The Effects of Accountability Incentives in Early Childhood Education

Daphna Bassok
Thomas S. Dee
Scott Latham

Abstract

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 about 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 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 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. © *2019 by the Association for Public Policy Analysis and Management.*

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[1] has grown rapidly, many programs are of low quality, particularly in low-income communities (Bassok & Galdo, 2016; Burchinal et al., 2010). 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

[1] Here and throughout the article, we use the term "ECE" broadly, to include any center-based child care provider, including those that are independently operated, part of a franchise, affiliated with Head Start, or any other local/state/federal initiative, religious-sponsored, or operated by any other agency.

Journal of Policy Analysis and Management, Vol. 0, No. 0, 1–29 (2019)

Published by Wiley Periodicals, Inc. View this article online at wileyonlinelibrary.com/journal/pam
DOI:10.1002/pam.22149

common 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; 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.

QRIS are similar to "organizational report cards" that have been employed in other markets for decades (Coe & Brunet, 2006; Gormley & Weimer, 1999). These programs are generally designed to provide simplified information about organization quality to the public. They have been shown to affect consumer demand in healthcare (Mukamel, Haeder, & Weimer, 2014), in the nonprofit sector (Grant & Potoski, 2015), and in the restaurant industry (Jin & Leslie, 2003).

Like accountability reforms in these 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. QRIS typically provide financial rewards for meeting standards, and many also offer technical assistance or professional development to help programs improve.

In addition, these accountability systems 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., one to five 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 subsequent program outcomes, including the program's overall quality scores, independent ratings of classroom quality as measured through observations, and subsequent program enrollments. 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 this credibly random assignment to a lower rating implied by North Carolina's QRIS 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 = 0.36). We also find that a lower QRIS rating led to reductions in program enrollment (effect size = 0.18). 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 underscore the importance of policy design that mitigates the possibly unintended consequences of such accountability systems. For instance, our findings show that centers that received a quality rating below the state-determined threshold made 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. This suggests 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 centers that received ratings below the RD threshold were more likely to opt 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 competitively allocated \$1.75 billion to states from 2011 through 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). It is notoriously challenging to measure quality in ECE settings in ways that are systematically and strongly related to children's learning gains (Burchinal, 2018). Still, existing research suggests that, although they are costlier to collect, measures of process quality (e.g., the Classroom Assessment Scoring System [CLASS]), are more consistent, though modest, predictors of children's learning than are structural measures (Araujo et al., 2016; Hamre & Pianta, 2005; Howes et al., 2008; Mashburn et al., 2008; Perlman et al., 2016; Sabol et al., 2013).

QRIS typically include measures of both structural and process quality. QRIS establish multiple "tiers" of quality (e.g., one to five stars) with benchmarks for each. They then rate programs based on their adherence to these measures. Programs often receive direct financial incentives for meeting higher quality benchmarks (e.g., subsidy reimbursement rates; merit awards), and states or local organizations may also provide support such as professional development and technical assistance (National Center on Child Care Quality Improvement, 2015). The ratings are also publicly available to parents and other stakeholders, who often struggle to discern program quality on their own (Bassok et al., 2018; Mocan, 2007).

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 suggest that QRIS policies may be effective tools for driving improvements in ECE quality at scale.

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. Research on the effects of information efforts in K-12 indicates that simply providing schools with information about quality did not lead to improvements in performance (Hanushek & Raymond, 2005), or to increased pressure on elected officials to improve low-performing schools (Kogan, Lavertu, & Peskowitz, 2016). 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 (e.g., Friesen et al., 2012; Koning & van der Wiel, 2013). Hastings and 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.

Compared to their choice set when making K-12 selections, many parents have a larger and more diverse set of options to consider when making early childhood choices, including providers with quite distinct foci, services, and costs. For this reason, information about program quality may have an even larger effect in the ECE context. Existing research shows that 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 (Bassok et al., 2018; Cryer & Burchinal, 1997; Mocan, 2007). 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.

Despite a plausible theoretical rationale and compelling evidence from the K-12 context, there is scant empirical 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 powerful enough to change the performance of ECE programs or the choices of parents 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 about $16 million annually to administer its QRIS (The BUILD Initiative and Child Trends, 2015), 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 (The BUILD Initiative and Child Trends, 2015; Tout et al., 2009; Zellman & Perlman, 2008).

Furthermore, while most state QRIS are voluntary, in North Carolina, all nonreligious ECE programs are automatically enrolled at the lowest (i.e., one star) level when they become licensed, and many religious-sponsored programs elect to participate as well. Thus, the vast majority of licensed ECE programs participate in the SRL program, including all Head Start programs, all state prekindergarten 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 percent of licensed center-based programs received star ratings in any given year. The

12 percent that do not receive star ratings consist primarily of religious-sponsored facilities (10 percent), with a smaller number having temporary/provisional licenses (2 percent). 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., three versus four 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 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.

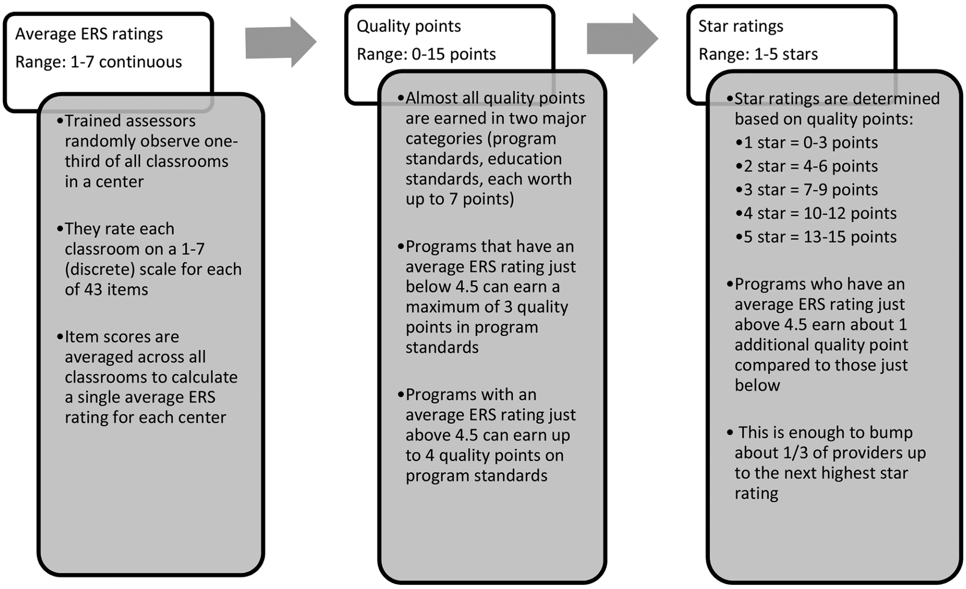
The Star Rated License (SRL) System[2]

North Carolina's Division of Child Development and Early Education rates ECE programs on a scale of one to five stars. The number of stars that a program receives is based on an underlying 15-point integer scale. These points map onto star ratings as follows: one star (zero to three points), two stars (four to six points), three stars (seven to nine points), four stars (10 to 12 points), and five stars (13 to 15 points). Both center-based and home-based child care programs can be rated as part of the SRL, but everything we describe below pertains to center-based programs, since they are the focus of the current study.

Programs can acquire up to 15 points through two subscales, each worth up to seven points, and an additional quality metric worth one point. The first subscale, "education standards" (worth between zero and seven discrete 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. The second subscale, "program standards" (also worth up to seven points), includes measures of quality such as staff-child ratios and square footage requirements. As described in detail below, the program standards subscale also includes an observational component, the Environment Rating Scale (ERS), scored on a continuous scale. The ERS is a widely used observation tool, currently included in 30 QRIS throughout the country. It is a broad measure of classroom quality, and it incorporates both structural features of the classroom (e.g., space and layout, daily schedules) as well as measures of "process" quality such as student-teacher interactions and classroom activities.

In addition to the "education standards" and "program standards," each program can receive one additional "quality point" by meeting at least one of a variety of other education or programmatic criteria (e.g., using a developmentally appropriate curriculum, having a combined staff turnover of no more than 20 percent, having 75 percent of teachers/lead teachers with at least 10 years of ECE experience).

[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 (ncchildcare.nc.gov).



