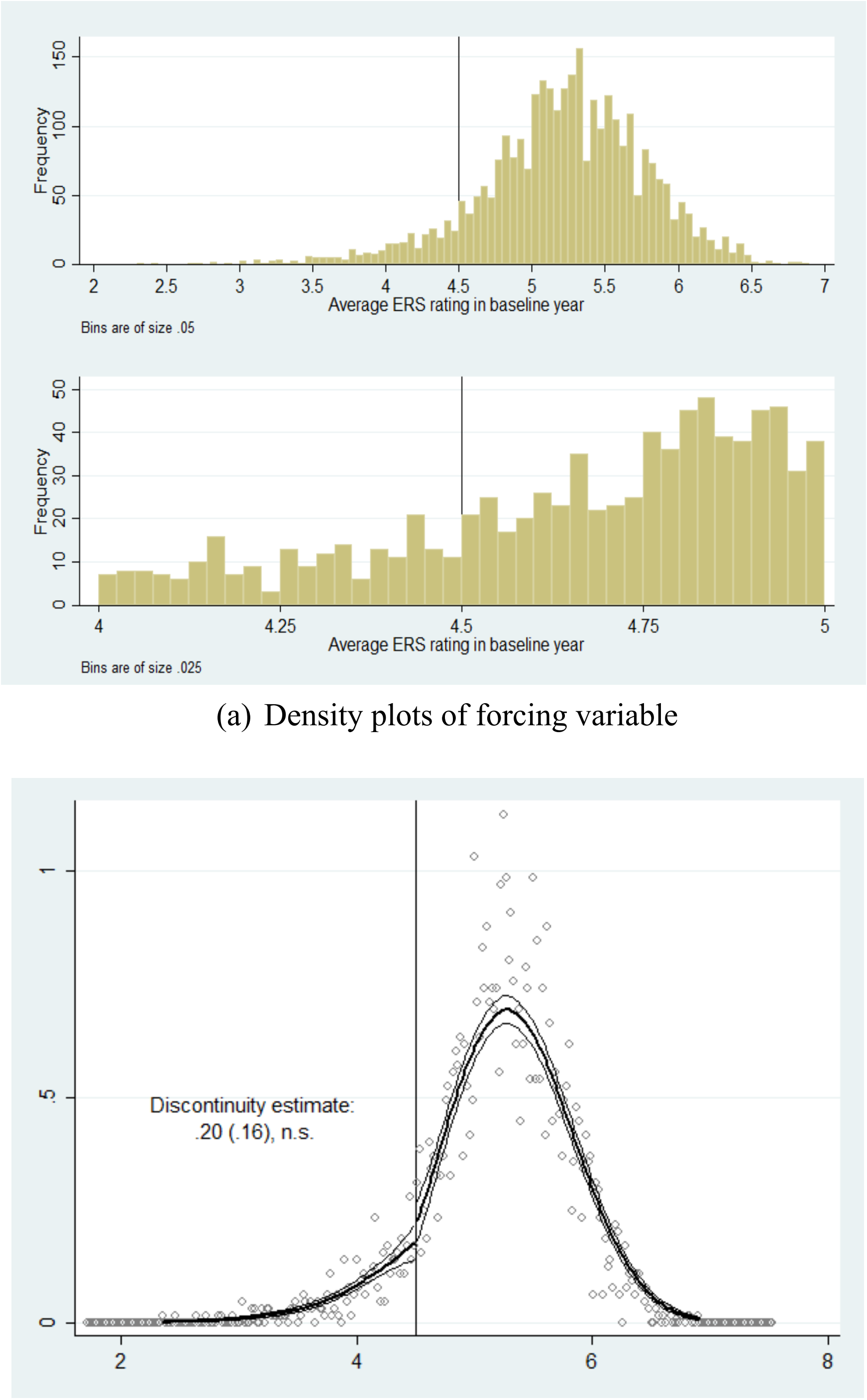
Zellman, G. L., & Perlman, M. (2008). Child-care quality rating and improvement systems in five pioneer states. Rand Corporation. Retrieved May 12, 2017 from http://www.rand.org/pubs/monographs/MG795/



