Stanford cepa | Center for Education Policy Analysis

CEPA Working Paper No. 16-07

# Transitional Kindergarten vs. Prekindergarten: A Fuzzy Regression Discontinuity Analysis of Student Literacy Skills

AUTHORS

Christopher Doss

Stanford University

ABSTRACT

A growing body of research provides evidence that quality early childhood experiences can affect a host of life outcomes. Equally well documented is the variation in the quality of prekindergarten programs (pre-K) offered to children. In this study I employ a fuzzy regression discontinuity approach to evaluate the efficacy of Transitional Kindergarten (TK) on student outcomes in a large, urban district in California. Importantly, universal prekindergarten was already established in the city which the district serves, making this study a comparison of different prekindergarten opportunities. TK is a highly regulated, state funded, early education program meant to provide a more developmentally appropriate kindergarten curriculum. This study is a test of whether a more highly regulated and academically oriented pre-K program can provide benefits over a more traditional pre-K approach for young five year olds. I find that students who attended TK outperform their peers on a variety of foundational literacy skills. In addition I find some evidence that the gains are larger for minority children.

**Acknowledgements**: I am grateful to Carla Bryant, Pamela Geisler, Meenoo Yashar, Laura Wentworth, Michelle Maghes, Norma Ming, and all other employees of San Francisco Unified School District who provided contextual details and answered all my questions. I am also grateful to Susanna Loeb, Thomas Dee, and Benjamin York for their guidance and support. I thank the participants of the Stanford Center for Education Policy Analysis seminar and the participants of the Association for Education Finance and Policy conference session for their suggestions. The research reported here was supported in part by the Institute of Education Sciences, U.S. Department of Education, through Grant R305B090016 to Stanford University. The opinions expressed are those of the author and do not necessarily represent views of the Institute or the U.S. Department of Education.

VERSION

September 2016

**Suggested citation:** Doss C. (2016). Transitional Kindergarten vs. Prekindergarten: A Fuzzy Regression Discontinuity Analysis of Student Literacy Skills (CEPA Working Paper No.16-07). Retrieved from Stanford Center for Education Policy Analysis: http://cepa.stanford.edu/wp16-07

# *Transitional Kindergarten vs. Prekindergarten: A Fuzzy Regression Discontinuity Analysis of Student Literacy Skills*

**Christopher Doss**[i]

**Stanford Graduate School of Education**

**September 2016**

Abstract: A growing body of research provides evidence that quality early childhood experiences can affect a host of life outcomes. Equally well documented is the variation in the quality of prekindergarten programs (pre-K) offered to children. In this study I employ a fuzzy regression discontinuity approach to evaluate the efficacy of Transitional Kindergarten (TK) on student outcomes in the San Francisco Unified School District. TK is a highly regulated, state funded, early education program. Importantly, universal prekindergarten was already established in San Francisco, making this study a comparison of prekindergarten opportunities. This study tests whether a more highly regulated pre-K program, situated solely in schools, can provide benefits to young five year olds over a modern, robust universal pre-K market. I find that students who attended TK outperform their peers on a variety of foundational literacy skills. I find some evidence that the gains are larger for minority children.

[i] Center for Education Policy Analysis, 520 Galvez Mall, CERAS Building, Stanford, CA 94305 cdoss@stanford.edu. I am grateful to Carla Bryant, Pamela Geisler, Meenoo Yashar, Laura Wentworth, Michelle Maghes, Norma Ming, and all other employees of San Francisco Unified School District who provided contextual details and answered all my questions. I am also grateful to Susanna Loeb, Thomas Dee, and Benjamin York for their guidance and support. I thank the participants of the Stanford Center for Education Policy Analysis seminar and the participants of the Association for Education Finance and Policy conference session for their suggestions. The research reported here was supported in part by the Institute of Education Sciences, U.S. Department of Education, through Grant R305B090016 to Stanford University. The opinions expressed are those of the author and do not necessarily represent views of the Institute or the U.S. Department of Education.

## 1. Introduction

The importance of providing a high quality early childhood education to young children has become increasingly clear over the past few decades. Researchers have shown that early childhood education programs can lead to short and medium term academic and socio-emotional gains and potentially improved long term outcomes (Deming, 2009; Currie & Thomas, 1995, 2000; Garces, Thomas & Currie, 2002; Gormley et al., 2005; Ludwig & Miller, 2007; Puma et al., 2010; Heckman et al. 2010; Belfield et al., 2006; Campbell et al. 2012). The results of these and other studies have spurred states and localities to invest in prekindergarten (pre-K) programs.

With the proliferation of pre-K services available to families, the conversation has now shifted to identifying the types of programs and pedagogical approaches that are most effective for our youngest students. From a programmatic standpoint, the pre-K sector is currently marked with a dramatic variation in the quality of programs and in the qualifications, compensation, and stability of the teaching staff (Bassok et al., 2013). Low-income and minority families often choose less effective programs, or fewer hours of instruction, leading to weaker academic outcomes (Magnuson et al., 2004; Phillips & Lowenstein, 2011). Pedagogically, researchers and practitioners are debating what level of academic instruction is appropriate for young children, with many pushing back at the increasing academic nature of early childhood education (Bassok & Rorem, 2014; Stipek, 2006; Elkind, & Whitehurst, 2001; Zigler & Bishop, 2006).

The institution of a state-mandated pre-K program in California provides me an opportunity to evaluate a large early childhood education policy while speaking to these pressing issues surrounding modern pre-K programs and markets. In 2010, Governor Schwarzenegger signed the Kindergarten Readiness Act into law in California. Previously, all children who turned five on or before December 2 were eligible for kindergarten. Stakeholders were concerned that the youngest of these children were not ready for the demands of kindergarten (Governor's State

Advisory Council, 2013). Beginning in 2012-2013, the law gradually moved the cutoff date to September 2 and established Transitional Kindergarten (TK) for students who turn five between September 2 and December 2. The state considers TK to be the first year in a two-year kindergarten sequence whose goal is to prepare children for kindergarten (Governor's Advisory Council, 2013). TK is therefore a state-mandated pre-K program for age-eligible children.

It distinguishes itself from other pre-K programs in that it is funded and governed in the same manner as the K-12 system, is situated solely within schools, and is completely free to families. TK is more highly regulated than typical prekindergarten programs and eliminates the variation in types of programs offered to families and reduced the variation in education and compensation of the teaching force. Further, the San Francisco Unified School District (SFUSD) created a curriculum that is academically and structurally a middle ground between pre-K and kindergarten, in keeping with the trend of increasing the academic focus of early childhood programs. Statewide, TK was projected to cost $675 million a year when fully implemented (Legislative Analyst Office, 2012), though a recent expansion will likely increase that amount.

In this study I leverage a fuzzy regression discontinuity (FRD) design to causally evaluate the efficacy of TK in raising student literacy skills in SFUSD. The San Francisco context provides an opportunity to compare the more regulated and academic TK program to traditional programs in a robust pre-K market because in 2004 San Francisco established universal pre-K. A child turning five years old on December 2 can enroll in TK (or choose from any pre-K program in San Francisco), while a child turning five years old on December 3 can only enroll in pre-K programs offered in the city. Bot sets of children enter kindergarten the following year. Figure 1 illustrates this assignment mechanism for the second cohort.

The unique eligibility requirements detailed in Figure 1 also provide the opportunity to address weaknesses in previous birthday RD studies of early childhood programs. Lipsey et al.

(2014) argue that these weaknesses stem from the fact that previous birthday RD studies compare children from different cohorts. This cross cohort comparison may not be capturing an accurate counterfactual and may result in biased estimates if children are subject to different assessment rules. A within cohort comparison is ideal because all children are assessed in the same way and the efficacy of a specific program can be compared with other educational opportunities available to children in the same cohort in the same year. The TK program eligibility requirements allow me to make exactly this type of comparison. The robust nature of the San Francisco universal pre-K market also means that the alternate pre-K experiences available to children are of high quality. Program effectiveness can vary significantly based on the quality of the counterfactual pre-K experiences (Shager et al., 2012; Zhai et al., 2014; Feller et al., 2015), making the results of this comparison especially relevant and timely.

I analyze 6,739 kindergarteners enrolled in SFUSD in the 2013-2014 and 2014-2015 school years. These classes contain the first two TK cohorts in the district. Of the students in the sample, 946 were eligible for TK in the previous year and 335 enrolled in the program. The primary outcomes are the fall kindergarten and fall first grade administrations of the Fountas and Pinnell Benchmark Assessment System (BAS) and the California English Language Development Test (CELDT). The BAS measures pre-literacy skills and the ability to read books of increasing difficulty. The CELDT is given to all Limited English Proficient (LEP) students and measures reading, listening, speaking, and writing. I find that, in the fall of kindergarten, former-TK students outperform their peers on both assessments. Fall first grade results show that the advantages in CELDT remain, but former-TK students do not read more advanced books on the BAS. There is some evidence that the effects are highest for minority children, consistent with the notion that TK reduced the sorting of children to less effective programs.

## 2. Literature Review and the District Context

### *2.1 Prior Early Education Literature*

Researchers have put considerable effort in estimating the effects of specific early childhood interventions. The Perry-Preschool experiment, the Abecedarian study, and studies on the efficacy of Head Start are among the most widely cited prekindergarten studies. The Perry-Preschool and Abecedarian programs are examples of intensive programs that have been found to have large, short to medium term effects on IQ, reading, and math scores (Campbell et al., 2012; Heckman et al., 2010). Head Start is a quintessential example of a large, federally funded program meant to provide services to economically disadvantaged children. Though less intensive than the Perry-Preschool and Abecedarian programs, Head Start has positive effects on language, literacy, and math (Deming, 2009; Currie & Thomas, 1995; Puma et al., 2010).

The establishment of TK fits into a larger trend of state and localities investing in their own pre-K programs as a response to this encouraging evidence. Causal evaluations of state and local programs in Oklahoma (Gormely et al., 2005), Tennessee (Lipsey et al., 2013), Boston (Weiland & Yoshikawa, 2013), North Carolina (Ladd et al., 2015), and five other states (Wong et al., 2008) have shown short-term benefits in academic outcomes. Wong et al. note that there is considerable variation in program effectiveness, making continued causal evaluations important.

Recent scholarship has posited that this variation in effectiveness can be explained, in part, by variation in the counterfactual. As pre-K markets expand, more families may choose to enroll their children in some form of early childhood education. Programs such as Head Start may seem less effective because the control group is receiving more services. In support of this hypothesis, studies have found that the benefits of Head Start are concentrated on students who, in the counterfactual, do not experience center care (Shager et al., 2012; Zhai et al., 2014; Feller et al., 2015). The counterfactual in this study is similarly mixed, however, San Francisco has an

unusually robust pre-K market that has been in existence for over a decade. This study tests whether school-based programs provide benefits over a mature, modern pre-K market where children have access to quality alternative care.

The school-based nature of TK may provide greater benefits because TK falls under the same regulations as the broader K-12 system. Economically disadvantaged and minority families are typically less likely to opt into formal early childhood programs or enroll in less effective programs (Magnuson et al., 2004; Phillips & Lowenstein, 2011). These sorting patterns are also related to academic outcomes (Bassok et al., forthcoming; Lee et al., 1998; Loeb et al., 2004). Layered on this sorting issue is a dramatic variation in the stability, education, and compensation of the teachers, even in the formal early childhood education sector (Bassok et al., 2013).

There is some evidence that addressing these factors can be beneficial for children. Rigby et al. (2007) showed that subsidies are associated with an increase in the quality of care provided to children and an increase in the uptake of center care. Meanwhile, pre-K programs in markets that more highly regulate the early childhood services and its labor market are associated with better outcomes (Hotz & Xiao, 2011; Rigby et al. 2007; Bassok et al., *forthcoming;* Fuller et al., 2004). The free nature of TK and the strict regulation of its labor force represents a new level of regulation of a pre-K program. If, despite the robust universal pre-K market in San Francisco, low-income and minority children still attend prekindergarten programs of relatively lower quality, combatting these selection effects can result in greater outcomes for minority children.

Additionally, the academic underpinnings of TK finds itself relevant to a current debate in the literature as to what a developmentally appropriate curriculum looks like for young children. Recent studies have shown that kindergarten is becoming increasingly focused on building academic behavior in reading and math (Bassok & Rorem, 2014). This trend has caused parents, researchers, and practitioners to debate whether we are asking too much of children too soon

(Elkind, & Whitehurst, 2001; Zigler & Bishop, 2006; Stipek, 2006; Hatch, 2002). This study measures the combined effect of higher regulation and a stronger focus on academic skills. However, if an academic curriculum and a structured classroom were too advanced for these children, their academic performance would suffer. Large academic gains would be consistent with notion that this mismatch did not occur to an appreciable extent.

The unique enrollment criteria of TK allow this study to address weaknesses inherent in previous birthday RD evaluations of pre-K programs (Lipsey et al., 2014) because the TK eligibility requirements allows for comparisons of students in the same cohort. As figure 1 illustrates, in year T students born on December 3 must attend pre-K while students born on December 2 have the same exact pre-K opportunities in San Francisco, but also have the option to attend TK. In year T+1 both sets of children attend kindergarten. This is in contrast to many pre-K birthday RD studies which compare students from different cohorts. Students in pre-K in year T (cohort 1) are compared to students who are ineligible for pre-K in year T (cohort 2). In year T+1 cohort 1 will advance to kindergarten while cohort 2 will begin pre-K.

Lipsey et al. argue that comparing different cohorts has many disadvantages. The aim of these evaluations is to estimate the effect of pre-K over the alternative child care arrangements parents would make for the same cohort. Parents of children in cohort 2 are not an accurate counterfactual because they are likely to make different arrangements knowing that their children are eligible for pre-K the next year. Furthermore, a change in the supply of pre-K programs in year T can change pre-K enrollment patterns in cohort 2 in year T+1. This would affect the types of students observed and assessed in the control group. Cohort differences can even complicate the assessment process. Many assessments have different start rules based on age or grade. If the two cohorts start at different points in the assessment the results may be biased.

This study sidesteps these issues. The children are in the same cohort and enter kindergarten at the same time. A changing supply of prekindergarten programs should not affect who enters kindergarten in the following year. All children are concurrently assessed in the same manner, in the same classrooms, with the same rules. One issue that the authors broach, and that I cannot fully resolve, is that only children who are in SFUSD are observed and assessed. Ideally one would identify the sample in the previous year and follow the students so as to ensure that attrition from, or entrance into, the sample do not bias the results. While I cannot take this approach, I have the universe of students in public kindergarten in San Francisco and leverage an extensive set RD checks to ensure the internal validity of the study is not compromised.

Finally, this study is similar in design and focus to an independent study that was concurrently fielded by a contractor and that looked at TK statewide (Manship, K. et al., 2015). The results of their unpublished study are broadly similar to the ones here. This study distinguishes itself from their report in a few ways. The authors sampled districts throughout the state while I use population data for a single diverse urban area. This area, SFUSD, was not included in the report sample. By focusing on the population of students, I have one, geographically consistent counterfactual pre-K condition. Given the great variation in counterfactual pre-K experiences seen in the literature, and their effects on estimates, this makes interpretation of results cleaner. The counterfactual is especially relevant when looking at subgroups because subgroups are likely sorted to different geographical areas with different TK programs and counterfactual pre-K experiences. Having a defined population off which to judge heterogeneity will greatly help in determining if results are larger for minority students, which is consistent with notion that TK mitigated the sorting of low income and minority students to less effective pre-K programs. Further, the report does not include heterogeneity analysis.

*2.2 Prekindergarten vs. Transitional Kindergarten, The District Context*

San Francisco has a voter-approved universal pre-K market that served about 83 percent of the city's four year olds in 2011-2012 (EED, 2012). The city funds an umbrella organization which establishes minimum criteria that all participating pre-K programs must meet. The pre-K market, thus, is regulated to an extent that is not typical in the country. There is evidence that San Francisco's efforts have been successful. In 2013, Applied Survey Research leveraged a regression discontinuity design to evaluate the umbrella organization's programs. They found that the program produced a three-month gain in letter and word recognition, a three- to four-month gain in problem solving and gains in self-regulation (Applied Survey Research, 2013).

This type of regulation is likely to establish a floor with regard to the quality of services provided to children in the city. Even in this regime the opportunity for sorting of children to settings remains. City providers must be licensed by the state; however, providers range from school-based programs, to Head Start, to home-based care. The teachers they employ must have 24 early childhood or child development credits and 16 general education credits, but providers can employ more highly educated teachers. Additionally, there is no minimum compensation for teachers. Programs can attract teachers of varying quality, partially through compensation.

Between 2013 and 2015, 142 of the current 147 programs in the universal pre-K market volunteered to be rated with the Quality Rating and Improvement System (QRIS). QRIS is an increasingly common tool used to measure the quality of pre-K services. Table 1 presents the average QRIS scores for SFUSD pre-K centers, Head Start centers, other center-based care, and home-based care.[1] There are differences in quality across the pre-K sector with the overall rating ranging from 3.35 to 4.1 stars (of 5 stars). This variation may be smaller than expected. Home-based programs, which typically produce weaker outcomes, were rated an average of 3.69 stars.

[1] Averages were calculated by the author. Source data is from First Five, 2015.

Despite the strength of the pre-K programs, variation remains among programs within a sector and in the components of care provided among sectors. Head Start has a comparative advantage in providing health screenings, teacher qualifications, and child interactions. SFUSD centers have a comparative advantage in director qualifications, child/teacher ratios, and program environment. The remaining variation in the market leaves the door open to the sorting of families to programs. The city also provides funding for only 612.5 hours of instruction spread through 175 to 245 days. This amounts to 3.5 to 2.5 hour school days. The organization does not subsidize more time, meaning disadvantaged families may select into fewer hours of instruction.

The highly regulated nature of TK can mitigate many of these lingering selection effects. TK is strictly school-based and completely eliminates the variation in types of programs offered to families. The state requires teachers to hold a bachelor's degree and the same credentials as other elementary school teachers. The district also compensates TK teachers at the same rate as other teachers. This raises the floor of, and reduces the variation in, provider qualifications, education, and compensation. TK is also open to all residents of the city and is completely free. In SFUSD, all eligible families can enroll in a full day early childhood program at no cost. Some variation certainly remains. There is likely variation in quality of TK classrooms across the city and selection to these classrooms is likely correlated with demographic and economic variables. On the balance, these selection effects are likely muted in comparison to the larger pre-K market.