Average ERS ratings
Range: 1-7 continuous

- Trained assessors randomly observe one-third of all classrooms in a center
- They rate each classroom on a 1-7 (discrete) scale for each of 43 items
- Item scores are averaged across all classrooms to calculate a single average ERS rating for each center

Quality points
Range: 0-15 points

- Almost all quality points are earned in two major categories (program standards, education standards, each worth up to 7 points)
- Programs that have an average ERS rating just below 4.5 can earn a maximum of 3 quality points in program standards
- Programs with an average ERS rating just above 4.5 can earn up to 4 quality points on program standards

Star ratings
Range: 1-5 stars

- Star ratings are determined based on quality points:
  - 1 star = 0-3 points
  - 2 star = 4-6 points
  - 3 star = 7-9 points
  - 4 star = 10-12 points
  - 5 star = 13-15 points
- Programs who have an average ERS rating just above 4.5 earn about 1 additional quality point compared to those just below
- This is enough to bump about 1/3 of providers up to the next highest star rating

Figure 1. Summary of How Average ERS Ratings Greater Than or Equal To 4.5 Relate to Higher Star Ratings.

A feature of the SRL system that is central for this study is the relationship between programs' ERS ratings, their points, and, ultimately, their star ratings. Specifically, programs that exceed specified thresholds on the ERS scale are eligible for more "program standards" points, and, in turn, they receive higher star ratings.

For instance, a program with an *average* ERS rating of 4.5 is eligible for up to four points on the program standards subscale, whereas a program that has an average ERS score just under 4.5 is only eligible for three points (see the Appendix[3] 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 one-third of a star, these small differences in ERS ratings lead to meaningful discontinuities in the probability of earning a higher versus lower star rating. The relationship between ERS ratings and star ratings is summarized in Figure 1.

Programs are not required to receive ERS ratings, but they face strong incentives to do so. Specifically, programs that opt to forego an ERS rating can earn a maximum of just 10 out of 15 total QRIS points. This makes it impossible to earn a five-star rating (which requires 13 points) and means that a program would need to earn every other possible point to earn a four-star rating (which requires 10 points). In practice, most programs opt to receive ERS ratings, and the percentage has increased over time, from 52 percent in 2008 to 66 percent by 2014. The decision to opt out of receiving an ERS rating is one of the policy-relevant outcomes we study.

In North Carolina, the Division of Child Development contracts with the North Carolina Rated License Assessment Project (NCRLAP) to conduct ERS assessments. These assessments are in addition to and separate from unannounced health and

[3] All appendices are available at the end of this article as it appears in JPAM online. Go to the publisher's website and use the search engine to locate the article at http://onlinelibrary.wiley.com.

safety inspections, which all licensed centers must undergo at least once every six months, without exception. Programs that wish to receive an ERS rating 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 be re-rated sooner must wait a minimum of six months after their previous rating, and they must cover the cost of assessment on their own (North Carolina Rated License Assessment Project, n.d.).

During the rating process, trained 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- to 4-year-olds, school-aged children). Prior to conducting formal observations, all observers must achieve 85 percent reliability in training (i.e., be within one point in their evaluation of 85 percent of the ERS items). Assessors spend a minimum of three hours in each classroom, recording notes on a wide variety of interactions, activities, and materials. They also spend 30 to 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.[4] Each item is scored either a one, three, five, or seven, indicating "inadequate," "minimal," "good," or "excellent" quality, respectively. The scores are then averaged across items to determine each program's overall ERS rating (The BUILD Initiative and Child Trends, 2015). These ratings serve as a continuous measure of program quality. In our data they are limited to two decimal places.

The Treatment Contrast

In the regression-discontinuity design we describe below, each program's baseline ERS rating serves as an assignment variable that creates plausibly exogenous variation in 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 four or more points on the program standards subscale.[5] Our research design makes use of the fact that small differences in programs' average ratings, driven in part by the measurement error in the underlying ERS tool rather than true quality differences across programs, can make a meaningful difference in programs' ratings. We show that programs' baseline scores relative to the 4.5 threshold generate a discontinuous "jump" in the likelihood a program earns more stars.

The treatment contrast defined by this "intent to treat" (ITT) merits careful scrutiny. The star ratings received by ECE programs are critical components in

[4] Four different versions of the ERS are used in North Carolina depending on the age of the 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 same continuous one to seven scale, and each contains measures of basic care provision, physical environment, curriculum, interactions, schedule/program structure, and parent/staff education.

[5] The SRL system also implies other candidate thresholds that could potentially be leveraged using a regression discontinuity. For instance, centers are also 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. In the current study, we focus on the 4.5 cutoff primarily because it offers the strongest "first stage" relationship (i.e., this cutoff is most strongly related to star ratings).

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 (the baseline year of our sample), the average five-star program received $642 per subsidy-eligible student, compared with $587 per student for a four-star program. An average three-star program received $560 per subsidy-eligible student and a two-star program received only $407 per student (NC Division of Child Development and Early Education, 2007). These performance-defined differences in subsidy rates may encourage lower-rated programs, particularly those that 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 on Child Care Quality Improvement, 2015; see Figure A1[6] for an example of a star-rated license).

Because North Carolina's QRIS includes non-trivial financial incentives as well as plausible market incentives created by publicizing the ratings, it provides a compelling context 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 that earn ERS ratings just below the RD threshold may focus on improving their ERS ratings, because small improvements along this dimension are likely to lead to higher star ratings. We expect to see improvements along this measure three years after the initial ratings occurred, because ERS ratings are valid for three years. In practice however, about 12 percent of programs did not receive new ratings until at least four years after the initial rating, so improvements may not be apparent until even later.[7] We also hypothesize that programs that fall just below the RD threshold 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. Finally, we expect that the effects of QRIS incentives will vary. In markets where providers face high levels of competition, QRIS incentives are likely to be more salient and powerful than in markets with low levels of competition.

DATA

Our analysis leverages program-by-year data for all licensed center-based ECE programs in the state of North Carolina in the years 2007 to 2014 (N = 6,929 unique

[6] All appendices are available at the end of this article as it appears in JPAM online. Go to the publisher's website and use the search engine to locate the article at http://onlinelibrary.wiley.com.

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

center-based programs across the entire panel).[8] 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, information about program sector (e.g., independently operated, Head Start, local public school), enrollment, and licensed capacity.[9] The data also include 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.

To construct a valid ITT sample, we include only ERS ratings that occurred after the 2005 revision to North Carolina's QRIS, because this revision changed the relationship between ERS ratings (the assignment variable) and star ratings (the first-stage outcomes). Our data begin in 2007, but because many of the programs were in operation prior to the regime change, and since ERS ratings are valid for three years, some of the ratings we observe in the earliest years of our panel occurred under the original QRIS regime. To determine each program's first rating after the regime change, we rely on recorded ERS visit dates where available (about 47 percent of observations). For the remaining observations, we use the following decision rules:

1. If we observed the same rating for a program throughout the years 2007 to 2009, we assumed the rating occurred in 2007, and thus came from the new regime.
2. In cases where the 2008 or 2009 rating differed from the 2007 rating, we assumed that the *changed* rating was the first under the new regime.

We limit our ITT sample to programs observed at some point in the three-year window 2007 through 2009, which allows us to track program outcomes for five years after the baseline observation. Our data include 5,866 unique center-based programs that were observed at baseline. However, we exclude 844 programs that never had a star rating (i.e., religious programs that chose not to participate or programs with temporary or provisional licenses), 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. Finally, we exclude 207 providers that served only school-aged children (i.e., no children ages 5 or below).

Our final ITT sample includes 2,950 unique center-based ECE programs. Table 1 presents descriptive statistics for this sample in the baseline year (T) and for subsequent years through T+5. At baseline, nearly all programs (97 percent) had earned at least a three-star rating, 80 percent had at least a four-star rating, and 42 percent had earned a five-star rating. The average ERS rating was 5.21, indicating relatively high quality across the sample. The average enrollment was about 52 children, and programs were operating, on average, at 72 percent of their licensed capacity. As mentioned above, this does not necessarily mean that there were many unfilled slots in these programs, as capacity is a measure of available square footage rather than an estimate of the number of children a program aims to serve.

[8] We include only center-based programs in our sample (i.e., we exclude home-based programs) because a much lower portion of home-based providers participated in the QRIS, and this group constitutes a small and self-selected portion of available home-based care.

[9] We use capacity as a proxy for the availability of ECE slots. In practice, however, licensed capacity is a measure of the physical space available within a center, rather than an estimate of the number of children a program actually aims to serve. For instance, programs must have at least 25 square feet of indoor space and 75 square feet of outdoor space per child.

