2021-06

Is intervention fadeout a scaling artefact?

Is Intervention Fadeout a Scaling Artefact?

Sirui Wan[1], Timothy N. Bond[2], Kevin Lang[3], Douglas H. Clements[4], Julie Sarama[4], and Drew H. Bailey[1]

[1] University of California, Irvine

[2] Purdue University

[3] Boston University

[4] University of Denver

Accepted at *Economics of Education Review* 2/2/2021

Updated version 2/21/2021

Corresponding Author: Drew H. Bailey, School of Education, 3200 Education, University of California, Irvine, Irvine, CA 92697-5500; dhbailey@uci.edu

Acknowledgments: We thank Greg Duncan and Jade Jenkins for feedback on prior drafts and presentations of this project. This research was supported by a Jacobs Foundation Fellowship to D. H. Bailey, by the Institute of Education Sciences, U.S. Department of Education, through Grants R305K05157 and R305A120813 to D. H. Clements and J. Sarama. The content and opinions expressed are those solely of the authors and do not necessarily represent the official views of the U.S. Department of Education.

Abstract

To determine whether scaling decisions might account for fadeout of impacts in early education interventions, we reanalyze data from a well-known early mathematics RCT intervention that showed substantial fadeout in the two years after the intervention ended. We examine how various order-preserving transformations of the scale affect the relative mathematics achievement of the control and experimental groups by age. Although fadeout was robust to most transformations, we were able to eliminate or even reverse fadeout by emphasizing differences in scores near typical levels of first-graders while treating differences elsewhere as unimportant. Such a transformation lowers treatment effects at preschool age and raises them in first grade, relative to the original scale. The findings suggest substantial implications for interpreting the effects of educational interventions.

Keywords: interventions; fadeout; scaling

Is Intervention Fadeout a Scaling Artefact?

I. Introduction

The impacts of early educational interventions on cognitive scores often fade over time, such that there are smaller or no discernible differences between treatment and control children a year or more following the end of treatment (for a review, see Bailey et al., 2020). Such fadeout occurs despite strong theoretical reasons to expect persistent effects, and, in some cases, evidence for beneficial effects on adult outcomes (e.g., Deming, 2009). One explanation for this pattern is that fadeout of cognitive effects is a statistical artefact of the way the test performance is translated to a numerical scale. Perhaps a different way of scaling the test would produce constant or even increasing advantages for treatment-group children, thus showing no fadeout or even amplification of the treatment effect.

Consider an experimental preschool mathematics curriculum that has a persistent effect on children's mathematics learning after the intervention ends. To fix ideas, suppose that children who receive the curriculum learn two more mathematics skills than children in the control group at the end of the one-year intervention, and they remain two skills ahead of children in the control group in each of the subsequent years.[1] If a "two skill" advantage is worth fewer points on later achievement tests relative to earlier years, it will create the illusion of fadeout. For instance, a two-skill deficit may appear large on a kindergarten exam which tests

[1] This of course presupposes that we can clearly quantify the stock of skills a child possesses, which faces many of the same scaling problems we focus on, but makes the example much easier to understand.

only three skills, but small on a graduation exam which tests over three hundred. On the other hand, the two most recently acquired skills could be the margin that determines whether a student passes or fails the graduation exam. A scale emphasizing the former would show strong fadeout, while the latter could conceivably show amplification, and reasonable people might disagree over which metric is more appropriate to measure treatment effects.

Previous studies have suggested that scale choice may cause artificial fadeout. Lang (2010) points out that fadeout can be a mechanical result of the convention of renormalizing each year's scores to have mean zero and variance one. Cascio & Staiger (2012) find evidence of such an effect but conclude that it is only of modest importance. Outside the intervention literature, Bond & Lang (2013, 2018) find that changes in the black-white reading test score gap across grade are highly sensitive to how tests are scaled. They propose that scaling matters when comparing changes across groups, which of course includes studying fadeout of intervention effects.