TK further distinguishes itself from pre-K in regards to the structure of the day and the focus of the curriculum. The city offers no set pre-K curriculum, but all providers must align their curriculums to the California Preschool Curriculum Frameworks. Perhaps the best way of illustrating the contrast in programs is to distinguish the key differences between SFUSD's pre-K program, which is part the universal pre-K system, and SFUSD's TK program. Table 1 indicates that the vast majority of available programs are provided by SFUSD and other center-based care.

In comparison to other center-based care, SFUSD performs about as well, or better, on almost all dimensions of QRIS. SFUSD's pre-K curriculum is therefore likely to be a valid approximation of the types of instruction the vast majority of students receive in the universal pre-K system.

Figure 2 compares the key elements of the SFUSD's TK and pre-K programs. The district structures the TK day to mirror that of kindergarten. In pre-K, children start the school day at different times and parents select the number of hours of instruction. In TK all children start the day at the same time and attend for six hours. The district uses a homegrown TK curriculum designed to be the middle ground between their pre-K and kindergarten curriculums. District officials emphasized literacy skills and socio-emotional skills and began to emphasize math skills. In many ways, pre-K represents a student centered and play-based approach while TK represents an academic and structured approach. In pre-K, students are allowed to guide the activities and instruction, no curriculum map or timeline exists, and students are given ample naptime and outdoor time. In TK, naptime is eliminated, outdoor time is limited, and teachers, who stay on a curriculum map and timeline, guide the activities. In both programs each session of whole group instruction lasts no more than 10 minutes, but TK utilizes it more often.

TK also differs from pre-K in the composition of the classroom. TK classrooms contain students of a relatively small age range, which may make it easier for teachers to teach to students' skill level. This advantage is moderated by the fact that there are less adults in the room. Qualified pre-K programs must have a maximum class size of 24 and a child-adult ratio of 8:1. In contrast TK is a modified kindergarten classroom with a maximum class size of 22 children, but only one paraprofessional is available for the first six weeks of class. This makes the overall child-adult ratio significantly larger in TK, though still less than that of kindergarten where there are no paraprofessionals and the maximum class size is still 22 students.

## 3. Data

This study examines the first two cohorts of TK students in SFUSD. The TK program was phased in over three years. In the first year children were eligible for TK if they turned five years old between November 2 and December 2. In the second year, children turning five between October 2 and December 2 were eligible. Enrollment into TK was not mandatory, and families also had all other pre-K opportunities in San Francisco available to them. Children born after December 2 were eligible the same pre-K opportunities in San Francisco, less TK. Children born before November 2 (or October 2 in the second year) enrolled in kindergarten and are not in the study.[2] The structure of the program means that a plausibly exogenous cut point, based solely on birthdate, dictates potentially very different educational experiences for children. Children born around the cutoff should, on average, be similar except for the probability of enrolling in TK. A FRD design can leverage this cut point to estimate the effect of TK on outcomes.

SFUSD provided administrative data on the universe of kindergarten students for the 2013-2014 and 2014-2015 school years. The administrative data included student background characteristics, detailed in Table 2, as well as each student's birthdate. I match kindergarten administrative data to the previous year's TK rosters to identify students who enrolled in TK. I repeat the process with pre-K rosters to identify students who attended pre-K in the district.

The district uses the Fountas and Pinnell Benchmark Assessment System (BAS) to measure literacy skills of every student in TK to third grade. The BAS is a formative assessment tool that has been shown to be a valid assessment of literacy development in children (Fountas and Pinnell,

[2] I can also compare students born on November 1 (October 1 in the second year), and therefore in kindergarten, to students born on November 2 (October 2) and therefore in TK. From a policy standpoint this contrast would be less relevant because TK is meant not meant to replace kindergarten, but to better prepare students for kindergarten. From a methodological standpoint I found significant covariate imbalance across this threshold, undermining the causal warrant of this approach.

2012). In the fall, all teachers are required to assess their children on the foundational skills. In 2013-2014, these skills were: upper- and lower-case letter recognition, letter sounds, initial word sounds, early literacy behaviors, rhyming, blending, 25 high frequency words, 50 high frequency words, and segmenting. If students mastered eight of the ten skills they read books of increasing difficulty. Students started with the easiest books (level A) and after reading with enough accuracy and comprehension they progressed to harder books (levels B-Z).

In 2014-2015, the district made segmenting and the 50 high frequency word skills optional. To advance to the leveled books, students needed to master six of the remaining eight foundational skills. For consistency, the fall kindergarten BAS outcomes in this paper are the eight foundational skills common to both years, the probability of mastering enough skills to move on to the leveled reading assessment, and the probability of reading at least at level A. The test could be administered in either English or Spanish. My main specification includes controls for test language. By first grade almost all children (98 percent) were assessed on their ability to read. The fall first grade results are whether TK students are reading more advanced books.

Because almost half the students in the district are English learners, I assess the effects of TK on the performance of LEP students on the CELDT. Students are identified as LEP if the family indicates they speak a language other than English in the home. Any student who is identified as LEP is required to take the CELDT the first year they enter the district and every year until they are reclassified as English proficient. The results of the CELDT are especially consequential for this group of students because reclassification as English proficient depends, in part, on their performance on the test. Students are assessed in listening, speaking, reading, and writing. The listening portion of the exam tests students' ability to follow directions and comprehend oral stories. The speaking section tests students on oral vocabulary, speech, ability to construct stories from pictures, and ability to communicate reasoning skills. The reading section tests many similar

skills as the BAS including the ability to identify letter sounds, pictures associated with words, and parts of a book. In the writing section, students copy letters and words, write words based on pictures, and recognize punctuation and capitalization.

The CELDT compliments the BAS in a few ways. Whereas the BAS is administered by teachers, the CELDT is administered by trained outside assessors. This mitigates any concern that the teachers expect differences in performance from former TK students and grade accordingly. In addition, the CELDT outcomes are expressed in traditional scale scores, which lends itself to a traditional interpretation of the estimates. Finally, because both assessments test many of the same skills, similar results reinforce our confidence in the estimates.

One caveat to the kindergarten results is that that TK students were exposed to the CELDT and BAS in their TK year (the year prior to K) while students in pre-K were not. The district uses the BAS as a formative assessment tool in TK and the state requires that all entering LEP TK students are assessed on the CELDT. The fall kindergarten results therefore contain any true learning in TK as well as any practice effects of having taken the test before. In the fall of first grade all students were exposed to the assessments, thereby eliminating any practice effects.

Across the two years 8,717 kindergarten students matched to the fall kindergarten administrations of the BAS. Teachers varied in the extent to which they followed district assessment guidelines in administering the BAS. Many students were missing individual skills scores and some teachers assessed the child's reading level if they were close to mastering the required number of skills. The final analytical sample consists of 6,739 out of the original 8,717 students. These students had scores for all skills except rhyming and blending. The missing data was largest for those two domains and the sample sizes are smaller. If the missing data is not the

same for students around the birthday threshold, comparisons of outcomes may be biased. Table A1 shows that missing scores are not related to the birthday threshold, making bias unlikely.[3]

Of the 6,739 students in the analytical sample, 3,310 are LEP and were tested with the CELDT in the fall of kindergarten, 6,219 continued to first grade and were assessed in the fall with the BAS, and 2,663 LEP students progressed to first grade and were assessed. Again the results for the LEP and first grade samples would be biased if the probability of being in those samples is discontinuous across the threshold. Table A1 indicates that this is not the case.[4]

Table 2 presents the descriptive statistics for the analytical sample, former TK students, and students who did not attend TK. The students are mostly Asian (31.1 percent) and Hispanic (25.0 percent), with fewer whites (16.5 percent). African Americans (6.3 percent) make up a small part of the sample and are contained in the other category (17.5 percent). Special education students compose 7.6 percent of the sample, while 49.1 percent has been classified as LEP.

The alternative pre-K experience of students who did not attend TK or the district's pre-K program is not fully known. However, 16.9 percent of the analytical sample attended SFUSD pre-K and 5 percent attended TK. In total, 22 percent of the sample was enrolled in the district in the prior year. Table 1 indicates that the vast majority of programs in the universal pre-K market are center-based. SFUSD centers compose 22 percent of that sample (containing 142 of the 147 programs), Head Start centers compose 12 percent, and the remaining 57 percent are composed of other center-based care. With only 9 percent of programs situated in the home, the vast majority of universal pre-K participants experience some sort of center care.

Compared to the former pre-K students, former TK students differ in some important ways. Due to the eligibility criteria, they are mechanically older. TK students were also more likely to

[3] Furthermore, results are robust to including all students in the sample.
[4] In the analytical sample only 1 student who was designated LEP in kindergarten was reclassified in first grade

be minority and LEP and less likely to be special education. Overall TK students, on average, significantly outperformed non-TK students in all administrations of the assessments.

## 4 Empirical Strategy

### *4.1 Identification Strategy*

The differences in age and background characteristics between former TK students and their kindergarten peers make clear the need for quasi-experimental techniques such as a FRD approach. One challenge in working with the BAS foundational skills is the left skewed nature of the distribution. In the fall assessment 6.5 percent to 48.5 percent of the sample achieved the highest score on the foundational skills. The non-normal distribution of the outcomes make OLS inappropriate.[5] I therefore backwards code each skill so that I have a count of how many items a student missed and treat each variable as a count variable. I can then use a family of parametric regressions based on the poisson distribution that include poisson regression, negative binomial regression, and their zero-inflated versions. I present estimates from negative binomial models.[6]

When analyzing the ability of students to read books of increasing difficulty, I use ordinal logit models due to the ordinal nature of the book levels. In addition I present linear probability models of the probability of reading at levels C, E, and I or above. I choose these levels because they represent approximately the 20th, 50th, and 80th percentiles of the sample's distribution in the fall of first grade. This strategy allows me to present an overall measure of a group's ability to read books of increasing difficulty, as well as probe points in the distribution for effects.

[5] All inferences are consistent when using OLS models.

[6] In choosing from among the models I follow Long and Freese (2014) and compare the Akaike Information Criterion (AIC), the Bayesian Information Criteria (BIC) and the Vuong statistic (1989) via Stata's - countfit- command. In all cases the negative binomial model was preferred to poisson model and the zero inflated negative binomial model was preferred to negative binomial model. I choose the negative binomial model because it is more easily interpretable. All inferences are consistent when using the zero-inflated negative binomial models.

Equations (1) and (2) model my fuzzy regression discontinuity approach:

$$TK_{ict} = \beta_0 + \beta_1 \mathbf{1}\{B_{ict} \geq 0\} + \beta_2 f(B_{ict}) + \boldsymbol{X_{ict}}\beta_3 + \delta_{at} + \varepsilon_{ict} \quad (1)$$

$$Y_{ict} = \gamma_0 + \gamma_1 \mathbf{1}\{B_{ict} \geq 0\} + \gamma_2 f(B_{ict}) + \boldsymbol{X_{ict}}\gamma_3 + \ \delta_{at} + \epsilon_{ict} \quad (2)$$

Equation (1) regresses $TK_{ict}$, an indicator for whether student, *i*, in classroom, *c*, in year, *t*, enrolled in TK in the previous year, on the following: an indicator for TK eligibility in the previous year, a flexible polynomial, *f*, of the rating birthday rating variable, $B_{ict}$,, a vector of student characteristics, $\boldsymbol{X_{ict}}$, and assessor-by-year fixed effects, $\delta_{at.}$[7] The rating variable, $B_{ict}$, is the distance, in days, a child is born from December 2. Following Lee and Lemieux (2008), I cluster standard errors on the rating variable because it may be considered a coarse rating variable. The coefficient of interest is $\beta_1$, the TK eligibility requirement compliance rate.

Equation (2) presents reduced form intent-to-treat (ITT) estimates of the effect of being eligible for TK on student outcomes. $Y_{ict}$ is now the literacy outcomes of the child. $\gamma_1$ in equation (2) is the coefficient of interest and represents the ITT estimate of being TK-eligible on student literacy outcomes. In both equations the vector $\boldsymbol{X_{ict}}$ includes all student characteristic variables in Table 2 and an indicator for kindergarten year. For the BAS outcome, the assessor-by-year fixed effect would account for differences among teachers in how they assess their students in a given year. I cannot identify CELDT assessors, but one to three assessors were deployed to a school depending on the size of the school. $\delta_{at}$ in these cases are the school-by-year fixed effects. Once again standard errors are clustered on the birthday rating variable.[8] Finally, I leverage Akaike's Information Criterion (AIC) to determine the optimal functional form of *f* (Schochet et al., 2010).

[7] If the TK program causes students sort to higher or lower performing classrooms or schools, the assessor-by-year fixed effects may not be appropriate. My preferred estimates include the fixed effects to account for any stable differences among assessors. To be inclusive, Table 5 presents my main results with and without covariates and fixed effects. Results are robust to both specifications.

[8] In the conditional negative binomial and ordinal logit models standard errors must be clustered on the assessor-by-year fixed effect.

The test indicates a linear spline, which allows the slope to differ across the discontinuity, is optimal in all cases. As a robustness check I present results from many bandwidths.

*4.2 Manipulation of the Threshold*

A key identifying assumption is that the potential outcomes, $Y_{ict,}$ are independent of the treatment assignment, conditional on the forcing variable, $B_{ict}$. That is, the cut point of December 2 threshold is plausibly exogenous such that, students near the threshold are, on average, similar. Any attempt to sort children to either side of the threshold undermines this identification strategy. The first two cohorts of TK students were born two to three years before Governor Schwarzenegger signed the law. Parents were unable to make family planning decisions based on the law. It is possible that the TK program affects enrollment into kindergarten. Figures 3(a) and (b) present visual depictions of the distribution of observations around the threshold. Figure 3(a) shows that there could be a drop in observations in crossing the threshold, however, fluctuations exist throughout the range of the rating variable. I follow McCrary (2008) and test whether a change in the density of observations around the threshold is significant. Figure 3(b) presents the graphical results. I cannot reject the null hypothesis that there is no change in density at the threshold. The point estimate and standard error of the density discontinuity is 0.110 (0.089).[9]

These natural fluctuations are indicative of regular heaping often found in birthday rating variables. Recent work by Barreca et al. (2015) shows that heaping can cause bias in RD point estimates if observations in the heaps are systematically different from other observations. To test for bias they recommend estimating the effects on heaped and non-heaped data separately. As shown in the histogram in Figure 3(a), 15 to 32 students are concentrated on some values of the

[9] To further ensure that the density of observations is continuous across the threshold, I perform the McCrary density test on each baseline covariate. Table A2 shows that the density of observations is continuous for virtually all covariates. Only one is marginally significant, which may occur by chance.

rating variable. In Section 7 I test for bias by eliminating observations in values of the rating variable that contain 15 or more students. My results are robust to eliminating these heaps.

The regression discontinuity technique additionally assumes that nothing that affects the outcomes, except for the probability of enrolling in TK, is discontinuous across the threshold. I partially test this assumption by running RD regressions to see if the covariates are discontinuous around the threshold. Table 3 presents these results for the full sample and with a bandwidth restriction of 60 days and 30 days on either side of the cutoff. The covariates tested are balanced across the threshold. No covariate is consistently unbalanced across all the bandwidths tested.

Finally, to be a valid FRD the December 2 threshold must predict a strong treatment contrast. Figure 4 presents the first stage results graphically. Virtually nobody who was TK-ineligible enrolled in TK. Only 1 child, who was born on December 3, managed to enroll into the program in the two years of the study. For those children born before December 2, the probability of enrollment increases considerably. Table 4 presents statistical estimates of this compliance rate for the full sample, and for the sample that lies in bandwidths of 60 and 30 days. I find a robust compliance rate of about 30 to 33 percent across models.

**5. Main Results**

Students who have previously experienced TK outperformed their peers on the foundational literacy skills. Figure 5 graphically presents the main fall kindergarten BAS results. After aggregating all foundational skills together, the number of items missed drops as one crosses the December 2 threshold. Figure 5(a) indicates that TK-eligible students missed about 8 items less than their peers, or a14 percent decrease from a base of about 56 items missed by TK-ineligible students at the threshold. For the individual skills, drops are present for upper- and lower-case letters, letter sounds, high frequency words, early literacy behaviors, and rhyming. Figure A1 in the appendix illustrates these results. There is also a jump in the probability of mastering enough

skills to be assessed in reading and the probability of reading at level A or above. For LEP students, Figure 5(d) shows a jump in the overall CELDT performance. Figure A2 shows similar jumps for the listening, reading, and writing subtests of the CELDT.

The picture changes somewhat by the fall of first grade. Figure 6 shows the advantage seen in foundational skills does not translate to the ability to read more advanced books. There are small, but insignificant, jumps in the probability of reading at levels C, E, and I or above. However, the advantages in CELDT remain and former-TK students still outperform their peers.

Table 5 presents the results from the statistical models. I report the coefficients for the unconditional FRD results, as well as results from my preferred specification that includes covariates and assessor-by-year fixed effects. Though this specification relies heavily on the validity of the linear functional form, I show in Section 7 that results are robust to a variety of bandwidths.[10] Columns 1 and 2 of panel A show that the intent to treat estimates are significant ($p<0.05$) for nine of the eleven kindergarten BAS outcomes. TK students benefited on all foundational skills but were not more likely to master the requisite number of skills to move on to the reading assessment, nor were they more likely to have been reading at level A or beyond.

The coefficients on the negative binomial models are difficult to interpret. Table 6 therefore presents incidence rate ratios versions of the coefficients in column 2 of Table 5. These estimates are obtained by: $e^{\gamma_1}$. Incidence rate ratios will indicate the rate at which TK-eligible students, on average, miss an outcome compared to TK-ineligible students. Table 6 indicates that TK-eligible students were less likely to miss foundational skills by factors of about 0.91 to 0.72. This translates to a 9 percent to 28 percent decrease in items missed respectively. To make these results more

[10] In an effort to find the optimal bandwidth I also implement the procedure recommended by Imbens and Kalyanaraman (2011). For most literacy outcomes, the procedure recommended bandwidth of about 2-11 days. This highly localized bandwidth only encompasses 2.1 to 7.4 percent of the data. Instead of using this restrictive slice of data I present the results using all observations and show robustness to a variety of bandwidth restrictions.

meaningful I calculate the average number of items missed by students in the control group born within 30 days of the threshold. I multiply the percent decrease in missed items by the control group mean. On average TK students missed nine fewer items, knew about two more upper case letters and letter sounds, and knew one more lower case letter. They could also recognize about 2 more words out of 25. Of the remaining skills, measured on a one to ten scale, TK students performed better by about half of a point. With about a 33 percent compliance rate, the treatment-on-the treated estimates will be roughly three times as big.