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	(0.10)	0.99	(0.11)
4+ star rating	0.80	(0.40)	0.83	(0.37)	0.85	(0.36)	0.87	(0.34)	0.89	(0.32)	0.90	(0.30)
5 star rating	0.42	(0.49)	0.46	(0.50)	0.47	(0.50)	0.54	(0.50)	0.58	(0.49)	0.60	(0.49)
N	2950		2789		2617		2475		2341		2239	
Average ERS rating (conditional on having ERS)	5.21	(0.58)	5.23	(0.57)	5.25	(0.55)	5.35	(0.52)	5.39	(0.48)	5.42	(0.46)
ERS rating below 4.5	0.10	(0.30)	0.09	(0.28)	0.08	(0.27)	0.05	(0.22)	0.03	(0.17)	0.03	(0.16)
N	2950		2737		2532		2316		2171		2068	
Total enrollment	52.3	(43.8)	53.8	(44.5)	54.0	(44.9)	53.7	(44.9)	54.8	(45.2)	54.5	(45.1)
Proportion of capacity filled	0.72	(0.25)	0.73	(0.24)	0.72	(0.25)	0.70	(0.25)	0.70	(0.25)	0.70	(0.25)
Num. providers within 5 mi.	41.0	(48.7)	44.1	(50.1)	45.8	(50.2)	45.3	(49.6)	44.4	(48.0)	44.3	(48.1)
N	2950		2789		2617		2475		2341		2239	

Notes: Year T includes observations from the years 2007 to 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. Standard deviations are in parentheses.

Not surprisingly, our analytic sample differs from the excluded programs in several ways (see Table A1[10]). For example, in 2007, the excluded programs were more likely to have religious sponsorship (e.g., 20 percent versus 9 percent in our study sample), which was expected since religious programs have the option not to participate in the QRIS. Excluded programs were also somewhat more likely (50 percent versus 46 percent) to be independently operated (i.e., not affiliated with any local/state/federal program or with a franchise). Only 1 percent of excluded programs were Head Start programs, compared with 10 percent 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.

These systematic differences have implications for the external validity of our findings. Specifically, our findings are most relevant for the types of programs that opt to fully participate in the QRIS and receive an ERS observation score. Still, our sample includes over 60 percent of center-based providers in the years we study, which is a larger portion of providers than participate in most state QRIS (The BUILD Initiative & Child Trends, 2015). This relatively broad coverage is a strength of the current study. Importantly, these necessary sample restrictions do not affect the internal validity of our estimates.

REGRESSION DISCONTINUITY DESIGN

Our RD analysis compares outcomes among programs whose average ERS rating at baseline is just above or below the 4.5 threshold. 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.

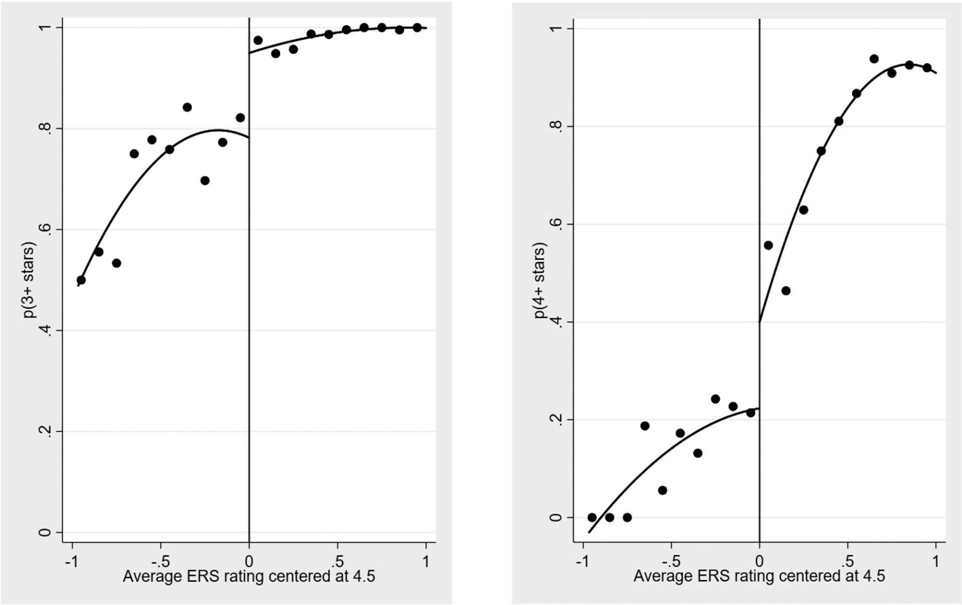
As is common practice (e.g., Lee & Lemieux, 2010; Schochet et al., 2010), we employ a combination of graphical and statistical evidence in our analysis. We estimate the magnitude and statistical significance of receiving a higher vs. lower star rating using reduced-form specifications that take the following form for outcome Y_i associated with program i:

$$Y_i = \gamma I(S_i < 0) + k(S_i) + \alpha_i + \varepsilon_i. \tag{1}$$

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. We condition on a fixed effect, $\alpha_{\mathrm{i},}$ for the specific year in which a program's ERS rating occurred (2007 through 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

[10] All appendices are available at the end of this article as it appears in JPAM online. Go to the publisher's website and use the search engine to locate the article at http://onlinelibrary.wiley.com.



(a) 3+ stars in year T

(b) 4+ stars in year T

Figure 2. First-Stage Relationships Between Average ERS Ratings and Star Ratings in Baseline Year.

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 five 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 we 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 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 2, 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 that earned at least a three-star rating or at least a four-star rating in each bin. We restrict these figures to a bandwidth of one around the focal RD threshold and superimpose regression lines from parametric estimates with quadratic splines.

Figure 2 shows that in North Carolina, programs whose average ERS rating was under 4.5 were significantly less likely to receive at least a three-star rating than those just at or above 4.5. These programs were also significantly less likely to receive at least a four-star rating. In Table 2, we present analogous regression estimates. These

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^{**} (0.05)	-0.15^{***} (0.04)	-0.15^{***} (0.04)	-0.14^{***} (0.04)	-0.13^{**} (0.04)	-0.14^{**} (0.05)
4+ stars	-0.27^{***} (0.05)	-0.45^{***} (0.04)	-0.41^{***} (0.04)	-0.37^{***} (0.04)	-0.31^{***} (0.05)	-0.26^{***} (0.05)
N	2950	2950	2753	2448	2004	1982