(a) 3+ stars in year T

(b) 4+ stars in year T

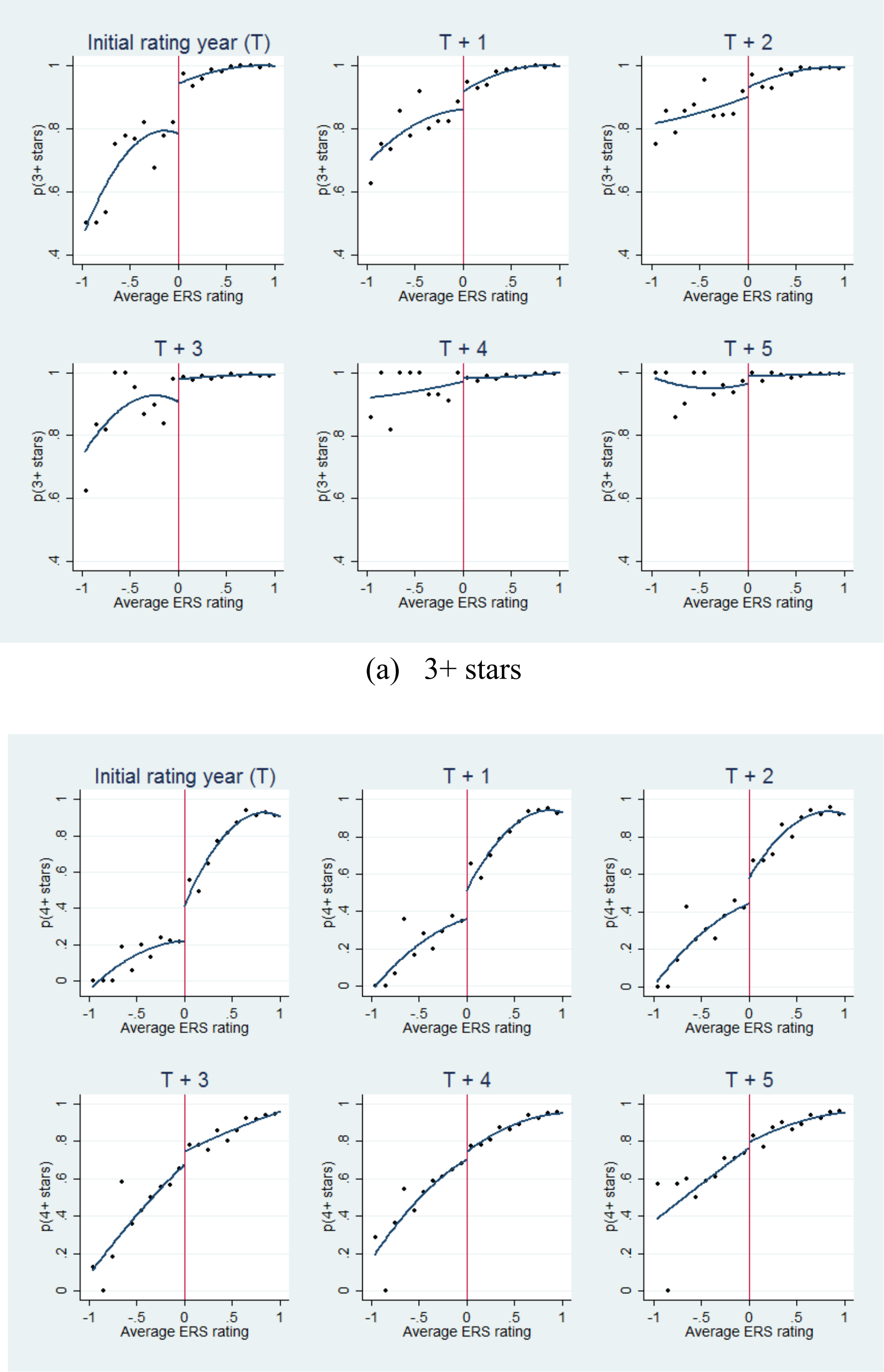
Figure 1 – First-stage relationships between average ERS ratings and star ratings in baseline year



(a) Density plots of forcing variable

(b) Density test (McCrary 2008)