Turning our attention to the performance of LEP students in kindergarten, column 4 of panel A in Table 5 indicates that overall students performed 0.176 standard deviations (SD) better on the CELDT exam ($p<0.05$). All subtests except speaking were also significantly better and estimates range from 0.132 SD to 0.221 SD. Overall the CELDT results corroborate the BAS results and indicate that TK students outperformed their peers on most literacy outcomes.

Because TK students entered the district a year earlier and were exposed to the tests, some of the gains could be from practice. The first grade CELDT outcomes seen in Column 4 of Panel B in Table 5 indicate that practice is not likely biasing the results. At this point all LEP students have been assessed at least once and the results remain similar. LEP students still outperform their peers by 0.231SD ($p<0.01$), estimates for the listening subsection are significant at the 1% level, and the speaking and writing estimates are significant at the 10% level.

The results differ for the first grade results of the BAS. Column 2 of panel B of Table 5 indicate that TK students are not reading more difficult books. The coefficient on the ordinal logit is slightly negative and insignificant, while the coefficients on the linear probability models are slightly positive and insignificant. There is robust evidence that TK improved pre-literacy skills, but it did not improve children's reading ability as measured by the BAS.

## 6. Heterogeneity of Results

Aggregate results can be hiding important heterogeneity based on gender, ethnicity and English proficiency status. Despite the regulation of the universal pre-K market, sorting of families to programs of varying quality can remain. TK can mitigate these trends because it is free to families and decreases variation in the credentials, compensation, and the curriculum offered. In this circumstance minority students can particularly benefit from the program.

Columns 1 and 3 of Table 7 indicate that the kindergarten advantages in the BAS are seen in both genders as well as the Asian, Hispanic, LEP, and English proficient subgroups. For brevity, I present intent to treat estimates from my preferred specification for the total number of items missed, the probability of mastering the requisite foundational skills, and the probability of reading at level A or beyond. Looking at the total items missed, all subgroups, except for the white and other subgroups, benefit in the kindergarten administration of the BAS. There is some indication that the Asian subgroup benefitted the most, with the most negative coefficient on the negative binomial portion of the model at -0.381 (or missing 32 percent less items). However I cannot reject the null hypothesis that all coefficients on the four racial subgroups are equal ($\chi_3^2 = 5.54$, p<0.1364). Looking at the probability of mastering the requisite number of foundational skills, only male and Asian students were more likely to move onto the leveled reading assessments. Males were 4.7 percentage points more likely to do so (p<0.10) and Asian students were 12.6 percentage points more likely to do so (p<0.01), and white students were actually 11.6 percentage points less likely to do so. Here I am able to reject the null hypothesis that the effects on the racial subgroups are equal ($\chi_3^2 = 13.71$, p<0.003). Finally males were also 4.6 percentage points more likely to read (p<0.05). They are the only subgroup more likely to reach the leveled reading assessment and read at level A or above.

Little heterogeneity is found in the fall first grade BAS results. Here, no subgroup has an advantage in reading books of increasing difficulty. The only estimate that is significant is the probability of reading level E books or above for English proficient students (9.3 percentage points, p<0.05), though with so many first grade outcomes this may occur by chance.

Table 8 presents subgroup results for the overall CELDT assessment. The white and other subgroup results are not reported due to small sample sizes. Here the female and minority subgroups are driving the results. Column 1 presents the kindergarten results where Hispanic TK-eligible students particularly benefit by 0.356SD (p<0.05) and female TK-eligible students outperform their peers by 0.241SD (p<0.05). It is worth noting that the point estimates on the male and Asian subgroups are also positive and large, but the smaller sample makes it harder to detect a significant effect. I cannot reject the null hypothesis that the male and female effects are equal ($\chi_1^2 = 0.42, p < 0.5181$), nor that the effects on the Asian and Hispanic subgroups are equal ($\chi_1^2 = 1.81, p < 0.1780$). Column 2 of Table 8 indicates that in the fall of first grade the female advantage remains at 0.199SD, though the slightly smaller point estimate results in a 10 percent significance level. The Hispanic effect is now half as large and insignificant, and the Asian subgroup now has a 0.279SD (p<0.01) advantage. The male and Hispanic subgroup point estimates are again relatively large, but imprecisely estimated due to smaller sample sizes. I cannot reject the null hypothesis that the male and female effects are equal ($\chi_1^2 = 0.16\, p < 0.6903$), nor that the Asian and Hispanic effects are equal ($\chi_1^2 = 0.39, p < 0.5340$).

Taken together the data indicate that TK increased the pre-literacy skills of most subgroups, though this did not translate to a higher reading level in first grade. There is some evidence that the Asian subgroup benefitted the most on the BAS and that the white subgroup benefitted the least. The CELDT and BAS results reinforce each other with the Hispanic and Asian subgroups experiencing advantages on both assessments. In SFUSD the Asian subgroup is a socio-

economically diverse community with many immigrants and first generation Americans. These results are consistent with the notion that the regulation of the TK attenuates selection effects that disadvantage minority students.

**7. Robustness Checks**

The results thus far employ the full set of data. While employing the full data maximizes the precision of my estimates, I am relying heavily on the assumption that a linear spline accurately models the relationship between the outcomes and the rating variable. As is standard practice (Schochet et al., 2010), I present evidence that the results are robust to different bandwidths. Figure 7 presents these robustness checks for the main outcomes. Figures A4 through A7 in the appendix present robustness checks for all other results. Each plot presents ITT estimates and their 95 percent confidence intervals for bandwidths from 30 days to 300 days. Figure 7 presents results for the total number of items missed in kindergarten and the overall CELDT scores in kindergarten and first grade. The point estimates are largely stable for bandwidths as small as 30 days, though the significance tends to decrease as the bandwidths get shorter. This is expected because the sample sizes significantly decrease.

I employ a second robustness check by running a series of placebo regression discontinuities. The effects previously seen should occur uniquely at the December 2 threshold. Moving the threshold to any other date should result in null effects. To test this proposition I move the threshold 30, 40, and 50 days on either side of December 2. Table 9 presents the results of this exercise for the total items missed in kindergarten and the overall CELDT results in both grades. The results from the original estimates, found in column 4, disappear in these placebo specifications, and the kindergarten and first grade effects are not present at other thresholds.

The last robustness check builds off by recent work by Barreca et al. (2015) who find that heaping can cause biased estimates if observations in the heaped portions of the data are

systematically different from observations in the non-heaped portion of the data. To investigate this bias they recommend estimating the effects on heaped and non-heaped data separately. The histogram in Figure 3(a) shows that there could be heaping in the birthday variable, with about 15 to 32 students concentrated in some values of the rating variable. These heaps are larger than the sample average of 18.5 students born in a day. I investigate whether this heaping is biasing the point estimates by re-estimating my main results on portions of the data that exclude successively smaller heaps. In Table 10 I present point estimates of the outcomes from portions of the data that exclude heaps with more than 25, 20, 18, and 15 students born on the same day.

The results indicate that heaping induced bias does not seem to be a concern in this study. Eliminating the biggest heaps containing more than 25 or 20 students does little to the point estimates. Point estimates are noticeably larger after heaps containing more than 18 or 15 students are eliminated, but less than half the sample remains. Even in these most restrictive situations the study's inferences remain: there are significant gains for TK-eligible students.

**8. Discussion and Policy Implications**

This paper presents evidence that Transitional Kindergarten produces large pre-literacy gains in students when compared to pre-K programs available to families as part of the San Francisco's universal pre-K program. Despite the causal nature of the study, one issue complicates the inference. The district uses the BAS as a formative assessment tool from TK to 3rd grade. If other pre-K programs in the city do not use the assessment, TK students were exposed to the BAS up to three times in the previous year. Similarly because LEP students are assessed every year they are in the district, TK students were exposed to the CELDT a year before non-TK students. The fall results may be biased due to a practice effect. The first grade CELDT results indicate that this practice effect is not likely an issue. In first grade all LEP students have been assessed with the

CELDT at least once and the advantages remain. These pre-literacy advantages, however, did not translate to the ability to read more difficult books.

In addition to providing evidence regarding the efficacy of a new and expensive state-mandated pre-K program, the unique enrollment criteria and structure of TK speak to many of the current issues in early childhood education. For example, the greater regulation that resulted from folding TK into the larger K-12 system could account for some of these gains. This regulation likely decreased the variation in the quality of programs offered, and the types of teachers available, to students. Parents who could only afford the subsidized half-day pre-K programs offered by the city can enroll their children in a free, full day program, staffed with teachers as well qualified and compensated as elementary school teachers. These features of TK may explain why the program provided benefits over alternative high-quality pre-K options -- a phenomenon not seen in other studies (Shager et al., 2012; Zhai et al., 2014; Feller et al., 2015).

Prior literature has also shown that minority and socio-economically disadvantaged families often select into less formal prekindergarten or lower quality prekindergarten experiences (Magnuson et al., 2004; Magnuson & Waldfogel, 2005; Phillips & Lowenstein, 2011; Capizzano, 2006). If TK provides these families with larger amounts of higher quality instruction, we would expect them to particularly benefit from this program. This study presents evidence that the Asian subgroup saw the greatest benefits in the BAS, while the white subgroup saw the least benefits. Furthermore the Asian and Hispanic subgroups saw benefits on both the BAS and CELDT assessments. These results support studies such as Hotz and Xiao (2011) and Rigby et al. (2007) who find that regulated markets lead to improved student outcomes.

Aligning the curriculum to the development of children in this age range may also have provided academic benefits. The district structured their TK classrooms and school days to be similar to those of kindergarteners and the curriculum contained less student-directed learning and

playtime than other pre-K programs. This study provides evidence that a more academically oriented curriculum can lead to increased student learning. Though the other aspects of TK confound any inference we can make along this dimension, if the greater academic demands of the program were detrimental to students, we would expect to see smaller or negative effects of the program. The academic gains seen in this study indicate this is not likely to be the case. Further, the link between student outcomes at a young age and improved longer-term outcomes (Chetty et al., 2011) means that TK students can enjoy significant and positive long-term effects.

The results of this study are somewhat smaller compared to evaluations of pre-K programs in other urban areas. Weiland and Yoshikawa (2013) find literacy effect of 0.45 SD – 0.62 SD in their evaluation of Boston's program and Gormley et al. (2005) find literacy effects of 0.64 SD - 0.79 SD in their evaluation of Tulsa's program. In this study, CELDT estimates and BAS estimates from OLS models are on the order of 0.15 SD – 0.3 SD. Differences in the type of estimates and the control group likely account for some of this discrepancy.

As Lipsey et al. (2014) point out, a shortcoming of previous studies is that students in the control group are part of a younger cohort and have yet to attend pre-K. The "treated students" consists of children who attended pre-K in the previous year and are starting their kindergarten year (cohort 1). The "control" students are those that are starting their pre-K year (cohort 2). This sampling strategy results in a type of "treatment-on-the treated" estimate because it excludes children who did not attend pre-K. In contrast, this study is a within-cohort comparison that includes all children in kindergarten, regardless of their pre-K experience. With a 33 percent take up the TK program, these intent-to-treat estimates will naturally be smaller. Two-stage least squares estimates in this study vary from 0.45 SD – 0.6SD. This order of magnitude is on par with Weiland & Yoshikawa's Boston study and but is still less than Gormley's Tulsa study.

The estimates from Gormley's study could be larger because the alternative pre-K experiences in this study are of high quality. Though I do not know the exact pre-K experience of each individual who did not attend TK, at least 83 percent of 4-year olds attend pre-K in San Francisco where about 91 percent of programs are center-based. The control group received services not typically seen in other studies. Indeed this study is a comparison of different pre-K opportunities with the goal of estimating the benefits of TK above the benefits a robust pre-K market imparts to children. From this perspective, smaller estimates should not be surprising.

A back-of-the-envelope calculation estimates that these literacy benefits may not come at a substantially greater cost. In 2012-2013 San Francisco spent $17.24 million on preschool subsidies, building early childhood education capacity, wages, and curriculum. The program served 3,225 students at a cost of $5,346 per student. The program provides 612.5 hours of instruction for a total cost of $8.73 per student per hour. In 2012-2013 the district spent $9,479 per pupil (California Department of Education, 2012). TK is funded at the same per pupil cost as the rest of the district and provides students with 6 hours of instruction a day for 180 days. As result, I estimate that TK costs SFUSD $8.78 per student per hour, just 5 cents per student per hour more. It is important to note that these calculations do not represent the complete costs of each program because they only include costs associated with the district or universal pre-K program. They do not include important opportunity costs that parents may regain by sending their child to a free, full day TK program. The calculations also likely understate the cost of providing pre-K services in San Francisco because the universal pre-K program provides subsidies only for those families that are financially in need. Nevertheless these calculations indicate the academic gains do not have to come at a significantly higher cost.

The TK program has recently been expanded with the introduction of Extended TK. Starting in 2015-2016, children who turn five after December 2, 2015 and before the end of school

year can either enter TK at the time they turn 5, or start TK at the beginning of the school year (Torlakson, 2015). This study cannot speak to whether extending TK to all four year olds, making it a form of universal pre-K, will benefit children. Offering free pre-K services to all four year olds would likely benefit families. However more scrutiny is needed to determine if the TK curriculums are appropriate for younger children. Like all RD studies, the results are valid only for children near the December 2 cutoff. This is especially pertinent in this case because children of this age develop rapidly in a small amount of time. This study indicates that for students near the December 2 threshold SFUSD's efforts to implement TK have been successful.

## References


Applied Survey Research (2013). Evaluating Preschool for All effectiveness: Research brief. Retrieved, June 26, 2015, first5sf.org/sites/default/files/page-files/Evaluating %20PFA%20Effectiveness%20-%20Research%20Brief.pdf

Barreca, A. I., Lindo, J.M., Waddell, G.R. (2015) Heaping induced bias in regression discontinuity designs. *Economic Inquiry*.

Bassok, D., Fitzpatrick, M., Greenberg, E., & Loeb, S., (forthcoming). Within-and between sector quality differences in early childhood education and care. *Child Development*.

Bassok, D., Fitzpatrick, M., Loeb, S., & Paglayan, A.S. (2013). The early childhood care and education workforce inthe United States: Understanding changes from 1990 through 2010. *Education Finance and Policy*, 8, 581–601.

Bassok, D., & Rorem, A. (2014). Is kindergarten the new first grade? The changing nature of kindergarten. EdPolicyWorks, Retrieved, Sept 14, 2015, from curry.virginia.edu/ uploads/resourceLibrary/20_Bassok_Is_Kindergarten_The_New_First_Grade.pdf

Belfield, C., Nores, M., Barnett, W., & Schweinhart, L. (2006). High/Scope Perry Preschool Program: Cost benefit analysis at age 40. *Journal of Human Resources*, 16, 162-190.

California Department of Education. (2012). Cost per average daily attendance. Retrieved, September 14, 2015, from http://www.cde.ca.gov/ds/fd/ec/currentexpense.asp

Campbell, F. A., Pungello, E. P., Burchinal, M., Kainz, K., Pan, Y., Wasik, B. H., Sparling, J. & Ramey, C. T. (2012). Adult outcomes as a function of an early childhood educational program: An Abecedarian Project follow-up. *Developmental Psychology*, 48, 1033-1043

Chetty, R., Friedman, J. N., Hilger, N., Saez, E., Schanzenbach, D. W., & Yagan, D. (2011). How does your kindergarten classroom affect your earnings? Evidence from Project Star. *The Quarterly Journal of Economics*, 126, 1593–1660.

Controller Office. (2013). City services measure performance report. Retrieved, October 7, 2015, from http://sfcontroller.org/Modules/ShowDocument.aspx?documentid=4957

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

Currie, J. & Thomas, D. (2000), School quality and the longer-term effects of Head Start. *The Journal of Human Resources*, 35, 755-774.

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

Early Education Department. (2012). PreK–3rd annual report. Retrieved, June 27, 2015, from sfusd.edu/en/assets/sfusd-staff/programs/files/Early%20Education/PreK-3rd%20Report%20Year%20One_7-18-13.pdf

Elkind, D., & Whitehurst, G. (2001). Young Einsteins: Should Head Start emphasize academics? Education Next, 1. Retrieved, Sept. 14, 2015 from: educationnext.org/young-einsteins/

Feller, A., Grindal, T., Miratrix, L., & Page, L. (2016). Compared to what? Variation in the impacts of early childhood education by alternative care-type settings, Working Paper, Retrieved June 17, 2016 from: papers.ssrn.com/sol3/papers.cfm?abstract_id=2534811

First Five (2015). Quality ratings of early care and education programs in San Francisco through December 7, 2015. Retrieved June 17, 2016 from: first5sf.org/programs/quality ratings-of-early-care-and-education-programs-in-san-francisco/

Fountas and Pinnell (2012). Field study of reliability and validity of the Fountas and Pinnell Benchmark Assessment Systems 1 and 2. Retrieved, July 8, 2015 from http://www.heinemann.com/fountasandpinnell/research/BASFieldStudyFullReport.pdf

Fuller, B., Kagan, S. L., Loeb, S., & Chang, Y.-W. (2004). Child care quality: Centers and home settings that serve poor families. *Early Childhood Research Quarterly*, 19, 505–527.

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

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

Governor's State Advisory Council on Early Learning and Care (2013). Transitional Kindergarten implementation guide. Retrieved, April 3, 2015 from http://www.cde.ca.gov/ci/gs/em/documents/tkguide.pdf

Hatch, J. A. (2002). Accountability shove down: Resisting the standards movement in early childhood education. *Phi Delta Kappa*, 83, 457–462.

Heckman, J.J., Seong, H. M., Pinto, R., Savelyev, P.A., Yavitz, A. (2010) Analyzing social experiments as implemented: A reexamination of the evidence from the High Scope Perry Preschool Program. *Quantitative Economics*, 1, 1-46

Hotz, V. J., & Xiao, M. (2011). The impact of regulations on the supply and quality of care in child care markets. *American Economic Review*, 101, 1775–1805.

Imbens, G.W., & Kalyanaraman, K. (2011). Optimal bandwidth choice for the regression discontinuity estimator. *The Review of Economic Studies*, 1-27

Imbens, G. W., & Lemieux, T. (2008). Regression discontinuity designs: A guide to practice. *Journal of Econometrics*, 142, 615-635

Ladd, H.F., Muschkin, C.G., Dodge, K.A. (2014). From birth to school: Early childhood initiatives and third-grade outcomes in North Carolina, *Journal of Policy Analysis and Management*, 33(1), 162-187.