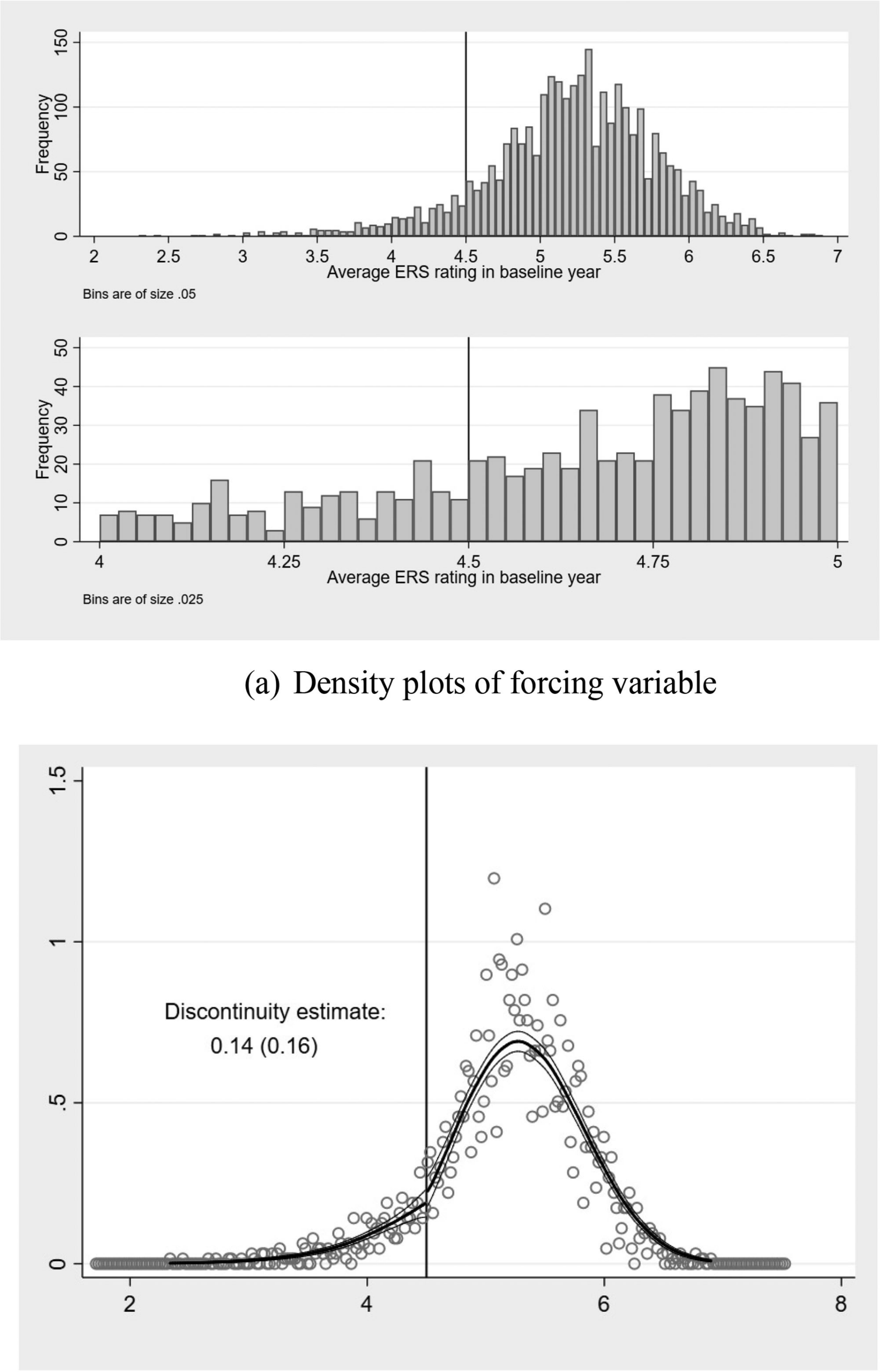
Notes: Each coefficient represents the results from a separate regression discontinuity estimate of the effect of a baseline average ERS rating below 4.5. In models based on the full sample, the Akaike information criterion privileges the quadratic specification, which also includes linear terms. Robust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

estimates show that, for the full sample, programs just below the RD threshold were 13 percentage points less likely to earn three or more stars and 27 percentage points less likely to earn four or more 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 and Kalyanaraman (2012) bandwidth of one. 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 the baseline average ERS rating relative to the RD threshold. This is unlikely to be a concern in the current context since the average ERS score for a program is an aggregation among multiple classrooms, and each classroom's ERS score is an aggregation of 38 or more individual items, making it difficult to predict or manipulate a program's exact rating. Still, in theory, either ECE programs or raters could be a source of such manipulation. Although programs have access to ERS rubrics and are encouraged to conduct self-assessments on the ERS, these self-assessments do not provide precise information about the ERS ratings that programs will ultimately receive, since the rating will depend on the specific experiences observed in the classroom at the time of the observation visit. Raters, who likely know the implications of receiving scores above or below particular thresholds, could, in theory, 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 most of the programs in our sample have two or more classrooms), a single classroom's rating cannot 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 3a). 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



(a) Density plots of forcing variable

(b) Density test (McCrary, 2008)

Figure 3. Density of the Forcing Variable Around the RD Threshold.

Table 3. Auxiliary regressions of baseline covariate balance.

Dependent variable	RD estimate
Independent center	−0.06 (0.06)
Local public school	0.03 (0.04)
Head Start	0.04 (0.04)
Religious sponsored	−0.03 (0.03)
Other center-based care	0.03 (0.04)
N	2950

Notes: 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 are in parentheses.
+ p < .10; *p < .05; **p < .01; ***p < .001.

(McCrary, 2008, Figure 3b) as well as a newly developed alternate procedure proposed by Cattaneo, Jansson, and Ma (2017).[11] Consistent with our visual inspection of the density function, we fail to reject the null hypothesis of no discontinuity with both tests. 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.

RESULTS

We begin illustrating our main findings graphically. Unless otherwise stated, all of these results reflect ITT estimates. Figure 4 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 that a program has three 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 three 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 three stars in T+5. By contrast, panel (b) of Figure 4 considers the probability that a program earned four 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 four or five 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 4 suggests, these RD results indicate that the baseline

[11] 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.

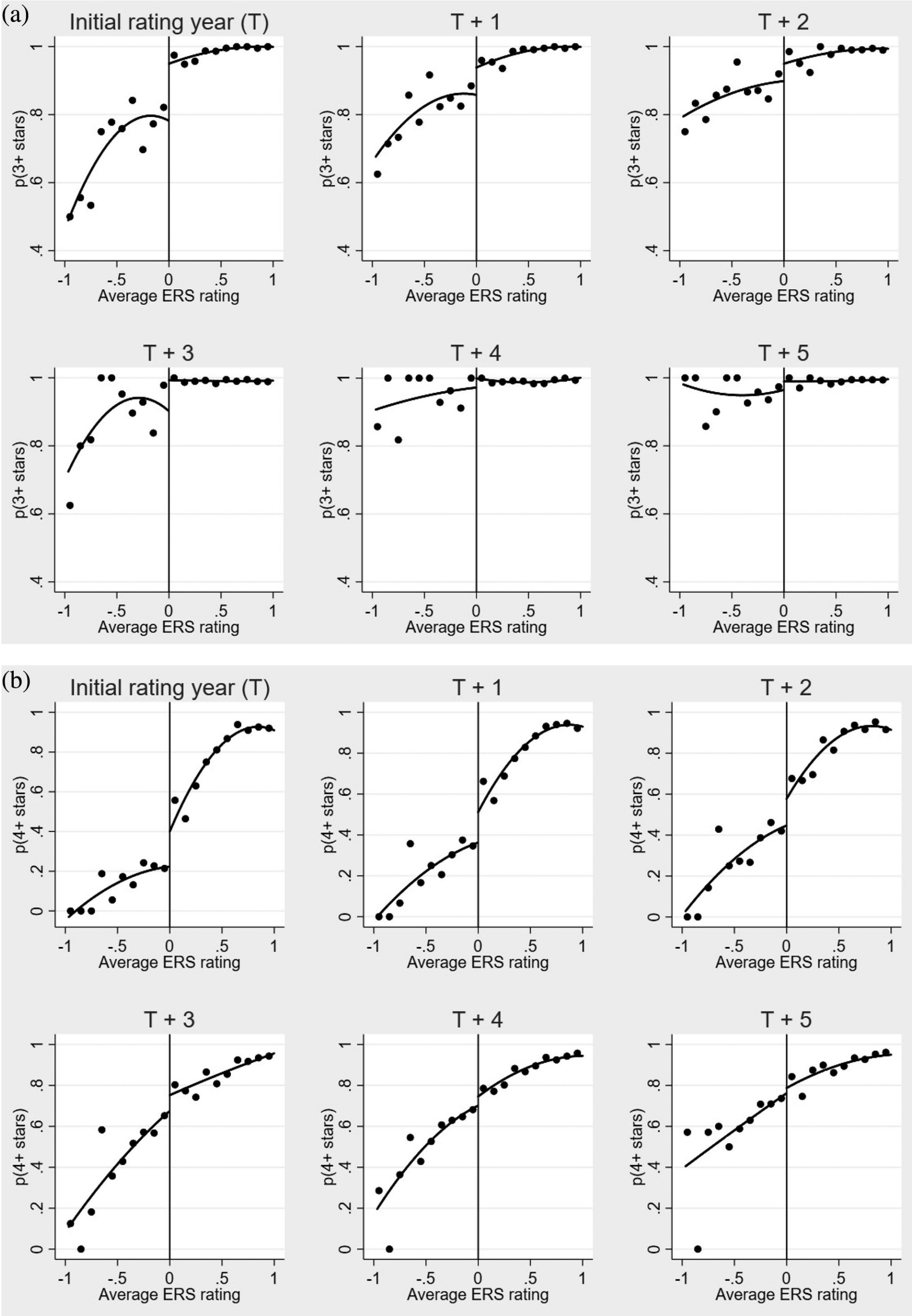


Figure 4. (a) Likelihood of Earning Three or More Stars in T through T+5 by Baseline ERS Rating. (b) Likelihood of Earning Four or More Stars in T through T+5 by Baseline ERS Rating.