In this study, we revisit the results of a well-known randomized controlled trial of an early mathematics intervention, the Technology-enhanced, Research-based, Instruction, Assessment, and professional Development (TRIAD) evaluation study, which in the original analysis showed substantial fadeout from spring of preschool through the spring of first grade (Clements et al., 2013). We examine how various order-preserving scale transformations affect the evolution of the mathematics achievement test-score gap between children in the control and treatment groups over this period. We apply both the scaling explanation of fadeout hypothesized by Cascio & Staiger (2012) and Lang (2010) and a complementary set of data-driven methods –

a fadeout minimizing and fadeout maximizing scale – that identify the theoretical bounds of fadeout and persistence, analogously to Bond & Lang (2013).

Fadeout is robust when we constrain scale variance to be constant across waves, and under the scale that maximizes the ability of early scores to predict later scores. We do however find scales within the theoretical bounds that eliminate fadeout. These scales largely shrink differences in performance levels commonly found in preschoolers and enlarge differences for performance at levels typical of first graders, regardless of the age at which the child is tested. As a result, treatment effects at preschool age are lowered while those in first grade are raised, compared with the original scale. We cannot rule out the hypothesis that fadeout is a scaling artefact because we do not know which transformation of the scale is correct (in some sense, they may all be correct). Results across transformations do not consistently point to a specific measurement problem (e.g., measurement error in very high scores) that would make fadeout appear larger than it is.

Perhaps most significantly, the fadeout maximizing transformation focuses our attention on the lack of lasting effects at the lower end of the test-score distribution while the fadeout minimizing transformation focuses on the positive effects at the top of the distribution. We suggest that rather than seeking to determine the proper weighting of these two effects, we should recognize that, in some cases, the proper weighting may follow from the objective of the measurement exercise. In others, it may simply be unknowable.

II. Evidence of and Explanations for Fadeout

Fadeout is common in longitudinal studies of early interventions. For example, many studies

have found fadeout in early mathematics interventions, despite impressive initial effects (Bailey et al., 2018; Clarke et al., 2016; Clements et al., 2013; Hassler Hallstedt et al., 2018; Smith et al., 2013). In a meta-analysis of 67 early childhood education interventions published between 1960 and 2007, impacts on cognitive outcomes fell, on average, by over half in the year after treatment ended, and the meta-analytic estimate was statistically insignificant 2-4 years after treatment ended (Li et al., 2017). The Head Start Impact Study, perhaps the early childhood intervention RCT (randomized controlled trial) best known to economists, also shows little or no effect of Head Start on either cognitive or noncognitive measures in the early school years after the program ended (Puma et al., 2012).

Psychologists have proposed several explanations for fadeout of the effects of initially successful interventions. Cognitive-processing theoretic explanations suggest fadeout results, in part, from children in the treatment group forgetting information they learned from the treatment (Campbell & Frey, 1970; Kang et al., 2019). Alternatively, environmental explanations suggest that, after a successful intervention, children are not exposed to content sufficiently advanced to allow them to build on the extra knowledge they gained (Engel et al., 2013; McCormick et al., 2017).[2]

On the other hand, there are reasons to suspect that cognitive test-score fadeout is misleading. Despite such fadeout, there is good quasi-experimental evidence of long-term effects

[2] Currie and Thomas (1998) make a similar argument for the absence of cognitive effects of Head Start among African American children.

for Head Start on educational attainment and other employment relevant outcomes (Deming, 2009; Garces et al., 2002; Gibbs et al., 2013; Johnson & Jackson, 2017). In several other classic studies of early educational programs such as Abecedarian and Perry, initial fadeout is also followed by long-term impacts on adult outcomes such as educational attainment and reduced incarceration rates (Campbell et al., 2014; Schweinhart et al., 2005). One explanation is that these interventions instead affect noncognitive skills (Deming, 2009; Heckman, 2006) through which these interventions produce long-run impacts.[3] Another possibility is that the interventions influenced cognitive or noncognitive skills in ways that are not reflected on standardized cognitive tests but persist into adulthood.