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

Lee, V. E., Loeb, S., & Lubeck, S. (1998). Contextual effects of pre-K classrooms for disadvantaged children on cognitive development. *Child Development*, 69, 479-494.

Legislative Analyst Office. (2012). Preschool and Transitional Kindergarten. Retrieved, September 14, 2015 from: lao.ca.gov/handouts/education/2012/ Preschool_and_Transitional_Kindergarten_41212.pdf

Lipsey, M.W., Hofer, K.G., Dong, N., Farran, D.C., & Bilbrey, C (2013). Evaluation of the Tennessee Voluntary Prekindergarten Program. Research Report. Retrieved June 17, 2016 from: my.vanderbilt.edu/tnprekevaluation/files/2013/10/ August2013_PRI_Kand1stFollowup_TN-VPK_RCT_ProjectResults_FullReport1.pdf

Lipsey, M.W., Weiland, C., Yoshikawa, H., Wilson, S.J., & Hofer, K.G. (2014). Kindergarten age-cutoff regression-discontinuity design: Methodological issues and implications for application. *Educational Evaluation and Policy Analysis*, 1-18.

Loeb, S., Fuller, B., Kagan, S. L., & Carrol, B. (2004). Child care in poor communities: Early learning effects of type, quality and stability. *Child Development*, 75, 47-65.

Long, J. S. & Freese, J. (2014). Regression models for categorical dependent variables using Stata. College Station, TX: Stata Press

Ludgwig, J., Miller, D.L. (2007). Does Head Start improve children's life chances? Evidence from a regression discontinuity design. *The Quarterly Journal of Economics*, 159-208.

Manship, K., Quick, H, Holod, A., Mill, N., Ogut, B., Chernoff, J., Blum, J., Hauser, A., Anthony, J., Gonzalez R. (2015). Impact of California's Transitional Kindergarten program, 2013-2014. AIR. Retrieved, December 29, 2015 from: air.org/sites/default/files/downloads/report/Impact-of-Californias-Transitional-Kindergarten-Program-Dec-15.pdf

Magnuson, K.A, Meyers, M.K., Ruhm, C.J., Waldfogel, J. (2004). Inequality in preschool education and school readiness. *American Education Research Journal*, 41, 115-157.

McCrary, J. (2008). Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of Econometrics*, 142(2), 698-714.

Phillips, D.A, Lowenstein, A.E, (2011). Early care, education, and child development. *Annual Review of Psychology*, 62, 483-500.

Puma, M. et al. (2010). Head start impact study: final report. U.S. Department of Health and HumanServices. Administration for Children and Families Retrieved, July 2, 2015 from http://www.acf.hhs.gov/sites/default/files/opre/hs_impact_study_final.pdf

Rigby, E., Ryan, R.M., Brooks-Gunn, J. (2007). Child care quality in different state policy contexts. *Journal of Policy Analysis and Management*, 26, 887-907.

Schochet, P., Cook, T., Deke, J., Imbens, G., Lockwood, J. R., Porter, J., & Smith, J. (2010). Standards for regression discontinuity designs. What Works Clearinghouse.

Shager, H.M., Schindler, H., Magnuson, K.A., Duncan, G., Yoshikawa, H., Hart, C.M.D. (2012). Can research design explain variation in Head Start research results? A meta-analysis of cognitive and achievement outcomes. *Educational Evaluation and Policy Analysis*.

Stipek, D. (2006). No child left behind comes to preschool. *Elementary School Journal*, 106.

Torlakson, T. (2015). Amendement to California education code 48000(c). Retrieved, September 14, 2015 from http://www.cde.ca.gov/nr/el/le/yr15ltr0717.asp

Vuong, Q.H (1989). Likelihood ratio tests for model selection & non-nested hypotheses. *Econometrica*, 57, 307-33.

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

Wong, V.C., Cook, T.D, Barnett, W. S., Jung, K. (2008). An evaluation of five state prekindergarten programs. *Journal of Policy Analysis and Management*, 27, 122-154.

Zhai, F., Brooks-Gunn, J., & Waldfogel, J. (2014). Head Start's impact is contingent on alternative type of care in comparison group. *American Psychological Association*, 50(12), 2572-2586

Zigler, E. F., & Bishop-Josef, S. J. (2006). The cognitive child versus the whole child: Lessons from 40 years of Head Start. Play= Learning: How Play Motivates and Enhances Children's Cognitive and Social Emotional Growth, 15–35.

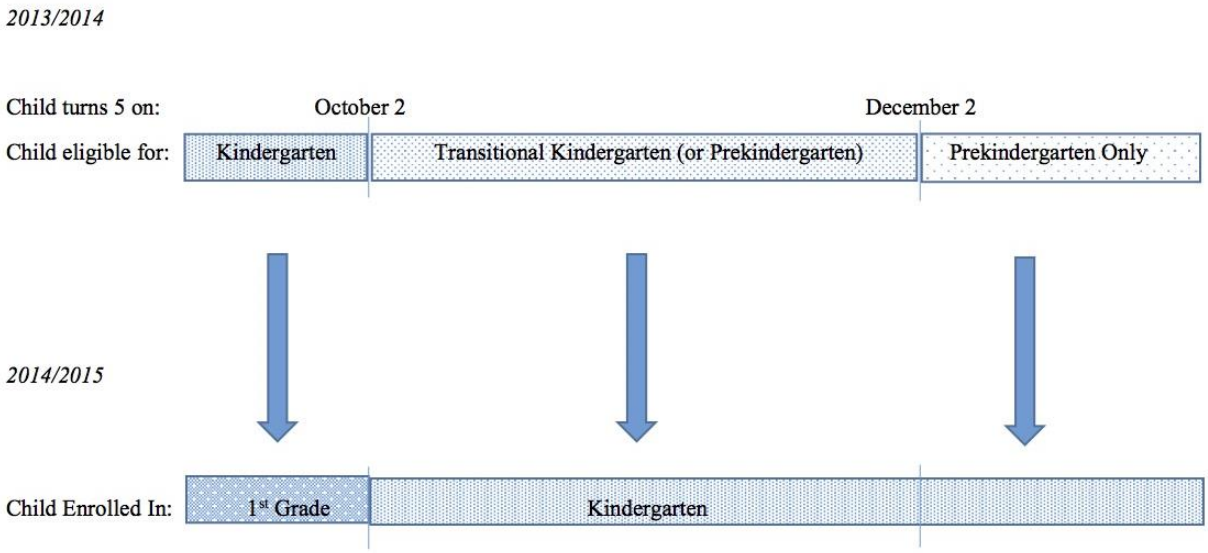
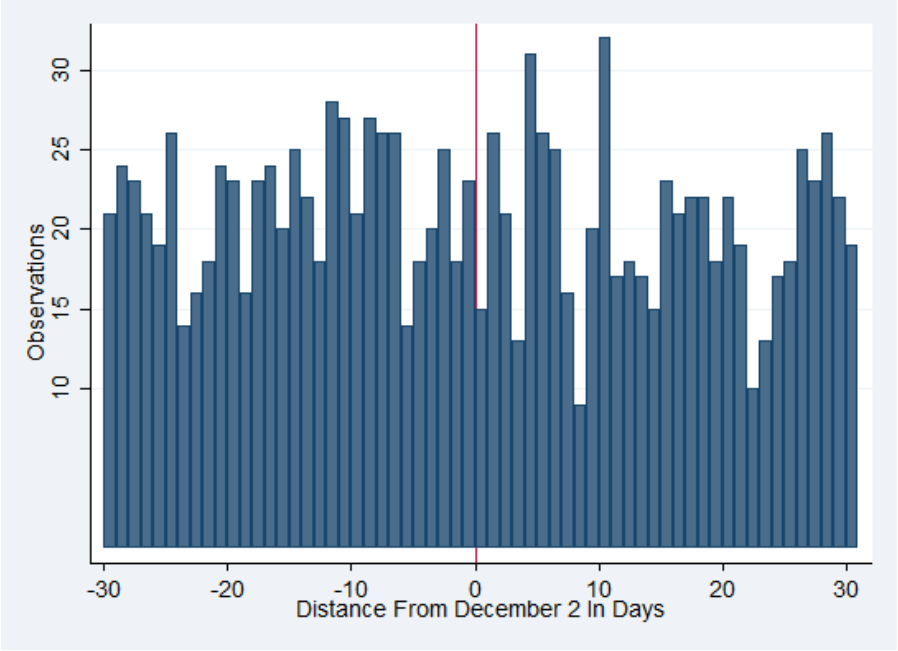


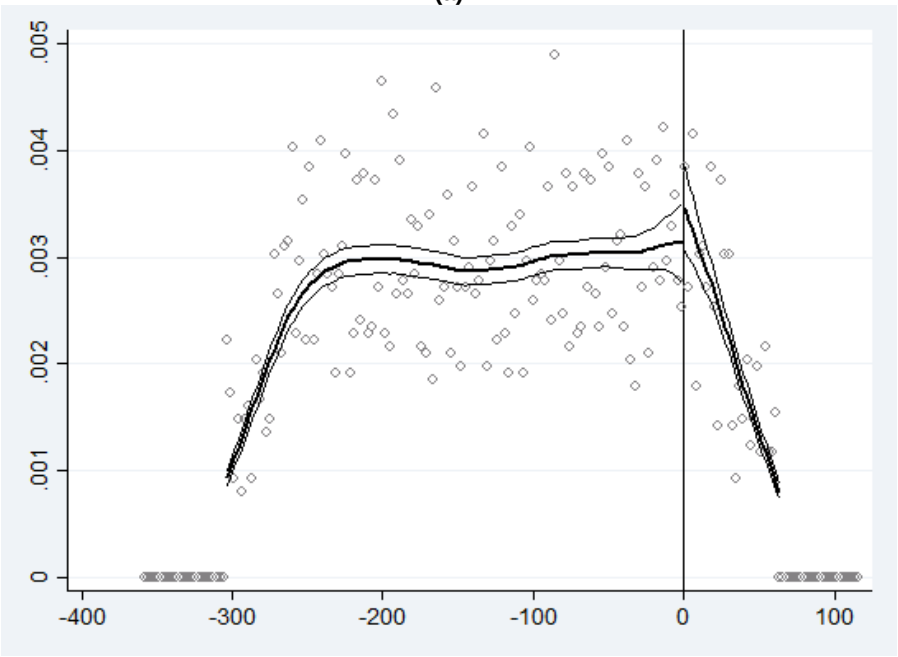

**Figure 1:** Early childhood education experience based on birthdate cut point for cohort 2

| SFUSD Prekindergarten | SFUSD Transitional Kindergarten |
|---|---|
| **Structure of Day** | |
| Children start at different times based on contract | Academic day starts at same time for all children |
| Families select hours of instruction | 6 hour program |
| Breakfast provided | No breakfast but may have morning snack |
| Nap time | No nap time |
| 1 hour of outdoor time | 15-20 minutes of outdoor time |
| **Curriculum** | |
| Activities and pace are based on child's skill | Activities and pace more structured |
| No curriculum map or timeline | Curriculum map and timeline exist |
| Whole group instruction lasts no more than 10 minutes | Whole group instruction lasts no more than 10 minutes |
| Whole group instruction used less frequently | Whole group instruction used more frequently |
| **Class Size** | |
| Maximum class size of 24 students | Maximum class size of 22 students |
| 1 adult for every 8 children | 1 paraprofessional for first 6 weeks |

**Figure 2:** Differences in SFUSD Transitional Kindergarten and prekindergarten programs




(a)




(b)

**Figure 3:** Histogram of observations by birthday and McCrary density test. Birthdays are centered at December 2 such that the x-axis represents the distance in days from December 2. TK ineligible students are to the left of the threshold and TK eligible students are to the right of the threshold. Figure (a) presents birthdays ranging from -30 to 30 days. Each bar indicates the number of observations born in a 1 day bin. Figure (b) presents the results from a McCrary density test. The point estimate and standard error of the discontinuity is 0.110 (0.089). Vertical lines indicate the December 2 threshold.

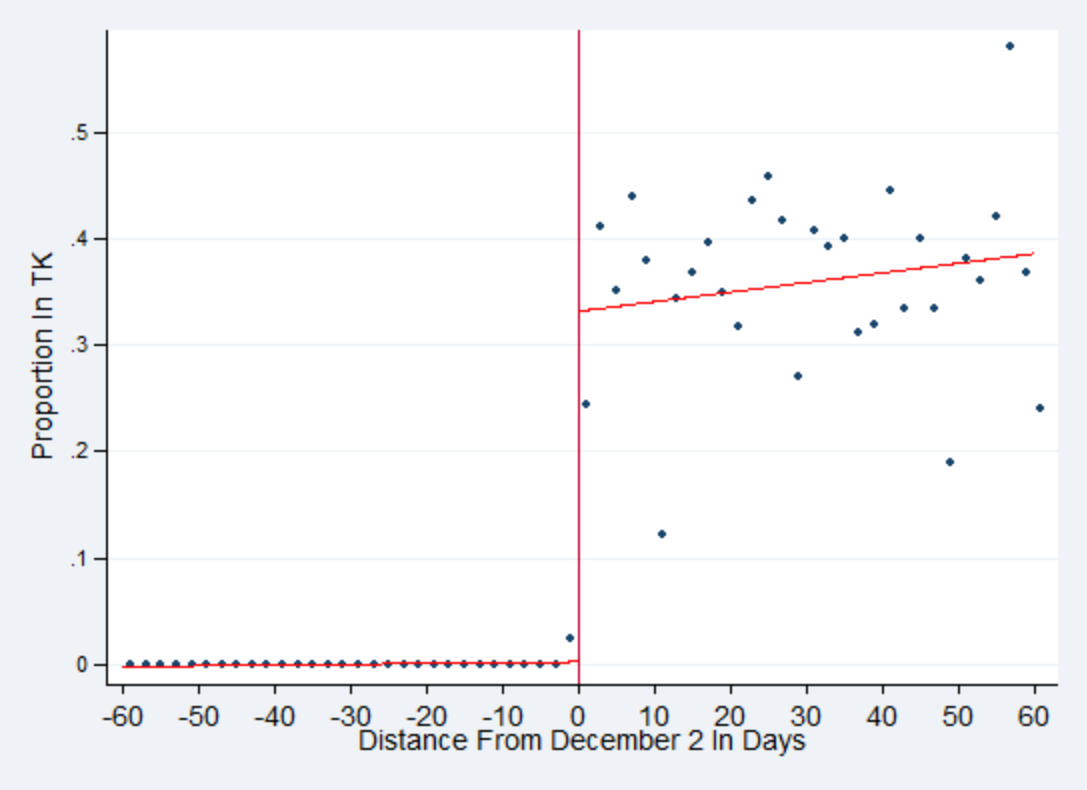
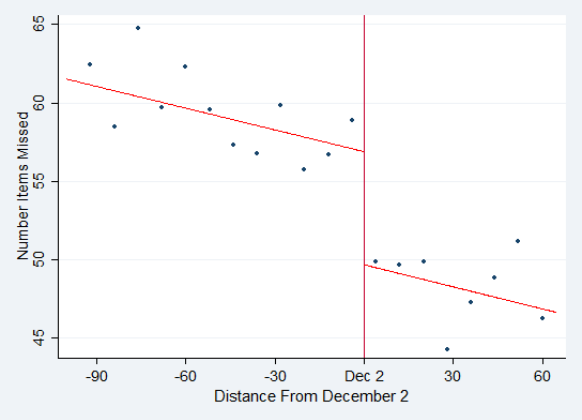
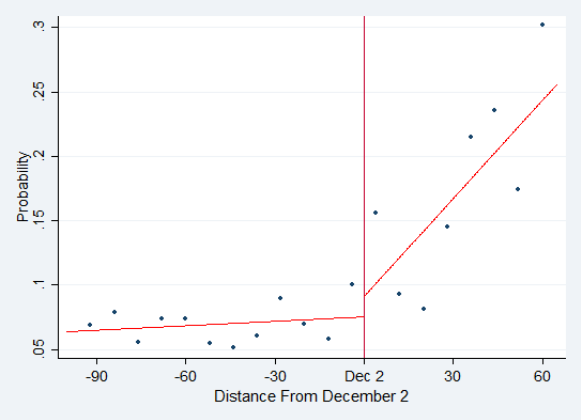


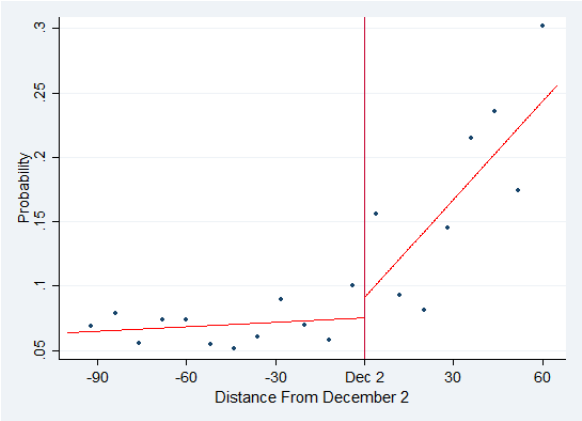

**Figure 4**: First Stage: Enrollment in TK in prior year by birthday. Each dot represents the proportion of students that enrolled in TK in the previous year within a bin of 2 days. The vertical line represents the December 2 threshold. Regression lines are estimated using local linear regression with a rectangular kernel on a bandwith of 60 days.



