**Response to Reviewers**

We thank the editor and the four reviewers for a very helpful set of comments. We were excited to read that you and the reviewers thought the paper is “clearly written,” “addresses an important topic” (changes in school readiness over time), and has “potential for a unique contribution.” We were also grateful for the thoughtful and detailed suggestions. We have completed a major revision to address all the issues raised, and believe the manuscript is far stronger.

Below we provide a point-by-point response to reviewer suggestions. However, let me summarize the main changes we have made. First, the reviewers provided a variety of helpful suggestions about ways to better situate our work in the existing literature, ways to motivate our choice of covariates, and ways to discuss the policy relevance of our findings. To address these issues, we have expanded our discussion of the extant knowledge base, highlighting evidence from state-level reports that mirror the findings in our current analysis. We have provided much more discussion of our multivariate analyses and the motivation for including each set of covariates (e.g. child care variables, home environment measures, etc.). We have also expanded our discussion of the results, and in particular the surprising finding that including measures of public preschool participation (which rose substantially over our study paper) did virtually nothing to our estimates of changes in school readiness over time. We now describe other potential mechanisms driving the observed changes in readiness and also discuss the policy implications of children entering kindergarten with more academic skills than they once did.

In addition, the reviewers provided some important suggestions around alternative model specifications and other sensitivity checks. For instance, Reviewer 4 appropriately raised concern about analyzing highly skewed behavioral outcomes using OLS regressions. We reviewed published studies that use the same behavioral measures and ultimately decided to dichotomize the behavioral measures as is common in this literature. We then ran all analyses involving dichotomous outcomes using Linear Probability Models (LPM) as well as logistic regressions. Given the very high level of correspondence across these two types of models, we present LPM results throughout for their ease of interpretation.

One limitation raised by the reviewers is that our analysis relies on teacher-reported rather than directly assessed measures of academic skills and behavior. We share this concern, and believe a similar analysis to the one we have conducted here leveraging direct assessments would be welcome, as it would eliminate concerns about changes in the underlying rating process over time. Unfortunately, no such data exists. The Early Childhood Longitudinal Studies that we use for the current study provide direct assessments in both waves, but they are not comparable over time and no cross-walk has been released. While we do agree that the use of teacher-directed assessments has notable limitations, and we now articulate those more directly in the manuscript, we also provide strong evidence both from the ECLS data and from the extant literature around the validity of these measures. In addition, we provide several theoretical and empirical reasons to believe the teacher measures are capturing meaningful changes in children’s skills over time. Given the lack of direct assessment data, and the importance of the research question, we believe it makes sense to use the teacher-reported data to provide compelling new evidence of these changes over time.

Finally, in your letter you note the need for: *“clarifying information regarding the measures (R1, R2, R4), reporting correlations among measures (R1) as well as providing statistically significant differences between measures across cohorts (R3), clarifying regression analyses (e.g., type, sample sizes, use of weights, covariates; R4), and careful copy editing (R1).”* We believe the current manuscript has been responsive to all of these requests, and provide specific details below.

Again, we thank you and the reviewers for the very helpful feedback. We believe the revised manuscript is much improved. We continue to believe this paper addresses a timely education issue and that the results are striking, especially in light of debates around “academic kindergarten” and the role of Common Core Standards for kindergarten. Our study suggests children are entering kindergarten with a different skill set than they once were, and this finding has meaningful implications for how we structure early learning. It also raises intriguing questions for future research. We look forward to hearing your reactions to this revision

**Reviewer 1**

*We thank reviewer 1 for these helpful suggestions.*

**Comments to the Author:**

**It would be useful to add the correlations between your summary measures and the direct assessments.  I understand you can’t compare across tests, but it would be useful in bolstering your argument about teacher ratings capturing skills.**

*Yes, we agree that these correlations are important to show that the measures we construct are related to direct measures of math and reading. These correlations are included in Appendix A.*

**The ECLS-K adapted the SSRS and uses their version of the “SRS” in the 1998 data.  This might not be the case in 2011, but please confirm.**

*The same behavioral items were used across both cohorts for four of the five behavioral outcomes examined (and indeed these outcomes are the “SRS” adapted from the original SSRS). For the fifth outcome (approaches to learning), the scale was adjusted slightly across cohorts (6 items in 1998, 7 items in 2010). We have updated the manuscript to state this information explicitly.*

**The manuscript could use a careful copy editing.**

*Thank you for this feedback. We carefully edited the manuscript.*

**Reviewer 2**  
  
**Comments to the Author:**

**Thank you for the opportunity to review the paper “Kids today: Changes in school-readiness in an early childhood era”. The authors take on an important topic, namely school readiness, and utilize two robust datasets. As my comments below note, I believe this paper has potential to make important contributions to our understanding of school readiness but that the current iteration faces a number of areas for improvement.**

*We thank Reviewer 2 for providing us terrific suggestions for improving our paper. As outlined below, we have addressed these comments and believe the paper is much stronger for it.*

**In motivating the paper, the authors state that “until now we have had no data that would allow for comparisons across cohorts of children at kindergarten entry”.  I am not sure this is factually accurate.  A number of states administer kindergarten readiness assessments, and some, such as Maryland, have been doing so for over a decade.  As an example, the Minnesota Department of Education published longitudinal data on this question in 2012 as part of the Minnesota School Readiness Study: Developmental Assessment at Kindergarten Entrance.  While examining this question at the national level may be unique, I would like to see the authors provide a more thorough consideration of what data already exists, what analyses have been conducted on it, and what such state level data might suggest regarding expected findings at a national level.**

*Thank you for the suggestion that we discuss trends in state kindergarten readiness assessments over time. In particular, you pointed out that Maryland and Minnesota both have been measuring indicators of school readiness over time. We were not familiar with these interesting data .The trends in these readiness assessments certainly suggest large changes over time in school readiness (which ultimately mirrors our results). We have added citations to several state reports that document such trends. Your comment also led us to carefully search for any additional evidence on this topic. We have added discussion of a recent Child Trends report that finds increases over time in four measures of academic knowledge (e.g. can the child count to 20, can they recognize the alphabet) among children ages 3-6 using repeated iterations of the National Household Education Surveys program, which is another survey administered by NCES.*