This raises the possibility that fadeout is at least partially an artefact of measurement and requires no substantive explanation. Perhaps the impact of early interventions does not really fade over time. Instead, fadeout merely reflects defects in how researchers measure learning across development. Specifically, fadeout may be partially an artefact of the way achievement tests are scaled. If we measure the treatment-control test score gaps in standard deviation units (rescaling scores to have mean zero and variance one), which implicitly assumes that the standard deviation of learning is constant over time, we will almost certainly observe fadeout (Lang, 2010). To see this, suppose that on a scale, after the intervention children in the experimental group have cumulative learning with mean μ_0+1 while the controls have mean μ_0.

[3] See also Cunha, Heckman, Lochner, and Masterov (2006), Cunha and Heckman (2007), and Cunha, Heckman, and Schennach (2010).

Assume the variance in learning of the control group is σ^2 and, to keep the example simple, that the control group is much larger so that the variance of the full sample is approximately the control-group variance (i.e., σ^2). A year after the intervention ends, all individuals have obtained additional skills with a mean gain of μ_1 in both groups. Now children in the experimental group have a mean learning of $\mu_0+\mu_1+1$ while the controls have mean $\mu_0+\mu_1$. Because both initial levels and learning vary across students, variance in the full sample one year after the intervention has doubled (i.e., $2*\sigma^2$). Scaling effect sizes to the new population standard deviation will yield a much smaller standardized effect size. Thus, the initial gap measured in standard deviations is $1/\sigma$ while the later gap falls by a factor of $2^{-.5}$, indicating some degree of fadeout.

Two empirical regularities are consistent with the claim that fadeout is partially a scaling artefact. First, Cascio and Staiger (2012) found that variance in knowledge rose as children progress through school, particularly in the early school years. On two different tests, the North Carolina end of grade exams in math for grades 3 to 8 and the Peabody Individual Achievement Tests (PIAT) of math for children age 5 through 14, they find scaling the earlier tests to have the later test's standard deviation reduced fadeout by approximately 20%, arguably a non-trivial amount. However, they did not use a test specially designed for use in early childhood, and so changes in variance across grade may not generalize to measures more sensitive to differences in young children's knowledge.

Second, the natural growth in vertically scaled test scores of both reading and math declines as students age. That is, children's learning, expressed in standard deviations, grows more slowly from year to year. Across a large set of nationally normed tests administered to U.S.

students, the average annual math gain was approximately 1 standard deviation for Grades 1-2, but only .4 standard deviations for Grades 5-6 (Hill et al., 2008). If children actually learn the same amount each year, and gains each year are equally variable and uncorrelated with previous knowledge, a standard deviation encompasses more knowledge among older than younger children. In our example above, if we assume $\mu_0 = \mu_1$, learning is constant between the two periods, but in standardized form children learn μ_0/σ in the first period and $\mu_0/(2^{.5}\sigma)$ in the second.

Researchers have hypothesized that changes in gaps between groups may also be scaling artefacts. Bond and Lang (2013) show that scaling matters for assessing the growth of the black-white achievement gap, which had previously been estimated to emerge only in the early school years, after controlling for other factors, rather than in early childhood (Fryer & Levitt, 2006). They used transformations of the original scale in two large national U.S. samples, the ECLS-K, which Fryer and Levitt (2006) used, and the CNLSY, which Cascio and Staiger (2012) used, to re-estimate black-white gaps. They maximized the growth of the black-white gap by compressing the middle of the distribution of kindergarten achievement. Under that transformation, the maximum possible test gap, computed by assuming that black children had all the lowest scores and white children all the highest, is much larger in 3rd grade than in kindergarten. This contrasts with the baseline scale for which the observed gap as a percentage of the maximum possible gap is similar in kindergarten and 3rd grade. Perhaps most importantly, an *a priori* plausible method for selecting a scale that is comparable across years, choosing the transformation maximizing the correlation between kindergarten and grade 3 achievement

scores, decreases the growth of the gap in both datasets and reverses it in one.