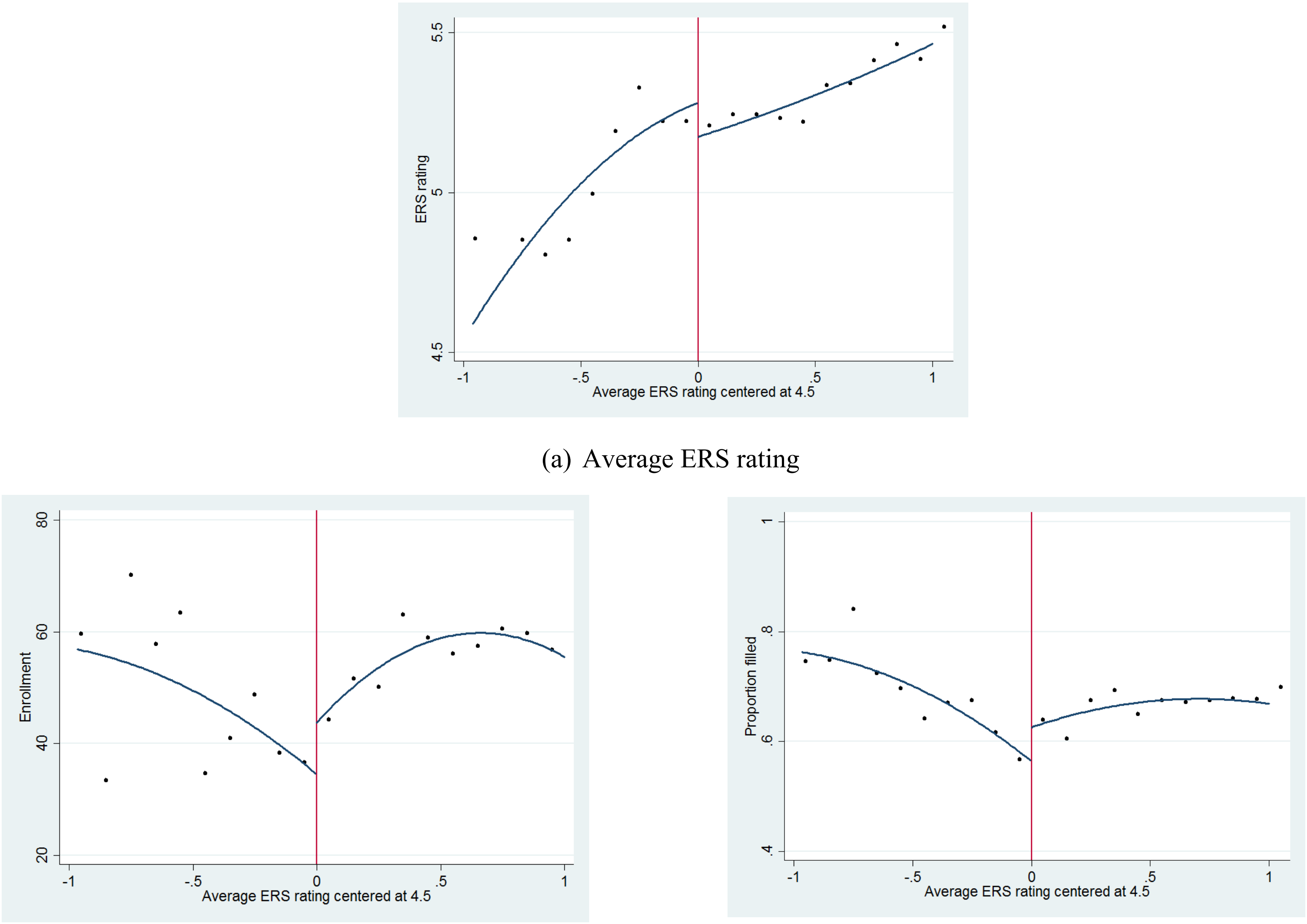
Figure 2 - Density of the forcing variable around the RD threshold