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	−0.08^{+}	−0.06	−0.05	−0.01	−0.03
	(0.04)	(0.04)	(0.03)	(0.02)	(0.03)
4+ stars	−0.22***	−0.21***	−0.06	−0.06	−0.07
	(0.06)	(0.06)	(0.07)	(0.07)	(0.07)
N	2789	2617	2475	2341	2239
Average ERS rating	0.02	0.01	0.15	0.25**	0.21*
	(0.04)	(0.06)	(0.10)	(0.09)	(0.08)
N	2737	2532	2316	2171	2068
Panel B. Enrollment					
Total enrollment	−0.79	−0.99	−5.32*	−4.11	−7.95**
	(1.73)	(1.94)	(2.48)	(2.50)	(3.03)
Proportion of capacity filled	0.01	0.02	−0.04	−0.02	−0.08*
	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
N	2789	2617	2475	2341	2239

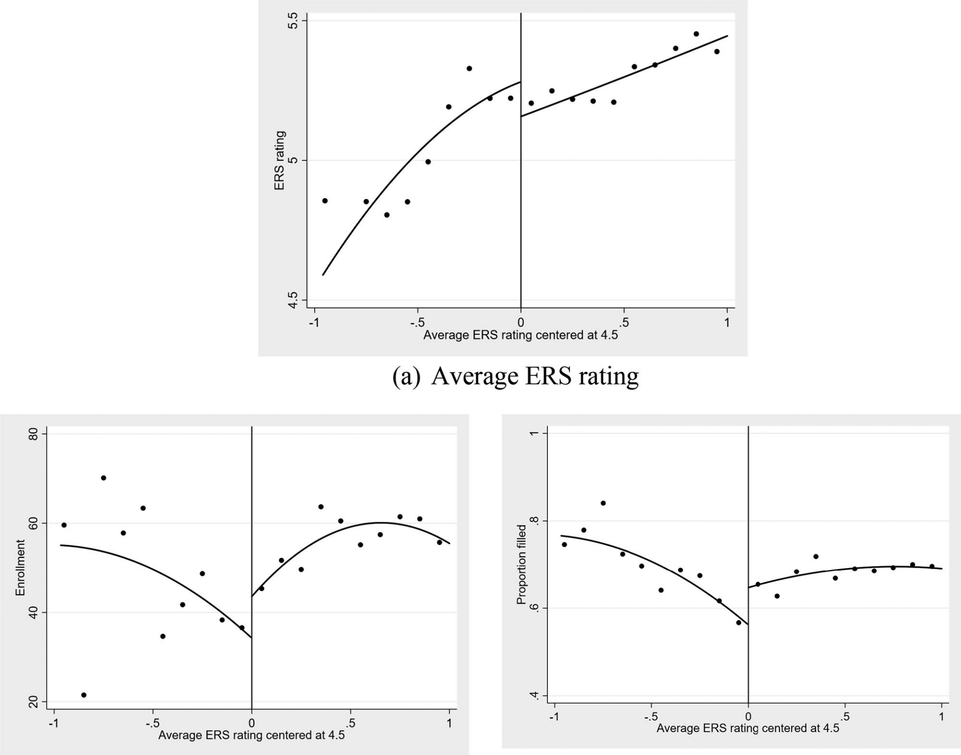
Notes: Each coefficient represents a separate RD estimate of the effect of a baseline average ERS below 4.5. Each estimate conditions on linear and quadratic splines of the assignment variables. Estimates for "total enrollment" and "proportion of capacity filled" control for the baseline values of these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

ratings gap created by a program's position relative to the 4.5 threshold shrunk and was no longer statistically significant within three years of the initial ratings assignment. These results suggest that credibly 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 three years, we are most interested in estimates from periods T+3, T+4, and T+5. We find that in T+3, which is the first point at which we would expect an increase in ERS scores given the timing of observations, programs below the 4.5 threshold had somewhat higher ERS ratings (i.e., an increase of 0.13) than programs just above the threshold. However, this difference was not statistically significant.[12] By T+4 and T+5, we do find that average ERS ratings jumped by 0.25 and 0.21, respectively, among programs to the left of the threshold. Figure 5(a) illustrates this relationship graphically in T+5. An ERS gain of 0.21 constitutes a 0.36 effect size with respect to the standard deviation observed at baseline (i.e., 0.21/0.58, where the denominator comes from Table 1).[13] Given our first-stage estimates (Table 2), this ITT estimate implies that

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

[13] 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



(a) Average ERS rating

(b) Total enrollment

(c) Proportion of capacity filled

Figure 5. Full Sample Outcomes in T+5.

the estimated effect of receiving a three-star rating instead of a four-star rating is over 1.2 *program-level* standard deviations (i.e., 0.36/0.27). 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 explored whether programs to the left of the cutoff were more likely to pay for an earlier ERS observation, rather than wait for the free rating that occurs every three years. 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 A2[14]). 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.

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.

[14] All appendices are available at the end of this article as it appears in JPAM online. Go to the publisher's website and use the search engine to locate the article at http://onlinelibrary.wiley.com.

We also examined the effect of lower quality ratings on other aspects of program quality collected by North Carolina as part of its SRL program, including staff education and experience, space requirements, and staff-child ratios (results available upon request). 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.

Next, we examined the effects 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 about five 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 less than eight children (effect size = 0.18). We also find that by T+5, programs that were initially below the 4.5 threshold had a reduction in their capacity utilization of 8 percentage points. We illustrate these findings graphically in Figures 5(b) and 5(c).

These results are consistent with the hypothesis that parents were less willing to enroll children in programs assigned to a lower rating. We note that reduced enrollment could potentially reflect a center's efforts to intentionally reduce scale, either to improve quality or in response to the lower state subsidy rate associated with a lower star rating. However, the lagged response of enrollment to a lower rating (i.e., several years) seems more in line with parents' enrollment decisions than the more immediate response we might expect from centers that were assigned a lower subsidy rate.

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 lag relative to program ratings. First, parents may be somewhat unwilling to transfer children who are already enrolled. 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 check of internal validity, we estimate our RD model on two measures of attrition from the sample. Specifically, we examine the proportion of programs that closed, and the proportion that remained open but opted out of the ERS rating process. The threat here is that if providers above and below the RD threshold are differentially likely to disappear from the sample, our estimates may be biased. First, we estimate the effect of receiving an ERS rating below 4.5 on rates of closure (Table A3[15]). We find no evidence that programs on different sides of the threshold were differentially likely to close in any year. This finding strongly suggests that program closure does not constitute an empirically relevant internal-validity threat.

As a second measure of attrition, we examine the proportion of programs that remained open, but chose not to receive an ERS rating (Table A4). Although ERS ratings are provided for free, and cannot lower a program's overall star rating, programs may decide that they prefer no public ERS rating to a low rating. We find weakly significant evidence that programs receiving an ERS rating below 4.5 were indeed more likely to opt out. Importantly, this source of attrition does not affect the analysis of future star ratings or program enrollment, because those outcomes are defined for *all* open programs in our ITT sample (i.e., *including* those that opted out

[15] All appendices are available at the end of this article as it appears in JPAM online. Go to the publisher's website and use the search engine to locate the article at http://onlinelibrary.wiley.com.

of ERS assessments). Instead, it suggests that the ERS assessment gains we observe among programs assigned to lower ratings could reflect a combination of some lower-rated programs improving and others opting out in response to the treatment contrast. We return to this finding when discussing the normative and policy-design implications of our results.

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). 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 one, a value chosen by the Imbens and Kalyanaraman (2012) procedure. We also show the results from RD specifications weighted by a triangular kernel. The results of these estimates are quite similar when we also condition on the baseline covariates from Table 3 (results not shown). The consistency of the findings across these specification choices suggests that our findings are not an artifact of functional form or omitted variable biases.

As an additional test of the robustness of our findings, we examined the effect of multiple "placebo" thresholds on our outcomes of interest (results available upon request). Specifically, we estimated the models from Table 4 using alternate thresholds at 5.25, 5.5, 5.75, and 6.0. We find no evidence of an effect of any of these thresholds on either our first-stage outcomes or our outcomes of interest, which lends additional credibility to the causal warrant of our estimates.

Finally, recent work by Calonico, Cattaneo, and Titiunik (2014) argues that standard regression discontinuity procedures may produce biased confidence intervals. In Tables A6 and A7,[16] we replicate results from Tables 4 and 5 and implement their bias-corrected RD estimator. We find that the enrollment results are quite robust to this alternate specification. The ERS results are not significant in the full sample, but still apparent in the high competition sample.