*While the existing data certainly aligns with our overarching hypothesis that academic readiness increased over our study period, we think the richness of the ECLS data is a major strength, and that our analysis fills several important gaps. In particular, our study explores heterogeneity in the patterns by race and socioeconomic status. We also look at changes over time in children’s behavioral scores at school entry. In addition, we do this in a regression context and examine the extent to which observed changes in a host of characteristics explain the trends observed.*

**A major limitation of this study appears to be its use of teacher-rated measures of literacy and mathematics knowledge. As the author’s acknowledge, such measures may reflect changes in teachers’ perceptions of readiness rather than true changes in incoming kindergartener knowledge.  The authors appear to try to address this issue by stating that “research demonstrates that teachers’ assessments of cognitive skills are strongly correlated with… direct assessments”.  The issue of concern, however, is not whether these correlate with direct assessments but, rather, whether or not they are appropriately aligned over time.  A set of teacher-rated measures with a mean of 2 and another set with a mean of 4 could both have the same correlation with a set of direct assessment scores; therefore, merely showing that teacher-reported ability is correlated with direct assessment scores within each year does not provide evidence that they are appropriately scaled over time.  Given that direct assessments were conducted in each iteration of the ECLS and that a cross-walk is anticipated to be released to allow for comparisons of these scores, I wonder if the authors would be better to wait for the release of the cross-walk and then present this paper utilizing the more objective direct assessment scores.**

*Thank you for highlighting this important point. We certainly agree that systematic changes in the way teachers assigned ratings is a concern for our study. We also fully agree with your point that similar correlations across time do not imply that teachers have not changed the way they assign ratings.*

*That said, we agree with Reviewer 1 that the correlations are still worth mentioning and including. For one, the fact that the correlations are strong in both periods (irrespective of how they compare over time) helps make the case that teacher ratings are capturing something valuable about children’s early reading and math skills. The strong correlations between direct and teacher-reported assessments of child skills are often used to justify the utilization of teacher-directed measures.*

*Further, while we completely agree that similar correlations over time DO NOT necessarily imply that they are rating children in the same way over time, IF we saw large changes we would have even greater reasons for concern about the comparability of these measures over time.*

*In addition, we note several theoretical and empirical points that strengthen our confidence that our results are capturing “real” changes in children’s learning over time:*

* *First, we know that over the period we examine, kindergarten classrooms became increasingly focused on literacy and math and that kindergarten teachers expressed much higher academic expectations, both for incoming kindergarteners and for learning during the kindergarten year (Bassok, Latham, & Rorem, 2016). We also know that teachers assessed children more frequently in the more recent period. It is not clear exactly how these changes would impact teachers’ evaluations of children’s skills. However, we think one plausible hypothesize is that, the much higher expectations combined with the need to teach more advanced reading and math content would lead teachers in the later period to rate children more harshly, all else equal. The fact that instead teachers report* higher *levels of skills in the second period is counter to the plausible narrative that teachers’ heightened expectations might be driving changes in their evaluation of children’s skills.*
* *Second, our findings of interactions by race and by income are not likely to be explained by systematic changes in the way that teachers assign ratings. In particular, teachers would have had to disproportionately change the way they assign ratings for black and low income students. We have no theoretical reason to believe that in the later period teachers would somehow become systematically “less harsh” or “more forgiving” in their ratings of low-income and non-white children.*
* *Relatedly, our results suggest a narrowing of racial and socio-economic school readiness achievement gaps. This same finding has been found in two recent studies that look at achievement gaps over the same period using direct assessments (Magnuson & Duncan, 2014; Reardon & Portilla, 2015). This alignment between the direct and teacher-reported measures further raises our confidence that our measures are capturing “real” changes over time.*

*All these points notwithstanding, we still agree with the reviewer that a comparison using the direct assessments would be stronger. However we are quite hesitant to wait for a crosswalk that will allow for this comparison as it is very unclear if/when such a crosswalk will be made available. There is still no specific time frame for the release of this crosswalk, and the release has already been pushed off several times. As such, and because we think the question about changes over time in school readiness is an important one, we think the best path forward is to proceed with these measures, but to carefully acknowledge their limitations.*

**A final consideration with this paper is the extent to which it informs policy and/or practice.  While the authors frame the paper in the context of increased investment in preschool, their tests of whether these or other investments by states, schools, or parents contribute to the observed gains appear limited.  If controlling for these characteristics does not impact the estimated coefficients, then what is another plausible explanation?  Does this suggest that the observed gains represent changes in the way teachers rate these skills rather than a policy amenable characteristic of the students’ experiences?  In the absence of such a connection, the reader seems to be left with little information on the interesting policy levers for continuing improvement in school readiness.**

*Thank you for these comments. The primary goal of this paper is descriptive. To date, there have been no large-scale empirical studies tracking changes over time in children’s school readiness. We consider this to be a troubling gap, given all the investment in early childhood education, and we think carefully documenting these changes, both overall and for subgroups is an important contribution.*

*That said, the reviewer’s point is that the controls we include explain a relatively small portion of the changes over time. We were particularly surprised at how little difference the inclusion of variables tracking changes in public preschool participation into our model made. We find the* lack *of a relationship here intriguing and potentially policy relevant.*

*We definitely do not think that this fact implies that teachers are changing the way they are rating children over time (see the discussion of this point above). However, we agree it is important to provide some plausible explanations. One plausible explanation is that the changes we are observing over time are driven by changes in the* quality *of the child care experiences children have. There has been a substantial public investment in improving quality both by moving children from informal to formal settings, and by investing in quality enhancements across sectors. Unfortunately, the ECLS-K data do not provide any information about quality measures. We were also unable to account for other potentially important shifts in this time period (e.g. shifts in access to health insurance).*