(a) 3+ stars

(b) 4+ stars

Figure 3 - Star ratings T through T+5 by baseline ERS rating

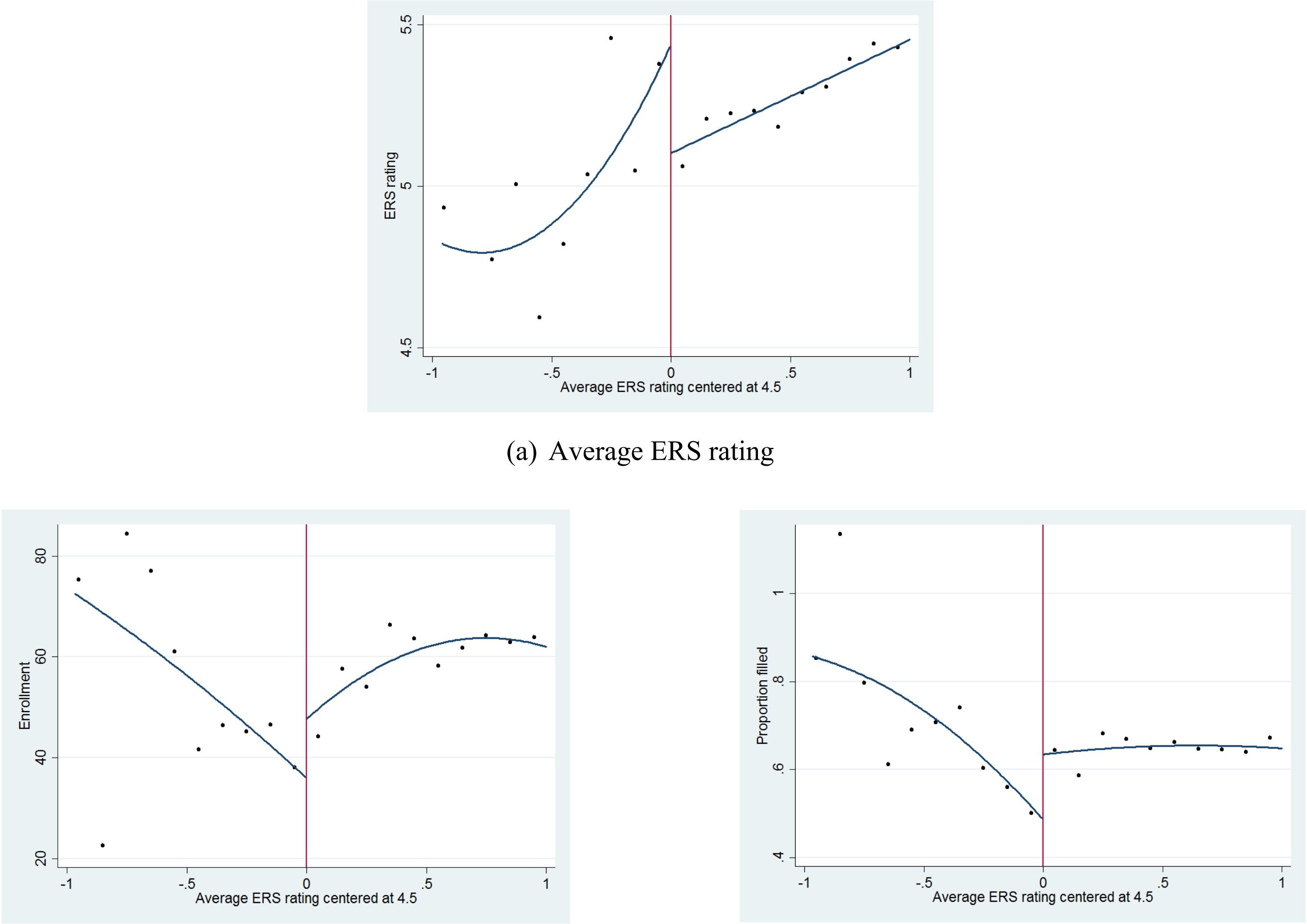


(a) Average ERS rating

(b) Total enrollment

(c) Proportion of capacity filled

Figure 4 - Full sample outcomes in T+5



(a) Average ERS rating

(b) Total enrollment

(c) Proportion of capacity filled

Figure 5 - High competition sample outcomes in T+5

Table 1 - Descriptive statistics for the analytic sample at baseline (T) through T+5

Center characteristic	T		T+1		T+2		T+3		T+4		T+5	
3+ star rating	0.97	(0.18)	0.97	(0.16)	0.98	(0.15)	0.98	(0.13)	0.99	(0.11)	0.99	(0.10)
4+ star rating	0.81	(0.39)	0.84	(0.37)	0.85	(0.36)	0.87	(0.34)	0.89	(0.32)	0.90	(0.30)
5 star rating	0.44	(0.50)	0.47	(0.50)	0.49	(0.50)	0.55	(0.50)	0.59	(0.49)	0.61	(0.49)
N	3157		2989		2809		2662		2520		2411	
Average ERS rating	5.21	(0.58)	5.23	(0.56)	5.26	(0.54)	5.36	(0.51)	5.40	(0.48)	5.43	(0.46)
ERS rating below 4.5	0.10	(0.30)	0.08	(0.28)	0.07	(0.26)	0.05	(0.21)	0.03	(0.17)	0.02	(0.15)
N	3157		2932		2718		2491		2336		2229	
Total enrollment	52.92	(43.44)	54.30	(44.11)	54.11	(44.20)	53.81	(44.31)	54.85	(44.50)	54.60	(44.53)
Proportion of capacity filled	0.71	(0.25)	0.72	(0.24)	0.70	(0.25)	0.69	(0.25)	0.69	(0.26)	0.68	(0.26)
Number of providers within 5 mi	40.72	(48.55)	43.79	(49.87)	45.45	(49.85)	45.01	(49.32)	44.05	(47.72)	43.88	(47.90)
N	3157		2989		2809		2662		2520		2411	

Note. Standard deviations in parenthesis. Year T includes observations from the years 2007-2009. Differences in sample sizes across years reflect providers that attrited from the sample, either because they closed or because they no longer had a valid ERS rating.

Table 2 – First-stage estimates across specifications and bandwidth restrictions

	Quadratic	Linear				
Dependent variable	Full sample	Full sample	1.5	1.25	1	Triangular kernel
3+ stars	-0.13**	-0.16***	-0.15***	-0.14***	-0.12**	-0.14**
	(0.05)	(0.04)	(0.04)	(0.04)	(0.04)	(0.05)
4+ stars	-0.29***	-0.47***	-0.43***	-0.38***	-0.33***	-0.28***
	(0.05)	(0.04)	(0.04)	(0.04)	(0.05)	(0.05)
N	3157	3157	2949	2619	2145	2122

Note. Each coefficient represents the results from a separate regression discontinuity estimate of the effect of a baseline average ERS rating below 4.5. In models based on the full sample, the Akaike information criterion privileges the quadratic specification, which also includes linear terms. Robust standard errors in parentheses.
+ *p* < .10; * *p* < .05; ** *p* < .01; *** *p* < .001.

Table 3 - Auxiliary regressions of baseline covariate balance

Dependent variable	RD estimate
Independent center	-0.02
	(0.06)
Local public school	-0.01
	(0.04)
Head Start	0.04
	(0.04)
Religious sponsored	-0.03
	(0.03)
Other center-based care	0.02
	(0.04)
N	3157

Note. Each row reports the RD estimate of the effect of a baseline average ERS rating below 4.5. Each estimate conditions on linear and quadratic splines of the assignment variables. Robust standard errors in parentheses.
+ *p* < .10; * *p* < .05; ** *p* < .01; *** *p* < .001.