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 is 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 0.26 and 0.27 points, respectively. This effect in T+5 is shown in Figure 6(a). Treated programs in the low-competition sample improved by 0.09 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 about 13 fewer students on average than untreated programs. By contrast, there was not a robust effect on enrollment among programs in the low-competition sample (there *is* a significant positive effect on enrollment in T+2 for this sample but given that this effect is no longer apparent in T+3 through T+5, we believe it is a statistical anomaly). The same pattern holds true when considering the proportion of capacity enrolled. These results are depicted for the high-competition sample in Figures 6(b) and 6(c). 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 influencing program performance.

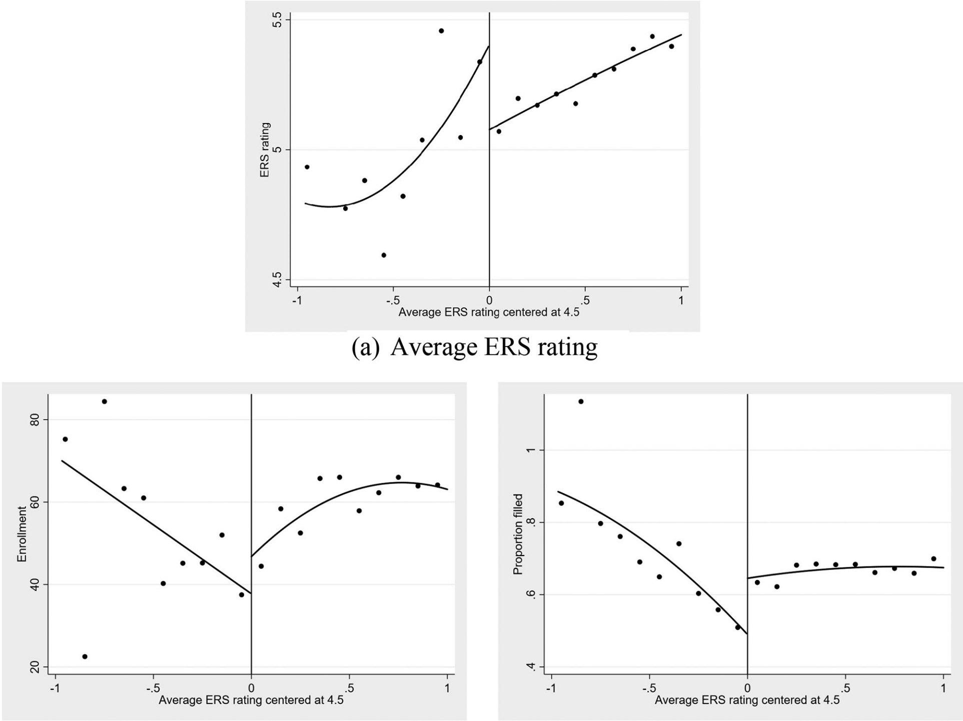
[16] All appendices are available at the end of this article as it appears in JPAM online. Go to the publisher's website and use the search engine to locate the article at http://onlinelibrary.wiley.com.

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	−0.12^{+}	−0.09	−0.12*	−0.02	−0.02	−0.04	−0.02	0.01	−0.00	−0.05
	(0.06)	(0.05)	(0.06)	(0.03)	(0.03)	(0.06)	(0.05)	(0.03)	(0.03)	(0.04)
4+ stars	−0.13	−0.08	0.01	0.01	−0.03	−0.31***	−0.34***	−0.12	−0.13	−0.10
	(0.09)	(0.09)	(0.09)	(0.09)	(0.10)	(0.08)	(0.09)	(0.09)	(0.09)	(0.09)
N	1327	1207	1136	1074	1032	1462	1410	1339	1267	1207
Average ERS rating	0.04	0.13^{+}	0.09	0.09	0.07	0.02	−0.08	0.19	0.26*	0.27**
	(0.05)	(0.08)	(0.13)	(0.14)	(0.15)	(0.06)	(0.08)	(0.14)	(0.12)	(0.10)
N	1301	1170	1069	1006	964	1436	1362	1247	1165	1104
Panel B. Enrollment										
Total enrollment	1.19	6.26*	−0.54	1.67	−1.93	−2.72	−7.55**	−9.70**	−9.23*	−13.41**
	(2.06)	(2.58)	(3.64)	(3.31)	(3.92)	(2.67)	(2.68)	(3.38)	(3.67)	(4.55)
Proportion of capacity filled	0.04	0.13***	0.02	0.06	−0.00	−0.02	−0.07*	−0.10*	−0.10*	−0.15**
	(0.04)	(0.04)	(0.05)	(0.04)	(0.05)	(0.04)	(0.03)	(0.04)	(0.04)	(0.05)
N	1327	1207	1136	1074	1032	1462	1410	1339	1267	1207

Notes: 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 are in parentheses.

$+ p < .10$; $^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$.



(a) Average ERS rating

(b) Total enrollment

(c) Proportion of capacity filled

Figure 6. High Competition Sample Outcomes in T+5.

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 and widely used policy levers 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. 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 ratings cause significant changes in both program quality and program enrollments.

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 cutoff for a star rating led to improvements in child-staff ratios or teacher/administrator credentials. The lack of improvement along these other dimensions is likely to be, at least in part, an artifact of our research design. Specifically, we leverage a treatment contrast in which treated programs stood to improve their overall star 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. Another possibility is that improving structural quality dimensions, such as child-staff ratios, may be more costly than improvements in ERS ratings, making them less amenable to the influence of incentives.

Although these specific incentive contrasts did not drive improvements in measures of quality other than ERS ratings, this does *not* necessarily imply that North Carolina's QRIS had no effect on these other dimensions of quality. For instance, between 2007 and 2014, North Carolina's licensed ECE programs made significant improvements on many of the quality indicators included in North Carolina's QRIS. These improvements may have been driven by aspects of the QRIS apart from the incentive contrasts that we examine here. Put another way, our study does *not* identify the average treatment effect of introducing a QRIS. Instead, our RD design studies the effects of specific incentive contrasts created by North Carolina's QRIS on ECE programs, all of whom are QRIS participants. The aggregate effects of introducing QRIS may differ from the effects of these incentive contrasts. Future studies may be able to leverage differences across states or across regimes to estimate the average treatment effect of a state QRIS on program quality more directly.

In addition to our findings related to quality improvements, we provide evidence of impacts on program enrollment. Here, our findings parallel findings by Hastings and Weinstein (2008), who found that parents responded to information about quality by selectively enrolling their children into higher-quality care. An important caveat is that we are not able to distinguish between supply and demand side effects on enrollment (i.e., are parents making enrollment decisions based on quality, or are providers changing enrollment rates to facilitate quality improvements?). One possibility for distinguishing between these possibilities is 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 four-star rating, which means that almost no Head Start centers fall below the 4.5 ERS threshold.

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.[17] This effect was not sufficiently large or common enough to nullify the performance gains among programs assigned to a lower rating. However, it suggests

[17] 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 & Jacob, 2006; Dee & Wyckoff, 2015) similarly find that incentives can encourage attrition as well as performance gains.

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. For instance, North Carolina could prohibit providers who opt out of ERS ratings from enrolling subsidy-eligible students.

Another key 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-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.

A related external validity issue is that providers that are part of North Carolina's preschool program are required to earn and maintain a minimum of four stars. This means that the RD threshold we consider here is generally not relevant for these programs. As a result, the effects we see on quality and enrollment are driven almost entirely by non-public centers (primarily those that are independently operated). More research is needed to explore the generalizability of our findings to other types of ECE providers including Head Start and state pre-kindergarten, whose ratings put them outside the range that was the focus for our RD analysis, and family child care homes that were not part of the current study but are nonetheless a major actor in the ECE market.

Aside from these generalizability concerns, several study limitations are worth highlighting. 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., Gordon et al., 2013; Perlman, Zellman, & Le, 2004). However, research suggests that overall ERS ratings are more strongly linked to student outcomes than individual subscales or factors (Brunsek et al., 2017), which provides further justification for our use of overall ratings as an outcome of interest. Recent work has also found that different ERS thresholds are more salient across different outcomes, and some outcomes don't lend themselves to any specific thresholds (Le, Schaack, & Setodji, 2015). Despite this, the thresholds examined here allow us to leverage the meaningful incentive contrasts embedded within North Carolina's QRIS.