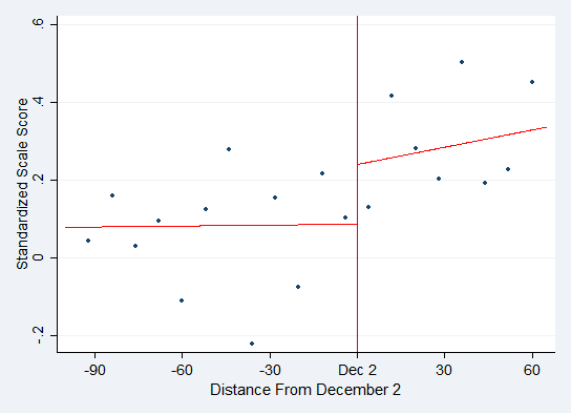

**(a) Total Items Missed**




**(b) Pr(Mastering Enough Foundational Skills)**

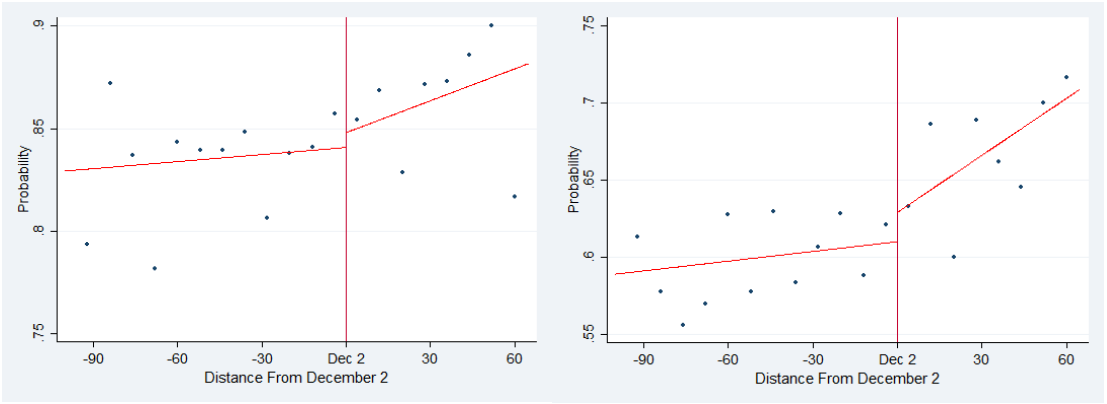



**(c) Pr(Reading At Level A or Above)**

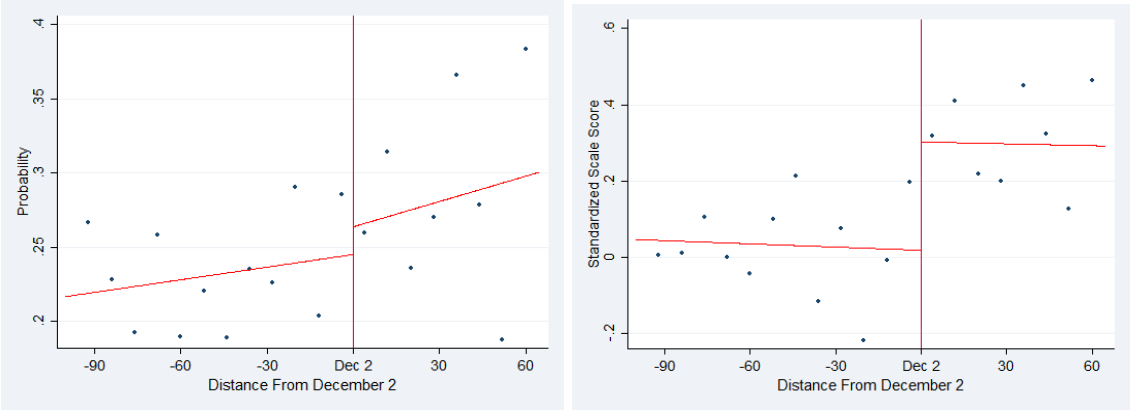



**(d) Overall CELDT Score**

**Figure 5:** Fall kindergarten literacy outcomes. Each dot represents the average outcome in an 8 day bin width. TK eligible students are to the right of the vertical line and TK ineligible students are to the left of the line. The x-axis represents distance of birthday in days from December 2. Birthdays are centered at December 2. The total items missed is the sum of items missed in the following skills: upper case letter recognition, lower case letter recognition, letter sounds, initial word sounds, high frequency words, early literacy behaviors, blending, and rhyming. Figures (a) – (c) are Fountas and Pinnell Outcomes. Figure (d) is the overall CELDT score.

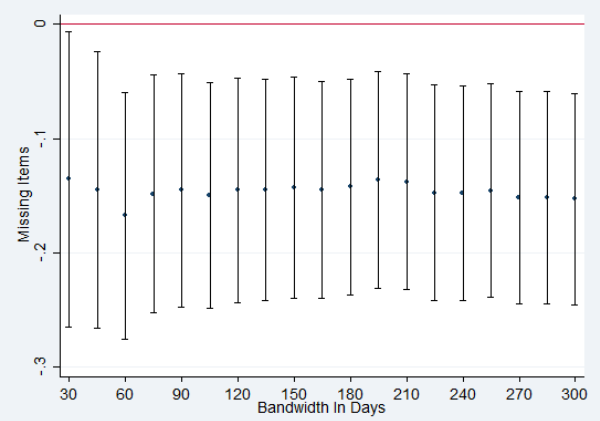



**(a) Pr(Reading at Level C or Above)** **(b) Pr(Reading At Level E or Above)**

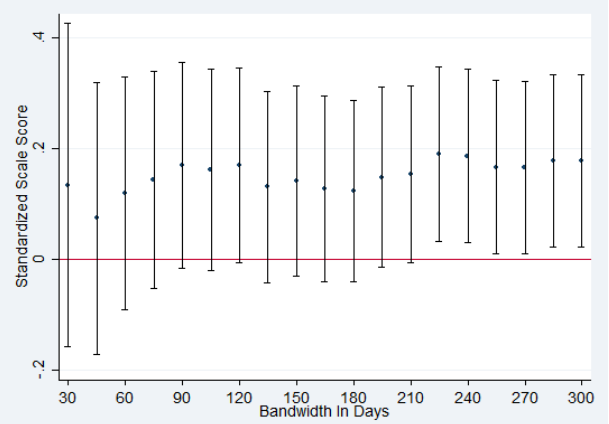



**(c) Pr(Reading at Level I or Above)** **(d) Overall CELDT Score**

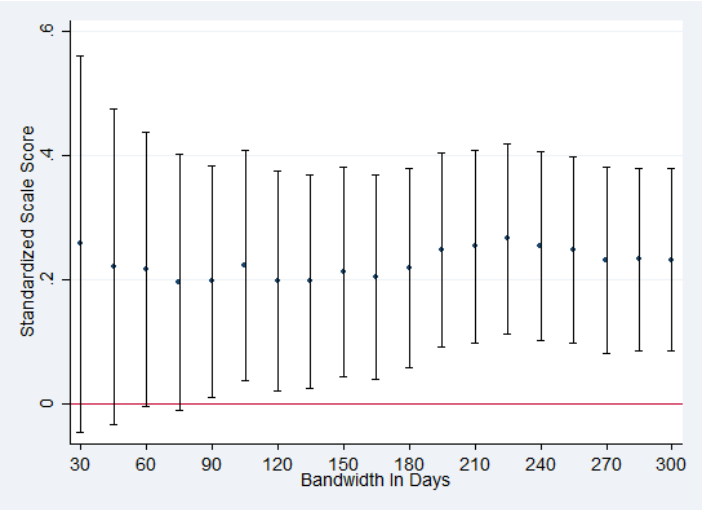
**Figure 6:** Fall first grade literacy outcomes. Each dot represents the average outcome in an 8 day bin width. TK eligible students are to the right of the vertical line and TK ineligible students are to the left of the line. The x-axis represents distance of birthday in days from December 2. Birthdays are centered at December 2. Figures (a) - (c) are Fountas and Pinnell outcomes. Figure (d) is the overall CELDT score.




**(a) Total Items Missed In Fall Kindergarten BAS**




**(b) Overall Fall Kindergarten CELDT Score**




**(c) Overall Fall First Grade CELDT Score**

**Figure 7:** Robustness checks of main BAS and CELDT outcomes. Each dot represents a regression discontinuity estimate of the effect of Transitional Kindergarten on the relevant outcome for observations in bandwidths between 30 and 300 days. Figure (a) employs a negative binomial model and represents the total items missed from the fall kindergarten administration of the BAS. The total items missed is the sum of items missed in the following skills: upper case letter recognition, lower case letter recognition, letter sounds, initial word sounds, high frequency words, early literacy behaviors, blending, and rhyming. Figures (b) and (c) employ OLS models and present results from fall kindergarten and first grade administration of the CELDT. Dots represent point estimates and vertical lines represent the 95 percent confidence inteval. All regressions employ a linear spline functional form with covariates detailed in Table 5. Standard errors are clustered on the birthday rating variable except in (a) where it must be clustered at the teacher-by-year cell.

Table 1: San Francisco universal pre-K Quality Rating and Improvement System results by sector

| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | N(Centers) |
|---|---|---|---|---|---|---|---|---|---|---|
| | Child Observation | Developmental & Health Screening | Minimum Qualifications of Lead Teacher | Child Interactions as Measured by CLASS | Ratio and Group Size | Program Environment Rating Scale | Director Qualifications | Total Points | Star Level | |
| SFUSD School-Based Centers | 3.32 | 0.42 | 4.03 | 3.29 | 4.45 | 4.45 | 4.90 | 24.87 | 3.35 | 31 |
| Head Start Centers | 4.06 | 5.00 | 4.35 | 3.94 | 4.29 | 3.88 | 3.82 | 29.35 | 4.12 | 17 |
| Other Center Care | 3.11 | 2.54 | 4.07 | 3.43 | 3.96 | 3.91 | 3.86 | 24.81 | 3.47 | 81 |
| Home Based Care | 2.69 | 2.85 | 4.69 | 3.38 | N/A | 4.46 | N/A | 18.08 | 3.69 | 13 |

*Note:* Each cell contains the average rating, calculated by the author, for programs in San Francisco's Universal Prekindergarten which opted to be evaluated on the Quality Rating and Improvement System (QRIS). This sample includes 142 of the 147 pre-K providers in the San Francisco universal pre-K market. These programs were evaluated between 2013 and 2015. Source data is from First Five, 2015.

Table 2: Descriptive Statistics

| | Analytical Sample | | | | | Former TK | | Former Non-TK | | p-value |
|---|---|---|---|---|---|---|---|---|---|---|
| Variable | Mean | St. Dev. | Min | Max | N (Total) | Mean | N | Mean | N | (TK-Non TK) |
| Programmatic Characteristics | | | | | | | | | | |
| TK Eligible | 0.140 | 0.347 | 0 | 1 | 6739 | 0.997 | 335 | 0.096 | 6404 | 0.000 |
| Attended TK In Year T-1 | 0.050 | 0.217 | 0 | 1 | 6739 | 1.000 | 335 | 0.000 | 6404 | --- |
| Attended District PreK in Year T-1 | 0.169 | 0.374 | 0 | 1 | 6739 | 0.000 | 335 | 0.177 | 6404 | 0.000 |
| Birthday (days from December 2) | -120.143 | 98.367 | -304 | 61 | 6739 | 26.188 | 335 | -127.798 | 6404 | 0.000 |
| Student Characteristics | | | | | | | | | | |
| Female | 0.492 | 0.500 | 0 | 1 | 6739 | 0.487 | 335 | 0.492 | 6404 | 0.837 |
| Asian | 0.311 | 0.463 | 0 | 1 | 6739 | 0.421 | 335 | 0.305 | 6404 | 0.000 |
| Hispanic | 0.250 | 0.433 | 0 | 1 | 6739 | 0.260 | 335 | 0.249 | 6404 | 0.666 |
| White | 0.165 | 0.371 | 0 | 1 | 6739 | 0.099 | 335 | 0.168 | 6404 | 0.001 |
| Other | 0.175 | 0.380 | 0 | 1 | 6739 | 0.179 | 335 | 0.175 | 6404 | 0.837 |
| Declined To State Ethnicity | 0.098 | 0.297 | 0 | 1 | 6739 | 0.042 | 335 | 0.101 | 6404 | 0.000 |
| Special Education | 0.076 | 0.265 | 0 | 1 | 6739 | 0.033 | 335 | 0.078 | 6404 | 0.002 |
| Limited English Proficient (LEP) | 0.491 | 0.500 | 0 | 1 | 6739 | 0.594 | 335 | 0.486 | 6404 | 0.000 |
| Home Language: | | | | | | | | | | |
| Chinese | 0.171 | 0.376 | 0 | 1 | 6739 | 0.296 | 335 | 0.164 | 6404 | 0.000 |
| Spanish | 0.149 | 0.356 | 0 | 1 | 6739 | 0.173 | 335 | 0.148 | 6404 | 0.206 |
| English | 0.597 | 0.491 | 0 | 1 | 6739 | 0.457 | 335 | 0.604 | 6404 | 0.000 |
| Other | 0.084 | 0.277 | 0 | 1 | 6739 | 0.075 | 335 | 0.084 | 6404 | 0.539 |
| Dominant Language: | | | | | | | | | | |
| Chinese | 0.206 | 0.404 | 0 | 1 | 6739 | 0.304 | 335 | 0.201 | 6404 | 0.000 |
| Spanish | 0.174 | 0.379 | 0 | 1 | 6739 | 0.182 | 335 | 0.173 | 6404 | 0.675 |
| English | 0.506 | 0.500 | 0 | 1 | 6739 | 0.418 | 335 | 0.511 | 6404 | 0.001 |
| Other | 0.114 | 0.318 | 0 | 1 | 6739 | 0.096 | 335 | 0.115 | 6404 | 0.267 |
| Kindergarten Fountas and Pinnell Outcomes | | | | | | | | | | |
| Upper Case Letters | 20.410 | 8.355 | 0 | 29 | 6739 | 22.499 | 335 | 20.300 | 6404 | 0.000 |
| Lower Case Letters | 18.804 | 8.596 | 0 | 29 | 6739 | 21.857 | 335 | 18.645 | 6404 | 0.000 |
| Letter Sounds | 12.679 | 9.137 | 0 | 29 | 6739 | 17.552 | 335 | 12.424 | 6404 | 0.000 |
| High Frequency Words | 6.912 | 7.815 | 0 | 25 | 6739 | 13.663 | 335 | 6.559 | 6404 | 0.000 |
| Initial Word Sounds | 5.293 | 3.219 | 0 | 8 | 6739 | 6.421 | 335 | 5.234 | 6404 | 0.000 |
| Early Literacy Behaviors | 6.915 | 3.049 | 0 | 11 | 6739 | 8.400 | 335 | 6.837 | 6404 | 0.000 |
| Blending | 6.915 | 3.049 | 0 | 10 | 6427 | 5.792 | 317 | 3.700 | 6110 | 0.000 |
| Rhyming | 6.915 | 3.049 | 0 | 10 | 5997 | 7.260 | 292 | 5.642 | 5705 | 0.000 |
| Mastered Required Found. Skills | 6.915 | 3.049 | 0 | 1 | 6739 | 0.239 | 335 | 0.061 | 6404 | 0.000 |
| Reading at Level A or Above | 6.915 | 3.049 | 0 | 1 | 6739 | 0.224 | 335 | 0.164 | 6404 | 0.004 |
| Test Given In Spanish | 0.140 | 0.347 | 0 | 1 | 6739 | 0.131 | 335 | 0.141 | 6404 | 0.631 |
| Kindergarten CELDT Outcomes | | | | | | | | | | |
| Listening | 374.863 | 86.019 | 220 | 570 | 3310 | 419.422 | 199 | 372.013 | 3111 | 0.000 |
| Speaking | 388.218 | 94.436 | 140 | 630 | 3310 | 428.211 | 199 | 385.659 | 3111 | 0.000 |
| Reading | 294.571 | 57.558 | 220 | 570 | 3310 | 343.297 | 199 | 291.455 | 3111 | 0.000 |
| Writing | 306.521 | 52.327 | 220 | 600 | 3310 | 352.688 | 199 | 303.567 | 3111 | 0.000 |
| Overall | 372.973 | 77.503 | 184 | 580 | 3310 | 415.759 | 199 | 370.236 | 3111 | 0.000 |
| First Grade Fountas and Pinnell Outcomes | | | | | | | | | | |
| Reading at Level C or Above | 0.819 | 0.385 | 0 | 1 | 6219 | 0.870 | 315 | 0.816 | 5904 | 0.016 |
| Reading at Level E or Above | 0.568 | 0.495 | 0 | 1 | 6219 | 0.692 | 315 | 0.562 | 5904 | 0.000 |
| Reading at Level I or Above | 0.211 | 0.408 | 0 | 1 | 6219 | 0.308 | 315 | 0.205 | 5904 | 0.000 |
| First Grade CELDT Outcomes | | | | | | | | | | |
| Listening | 454.807 | 62.608 | 220 | 570 | 2663 | 485.439 | 180 | 452.586 | 2483 | 0.000 |
| Speaking | 457.292 | 65.408 | 140 | 630 | 2663 | 483.778 | 180 | 455.372 | 2483 | 0.000 |
| Reading | 396.753 | 76.247 | 220 | 570 | 2663 | 426.289 | 180 | 394.612 | 2483 | 0.000 |
| Writing | 400.983 | 57.135 | 220 | 600 | 2663 | 430.872 | 180 | 398.816 | 2483 | 0.000 |
| Overall | 449.836 | 56.290 | 184 | 594 | 2663 | 478.500 | 180 | 447.758 | 2483 | 0.000 |

*Note:* Former TK students are students in the analytical sample who enrolled in the district's TK program in the previous year. Former prekindergarten students are students who enrolled in the district's pre-kindergarten program in the previous year. 2013-2014 and 2014-2015 kindergarten administrative data contained student characteristics, including exact birthdate. Administrative data were linked to district test files to obtain Fountas and Pinnell and CELDT outcome data. Students who experienced district TK and prekindergarten were identified by linking kindergarten administrative data to the district TK and pre-K administrative data sets from the previous school year. TK stands for Transitional Kindergarten, pre-K stands for prekindergarten, and CELDT stands for California English Langauge Development Test.