Bond and Lang (2018) rescale test scores in the CNLSY by tying them to an external metric so that a one unit change in the new scale corresponds to a one-year difference in predicted education. They show that the black-white gap based on this predicted outcome is constant from kindergarten through grade 7. Indeed, much of the variance in kindergarten scores on the test they use is measurement error so that the apparent growth in the gap is largely an artefact of dividing by a standard deviation of test scores that includes more measurement error in the early grades. This finding is consistent with the work by Cascio and Staiger (2012), which demonstrated that the reliability ratio for kindergarten and first grade scores is relatively low while measurement error is less of an issue in upper grades.

These findings have unclear implications for the possibility that fadeout is a measurement artefact. On one hand, they show that relying on a single scale for evidence of changes in test score gaps across time may yield misleading results. However, if low reliability of psychometric measures administered to young children is general rather than particular to the tests they used, initial treatment effects would be plausibly *under*- rather than over-estimated, leading to an underestimate, not an overestimate, of fadeout. Additionally, fadeout has been observed after mathematics interventions including children from preschool to grade 3 (Bailey et al., 2018; Clarke et al., 2016; Clements et al., 2013; Smith et al., 2013), and grades 3-6 (Jacob et al., 2010) and grades 6-8 (Taylor, 2014), grades for which Bond and Lang do not find decreasing measurement error.

However, the direction of bias in estimates of the change in the test score gap over time

may be test specific. Bond and Lang (2018) also show that, in these same data, there exists a large racial gap on the PPVT, a test of general aptitude given to children before they enrolled in kindergarten. They do not address whether measurement error in the PPVT changes as children age. One important factor to consider is the age range for which the test was developed: if a test is optimized to measure knowledge in grade *G*, measurement error might be higher in grade *G*-1 and grade *G*+1. Finally, differences in measurement error are far from the only reason to be interested in alternative test scales. The shape of the relation between number correct and the value of learning could vary wildly under different test designs.[4]

III. Data: The TRIAD Evaluation Study

We focus on one of the most well-known examples of intervention fadeout in the education literature, the RCT evaluation of the Technology-enhanced, Research-based, Instruction, Assessment, and professional Development (TRIAD) scale-up model (Clements, Sarama, Spitler et al., 2011; Clements et al., 2013; Sarama et al., 2012). The RCT evaluated the efficacy of the TRIAD scale-up model's implementation, in which the *Building Blocks* early mathematics curriculum was used.

The TRIAD scale-up model was created to help children develop conceptual understanding, procedural skill, and problem solving competencies in various foundational areas of mathematics (e.g., counting, comparing number, measurement, and geometry). It was

[4] One other scaling choice that economists might also be interested in is scales that reflect differences in the extent to which items predict, or better, *affect* outcomes of interest. This approach requires access to data on long-term outcomes unavailable for the TRIAD study.

designed to teach at the zone of proximal development for children who come in with low scores but move them quickly to more advanced levels—as quickly as possible for the group or the individual. Perhaps more important, even the so-called basic skills are taught within a conceptual and the problem-solving framework for that development. Impacts at the end of treatment were nearly identical for children with higher or lower achievement test scores (Clements, Sarama, Spitler et al., 2011). The *Building Blocks* curriculum included the *Building Blocks* software, which further helped teachers personalize instruction to each child's unique needs. The curriculum was designed to take approximately 15 to 30 minutes each day. Implementation of the curriculum was assessed with two instruments, the Building Blocks Fidelity of Implementation (Fidelity) and Classroom Observation of Early Mathematics Environment and Teaching (COEMET). Both instruments measured how mathematics was taught in each classroom. Previous studies using this dataset showed that the instruments have high reliability and validity, and teachers implemented the curriculum with adequate fidelity (Clements & Sarama, 2008; Clements, Sarama, Spitler et al., 2011).