Table 4 - Reduced-form RD estimates for outcomes at T+1 through T+5

Dependent variable	T+1	T+2	T+3	T+4	T+5
Panel A. Quality					
3+ stars	-0.07+	-0.05	-0.04	-0.00	-0.04
	(0.04)	(0.04)	(0.03)	(0.02)	(0.03)
4+ stars	-0.23***	-0.22***	-0.06	-0.06	-0.07
	(0.06)	(0.06)	(0.06)	(0.07)	(0.07)
N	2989	2809	2662	2520	2411
Average ERS rating	0.02	0.01	0.13	0.23*	0.20*
	(0.04)	(0.06)	(0.10)	(0.09)	(0.08)
N	2932	2718	2491	2336	2229
Panel B. Enrollment					
Total enrollment	-0.61	-0.64	-4.86*	-3.35	-7.20*
	(1.74)	(1.94)	(2.46)	(2.48)	(3.01)
Proportion of capacity filled	0.01	0.03	-0.04	-0.02	-0.07*
	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
N	2989	2809	2662	2520	2411

Note. Each coefficient represents a separate RD estimate of the effect of a baseline average ERS below 4.5. Each estimate conditions on linear and quadratic splines of the assignment variables. Estimates for "total enrollment" and "proportion of capacity filled" control for the baseline values of these outcomes. Robust standard errors in parentheses.

+ p < .10; * p < .05; ** p < .01; *** p < .001.

Table 5 – Reduced-form RD estimates by competition

	Below median competition (# of centers within 5 mi)					Above median competition (# of centers within 5 mi)				
Dependent variable	T+1	T+2	T+3	T+4	T+5	T+1	T+2	T+3	T+4	T+5
Panel A. Quality										
3+ stars	-0.09	-0.06	-0.10+	0.00	-0.03	-0.07	-0.05	-0.01	-0.02	-0.06
	(0.06)	(0.06)	(0.06)	(0.04)	(0.03)	(0.06)	(0.05)	(0.03)	(0.03)	(0.05)
4+ stars	-0.13	-0.07	0.03	0.03	-0.02	-0.30***	-0.32***	-0.13	-0.14	-0.13
	(0.09)	(0.09)	(0.09)	(0.09)	(0.10)	(0.08)	(0.09)	(0.10)	(0.10)	(0.10)
N	1424	1297	1222	1157	1114	1522	1472	1402	1326	1260
Average ERS rating	0.04	0.14+	0.07	0.08	0.07	0.01	-0.09	0.15	0.23+	0.27*
	(0.05)	(0.08)	(0.13)	(0.14)	(0.15)	(0.06)	(0.08)	(0.15)	(0.13)	(0.11)
N	1395	1255	1146	1079	1039	1494	1426	1310	1223	1156
Panel B. Enrollment										
Total enrollment	1.05	6.36*	-0.33	2.28	-1.11	-2.73	-7.34**	-9.27**	-7.88*	-11.84*
	(2.14)	(2.61)	(3.63)	(3.30)	(3.89)	(2.78)	(2.79)	(3.43)	(3.57)	(4.65)
Proportion of	0.05	0.13***	0.03	0.07	0.01	-0.02	-0.07*	-0.10*	-0.10*	-0.14**
capacity filled	(0.04)	(0.04)	(0.05)	(0.04)	(0.05)	(0.04)	(0.03)	(0.04)	(0.04)	(0.05)
N	1424	1297	1222	1157	1114	1522	1472	1402	1326	1260

Note. Each coefficient represents a separate RD estimate of the effect of a baseline average ERS below 4.5. Each estimate conditions on linear and quadratic splines of the assignment variables. Estimates for "total enrollment" and "proportion of capacity filled" control for the baseline values of these outcomes. Robust standard errors in parentheses.
+ p < .10; * p < .05; ** p < .01; *** p < .001.

Appendix

Calculation of program standards scores in North Carolina

In North Carolina, the program standards component of the QRIS accounts for nearly half of the total points that centers can receive (i.e. 7 out of a total 15). Criteria for the program standards component build on one another so that to receive a higher score a center must meet all requirements for each of the lower scores. Specifically, points are earned as follows. Many of these requirements refer to "enhanced standards," which are detailed in full immediately afterward.

Program standards score	Requirement
1	Meets minimum licensing requirements
2	Meets all enhanced standards except either staff-child ratios OR space requirements
3	Lowest classroom ERS score ≥ 4.0
4	Meets all enhanced standards except space requirements AND average ERS score ≥ 4.5 with no single score below 4.0
5	Average ERS score ≥ 4.75 with no single score below 4.0
6	Meets all enhanced standards AND average ERS score ≥5.0 with no single score below 4.0
7	Meets enhanced ratios minus 1 AND lowest classroom ERS score ≥ 5.0

Enhanced program standards (North Carolina Division of Child Development 2009):

Space requirements

- There must be at least 30 sq ft of inside space and 100 sq ft outside space per child per the licensed capacity, OR
- There must be at least 35 sq ft of inside space and 50 sq ft outside space per child per the licensed capacity
- There must be an area which can be arranged for administrative and private conference activities

Staff child ratios

- Staff-child ratios must be posted at all times in a prominent classroom area
- To meet enhance staff child ratio requirements, centers must meet the following criteria:

Age of children served	Staff child ratio	Maximum group size
0-12 months	1/5	10
1-2 years	1/6	12
2-3 years	1/9	18
3-4 years	1/10	20

Administrative policies:

- Selection and training of staff
- Communication with and opportunities for participation by parents
- Operational and fiscal management
- Objective evaluation of the program, management, and staff

Personnel policies

- Each center with 2 or more staff must have written personnel policies including job descriptions, minimum qualifications, health & medical requirements etc.
- Personnel policies must be discussed with each employee at the time of employment and copies must be available to staff
- Each employee's personnel file must contain an evaluation and development plan
- Personnel files must contain a signed statement verifying that the employee has received and reviewed personnel policies

Operational policies

- Must have written policies that describe the operation of the center and services which are available to children/parents, including days/hours of operation, age range of children served, parent fees, etc.
- Operational policies must be discussed with parents when they inquire about enrolling their child, and written copies must be provided
- Copies of operational policies must be distributed to all staff

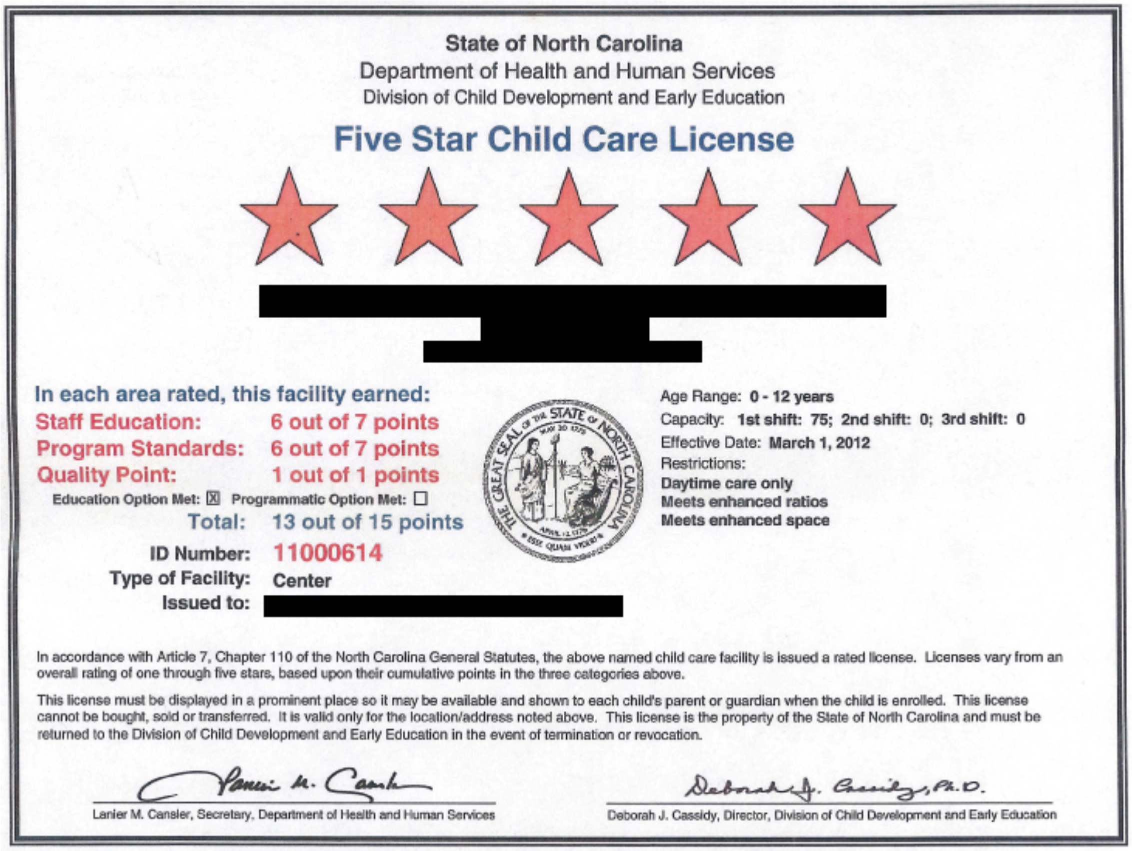
Caregiving activities for preschool aged children

- Each center providing care to preschool-age children 2 or older must provide all five of the following activity areas daily
    - Art/creative play
    - Children's books
    - Block & block building
    - Manipulatives
    - Family living & dramatic play
- The following activities must also be provided at least once per week
    - Music and rhythm
    - Science and nature
    - Sand/water play

Parent participation

- Each center must have a plan to encourage parent participation and inform parents about programs/services that includes the following
    - A procedure for encouraging parents to visit the center before their child starts attending
    - Opportunities for staff to meet with parents on a regular basis
    - Activities which provide parents opportunities to participate
    - A procedure for parents who need information or have complaints about the program
- The plan must be provided to and discussed with parents when the child is enrolled

Figure A1 - Sample five star rated license



State of North Carolina
Department of Health and Human Services
Division of Child Development and Early Education

Five Star Child Care License

★ ★ ★ ★ ★

[redacted]

In each area rated, this facility earned:

Staff Education:	**6 out of 7 points**
Program Standards:	**6 out of 7 points**
Quality Point:	**1 out of 1 points**

Education Option Met: ☒ **Programmatic Option Met:** □

Total:	**13 out of 15 points**
ID Number:	**11000614**
Type of Facility:	**Center**
Issued to:	[redacted]

THE GREAT SEAL OF THE STATE OF NORTH CAROLINA
MAY 20 1775
APRIL 12 1776
ESSE QUAM VIDERI

Age Range: **0 - 12 years**
Capacity: **1st shift: 75; 2nd shift: 0; 3rd shift: 0**
Effective Date: **March 1, 2012**
Restrictions:
Daytime care only
Meets enhanced ratios
Meets enhanced space

In accordance with Article 7, Chapter 110 of the North Carolina General Statutes, the above named child care facility is issued a rated license. Licenses vary from an overall rating of one through five stars, based upon their cumulative points in the three categories above.