Table 3: RD regressions of covariate balance

| Variable | Full Sample | $\|B_{ict}\| \leq 60$ | $\|B_{ict}\| \leq 30$ |
|---|---|---|---|
| Student Characteristics | | | |
| Female | 0.011 | -0.017 | -0.029 |
| | (0.029) | (0.037) | (0.050) |
| Asian | -0.016 | -0.044 | -0.034 |
| | (0.035) | (0.044) | (0.059) |
| Hispanic | 0.016 | 0.017 | -0.022 |
| | (0.028) | (0.036) | (0.046) |
| White | -0.028 | -0.032 | -0.001 |
| | (0.028) | (0.036) | (0.050) |
| Other | 0.047+ | 0.036 | 0.034 |
| | (0.025) | (0.035) | (0.055) |
| Declined To State Ethnicity | -0.019 | 0.021 | 0.018 |
| | (0.019) | (0.024) | (0.030) |
| Special Education | -0.011 | -0.013 | -0.002 |
| | (0.015) | (0.018) | (0.021) |
| Limited English Proficient (LEP) | -0.029 | -0.057 | -0.078 |
| | (0.038) | (0.047) | (0.066) |
| Home Language: | | | |
| Chinese | -0.000 | -0.018 | -0.036 |
| | (0.030) | (0.034) | (0.047) |
| Spanish | -0.005 | -0.014 | -0.024 |
| | (0.020) | (0.028) | (0.041) |
| English | -0.011 | -0.004 | 0.045 |
| | (0.035) | (0.041) | (0.061) |
| Other | 0.016 | 0.036+ | 0.015 |
| | (0.015) | (0.020) | (0.026) |
| Dominant Language: | | | |
| Chinese | -0.019 | -0.048 | -0.066 |
| | (0.028) | (0.034) | (0.046) |
| Spanish | -0.010 | 0.000 | -0.002 |
| | (0.021) | (0.027) | (0.038) |
| English | 0.029 | 0.049 | 0.072 |
| | (0.037) | (0.046) | (0.065) |
| Other | -0.000 | -0.001 | -0.004 |
| | (0.018) | (0.024) | (0.032) |
| Test Characteristic | | | |
| Test Given In Spanish | -0.026 | -0.012 | 0.027 |
| | (0.026) | (0.033) | (0.045) |
| N | 6,739 | 2,182 | 1,271 |

*Note:* Each cell represents the results of a separate regression discontinuity estimate of the covariate balance. Row headers indicate the appropriate covariate tested. Column headers indicate the bandwidth restriction. In all regressions the functional form is a linear spline. Akaike's Information Criterion indicates a linear spline is the optimal functional form for the majority of covariates. All standard errors are clustered on the day of birth running variable. +indicates p<0.10, *p<0.05, **p<0.01

Table 4: RD regressions of first stage

| *Dependent Variable: Enrolled In TK in Year T-1* | | | |
|---|---|---|---|
| | (1) | (2) | N |
| Full Sample | 0.335** | 0.321** | 6,739 |
| | (0.032) | (0.027) | |
| $\|B_{ict}\| \leq 60$ | 0.329** | 0.309** | 2,182 |
| | (0.032) | (0.031) | |
| $\|B_{ict}\| \leq 30$ | 0.312** | 0.284** | 1,271 |
| | (0.042) | (0.044) | |
| Covariates | | √ | |
| Fixed Effects | | √ | |

*Note:* Each cell represents the results of a separate first stage regression discontinuity estimate. The dependent variable in all regressions is an indicator for enrolling in TK in the previous year. Row headers indicate the bandwidth restriction. Covariates include all variables in Table 2. Covariates also include an indicator for kindergarten year, and teacher-by-year fixed effects. The functional form in all regressions is a linear spline. Akaike's Information Criterion indicates a linear spline is the optimal functional form. All standard errors are clustered on the day of birth running variable. +indicates p<0.10, *p<0.05, **p<0.01

Table 5: Reduced form estimates of fall kindergarten and first grade literacy outcomes

| | (1) | (2) | | | (3) | (4) | |
|---|---|---|---|---|---|---|---|
| Panel A: Fall Kindergarten Outcomes | | | | | | | |
| Fountas And Pinnell Outcomes | | | N | CELDT Outcomes | | | N |
| Total Items Missed | -0.141* | -0.181** | 6,739 | Overall Score | 0.118 | 0.176* | 3,310 |
| | (0.059) | (0.042) | | | (0.110) | (0.079) | |
| Upper Case Letters | -0.289* | -0.332** | 6,739 | Listening | 0.135 | 0.178* | 3,310 |
| | (0.133) | (0.087) | | | (0.105) | (0.080) | |
| Lower Case Letters | -0.229* | -0.163* | 6,739 | Speaking | 0.067 | 0.132+ | 3,310 |
| | (0.103) | (0.068) | | | (0.106) | (0.079) | |
| Letter Sounds | -0.130* | -0.184** | 6,739 | Reading | 0.195* | 0.216* | 3,310 |
| | (0.055) | (0.050) | | | (0.098) | (0.092) | |
| High Frequency Words | -0.099** | -0.141** | 6,739 | Writing | 0.199+ | 0.210** | 3,310 |
| | (0.035) | (0.038) | | | (0.103) | (0.078) | |
| Early Literacy Behaviors | -0.161 | -0.210** | 6,739 | | | | |
| | (0.099) | (0.060) | | | | | |
| Initial Word Sounds | -0.157 | -0.221* | 6,739 | | | | |
| | (0.110) | (0.091) | | | | | |
| Rhyming | -0.164 | -0.191* | 5,997 | | | | |
| | (0.103) | (0.080) | | | | | |
| Blending | -0.033 | -0.098* | 6,427 | | | | |
| | (0.053) | (0.050) | | | | | |
| Pr(Mastering Required Found. Skills) | 0.012 | 0.033 | 6,739 | | | | |
| | (0.022) | (0.021) | | | | | |
| Pr(Reading at Level A or Above) | 0.020 | 0.014 | 6,739 | | | | |
| | (0.028) | (0.016) | | | | | |
| Panel B: Fall First Grade Outcomes | | | | | | | |
| Fountas And Pinnell Outcomes | | | N | CELDT Outcomes | | | N |
| Reading Scale (Ordinal Logit) | -0.051 | -0.036 | 6,219 | Overall Score | 0.250** | 0.231** | 2,663 |
| | (0.120) | (0.120) | | | (0.092) | (0.075) | |
| Pr(Reading at Level C or Above) | 0.007 | 0.008 | 6,219 | Listening | 0.307** | 0.301** | 2,663 |
| | (0.027) | (0.023) | | | (0.087) | (0.079) | |
| Pr(Reading at Level E or Above) | 0.013 | 0.021 | 6,219 | Speaking | 0.145 | 0.128+ | 2,663 |
| | (0.038) | (0.030) | | | (0.093) | (0.076) | |
| Pr(Reading at Level I or Above) | 0.021 | 0.017 | 6,219 | Reading | 0.146 | 0.095 | 2,663 |
| | (0.031) | (0.028) | | | (0.115) | (0.090) | |
| | | | | Writing | 0.234* | 0.172+ | 2,663 |
| | | | | | (0.110) | (0.092) | |
| Covariates | | √ | | | | √ | |
| Fixed Effects | | √ | | | | √ | |

*Note:* Each cell represents the results of a separate regression discontinuity estimate of the effect of Transitional Kindergarten on the indicated literacy outcome. Row headers indicate the dependent variable. Columns 1 and 2 present estimates for Fountas and Pinnell outcomes. Columns 3 and 4 present estimates for CELDT outcomes. Covariates include an indicator for kindergarten year, teacher-by-year fixed effects, and all variables in Table 2. Negative binomial models are used to estimate the effect of Transitional Kindergarten on foundational literacy skills, ordinal logit models are used to estimate the effect of Transitional Kindergarten on literacy skills, and OLS is used in all other models. The functional form of all regressions is a linear spline. Akaike's Information Criteria indicates a linear spline is optimal. All standard errors are clustered on the day of birth running variable except for the conditional negative binomial and ordinal logit models which must be clustered on the teacher-by-year fixed effect. +indicates p<0.10, *p<0.05, **p<0.01

Table 6: Reduced form incidence rate ratio estimates of fall kindergarten literacy outcomes

| Literacy Outcome | (1) Incidence Rate Ratio | (2) Avg Number of Items Missed by Control Group | (3) Fewer Items Missed By TK Students |
|---|---|---|---|
| Total Items Missed | 0.835** | 57.311 | 9.456 |
| Upper Case Letters | 0.718** | 5.792 | 1.633 |
| Lower Case Letters | 0.850* | 7.023 | 1.053 |
| Letter Sounds | 0.832** | 12.92 | 2.171 |
| High Frequency Words | 0.869** | 17.337 | 2.271 |
| Early Literacy Behaviors | 0.811** | 2.705 | 0.511 |
| Initial Word Sounds | 0.802* | 2.311 | 0.458 |
| Rhyming | 0.826* | 4.120 | 0.717 |
| Blending | 0.907* | 5.844 | 0.543 |
| Covariates | √ | √ | √ |
| Fixed Effects | √ | √ | √ |

*Note:* Column 1 presents results of a separate regression discontinuity estimate of the effect of Transitional Kindergarten on the indicated literacy outcome. Row headers indicate the dependent variable. Point estimates in column 1 represents the incidence rate ratios of the point estimates in column 2 of Table 5. Column 3 represents the average number of items missed by the control group born within 30 days of the Transitional Kindergarten threshold. Included covariates are defined in Table 4. +indicates p<0.10, *p<0.05, **p<0.01

Table 7: ITT RD estimates of kindergarten and first grade Fountas and Pinnell outcomes by subgroup

| Kindergarten | | 1st Grade | | Kindergarten | | 1st Grade | |
|---|---|---|---|---|---|---|---|
| | (1) | | (2) | | (3) | | (4) |
| *Panel A: Full Sample, N= 6,739* | | *N=6,219* | | *Panel F: White N=1,111* | | *N=1,001* | |
| Total Items Missed | -0.181** | Reading Scale | -0.036 | Total Items Missed | -0.039 | Reading Scale | -0.122 |
| | (0.042) | | (0.120) | | (0.128) | | (0.331) |
| Pr(Mastering Required Found. Skills) | 0.033 | Pr(Level C or Above) | 0.008 | Pr(Mastering Required Found. Skills) | -0.116* | Pr(Level C or Above) | 0.031 |
| | (0.021) | | (0.023) | | (0.058) | | (0.052) |
| Pr(Reading at Level A or Above) | 0.014 | Pr(Level E or Above) | 0.021 | Pr(Reading at Level A or Above) | -0.033 | Pr(Level E or Above) | 0.039 |
| | (0.016) | | (0.030) | | (0.056) | | (0.089) |
| | | Pr(Level I or Above) | 0.017 | | | Pr(Level I or Above) | 0.151 |
| | | | (0.028) | | | | (0.097) |
| *Panel B: Male, N=3,423* | | *N=3,144* | | *Panel G: Other N=1,179* | | *N=1,068* | |
| Total Items Missed | -0.210** | Reading Scale | -0.136 | Total Items Missed | 0.018 | Reading Scale | -0.136 |
| | (0.060) | | (0.167) | | (0.115) | | (0.280) |
| Pr(Mastering Required Found. Skills) | 0.047+ | Pr(Level C or Above) | 0.018 | Pr(Mastering Required Found. Skills) | -0.038 | Pr(Level C or Above) | 0.055 |
| | (0.027) | | (0.034) | | (0.056) | | (0.072) |
| Pr(Reading at Level A or Above) | 0.046* | Pr(Level E or Above) | -0.021 | Pr(Reading at Level A or Above) | -0.023 | Pr(Level E or Above) | -0.016 |
| | (0.021) | | (0.043) | | (0.044) | | (0.090) |
| | | Pr(Level I or Above) | -0.010 | | | Pr(Level I or Above) | -0.145+ |
| | | | (0.041) | | | | (0.075) |
| *Panel C: Female, N=3,316* | | *N=3,075* | | *Panel H: Limited English Proficient (LEP), N=3,310* | | *N=3,115* | |
| Total Items Missed | -0.164** | Reading Scale | 0.078 | Total Items Missed | -0.166** | Reading Scale | -0.084 |
| | (0.061) | | (0.177) | | (0.056) | | (0.173) |
| Pr(Mastering Required Found. Skills) | 0.023 | Pr(Level C or Above) | -0.017 | Pr(Mastering Required Found. Skills) | 0.045 | Pr(Level C or Above) | -0.011 |
| | (0.031) | | (0.034) | | (0.029) | | (0.036) |
| Pr(Reading at Level A or Above) | -0.021 | Pr(Level E or Above) | 0.064 | Pr(Reading at Level A or Above) | 0.016 | Pr(Level E or Above) | -0.057 |
| | (0.024) | | (0.047) | | (0.019) | | (0.045) |
| | | Pr(Level I or Above) | 0.039 | | | Pr(Level I or Above) | -0.026 |
| | | | (0.042) | | | | (0.039) |
| *Panel D: Asian, N=2,095* | | *N=2,017* | | *Panel I: English Proficient N=3,429* | | *N=3,104* | |
| Total Items Missed | -0.381** | Reading Scale | 0.133 | Total Items Missed | -0.227** | Reading Scale | 0.067 |
| | (0.086) | | (0.215) | | (0.063) | | (0.170) |
| Pr(Mastering Required Found. Skills) | 0.126** | Pr(Level C or Above) | 0.049 | Pr(Mastering Required Found. Skills) | 0.019 | Pr(Level C or Above) | 0.027 |
| | (0.048) | | (0.035) | | (0.030) | | (0.032) |
| Pr(Reading at Level A or Above) | 0.023 | Pr(Level E or Above) | 0.004 | Pr(Reading at Level A or Above) | 0.012 | Pr(Level E or Above) | 0.093* |
| | (0.028) | | (0.054) | | (0.026) | | (0.043) |
| | | Pr(Level I or Above) | 0.028 | | | Pr(Level I or Above) | 0.056 |
| | | | (0.054) | | | | (0.041) |
| *Panel E: Hispanic, N=1,683* | | *N=1,546* | | | | | |
| Total Items Missed | -0.174** | Reading Scale | -0.146 | | | | |
| | (0.067) | | (0.241) | | | | |
| Pr(Mastering Required Found. Skills) | 0.028 | Pr(Level C or Above) | -0.091 | | | | |
| | (0.022) | | (0.065) | | | | |
| Pr(Reading at Level A or Above) | 0.024 | Pr(Level E or Above) | -0.022 | | | | |
| | (0.024) | | (0.070) | | | | |
| | | Pr(Level I or Above) | 0.018 | | | | |
| | | | (0.045) | | | | |

*Note:* Each cell represents the results of a separate regression discontinuity estimate of the effect of Transitional Kindergarten on the indicated literacy outcome. Row headers indicate the dependent variable and panel headers indicate the subsample. Negative binomial models were used to estimate the effect of Transitional Kindergarten on the total items missed, ordinal logit models were used to estimate the effet of Transitional Kindergarten on the reading level, and OLS models were used in all other cases. All functional forms include a linear spline and covariates defined in Table 5. Akaike's Information Criteria indicates a linear spline is optimal. All standard errors are clustered on the day of birth running variable, except for the conditional negative binomial and ordinal logit models which must be clustered on the teacher-by-year fixed effect. +indicates p<0.10, *p<0.05, **p<0.01

Table 8: ITT RD estimates of kindergarten and first grade CELDT outcomes by subgroup

| Dependent Variable: Overall Score | Kindergarten | | First Grade | |
|---|---|---|---|---|
| | (1) | N | (2) | N |
| All Limited English Proficient (LEP) | 0.176* | 3,310 | 0.231** | 2,663 |
| | (0.079) | | (0.075) | |
| Male | 0.135 | 1,662 | 0.212+ | 1,354 |
| | (0.120) | | (0.123) | |
| Female | 0.241* | 1,648 | 0.199+ | 1,309 |
| | (0.111) | | (0.106) | |
| Asian | 0.117 | 1,523 | 0.279** | 1,291 |
| | (0.117) | | (0.099) | |
| Hispanic | 0.356* | 1,159 | 0.159 | 950 |
| | (0.138) | | (0.139) | |

*Note:* Each cell represents the results of a separate regression discontinuity estimate of the effect of Transitional Kindergarten on the overall CELDT scale score. Row headers indicate the subsample. All functional forms include a linear spline and covariates defined in Table 5. Akaike's Information Criteria indicates a linear spline is optimal. All standard errors are clustered on the day of birth running variable. +indicates p<0.10, *p<0.05, **p<0.01

Table 9: Robustness check: Placebo estimates of fall and midyear literacy outcomes

| | (1) | (2) | (3) | **(4)** | (5) | (6) | (7) | |
|---|---|---|---|---|---|---|---|---|
| *Panel A: Kindergarten Outcomes* | $B_{ict}$-50 | $B_{ict}$-40 | $B_{ict}$-30 | **$B_{ict}$** | $B_{ict}$+30 | $B_{ict}$+40 | $B_{ict}$+50 | N |
| Total Items Missed | -0.075 | -0.085 | -0.138+ | **-0.181**** | 0.033 | 0.037 | 0.060+ | 6,739 |
| | (0.112) | (0.083) | (0.071) | **(0.042)** | (0.033) | (0.032) | (0.031) | |
| Overall CELDT Score | -0.248 | -0.100 | 0.157 | **0.176*** | 0.042 | -0.094 | -0.055 | 3,310 |
| | (0.253) | (0.123) | (0.118) | **(0.079)** | (0.075) | (0.073) | (0.069) | |
| *Panel B: First Grade Outcomes* | | | | | | | | |
| Overall CELDT Score | -0.034 | 0.151 | 0.194 | **0.231**** | -0.006 | -0.089 | -0.031 | 2,663 |
| | (0.225) | (0.137) | (0.122) | **(0.075)** | (0.077) | (0.077) | (0.078) | |
| Covariates | √ | √ | √ | √ | √ | √ | √ | |
| Fixed Effects | √ | √ | √ | √ | √ | √ | √ | |

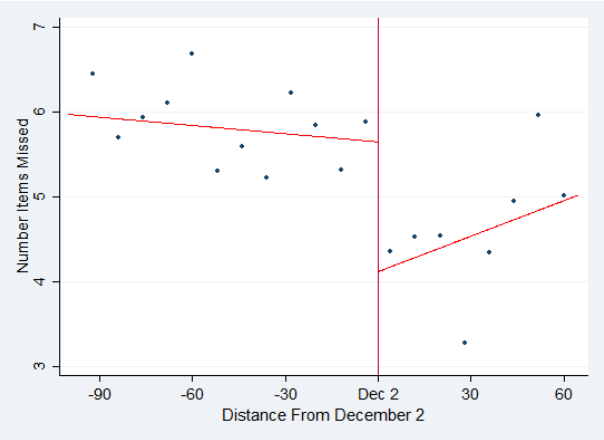
*Note:* Each cell represents the results of a separate regression discontinuity estimate of the effect of Transitional Kindergarten on the indicated literacy outcome. Row headers indicate the dependent variable. Column 4 contains estimates from the regression discontinuity found in Table 5, Columns 2 and 4. All other columns contain estimates from placebo RDs. Covariates are the same as those in Table 5. The functional form of all regressions is a linear spline. All standard errors are clustered on the day of birth running variable. +indicates p<0.10, *p<0.05, **p<0.01