**Specific Notes:**

**Page 1, paragraph 1: The authors cite a number of statistics on the increase in preschool funding and enrollment; however, proper citations are not present.  Please include citations in the text for these descriptive statements.**

*Thank you for noting this omission, we now include proper citations.*

**Page 1, paragraph 2: The authors state that “we have little empirical evidence on whether children today know more at school entry than they did two decades ago”.  This statement should be brought into alignment with the timeframe of the study, namely 1998 to 2010.**

*Thank you for the suggestion. We have revised the text accordingly.*  
  
**Page 8, paragraph 2: The authors note that they control for the month of assessment in order to deal with the fact that the fall of K scores may be picking up some skills acquired in the early weeks of kindergarten.  I would suggest that the authors may also consider controlling for the starting month of school (which presumably would be attainable through state identifiers) as this, in conjunction with the month of assessment, would represent the amount of time spent in kindergarten prior to the assessment.**

*Thank you for this thoughtful suggestion. We agree the issue of controlling for the amount of time spent in kindergarten prior to being assessed is an important one. We were not able to find reliable state-level documentation as to school starting months for both 1998 and 2010, so unfortunately we cannot follow your suggestion exactly. However, we have thought carefully about how best to control for the amount of time spent in kindergarten. We now control for both children’s age in August of the kindergarten year and their age at the time of assessment. Although imperfect, we believe these variables combined provide a good proxy for the amount of time a child spent in kindergarten before being assessed. In addition, our results are robust to various constructions of the amount of time children spent in kindergarten, and also to omitting this variable entirely.*

**Page 6, paragraph 2: Recent research on multiple imputation suggests that 5 imputed datasets may be insufficient depending on the percentage of missing data.  It would be useful to provide the reader with an indication of the percentage of data missing (by variable or at least as a range across imputed variables) and then, if necessary, to consider conducting a more robust imputation.  See Paul Allison’s comments on this and list of references (**[**http://www.statisticalhorizons.com/more-imputations**](http://www.statisticalhorizons.com/more-imputations)**)**

*In the past, we have come across contradictory advice regarding the number of imputed datasets that are appropriate, so we are thankful for the link to a clear and concise summary of current best practices.*

*Based on Paul Allison’s remarks, it seems that a good rule of thumb is to base the number of imputed datasets off the percent of data that is missing (so for example, 10 datasets if 10% of the data is missing). In our case the percent of data that is missing ranges from 0-17%, varying a bit across variables and across cohorts. As a result, all of our reported estimates now use 20 imputed datasets.*

**Page 9: When describing the social-scales rating, it might be useful to provide reliability measures.  These may be available in the ECLS-K documentation.**

*We agree that it is helpful to include these measures. Here we report these reliability coefficients separately by cohort.*



*In general, the reliabilities are quite similar across cohorts, and so we report pooled reliability coefficients in the text of the manuscript. The exception is that we report reliability coefficients separately for the “approaches to learning” scale, which differed slightly in the number of items across cohorts.*

**Page 18, final paragraph: The authors note a “15 year” difference.  Based on the data, it appears that a more accurate description would be 12 years.**

*We agree. We have changed the text to reflect the correct timeframe.*  
  
**Other comments:**

**-Would the authors be able to leverage any data from the ECLS-B dataset?  I believe this dataset includes fall of K measures that may align with those in both iterations of the ECLS-K and could provide a third year.**

*We definitely considered the idea of incorporating data from the ECLS-B in this study. Ultimately we decided against this for two reasons. First, the measures that we use to capture literacy and math ability were not collected as part of the ECLS-B. Instead, the only measures of reading and math were direct student assessments. These assessments are not directly comparable to the ECLS-K datasets, and NCES has not announced plans to provide a crosswalk allowing us to link them (this is in contrast to their stated plans to link direct assessments across the two ECLS-K waves).*

*Secondly, while the behavioral measures that we use were collected as part of the ECLS-B, we have concerns about the comparability of the ECLS-K samples to the ECLS-B. In particular, the two ECLS-K samples are nationally representative of students entering kindergarten in their respective years. By contrast, the ECLS-B is a nationally representative sample of infants born in 2001. Although these samples are likely to be similar, we prefer the “apples to apples” comparison of using two waves with identical sampling designs.*

**-I believe the authors could further develop the theoretical framework.  While they mention life-course theory briefly, I believe a fuller outline of the theoretical framework is in order.  Given that they consider aspects of preschool, parental investment, and so forth, they might consider work by Bronfenbrenner that frames the various ecological contexts that contribute to child development.**

*This is a nice suggestion. Due to space constraints, our theoretical framing is still relatively brief.*

*However, we have added more discussion of Bronfenbrenner and ecological contexts to our introduction.*

**Reviewer 3**

**This is a clearly written manuscript with analyses that respond to an important policy question regarding differences in teacher reported kindergarten readiness skills of children entering in either 1998 or 2010. The analyses clearly show significant differences between the two cohorts with the 2010 cohort showing higher literacy and math skills upon Kindergarten entry than the 1998 cohort. The analyses also show that there are mixed findings regarding differences in the behavioral outcomes of children in these two cohorts.**

*Thank you for this encouraging feedback and for the excellent suggestions.*

**I have just a couple of comments and suggestions for the author(s):**

**1) in the 3rd question on p. 2, change "non-academic" to "behavioral" as this is how the data are referred to throughout the manuscript**

**2) Under "Background" it may be prudent to qualify the first sentence beginning with "Some research indicates..."**

*Thank you for both of these comments. The text has been changed to reflect your suggestions.*

**It would be interesting to be able to tease apart race and SES by providing results of the interaction between these variables in predicting academic and behavioral outcomes across the two cohorts**

*Thank you for this suggestion. To get at this question, we examined whether our main set of results differed for low and high income children. Rather than running models that included three way interactions between race, SES, and 2010, we simply reran our analyses separately for children in the bottom two SES quintiles, and then the top three. These are presented below in tables L1 and L2 (L for response Letter). While this is not exactly the same as interacting race and SES, we think it answers a similar question. In general, patterns of changes over time are relatively similar among lower income and higher income students, by race, though our estimates get noisier and we have concerns about small cell sizes for particular race by SES combinations. For that reason, and due to space considerations, we have elected not to include these results in the manuscript.*