This license must be displayed in a prominent place so it may be available and shown to each child's parent or guardian when the child is enrolled. This license cannot be bought, sold or transferred. It is valid only for the location/address noted above. This license is the property of the State of North Carolina and must be returned to the Division of Child Development and Early Education in the event of termination or revocation.

Lanier M. Cansler
Lanier M. Cansler, Secretary, Department of Health and Human Services

Deborah J. Cassidy, Ph.D.
Deborah J. Cassidy, Director, Division of Child Development and Early Education

Table A1 - Comparison of average characteristics for included and excluded ECE programs, 2007-2009

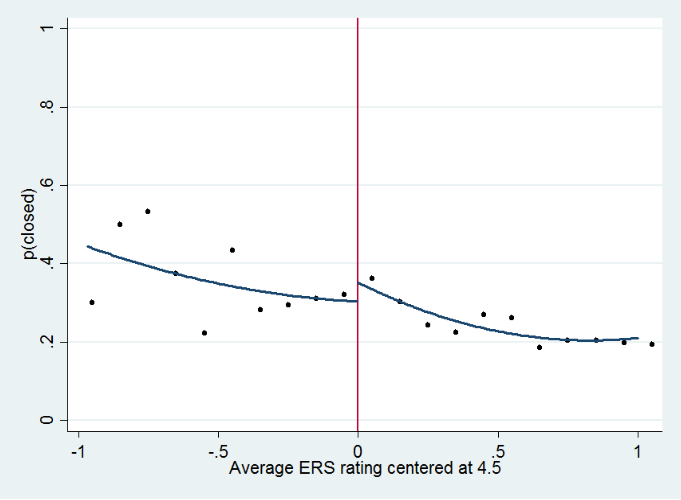
	2007		2008		2009	
Center characteristic	Sample	Non-sample	Sample	Non-sample	Sample	Non-sample
Independent center	0.44	0.53	0.44	0.52	0.45	0.53
Local public school	0.27	0.17	0.27	0.17	0.27	0.16
Head Start	0.10	0.01	0.09	0.01	0.09	0.02
Religious sponsored	0.08	0.21	0.08	0.22	0.08	0.22
3+ star rating	0.92	0.43	0.92	0.34	0.97	0.37
4+ star rating	0.73	0.10	0.76	0.07	0.83	0.07
5 star rating	0.38	0.01	0.42	0.00	0.46	0.00
ERS opt-out	0.45	1.00	0.14	1.00	0.02	1.00
Capacity	79.22	72.09	80.57	73.04	81.95	73.04
Total enrollment	54.16	43.39	53.36	42.78	53.33	40.41
Proportion of capacity filled	0.73	0.64	0.71	0.62	0.71	0.59
Number of providers within 5 miles	38.49	27.62	40.96	33.23	45.98	41.55
N	2970	2050	3053	1977	2952	2000

Note. This table compares mean values for child care centers in our sample to all other child care centers in North Carolina in the years 2007-2009. Centers were included in our sample if they received an ERS rating during the years 2007-2009, and excluded otherwise. The differences between sample and nonsample centers are significant at the .001 level for each variable in each year.

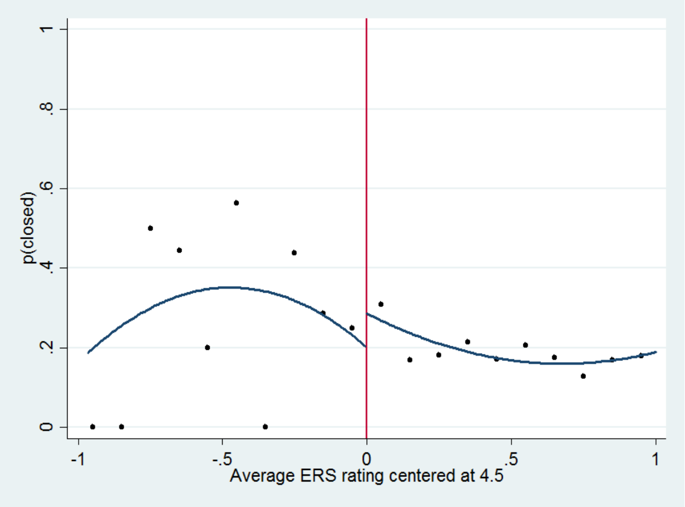
Table A2 – RD estimates for center closure