Table 10: Robustness check: Estimates after eliminating heaps

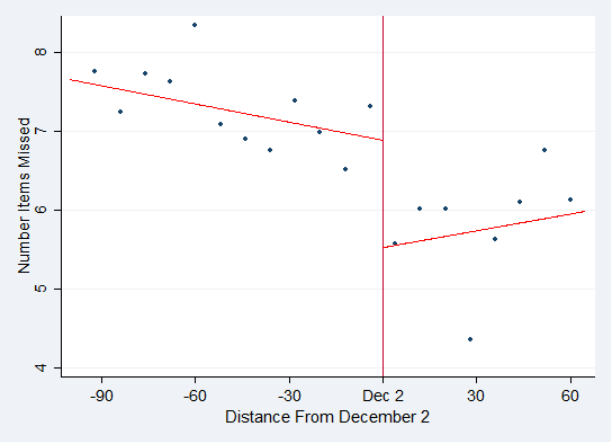
| | (1) | (2) | (3) | (4) | (5) |
|---|---|---|---|---|---|
| *Panel A: Fall Kindergarten Outcomes* | **Full Sample** | $H_B \leq 25$ | $H_B \leq 20$ | $H_B \leq 18$ | $H_B \leq 15$ |
| Total Items Missed | **-0.181**** | -0.253** | -0.298** | -0.364** | -0.337** |
| | **(0.042)** | (0.050) | (0.068) | (0.076) | (0.114) |
| N | **6,739** | 5,663 | 3,417 | 2,536 | 1,248 |
| Overall CELDT Score | **0.176*** | 0.220* | 0.179 | 0.287* | 0.381+ |
| | **(0.079)** | (0.092) | (0.120) | (0.134) | (0.212) |
| N | **3,310** | 2,794 | 1,703 | 1,263 | 661 |
| *Panel B: Fall First Grade Outcomes* | | | | | |
| Overall CELDT Score | **0.231**** | 0.268** | 0.191 | 0.400** | 0.296 |
| | **(0.075)** | (0.093) | (0.136) | (0.137) | (0.219) |
| N | **2,663** | 2,251 | 1,360 | 1,017 | 547 |

*Note:* Each cell represents the results of a separate regression discontinuity estimate of the effect of Transitional Kindergarten on the indicated literacy outcome. Row headers indicate the dependent variable. Column 1 contains estimates from regression discontinuity found in Table 5, Columns 2 and 4. All other columns contain estimates from samples obtained from by eliminating heaps of varying sizes. $H_B$ represents heaps at values of the running variable, $B_{ict}$ . Heaps greater than the value in the column headers were eliminated from the sample. Covariates include those used in Table 5. The functional form of all regressions is a linear spline. +indicates p<0.10, *p<0.05, **p<0.01

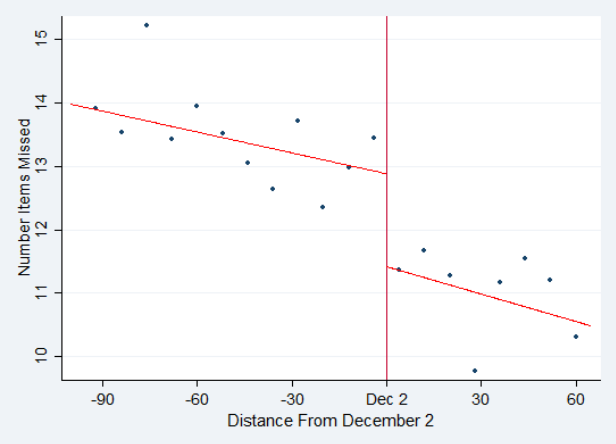
# Appendix



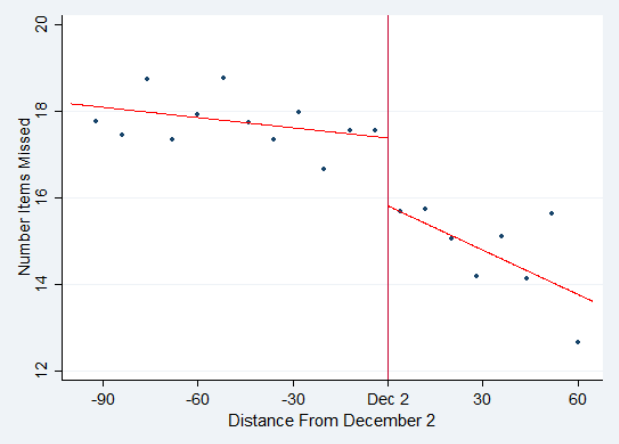

**(a) Upper Case Letter Recognition**



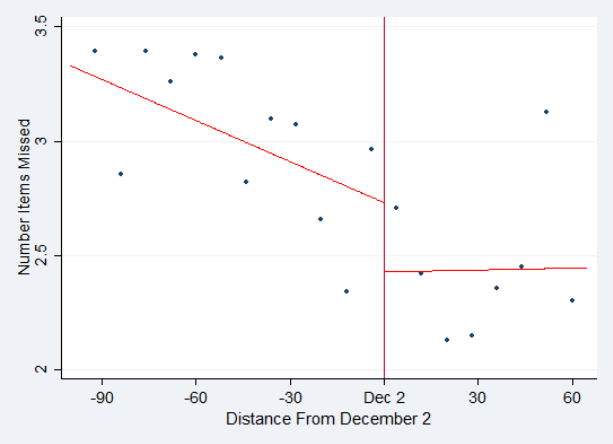

**(b) Lower Case Letter Recognition**



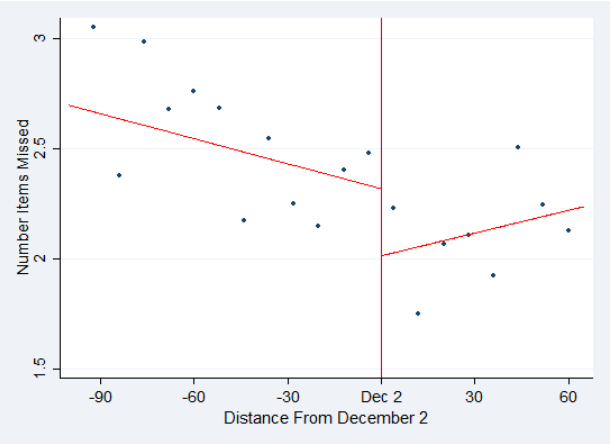

**(c) Letter Sounds**



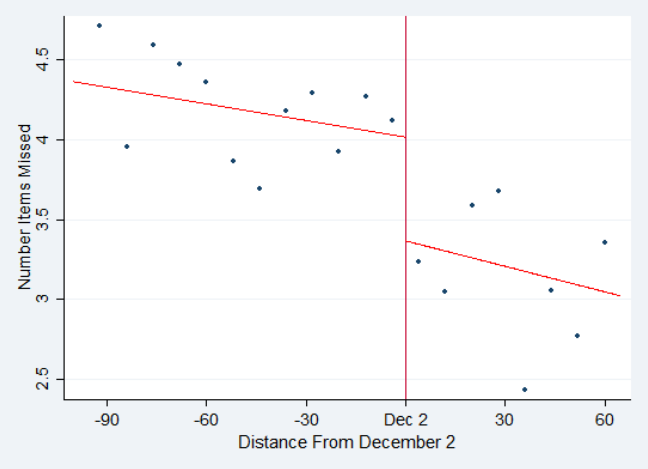

**(d) High Frequency Word Recognition**




**(e) Early Literacy Behaviors**

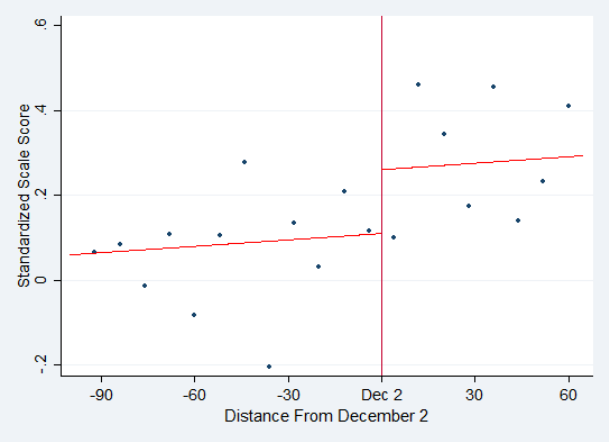



**(f) Initial Word Sounds**

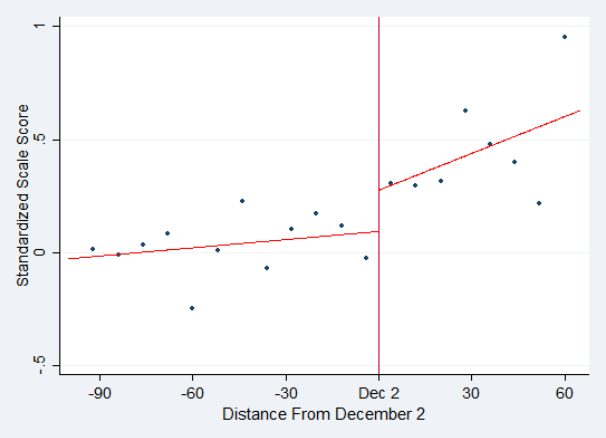



**(g) Rhyming**

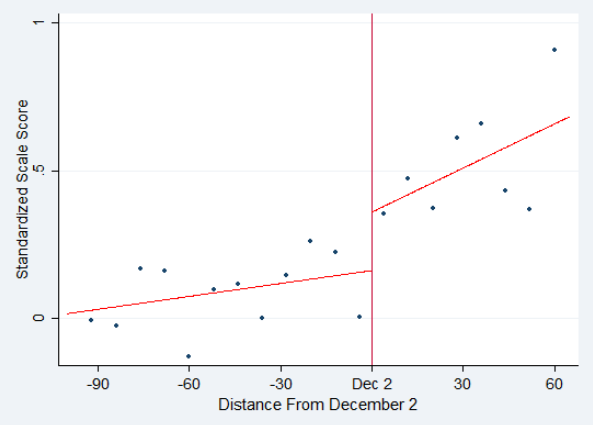
**Figure A1:** Fall kindergarten Fountas and Pinnell foundational literacy outcomes. Each dot represents the average outcome in an 8 day bin width. TK eligible students are to the right of the vertical line and TK ineligible students are to the left of the line. The x-axis represents distance of birthday in days from December 2. Birthdays are centered at December 2.



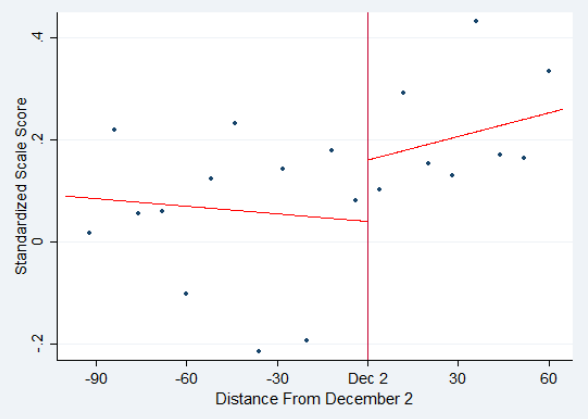

**(a) Listening**




**(b) Reading**

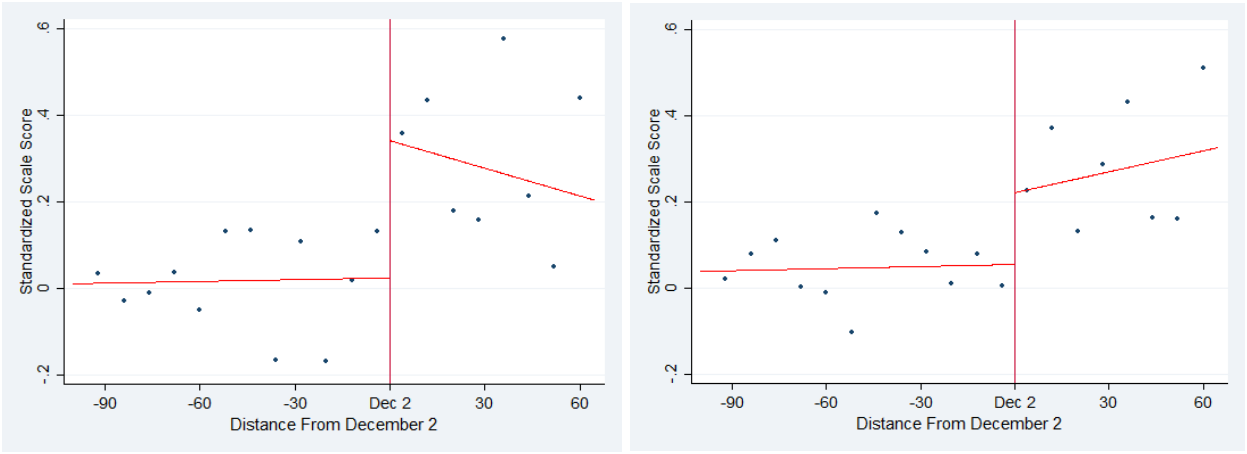



**(c) Writing**

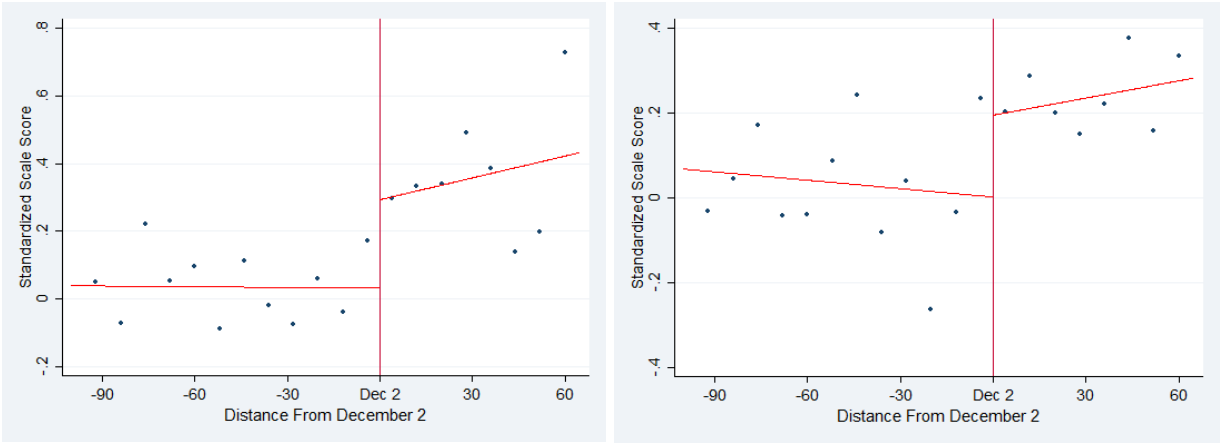



**(d) Speaking**

**Figure A2:** Fall kindergarten CELDT subtest outcomes. Each dot represents the average outcome in an 8 day bin width. TK eligible students are to the right of the vertical line and TK ineligible students are to the left of the line. The x-axis represents distance of birthday in days from December 2. Birthdays are centered at December 2. CELDT stands for the California English Language Development Test.

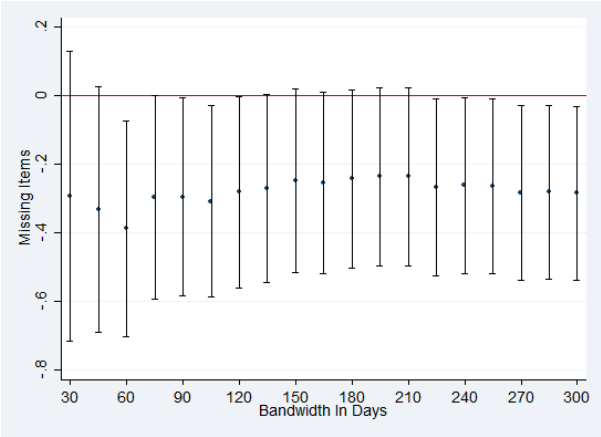



**(a) Listening** **(b) Reading**

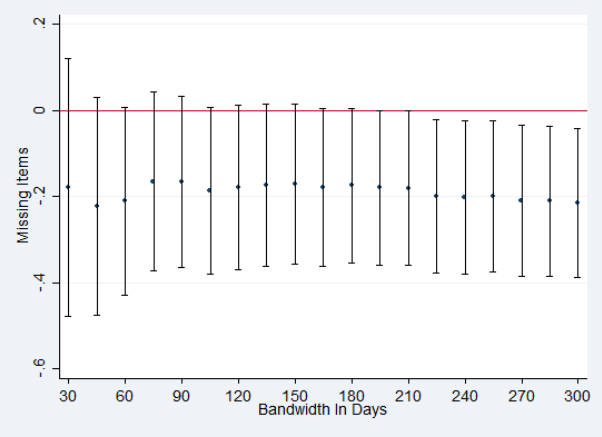



**(c) Writing** **(d) Speaking**

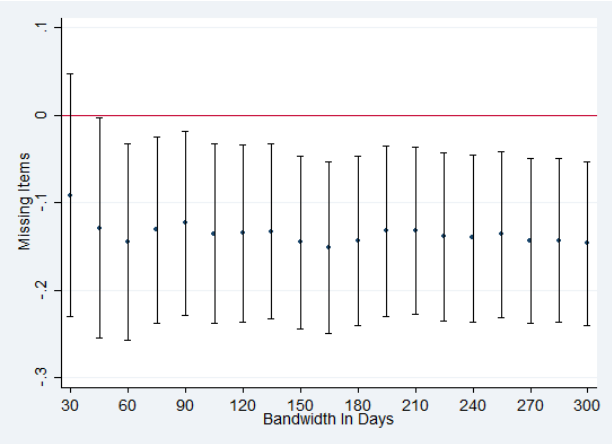
**Figure A3:** Fall first grade CELDT subtest outcomes. Each dot represents the average outcome in an 8 day bin width. TK eligible students are to the right of the vertical line and TK ineligible students are to the left of the line. The x-axis represents distance of birthday in days from December 2. Birthdays are centered at December 2. CELDT stands for the California English Language Development Test.