***On p.21, it may be interesting to compare the 4th grade NAEP scores of both cohorts when they become available for the 2010 cohort  in order to see if the differences are sustained into elementary school--I also appreciate the author(s) plan to use the ECLS data on the cohorts when available***

*We think this is an excellent suggestion. The “ideal” comparison we would like to make here is to compare 4th grade NAEP scores from 2002 to 4th grade scores in 2014, which would reflect scores from the samples that were surveyed in the two ECLS-K waves. Unfortunately, math scores were not collected in 2002, and neither math nor reading scores were collected in 2014. However, we can compare our estimates to NAEP scores across a similar time frame, to get a sense of whether fourth graders are scoring better on NAEP assessments over time (*[*http://www.nationsreportcard.gov/reading\_math\_2013/#/performance-overview*](http://www.nationsreportcard.gov/reading_math_2013/#/performance-overview)*). Fourth-grade NAEP math scores rose from 226 to 242 between 2000 and 2013, with the steepest increase (9 point) between 2000 and 2003. Reading scores also rose over this period from 213 to 222.*

*In the current paper we found across the board improvements in school readiness, but also consistent evidence of narrowing Black-white school readiness gaps. The fourth grade math NAEP results also suggest some narrowing in this gap from 31 in 2000 to 26 in 2013, from 34 to 26 on the fourth grade reading assessments (*[*http://www.nationsreportcard.gov/reading\_math\_2013/#/achievement-gaps*](http://www.nationsreportcard.gov/reading_math_2013/#/achievement-gaps)*)*

*Over this period there was also narrowing in the Hispanic-white test score gaps for both subjects. Socioeconomic gaps, proxied by school lunch eligibility and available since 2003, have been relatively stable.*

*Overall, the patterns in NAEP scores are generally consistent with the findings of the current study. We mention the overarching trends in our discussion, as well we the idea to track the ECLS children over time and assess whether the improvements observed at kindergarten entry are still observed later in elementary school.*

**On p. 22 the author(s) indicate they conducted sensitivity analyses to rule out competing reasons for the differences in academic outcomes, but teacher characteristics are not listed. In particular, there is a 10% difference in the teachers who had a grad degree in the 2010 cohort vs. 1998. It would be good to add this variable if not already included.**

*Thanks for this suggestion. We absolutely agree that changes in teacher characteristics over time are relevant, and for this reason we did in fact include teacher education as a control in the models that indicate they include “Kindergarten teacher characteristics.” In fact, all the variables that we list in Appendix B are included as controls in our models, and they are grouped in the table the same way that they are grouped for inclusion in the models. We regret that we did not explain this point clearly in the earlier draft and have now added text to clarify this issue.*

**In appendix B, descriptive statistics are listed for both cohorts, but the reader does not know which are statistically different. This would be good to add to this table.**

*This is an excellent suggestion. We now indicate whether variables are significantly different across cohorts.*

**Reviewer: 4**

**Comments to the Author**

**Thank you for the opportunity to review the article, “Kids today: Changes in school-readiness in an early childhood era “. This article uses data from the ECLS-K 1998 and 2010 cohorts to contrast teachers’ ratings of kindergartners’ cognitive and behavioral skills in the fall of kindergarten. The aim of the article is to test whether kindergarten readiness has increased and whether achievement gaps have narrowed. The article is well-written but would benefit from additional detail, particularly surrounding the measures and methods of analysis.**

**I have the following suggestions for how to make it better:**

**1.      The authors set up a series of hypotheses in the background section: children who had entered kindergarten in 2010 had more exposure to academic material and more advanced pre-literacy and mathematics skills than those who entered in 1998 because of increased attendance at pre-K and increased parental investment. Furthermore, low-income families may have benefitted more because of greater improvements to their access and quality of early childhood education opportunities. However, the authors do not measure whether preschool attendance and parental investment increased and whether either increased more for poor children than non-poor children. Providing a look at these mechanisms would give the paper more theoretical weight and make it more interesting to a broader readership.**

*This is a good point. We are somewhat constrained due to the table and figure restrictions but we do show how all of the variables included in our model have changed in Appendix B. For example, this appendix highlights the increase in public preschool participation, as well as some striking changes in children’s early home environments. We highlight key trends in the manuscript.*

*While we examined patterns disaggregated by income, and they are intriguing, we do not present these descriptive tables due to space limitations. We do note that a recent working paper provides a very detailed examination of exactly the issue you raise* (Bassok, Finch, Lee, Reardon, & Waldfogel, 2016)*. That paper uses the same datasets here to assess whether income-based gaps in early childhood experiences (both inside and outside the home) have narrowed across the two cohorts, and find substantial support for this hypothesis with respect to home environment. Interestingly, however, that study shows that low-income children actually experienced a drop in formal care participation over this period, perhaps due to the Great Recession. We mention this paper and the relevant findings in our current manuscript.*

**2.      The authors reported using a summary measure of teachers’ ratings of several cognitive skills for literacy and several for math. I would like to see a factor analysis to see whether these skills represent a single factor for literacy and a single factor for math. What is the Cronbach’s alpha on each summary measure?  What does this distribution of each measure look like?**

*The alpha levels of our overall math and literacy measures are extremely high (over .95 for both literacy and math). We conducted factor analyses separately by subject and using the “eigenvalues greater than one” rule found a single math factor and a single literacy factor. Figures L1 and L2 (at the end of this document) present factor analyses for each subject along with the distributions of these variables in each cohort.*

**3.      The authors say that they used a “regression framework”.  What type of regression did they use? What was the distribution of the outcome variables? If OLS regression was used, were the outcome variables normally distributed? I know, from prior experience with these data, that the behavioral measures are not normally distributed, but rather highly skewed.**