Finally, Cannon et al. (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 are relevant to the national conversation about the role of accountability in early childhood education. They 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.

DAPHNA BASSOK is an Associate Professor of Education and Public Policy at the University of Virginia, Curry School of Education and Human Development, 405 Emmet Street South, Charlottesville, VA 22904 (e-mail: dbassok@virginia.edu).

THOMAS S. DEE is the Barnett Family Professor of Education at Stanford University, Graduate School of Education, 520 Galvez Mall, CERAS Building, 5th Floor, Stanford, CA 94305-3084 (e-mail: tdee@stanford.edu). He is a Research Associate at the National Bureau of Economic Research.

SCOTT LATHAM is an Associate Research Scholar at Princeton University, Woodrow Wilson School of Public and International Affairs, 183 Wallace Hall, Princeton, NJ 08540 (e-mail: slatham@princeton.edu).

ACKNOWLEDGMENTS

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 Kimberly Boller, Chloe Gibbs, Robert Pianta, Vivian C. Wong, Jim Wyckoff, 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 Institute of Education Sciences (IES) grant R305B090002 to the University of Virginia, and by IES grant R305B130017 to Stanford University.

REFERENCES

Akaike, H. (1974). A new look at the statistical model identification. IEEE Control Systems Society, Transactions on Automatic Control, 19, 716–723.

Araujo, M., Carneiro, P., Cruz-Aguayo, Y., & Schady, N. (2016). Teacher quality and learning outcomes in kindergarten. The Quarterly Journal of Economics, 131, 1415–1453.

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. Available at 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, 87, 1627–1645.

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

Bassok, D., Markowitz, A., Player, D., & Zagardo, M. (2018). Are parents' ratings and satisfaction with preschools related to program features? AERA Open, 4(1).

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, 279–341.

Brunsek, A., Perlman, M., Falenchuk, O., McMullen, E., Fletcher, B., & Shah, P. S. (2017). The relationship between the Early Childhood Environment Rating Scale and its revised form and child outcomes: A systematic review and meta-analysis. PLOS One, 12(6), e0178512.

Burchinal, M. (2018). Measuring early care and education quality. Child Development Perspectives, 12, 3–9.

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

Calonico, S., Cattaneo, M., & Titiunik, R. (2014). Robust nonparametric confidence intervals for regression-discontinuity designs. Econometrica, 82, 2295–2326.

Camerer, C., & Hogarth, R. (1999). The effects of financial incentives in experiments: A review and capital-labor-production framework. Journal of Risk and Uncertainty, 29, 7–42.

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. RAND Corporation. Available at https://www.rand.org/pubs/perspectives/PE235.html.

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.

Coe, C., & Brunet, J. (2006). Organizational report cards: Significant impact or much ado about nothing? Public Administration Review, 66, 90–100.

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. A. (2006). Do high school exit exams influence educational attainment or labor market performance? NBER Working Paper No 12199. Cambridge, MA: National Bureau of Economic Research. Retrieved June 1, 2017, from http://www.nber.org/papers/w12199.

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.

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. Amsterdam, the Netherlands: 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.

Gormley, W. T., & Weimer, D. L. (1999). Organizational report cards. Cambridge, MA: Harvard University Press.

Grant, L. E., & Potoski, M. (2015). Collective reputations affect donations to nonprofits. Journal of Policy Analysis and Management, 34, 835–852.

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. New York, NY: Teachers College Press.

Harms, T., Cryer, D., & Clifford, R. (2003). Infant/Toddler Environment Rating Scale, Revised Edition. New York, NY: Teachers College Press.

Harms, T., Cryer, D., & Clifford, R. (2007). Family Child Care Environment Rating Scale, Revised Edition. New York, NY: Teachers College Press.

Harms, T., Jacobs, E., & White, D. (1996). School-Age Care Environment Rating Scale. New York, NY: Teachers College Press.

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.

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.

Hoxby, C. M. (2003). School choice and school productivity. Could school choice be a tide that lifts all boats? In C. M. Hoxby (Ed.), The Economics of School Choice. Chicago, IL: 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.

Jin, G. Z., & Leslie, P. (2003). The effect of information on product quality: Evidence from restaurant hygiene grade cards. The Quarterly Journal of Economics, 118, 409–451.

Kogan, V., Lavertu, S., & Peskowitz, Z. (2016). Do school report cards produce accountability through the ballot box? Journal of Policy Analysis and Management, 35, 639–661.

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.

Le, V. N., Schaack, D. D., & Setodji, C. M. (2015). Identifying baseline and ceiling thresholds within the Qualistar Early Learning Quality Rating and Improvement System. Early Childhood Research Quarterly, 30, 215–226.

Lee, D. S., & Lemieux, T. (2010). Regression discontinuity designs in economics. Journal of Economic Literature, 48, 281–355.

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.

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

Mukamel, D. B., Haeder, S. F., & Weimer, D. L. (2014). Top-down and bottom-up approaches to health care quality: The impacts of regulation and report cards. Annual Review of Public Health, 35, 477–497.

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, DC: National Academies Press. Retrieved April 1, 2017, from http://www.nap.edu/catalog/12521.

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., Falenchuk, O., Fletcher, B., McMullen, E., Beyene, J., & Shah, P. (2016). A systematic review and meta-analysis of a measure of staff/child interaction quality (the Classroom Assessment Scoring System) in early childhood education and care settings and child outcomes. PLOS One, 11(12), e0167660.

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.

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. Washington, DC: 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? Washington, DC: The National Academies Press.

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. Available at https://eric.ed.gov/?id=ED529579.

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. Available at 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. Foundation for Child Development. Available at https://www.fcd-us.org/the-evidence-base-on-preschool/.

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

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., seven out of a total of 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	Requirement score
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 square feet of inside space and 100 square feet of outside space per child per the licensed capacity, OR
- There must be at least 35 square feet of inside space and 50 square feet of 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 enhanced 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 two or more staff must have written personnel policies including job descriptions, minimum qualifications, health and 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:

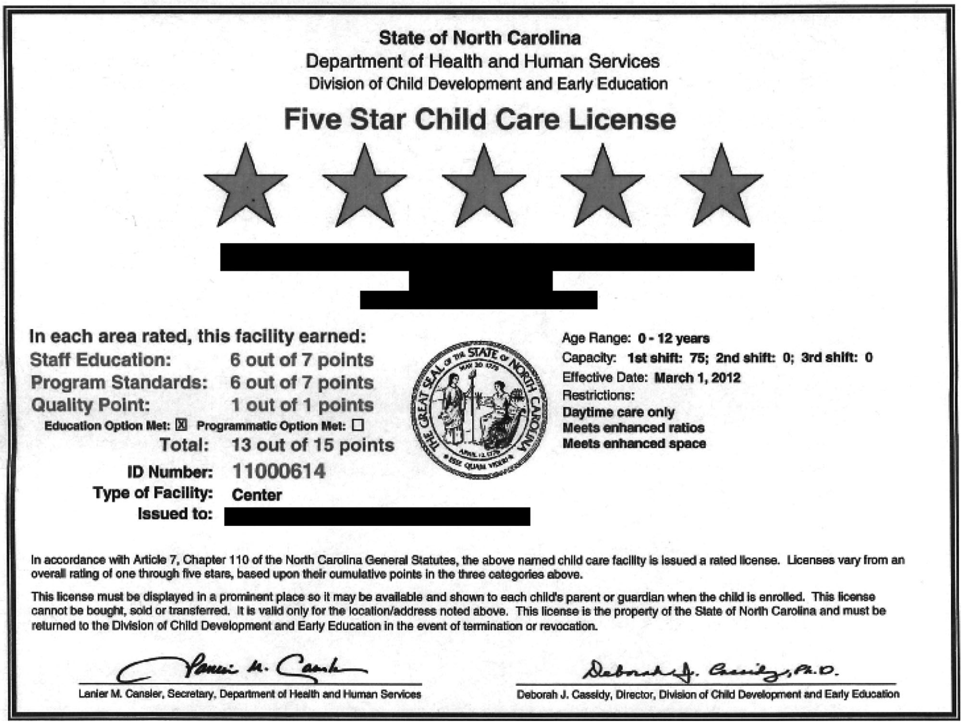
- Must have written policies that describe the operation of the center and services that are available to children/parents, including days/hours of operation, age range of children served, parent fees, etc.
- Operational policies must be discussed with parents when they inquire about enrolling their 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 two years old 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 that 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



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]

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, Secretary, Department of Health and Human Services

Deborah J. Cassidy, Ph.D.