		T+1	T+2	T+3	T+4	T+5
Full sample	Sample mean	0.05	0.11	0.16	0.20	0.24
	RD estimate	-0.00	-0.04	-0.06	-0.07	-0.04
		(0.03)	(0.04)	(0.05)	(0.05)	(0.06)
	N	3157	3157	3157	3157	3157
High competition	Sample mean	0.02	0.05	0.10	0.14	0.19
	RD estimate	-0.02	-0.04	-0.06	-0.07	-0.02
		(0.02)	(0.03)	(0.05)	(0.06)	(0.08)
	N	1594	1594	1594	1594	1594
Low competition	Sample mean	0.09	0.17	0.22	0.26	0.29
	RD estimate	-0.01	-0.07	-0.09	-0.10	-0.08
		(0.06)	(0.07)	(0.08)	(0.08)	(0.08)
	N	1563	1563	1563	1563	1563

Note. Each RD coefficient represents a separate estimate of the effect of a baseline average ERS below 4.5. Each estimate conditions on linear and quadratic splines of the assignment variable. Robust standard errors in parentheses.
+ p < .10; * p < .05; ** p < .01; *** p < .001.



(a) Full sample



(b) High competition sample

Figure A2 - Probability of closure in T+5

Table A3 - RD estimates for ERS opt-outs

		T+3	T+4	T+5
Full sample	Sample mean	0.06	0.07	0.08
	RD estimate	0.09 (0.06)	0.12+ (0.06)	0.12+ (0.07)
	N	2662	2520	2411
High competition	Sample mean	0.07	0.08	0.08
	RD estimate	0.19* (0.08)	0.20* (0.09)	0.13 (0.09)
	N	1440	1363	1297
Low competition	Sample mean	0.06	0.07	0.07
	RD estimate	-0.02 (0.07)	0.03 (0.08)	0.10 (0.09)
	N	1222	1157	1114

Note. Each RD coefficient represents a separate estimate of the effect of a baseline average ERS below 4.5. Each estimate conditions on linear and quadratic splines of the assignment variable. Robust standard errors in parentheses.

+ p < .10; * p < .05; ** p < .01; *** p < .001.

Table A4 – RD estimates for early ERS re-rating

		T+1	T+2
Full sample	Sample mean	0.10	0.20
	RD estimate	0.09+	0.02
		(0.05)	(0.06)
	N	2989	2809
High competition	Sample mean	0.12	0.22
	RD estimate	0.06	-0.09
		(0.07)	(0.08)
	N	1565	1512
Low competition	Sample mean	0.09	0.18
	RD estimate	0.12+	0.16+
		(0.07)	(0.09)
	N	1424	1297

Note. Each RD coefficient represents a separate estimate of the effect of a baseline average ERS below 4.5. Each estimate conditions on linear and quadratic splines of the assignment variable. Robust standard errors in parentheses.
+ p < .10; * p < .05; ** p < .01; *** p < .001.

Table A5 - Reduced-form RD estimates in T+5 across bandwidths and specifications

	Quadratic		Linear				
Dependent variable	Full sample	Full sample w/controls	Full sample	1.5	1.25	1	Triangular kernel
3+ stars	-0.04	-0.04	-0.04*	-0.04+	-0.04+	-0.03	-0.03
	(0.03)	(0.03)	(0.02)	(0.02)	(0.02)	(0.03)	(0.03)
4+ stars	-0.07	-0.07	-0.14**	-0.15**	-0.12*	-0.05	-0.04
	(0.07)	(0.07)	(0.05)	(0.05)	(0.06)	(0.06)	(0.07)
ERS opt-out	0.12+	0.13*	0.19***	0.19***	0.16**	0.09	0.10
	(0.07)	(0.06)	(0.05)	(0.05)	(0.05)	(0.06)	(0.06)
Total enrollment	-7.20*	-6.72*	-6.82**	-7.07**	-8.03**	-7.89**	-7.12*
	(3.01)	(3.02)	(2.11)	(2.32)	(2.51)	(2.77)	(2.96)
Proportion of capacity filled	-0.07*	-0.06+	-0.02	-0.04	-0.06*	-0.06*	-0.06+
	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
N	2411	2409	2411	2252	1997	1619	1602
Average ERS rating	0.20*	0.18*	0.14*	0.19**	0.17*	0.16*	0.14+
	(0.08)	(0.08)	(0.07)	(0.07)	(0.07)	(0.08)	(0.08)
N	2229	2227	2229	2079	1832	1470	1455

Note. Each coefficient represents the results from a separate regression discontinuity estimate. Each estimate conditions on a quadratic spline of the assignment variables as well as an indicator equal to 1 if a center score below the RD threshold. Robust standard errors in parenthesis. Estimates that include controls condition on provider auspice (i.e. independent center, local public school, Head Start, religious-sponsored) as well as a fixed effect for the initial ERS rating year (i.e. 2007, 2008, or 2009). Estimates for "total enrollment" and "proportion of capacity filled" control for the baseline values of these outcomes. We privilege the quadratic results based on the Akaike information criterion.

+ p < .10; * p < .05; ** p < .01; *** p < .001.