*Thank you for raising these questions. In the earlier version of the manuscript we did use Ordinary Least Squares regressions with the behavioral measures. You are correct to point out that the behavioral measures are highly skewed, which makes such an analysis of these measures problematic. To address this issue, in our revision, we follow Grimm et al. (2010) and construct dichotomous variables set to 1 to indicate potentially problematic behavior, which is defined as one standard deviation above the mean for behavioral measures where positive values equal problematic behaviors, and one standard deviation below the mean for those coded in the reverse.*

*We then ran all analyses involving dichotomous outcomes (both academic and behavior) using Linear Probability Models (LPM) as well as logistic regressions. Given the very high level of correspondence across these two types of models, we present results from the LPM throughout for their ease of interpretation. We make note of this in the manuscript and have included results from these logistic regressions in table L3 at the bottom of this letter. For our continuous academic measures, we use OLS.*

**4.      There is no clear description of the regression models for low and high proficiency students. The authors state that they divided the total number of skills at which the student was rated either highly proficient or not proficient by the number of skills. But how do the authors define the low and high proficiency samples? What are the sample sizes for these subsamples?**

*We apologize for not being clear in the text of the manuscript about these analyses. We have revised both the analyses themselves and our explanation in ways that we hope are much clearer. Note we do not define subsamples of students that are low and high proficiency. Instead, in the current manuscript we have constructed dichotomous variables that take on either a value of 0 or a value of 1 for ALL students in our analysis. Essentially, these are models to assess the likelihood that children will be on either end of the skill distribution.*

*We define “low proficiency” to mean that a student was rated either a 1 or 2 on at least half of the skills considered, and “high proficiency” to mean that a student was rated either a 4 or 5 on at least half of the skills. So for instance, a student who was rated a 1 on 4 or more math skills (of 7) would be considered low proficiency. We have described these variables carefully in the revised manuscript. Importantly, our results are robust to varied definitions of low and high proficiency.*

**5.      I want a better rationale for why models included each set of control variables. For example, why include the characteristics of the kindergarten teacher and classroom? I can think of reasons, but it would be helpful for the authors to state these rationales more explicitly.**

*This is a great point, and we now provide far more rationale for each set of controls (e.g. demographic factors, preschool factors, home environment, and teacher characteristics). The rationale for the preschool factors is most obvious, as changes in public investments in preschool over the study period was a key motivator for the analysis. We also know that between 1998 and 2010 the United States experienced some important demographic shifts. For instance, Hispanics made up 14 percent of the adult population in 2010, up from 11 percent in 2000, and accounted for more than half of the population growth in the country over the decade (Passel, Cohn, & Lopez, 2011). We therefore include a set of demographic covariates which may correlate with children’s school readiness outcomes. In a recent working paper, Bassok, Finch, et al. (2016) explore changes over time in children’s early home environments, and highlight substantial increases in home resources, time spent engaging with parents, and parents’ expectations around school readiness. These changes are plausibly linked with changes over time in school readiness, thus their inclusion. Finally, extant literature suggests that teacher characteristics such as education and experience explain some of the variation in teacher ratings of children’s abilities (Mashburn, Hamre, Downer, & Pianta, 2006). We therefore include these variables to assess whether they may partially explain the observed changes in school readiness.*

**6.      What does it mean that the regression coefficients do not change much after accounting for changes in preschool and home environment? It would be helpful to be able to see the coefficients for each of these control variables.**

*We have opted* not *to present all coefficients for all control variables in the analysis because of the large number of models and outcomes considered in the current analysis. For instance, just Panel A of Table 1 includes 24 different regressions, and our fully saturated model (Model 3) includes all the extensive controls presented in Appendix B. We worry that showing the full models would be confusing and difficult to interpret. But the overall question you raise is an important one. What should we make of the fact that an extensive set of observable measures explains very little about the observed findings? The coefficients on the covariates in our analysis are significantly associated with child outcomes in the ways we would expect (e.g. children who do not speak English in the home score lower on our academic measures, children who attended formal care in the year before kindergarten are rated as having stronger academic skills at kindergarten entry). However, these covariates explain little about* changes *in these readiness measures over time. In general, we interpret this as “an omitted variable” issue, and think we have failed to account for important factors that may be driving this result. In exploratory analyses not presented in the current paper, we find that children’s access to learning games on computers as well as parents’ beliefs about academic school readiness explained a non-trivial portion of the changes in literacy and math skills at kindergarten entry. Perhaps we are failing to adequately capture a heightened focus in young children’s homes around academic materials. Another possibility, mentioned in our response to Reviewer 2, is that changes in child care* quality *or child care curriculum* *(which we do not observe in our data) may be very relevant. More research is needed to explore these and other alternatives.*

**7.      The teachers’ ratings of students have correlations of about .5 with the direct measures of achievement. What else goes into teachers’ ratings of students beyond ability? Why do we see a divergence between teachers’ ratings and direct measures? I would like to see more discussion on this topic.**

*Thank you for raising this important point. We have now added more discussion of the literature assessing correspondence between direct and teacher assessments, particularly within an early childhood context. In the current manuscript we note that earlier studies have shown that a portion of the variation in teacher assessments of young children is explained by teacher rather than child characteristics, and further that both teacher characteristics (e.g. education levels, experience) and child characteristics (e.g. race, socio-economic status (SES)) explain some of the variation in teacher assessments* (Kilday, Kinzie, Mashburn, & Whittaker, 2012; Mashburn et al., 2006)*. On the other hand, teacher assessments are the most widely-used, cost-effective, and efficient method for assessing young children. In fact, teachers’ extensive interactions with children may allow them unique insights and knowledge not captured by direct assessments (Epstein, Schweinhart, DeBruin-Parecki, & Robin, 2004).*

**8.      I would like to see sample sizes on all regression models. In particular what are the sample sizes for the 1998 and 2010 samples on each model?**