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

Figure A1. Sample Five-Star Rated License.

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

	2007		2008		2009	
Center characteristic	Sample	Non-sample	Sample	Non-sample	Sample	Non-sample
Independent center	0.46	0.50	0.47	0.48	0.47	0.49
Local public school	0.23	0.22	0.23	0.23	0.23	0.22
Head Start	0.10	0.01	0.10	0.01	0.10	0.01
Religious sponsored	0.09	0.20	0.08	0.21	0.08	0.20
3+ star rating	0.92	0.47	0.92	0.40	0.97	0.43
4+ star rating	0.72	0.16	0.76	0.14	0.83	0.14
5-star rating	0.37	0.06	0.41	0.06	0.44	0.06
ERS opt-out	0.46	0.95	0.14	0.92	0.02	0.91
Total enrollment	53.62	45.02	52.83	44.47	52.97	42.01
Proportion of capacity filled	0.74	0.64	0.73	0.62	0.72	0.58
Number of providers within 5 miles	38.75	28.26	41.21	33.64	46.28	41.57
N	2770	2250	2848	2182	2755	2197

Notes: 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 to 2009. Centers were included in our sample if they received an ERS rating during the years 2007 to 2009 and served at least one child from 0 to 5 years old, and they were excluded otherwise. The differences between sample and non-sample centers are significant at the .001 level for each variable in each year, with the exceptions of whether providers were independent centers or operated out of a local public school.

Table A2. RD estimates for early ERS re-rating.

		T+1	T+2	T+3	T+4	T+5
Full sample	Sample mean	0.11	0.21	-	-	-
	RD estimate	0.08+	0.00	-	-	-
		(0.05)	(0.06)	-	-	-
	N	2789	2617	2475	2339	2234
High competition	Sample mean	0.12	0.23	-	-	-
	RD estimate	0.07	−0.09	-	-	-
		(0.07)	(0.08)	-	-	-
	N	1462	1410	1339	1266	1204
Low competition	Sample mean	0.08	0.19	-	-	-
	RD estimate	0.12	0.13	-	-	-
		(0.07)	(0.09)	-	-	-
	N	1327	1207	1136	1073	1030

Notes: 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 are in parentheses.
+ p < .10; *p < .05; **p < .01; ***p < .001.

Table A3. 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	0.21	0.24
	RD estimate	−0.01	−0.04	−0.07	−0.08	−0.05
		(0.03)	(0.04)	(0.05)	(0.05)	(0.06)
	N	2950	2950	2950	2950	2950
High competition	Sample mean	0.02	0.05	0.10	0.15	0.19
	RD estimate	−0.02	−0.04	−0.07	−0.07	−0.03
		(0.02)	(0.03)	(0.05)	(0.07)	(0.08)
	N	1489	1489	1489	1489	1489
Low competition	Sample mean	0.09	0.17	0.22	0.26	0.29
	RD estimate	−0.03	−0.08	−0.10	−0.11	−0.09
		(0.06)	(0.07)	(0.08)	(0.08)	(0.09)
	N	1461	1461	1461	1461	1461

Notes: 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 are in parentheses.
+ p < .10; *p < .05; **p < .01; ***p < .001.

Table A4. RD estimates for ERS opt-outs.

		T+1	T+2	T+3	T+4	T+5
Full sample	Sample mean	-	-	0.06	0.07	0.08
	RD estimate	-	-	0.10^{+}	0.13^{*}	0.13^{+}
		-	-	(0.06)	(0.06)	(0.07)
	N	2789	2617	2475	2341	2239
High competition	Sample mean	-	-	0.07	0.08	0.09
	RD estimate	-	-	0.18^{*}	0.19^{*}	0.13
		-	-	(0.09)	(0.09)	(0.09)
	N	1462	1410	1339	1267	1207
Low competition	Sample mean	-	-	0.06	0.06	0.07
	RD estimate	-	-	0.01	0.06	0.11
		-	-	(0.07)	(0.08)	(0.09)
	N	1327	1207	1136	1074	1032

Notes: 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 are 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	1.5	1.25	1	Triangular kernel
Panel A. Quality						
3+ stars	−0.03	-0.04^{*}	-0.04^{+}	-0.04^{+}	−0.03	−0.03
	(0.03)	(0.02)	(0.02)	(0.02)	(0.03)	(0.03)
4+ stars	−0.07	-0.14^{**}	-0.14^{**}	-0.11^{*}	−0.05	−0.04
	(0.07)	(0.05)	(0.05)	(0.06)	(0.06)	(0.07)
N	2239	2239	2086	1852	1501	1485
Average ERS rating	0.21^{*}	0.16^{*}	0.20^{**}	0.18^{*}	0.17^{*}	0.15^{+}
	(0.08)	(0.07)	(0.07)	(0.07)	(0.08)	(0.08)
N	2068	2068	1924	1698	1361	1347
Panel B. Enrollment						
Total enrollment	-7.95^{**}	-7.41^{***}	-7.62^{**}	-8.56^{***}	-8.23^{**}	-7.88^{**}
	(3.03)	(2.13)	(2.34)	(2.53)	(2.80)	(3.03)
Proportion of capacity filled	-0.08^{*}	−0.03	-0.05^{+}	-0.07^{*}	-0.07^{*}	-0.08^{*}
	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
N	2239	2239	2086	1852	1501	1485

Notes: 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 one if a center score is below the RD threshold. Robust standard errors are in parenthesis. 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.

Table A6. Bias-corrected RD estimates in T+1 through T+5.

Outcome	T+1	T+2	T+3	T+4	T+5
Panel A. Quality					
3+ stars	−0.05 (0.05)	−0.04 (0.04)	0.00 (0.03)	0.02 (0.01)	−0.03 (0.03)
N	2070	1705	1280	1083	1716
4+ stars	-0.17^{*} (0.09)	-0.17^{+} (0.09)	−0.07 (0.08)	−0.05 (0.08)	−0.08 (0.09)
N	2032	2033	1444	1955	1858
Average ERS rating	0.08 (0.06)	0.04 (0.08)	−0.04 (0.11)	−0.01 (0.11)	−0.02 (0.10)
N	1647	1698	1604	1706	1598
Panel B. Enrollment					
Total enrollment	-4.69^{+} (2.48)	-6.62^{*} (2.78)	-10.04^{**} (3.43)	-7.71^{*} (3.08)	-9.27^{*} (3.63)
N	1943	1308	1563	1758	1439
Proportion of capacity filled	−0.03 (0.04)	−0.02 (0.04)	-0.10^{*} (0.04)	-0.09^{*} (0.04)	-0.09^{*} (0.04)
N	2137	2067	1777	1892	1384

Notes: Bias-corrected estimates are implemented following Calonico, Cattaneo, and Titiunik (2014). 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 are in parentheses.
+ p < .10; *p < .05; **p < .01; ***p < .001.

Table A7. Bias-corrected RD estimates in T+1 through T+5, high competition sample.

Outcome	T+1	T+2	T+3	T+4	T+5
Panel A. Quality					
3+ stars	−0.10	−0.11	0.00	0.03	−0.06
	(0.08)	(0.07)	(0.00)	(0.02)	(0.05)
N	1146	922	584	919	787
4+ stars	−0.25*	−0.32**	−0.19+	−0.20+	−0.17
	(0.11)	(0.12)	(0.11)	(0.11)	(0.13)
N	1005	1048	797	1076	957
Average ERS rating	0.05	−0.07	0.02	0.25+	0.22*
	(0.09)	(0.10)	(0.16)	(0.14)	(0.11)
N	790	998	919	900	872
Panel B. Enrollment					
Total enrollment	−7.25*	−10.28**	−9.04*	−9.03*	−15.25**
	(3.40)	(3.17)	(4.59)	(4.21)	(5.41)
N	968	1008	947	836	855
Proportion of capacity filled	−0.08+	−0.08+	−0.12*	−0.11*	−0.16**
	(0.05)	(0.04)	(0.05)	(0.05)	(0.05)
N	1186	853	918	1146	973

Notes: Bias-corrected estimates are implemented following Calonico, Cattaneo, and Titiunik (2014). 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 are in parentheses.
+ p < .10; *p < .05; **p < .01; ***p < .001.