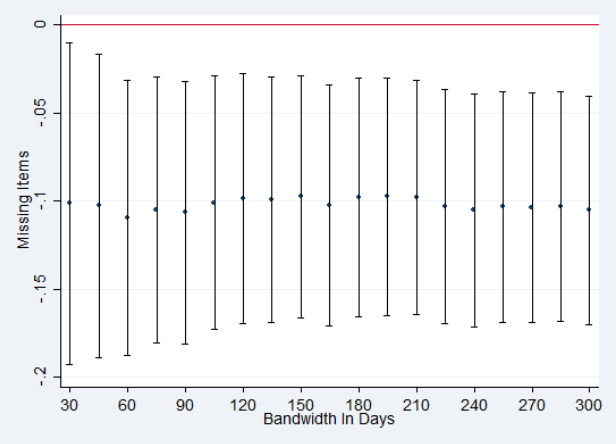

**(a) Upper Case Letter Recognition**



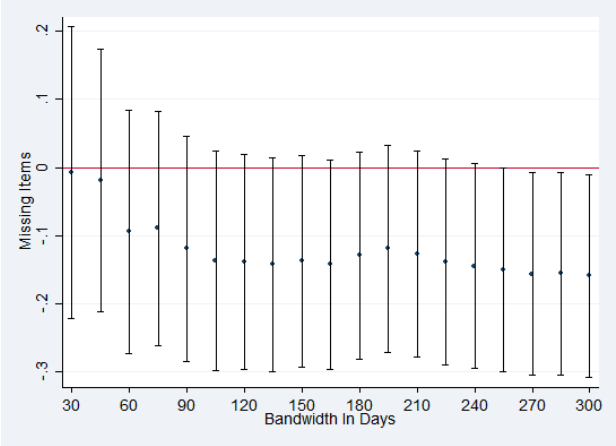

**(b) Lower Case Letter Recognition**



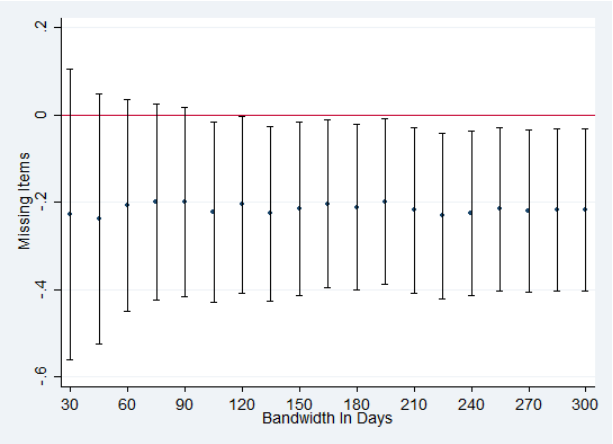

**(c) Letter Sounds**



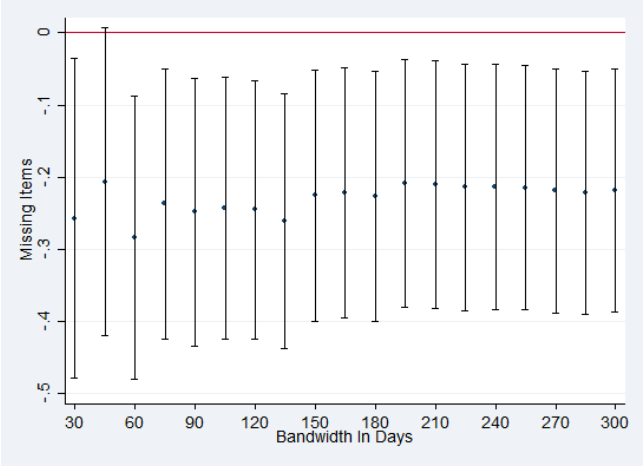

**(d) High Frequency Word Recognition**




**(e) Early Literacy Behaviors**

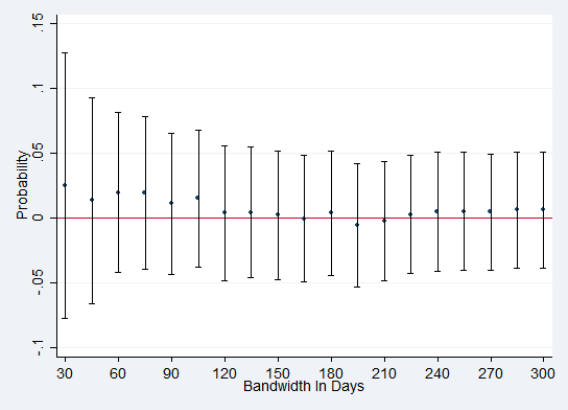



**(f) Initial Word Sounds**

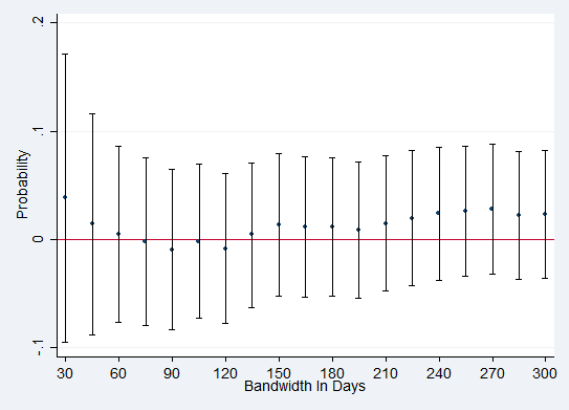



**(g) Rhyming**

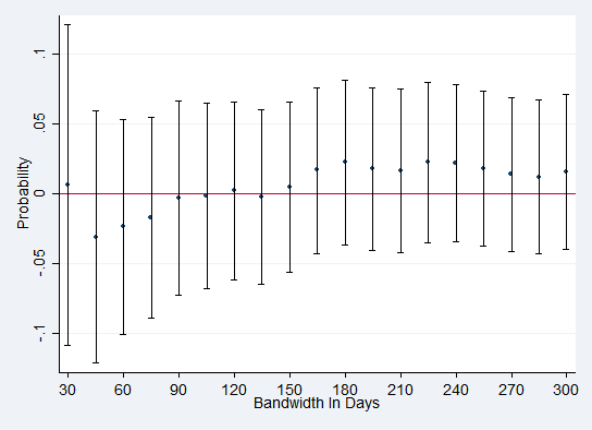
**Figure A4:** Auxiliary robustness checks of fall kindergarten Fountas and Pinnell foundational literacy outcomes. Each dot represents a regression discontinuity estimate of the effect of Transitional Kindergarten on the relevant outcome for observations in bandwidths between 30 and 300 days. Dots represent point estimates and vertical lines represent the 95 percent confidence inteval. All figures employ a negative binomial regression. Teacher-by-year fixed effects are not included because models would not converge for all banwidths. All regressions employ a linear spline functional form with covariates detailed in Table 5. Standard errors are clustered at the teacher-by-year cell.




**(a) Pr(Reading at Level C or Above**

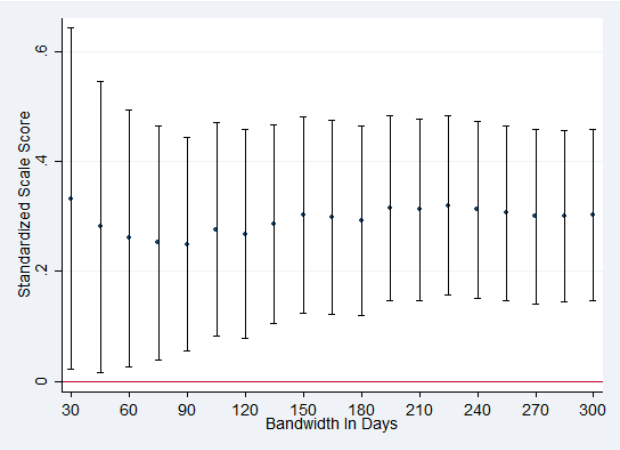



**(b) Pr(Reading at Level E or Above)**

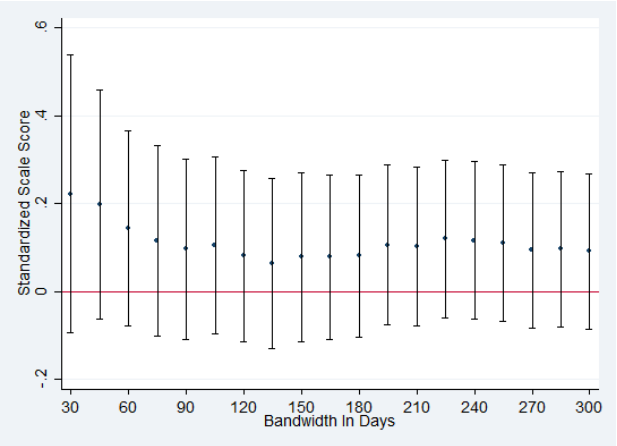



**(c)Pr(Reading at Level I or Above**

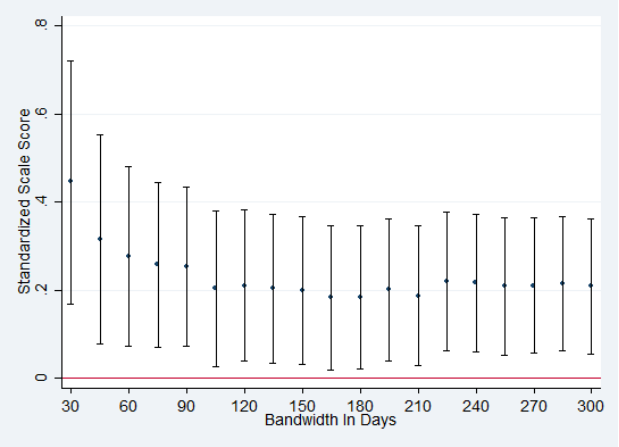
**Figure A5:** Robustness checks of fall first grade Fountas and Pinnell foundational literacy outcomes. Each dot represents a regression discontinuity estimate of the effect of Transitional Kindergarten on the relevant outcome for observations in bandwidths between 30 and 300 days. Dots represent point estimates and vertical lines represent the 95 percent confidence inteval. All regressions employ a linear spline functional form with covariates detailed in Table 5. Standard errors are clustered on the day of birth rating variable.



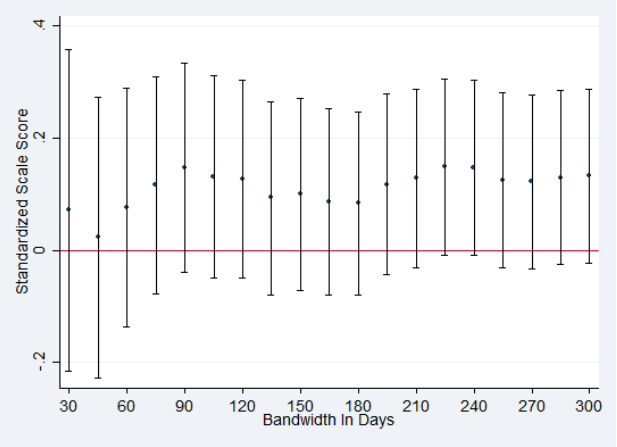

**(a) Listening**




**(b) Reading**

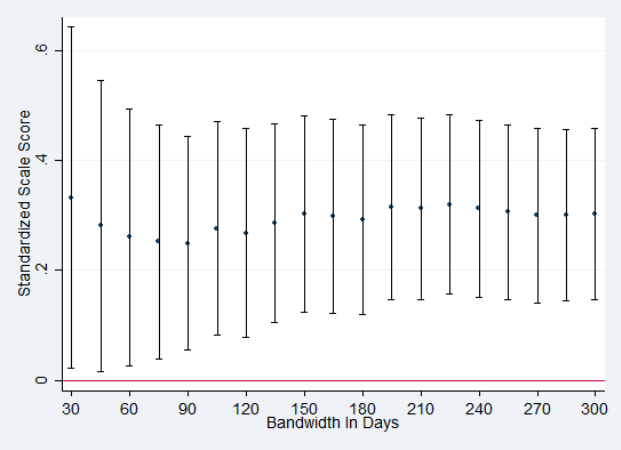



**(c) Writing**

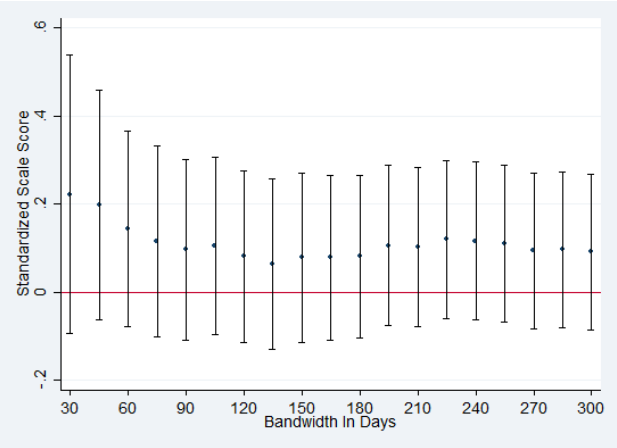



**(d) Speaking**

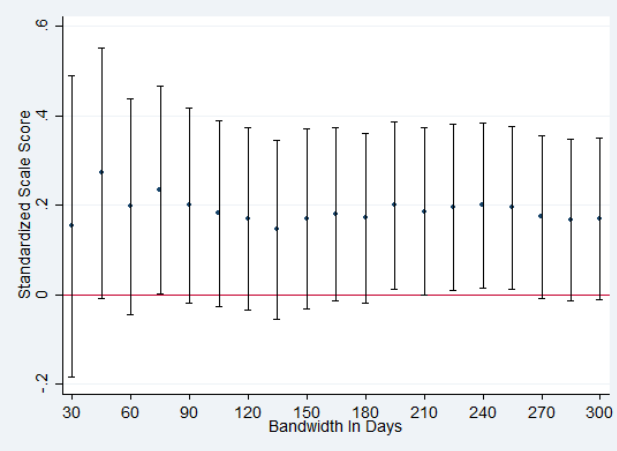
**Figure A6:** Auxiliary robustness checks of fall kindergarten CELDT subtest outcomes. Each dot represents a regression discontinuity estimate of the effect of Transitional Kindergarten on the relevant outcome for observations in bandwidths between 30 and 300 days. Dots represent point estimates and vertical lines represent the 95 percent confidence inteval. All regressions employ a linear spline functional form with covariates detailed in Table 5. Standard errors are clustered on the day of birth rating variable.



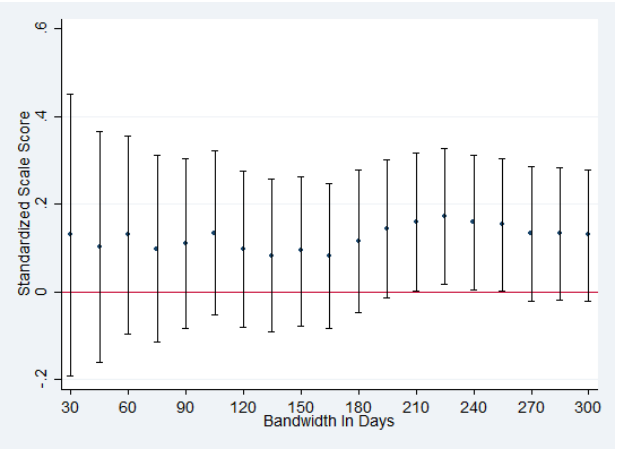

**(a) Listening**




**(b) Reading**




**(c) Writing**




**(d) Speaking**

**Figure A7:** Auxiliary robustness checks of fall first grade CELDT subtest outcomes. Each dot represents a regression discontinuity estimate of the effect of Transitional Kindergarten on the relevant outcome for observations in bandwidths between 30 and 300 days. Dots represent point estimates and vertical lines represent the 95 percent confidence inteval. All regressions employ a linear spline functional form with covariates detailed in Table 5. Standard errors are clustered on the day of birth rating variable.

Table A1: RD regressions of balance In sample restrictions

| | (1) | (3) | (5) | (5) |
|---|---|---|---|---|
| | Full Sample | $\|B_{ict}\| \leq 60$ | $\|B_{ict}\| \leq 30$ | $\|B_{ict}\| \leq 15$ |
| Missing Kindergarten Blending | 0.010 | 0.001 | 0.011 | 0.056 |
| | (0.017) | (0.019) | (0.028) | (0.037) |
| Missing Kindergarten Rhyming | -0.035 | -0.026 | 0.011 | -0.010 |
| | (0.023) | (0.028) | (0.035) | (0.047) |
| Missing First Grade Fountas and Pinnell | 0.019 | 0.035 | 0.070* | 0.034 |
| | (0.017) | (0.020) | (0.026) | (0.031) |
| Missing Kindergarten CELDT | 0.032 | 0.059 | 0.083 | -0.016 |
| | (0.038) | (0.047) | (0.066) | (0.089) |
| Missing First Grade CELDT | -0.007 | 0.021 | 0.037 | -0.037 |
| | (0.040) | (0.050) | (0.074) | (0.105) |
| N | 6,739 | 2,182 | 1,271 | 662 |

*Note:* Each cell represents the results of a separate regression discontinuity estimate on an indicator for not being in the sample defined in the row headers. Column headers indicate the bandwidth restriction. The functional form in all regressions is a linear spline. All standard errors are clustered on the day of birth running variable. +indicates p<0.10, *p<0.05, **p<0.01

Table A2: McCrary density test on baseline covariates

| | Point Estimate (Standard Error) | N |
|---|---|---|
| Student Characteristics | | |
| Female | 0.048 (0.134) | 3,316 |
| Asian | 0.026 (0.170) | 2,095 |
| Hispanic | 0.119 (0.183) | 1,683 |
| White | -0.006 (0.211) | 1,111 |
| Other | 0.254 (0.193) | 1,179 |
| Declined To State Ethnicity | 0.218 (0.287) | 660 |
| Special Education | 0.123 (0.339) | 510 |
| Limited English Proficient (LEP) | -0.019 (0.122) | 3,310 |
| Home Language: | | |
| Chinese | 0.000 (0184) | 1,150 |
| Spanish | 0.049 (0.213) | 1,005 |
| English | 0.148 (0.127) | 4,020 |
| Other | 0.388 (0.268) | 564 |
| Dominant Language: | | |
| Chinese | -0.075 (0.178) | 1,387 |
| Spanish | 0.188 (0.227) | 1,170 |
| English | 0.236+ (0.129) | 3,412 |
| Other | 0.009 (0.248) | 770 |
| Test Characteristic | | |
| Test Given In Spanish | 0.147 (0.257) | 945 |

*Note:* Each cell represents the results of a separate McCrary density test on the sample defined in the row headers. +indicates p<0.10, *p<0.05, **p<0.01