*Thank you for raising this issue of sample sizes. We use a fixed sample for all regression estimates relating to the academic outcomes, and a second fixed sample for all analyses related to the behavioral measures. The sample for the analyses related to math and literacy outcomes includes approximately 29,550 children, with 16,050 from the 1998 sample and 13,500 from the 2010 sample. For the analyses related to the behavioral measures, the sample includes about 26,650 children, with 14,750 from the 1998 sample and 11,900 from the 2010 sample. We now include sample sizes along with our estimates.*

**9.      The authors report that the teachers saw fewer internalizing behavior problems in the 2010 cohort but the coefficient on internalizing behavior is positive. Only after re-reading the methods section did I see that they had recoded internalizing behavior so that positive scores represent better behavior. This is not standard practice and will confuse readers familiar with these measures. I would at least remind readers of this on the table and in discussion of results.**

*Thanks for this suggestion. As discussed above, we have now revised the way we treat our behavioral outcomes. In particular, we construct indicators for “problem behavior,” which is indicated by HIGHER scores on some outcomes (e.g. externalizing behavior) and by LOWER scores on others (e.g. self control). We are careful to state this clearly in the manuscript.*

**10.     The authors should provide more discussion of why students might have worse Approaches to Learning and better Internalizing Behavior.  Why should we expect these results?**

*That is a good question. We did not expect this pattern, and do not have an obvious explanation. More broadly, we did not have strong directional hypotheses about the behavioral measures and could make plausible cases for changes in either direction (e.g. access to preschool may prepare children for the behavioral expectations of preschool, but the heightened expectations and structured of today’s kindergarten classrooms may be stressful for children).We did, in general, expect the behavioral measures to all move in the same direction.*

*The “approaches to learning” measure includes items such as “shows eagerness to learn,” “works independently,” “persists in completing tasks,” “pays attention well,” etc. We were not surprised to see the drop in this measure, and believe that in part the change here is driven by changes over this period to kindergarten classrooms. In 2010, kindergarteners did much more desk work (text books, worksheets, teacher-directed instruction). They also had far less opportunity for child-selected activities, art, music, and hands-on exploration. In some ways, then, the expectations for “focus” or “persistence” were being evaluated in very different contexts, and it strikes as plausible that the higher expectations in the 2010 challenged young children’s approaches to learning.*

*The improvement with respect to internalizing skills is a puzzle to us and we acknowledge this in the manuscript.*

**11.     Which weights were used in the analyses? How did the authors handle the weights, given that there were appended samples from 1998 and 2010?**

*Our analysis combines data from the fall kindergarten parent and teacher surveys. Ideally we would use weights in both datasets that adjust for non-response associated with both of these surveys. In 1998, we are able to do exactly this. We use the weight C1CPTW0, which is a child-level weight to be used for analysis of fall-kindergarten direct child assessment data combined with fall-kindergarten parent interview data and fall-kindergarten teacher data.*

*However, in 2010 there is no weight that adjusts for non-response associated with both teacher and parent surveys. Since the primary outcomes that we use are teacher reported, we use the weight W1T0, which is a child-level weight designed to be used in the analysis of fall kindergarten teacher data. We combined our weights in keeping with other studies that leverage the two cohorts of the ECLS data. Specifically, in the appended sample, we constructed a new “combined weight” that took on the C1CPTW0 value in 1998 and W1T0 value in 2010 and used this to weight all of our analysis.*

*We have also examined the sensitivity of our results to using different weights. We have found that our results are extremely robust to the use of different weights, and are also quite similar even when no weights are used.*

**12.     Can the authors anchor the changes to something concrete, perhaps by comparing it to effect size results from past research for interventions or by comparing it to normative expectations for growth over time in student achievement? What does it mean substantively that children are now .25 of a SD unit more proficient in math?**

*One way to think about the magnitude of the changes in academic readiness observed in the current study is relative to the growth observed from the fall to the spring of kindergarten for the 1998 cohort. In 1998 students average score on the overall math measure was 2.51 in the fall and 3.62 in the spring. Students in 2010 started kindergarten at an average score of 2.70 in the fall. This suggests that students arriving at kindergarten in 2010 had already learned about 17 percent of what they previously would have learned in kindergarten. For literacy the change is just slightly smaller.*

**13.     I would like to see a richer discussion of what children gain and lose by an earlier focus on academic readiness.**

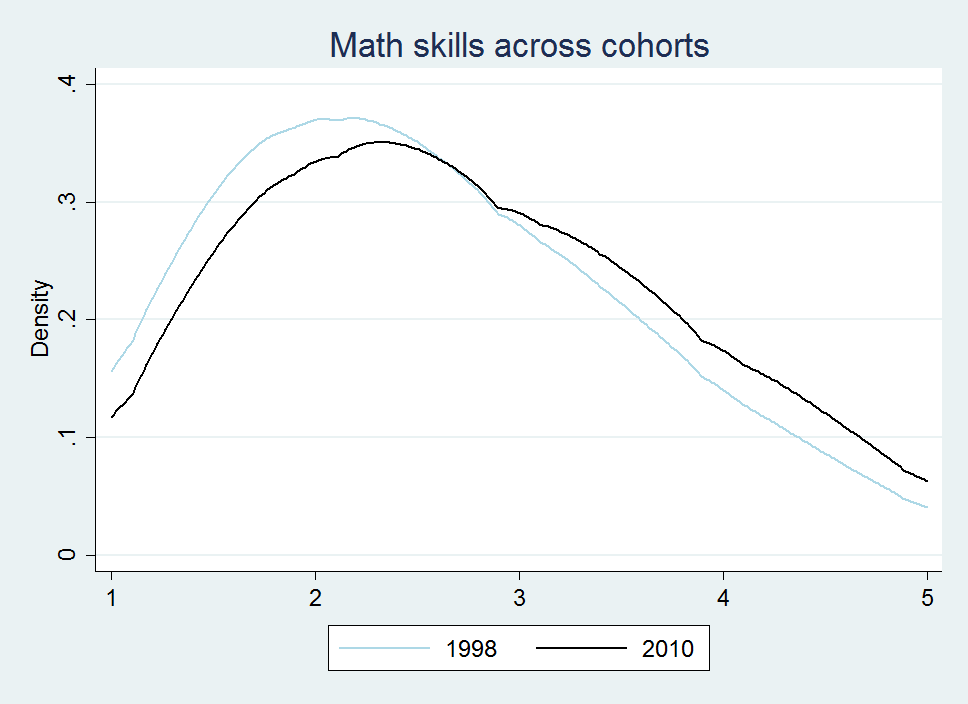
*Thank you for this suggestion. In the introduction we now raise the issue of potential trade-offs from an early focus on academic content. In fact, this notion of potential trade-offs is our motivation for examining behavioral outcomes. We wondered if a heightened focus on academic content in early childhood may be crowding out other important types of early experiences. We come back to this point in the discussion. Ultimately, there is not yet good evidence about the impacts of heightened focus on academics in early childhood, but we certainly agree that there may be some negative unintended consequences and are examining this issue in separate work.*

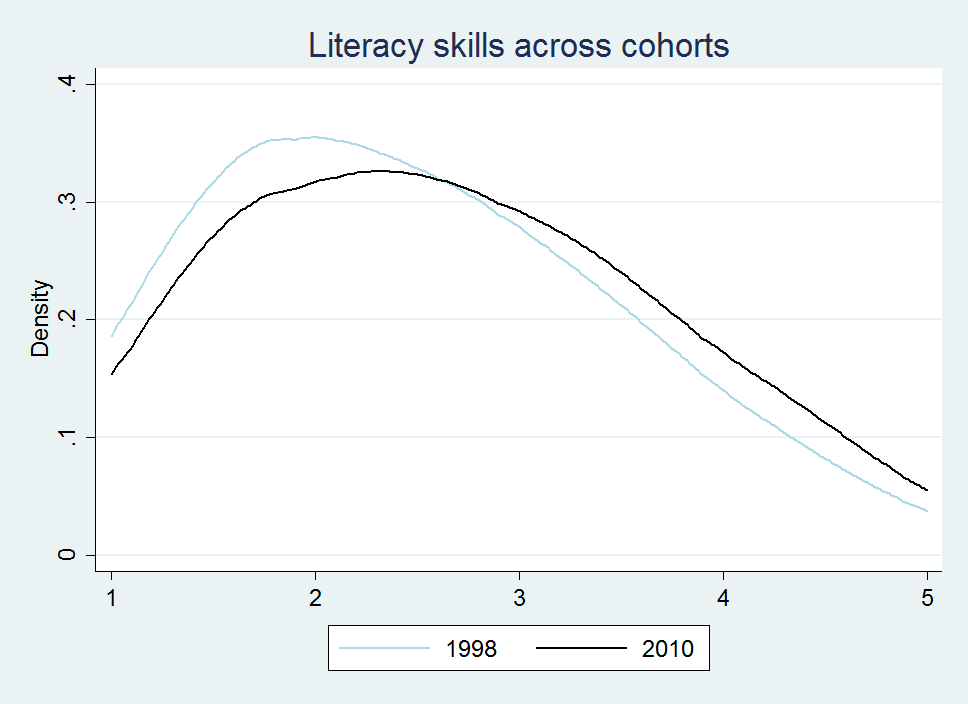
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Table L1. Differences in teacher-rated math and literacy proficiency, by race and SES (OLS estimates)** | | | | | | | | | |  |  |  |  |  |  |
|  | Low SES (bottom 2 quintiles) | | | | |  |  |  | High SES (top 3 quintiles) | | |  |  |  |  |
|  | Math |  |  |  | Literacy |  |  |  | Math |  |  |  | Literacy |  |  |
|  | Overall | Low | High |  | Overall | Low | High |  | Overall | Low | High |  | Overall | Low | High |
| 2010 cohort | 0.26\*\*\* | -0.10\*\*\* | 0.08\*\*\* |  | 0.21\*\*\* | -0.09\*\*\* | 0.04\* |  | 0.23\*\*\* | -0.09\*\*\* | 0.10\*\*\* |  | 0.24\*\*\* | -0.08\*\*\* | 0.06\*\*\* |
|  | (0.04) | (0.02) | (0.02) |  | (0.04) | (0.02) | (0.01) |  | (0.03) | (0.02) | (0.02) |  | (0.03) | (0.01) | (0.01) |
| Black\*2010 | 0.12+ | -0.07\* | 0.04+ |  | 0.14\* | -0.05+ | 0.06\* |  | 0.15\* | -0.04 | 0.05+ |  | 0.10 | -0.04 | 0.05 |
|  | (0.07) | (0.03) | (0.03) |  | (0.06) | (0.03) | (0.02) |  | (0.07) | (0.03) | (0.03) |  | (0.07) | (0.03) | (0.03) |
| Hispanic\*2010 | 0.05 | -0.02 | 0.00 |  | 0.06 | -0.02 | 0.01 |  | 0.10+ | -0.04 | 0.03 |  | 0.06 | -0.05+ | 0.01 |
|  | (0.06) | (0.03) | (0.02) |  | (0.05) | (0.03) | (0.02) |  | (0.06) | (0.03) | (0.02) |  | (0.06) | (0.03) | (0.02) |
| Asian\*2010 | 0.12 | -0.04 | 0.03 |  | 0.20+ | -0.04 | 0.05 |  | -0.01 | 0.02 | -0.02 |  | 0.02 | 0.02 | 0.00 |
|  | (0.13) | (0.06) | (0.05) |  | (0.11) | (0.06) | (0.05) |  | (0.11) | (0.05) | (0.04) |  | (0.10) | (0.04) | (0.04) |
| N | 11100 | 11100 | 11100 |  | 11100 | 11100 | 11100 |  | 17150 | 17150 | 17150 |  | 17150 | 17150 | 17150 |
| Note. Each coefficient comes from a separate regression where outcomes were regressed on an indicator for the 2010 cohort and interactions between this indicator and either race (omitting white) or SES indicators (omitting the highest quintile). Regressions that use race indicators also included "other race" as a category (results not shown). All regressions include controls for age, demographics, preschool attendance, aspects of the home environment, and kindergarten teacher characteristics. | | | | | | | | | | | | | | | |
| \* p<.05 \*\* p<.01 \*\*\*p<.001 | | | | | |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Table L2. Differences in teacher-rated behavioral outcomes, by race and SES (OLS estimates)** | | | | | | | | | | | |
|  | Low SES (bottom 2 quintiles) | | | | |  | High SES (top 3 quintiles) | | | | |
|  | Self control | Interpersonal behavior | Approaches to learning | Externalizing behavior | Internalizing behavior |  | Self control | Interpersonal behavior | Approaches to learning | Externalizing behavior | Internalizing behavior |
| 2010 cohort | 0.02 | 0.01 | 0.08\*\*\* | 0.01 | -0.03\*\* |  | 0.01 | -0.01 | 0.05\*\*\* | -0.01 | -0.02\* |
|  | (0.02) | (0.02) | (0.02) | (0.01) | (0.01) |  | (0.01) | (0.01) | (0.01) | (0.010) | (0.01) |
| Black\*2010 | 0.02 | -0.01 | -0.04 | 0.01 | -0.03 |  | -0.02 | -0.05\* | 0.00 | -0.02 | 0.04\* |
|  | (0.03) | (0.03) | (0.03) | (0.03) | (0.02) |  | (0.03) | (0.02) | (0.03) | (0.020) | (0.02) |
| Hispanic\*2010 | 0.01 | -0.02 | -0.03 | 0.00 | -0.01 |  | 0.00 | 0.00 | -0.01 | -0.02 | 0.00 |
|  | (0.02) | (0.02) | (0.02) | (0.02) | (0.02) |  | (0.02) | (0.02) | (0.02) | (0.020) | (0.02) |
| Asian\*2010 | 0.06 | 0.02 | 0.00 | 0.05 | 0.02 |  | 0.03 | 0.05 | 0.05 | -0.01 | 0.00 |
|  | (0.06) | (0.06) | (0.05) | (0.04) | (0.04) |  | (0.04) | (0.03) | (0.04) | (0.020) | (0.02) |
| N |  |  |  |  |  |  |  |  |  |  |  |
| Note. Each coefficient comes from a separate regression where outcomes were regressed on an indicator for the 2010 cohort and interactions between this indicator and either race (omitting white) or SES indicators (omitting the highest quintile). Regressions that use race indicators also included "other race" as a category (results not shown). All regressions include controls for age, demographics, preschool attendance, aspects of the home environment, and kindergarten teacher characteristics. | | | | | | | | | | | |
| \* p<.05 \*\* p<.01 \*\*\*p<.001 | | | |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Table L3. Differences in teacher-rated student outcomes, across cohorts (Logit estimates)** | | | | | | |  |  |  |  |
| Panel A - Math and literacy proficiency | | |  |  |  | Panel B - Behavioral outcomes | |  |  |  |
|  | (0) | (1) | (2) | (3) |  |  | (0) | (1) | (2) | (3) |
| Math |  |  |  |  |  | Poor self control | 1.11\* | 1.13\* | 1.07 | 1.09 |
| Overall+ | - | - | - | - |  |  | (0.05) | (0.06) | (0.05) | (0.06) |
|  |  |  |  |  |  |  |  |  |  |  |
| Low proficiency | 0.66\*\*\* | 0.61\*\*\* | 0.60\*\*\* | 0.63\*\*\* |  | Poor interpersonal behavior | 0.94 | 0.94 | 0.91+ | 0.97 |
|  | (0.03) | (0.03) | (0.03) | (0.04) |  |  | (0.05) | (0.05) | (0.05) | (0.05) |
| High proficiency | 1.65\*\*\* | 1.81\*\*\* | 1.82\*\*\* | 1.75\*\*\* |  |  |  |  |  |  |
|  | (0.09) | (0.11) | (0.11) | (0.11) |  | Poor approaches to learning | 1.42\*\*\* | 1.45\*\*\* | 1.37\*\*\* | 1.52\*\*\* |
| Literacy |  |  |  |  |  |  | (0.06) | (0.07) | (0.06) | (0.08) |
| Overall+ | - | - | - | - |  |  |  |  |  |  |
|  |  |  |  |  |  | High externalizing behavior | 0.96 | 1.00 | 0.95 | 1.02 |
| Low proficiency | 0.70\*\*\* | 0.63\*\*\* | 0.62\*\*\* | 0.66\*\*\* |  |  | (0.05) | (0.05) | (0.05) | (0.06) |
|  | (0.03) | (0.03) | (0.03) | (0.03) |  |  |  |  |  |  |
| High proficiency | 1.33\*\*\* | 1.45\*\*\* | 1.46\*\*\* | 1.35\*\*\* |  | High internalizing behavior | 0.77\*\*\* | 0.77\*\*\* | 0.76\*\*\* | 0.76\*\*\* |
|  | (0.06) | (0.07) | (0.08) | (0.08) |  |  | (0.04) | (0.04) | (0.04) | (0.05) |
| N | 29550 | 29550 | 29550 | 29550 |  | N | 26650 | 26650 | 26650 | 26650 |
| Age | X | X | X | X |  | Age | X | X | X | X |
| Demographics |  | X | X | X |  | Demographics |  | X | X | X |
| Preschool variables |  |  | X | X |  | Preschool variables |  |  | X | X |
| Home environment variables | |  |  | X |  | Home environment variables |  |  |  | X |
| Teacher/class characteristics | |  |  | X |  | Teacher/class characteristics |  |  |  | X |
| Note. Each coefficient comes from a separate regression where an outcome was regressed on an indicator for the 2010 cohort.  +Measure has been standardized to have mean 0 and SD 1.  \* *p*<.05 \*\* *p*<.01 \*\*\**p*<.001 | | | | | | | | | | |

**Figure L1. Math skills factor analysis and distribution of “overall math” variable**



**Figure L2. Literacy skills factor analysis and distribution of “overall literacy” variable**