In TRIAD, forty-two low-income schools in Buffalo, NY, and Boston, MA, were randomly assigned to one of three conditions: 1) control (n = 391; school N = 16), 2) treatment in preschool (n = 494; school N = 14), or 3) treatment in preschool with follow-through (n = 420; school N = 12). In both treatment conditions, teachers received pedagogical development and implemented the *Building Blocks* mathematics curriculum during preschool. For schools in the treatment with follow-through condition, the kindergarten and first grade teachers also received additional professional development on mathematics learning trajectories. Schools in the control

condition kept their pre-existing preschool mathematics programs. We use only the first two groups, because previous research shows the largest fadeout occurred between these two groups (Clements et al., 2013). [5]

The TRIAD study provides an ideal setting for assessing the extent in which intervention fadeout can be a scaling artefact. First, as a well-designed RCT, there are few plausible threats to internal validity. Second, there was substantial fadeout of the initial treatment effects. It is thus clear that based on the scale used in the initial evaluation that fadeout (and not, for example, publication bias or time-specific sampling error) occurred. [6] Third, it is a vertically scaled achievement test, and thus in line with the assessments on which prior work in this area has focused. Finally, it was an intervention during the early school years, the same period for which Bond and Lang (2013, 2018) and Cascio and Staiger (2012), as well as much of the fadeout literature focuses.

Further, understanding the nature of fadeout in the TRIAD study is of considerable policy

---

[5] After the initial RCT assignment, two schools crossed over from the treatment group to the control group and two schools crossed over from the control group to the treatment group. We use the original treatment assignments (intent-to-treat) for our main results; and then demonstrate their robustness to alternative formulations that acknowledge this crossover. After dropping the schools that crossed over to the treatment follow-through group (condition 3), there are 834 students who were originally assigned to one of three conditions: 1) control (n = 361; school N = 15), 2) treatment in preschool (n = 399; school N = 13), or 3) treatment in preschool with follow-through (n = 74; school N = 2). We combined the two treatment groups.

[6] This stands in surprising contrast to the Head Start Impact Study. Although these data are widely used by economists and policy researchers (e.g., Bitler et al., 2014; Feller et al., 2016; Kline & Walters, 2016) and the sample size in the Head Start Impact Study is large, the fadeout effect is surprisingly small (Gibbs et al., 2011; Puma et al., 2012), because it is limited by the size of the initial impact. We compare TRIAD with the Head Start Impact Study in Table S1 in the Appendix.

significance. As of May 5, 2020, *Building Blocks*, the curriculum used within TRIAD, is the most positively evaluated preschool mathematics curriculum in the Institute of Education Sciences What Works Clearinghouse, with an improvement index of +36 on a scale from -50 to +50. Yet, the TRIAD finding of substantial fadeout has also been a significant source of concern, triggering additional research to address the fadeout (Bailey et al., 2016; Kang et al., 2019). If there was no fadeout, then these attempts to understand and address fadeout are misguided. On the other hand, if fadeout was underestimated in the original TRIAD study, these attempts are even more necessary and urgent.

We limit the sample to children with mathematics achievement scores at spring of preschool, spring of kindergarten, and spring of first grade. This affects approximately 12% of the control and 15% of the treatment group, resulting in a final sample of 720 observations. There were no significant group differences in attrition (see Table S2 in the Appendix).[7] See Figure 1 for the cumulative distribution functions and Figure S1 in the Appendix for the density plots of scores for both groups. Table 1 gives descriptive statistics for this sample. The *p*-values are calculated using simple regression models with clustered standard errors at the school level (for math scores and age), and logit with clustered standard errors at the school level (for sex and ethnicity). The results are unchanged if we use a linear probability model instead of logit. The table shows that the original randomization, after attrition, matched the control groups well in

[7] By looking at the regression estimates of the interaction variable on spring preschool math scores, we can see that the difference between leavers and stayers is the same for the control and treatment groups.

terms of sex, age, and ethnicity.

During the spring of preschool, spring of kindergarten, and spring of first grade, mathematics achievement was assessed using the Research-based Early Math Assessment (REMA). The REMA was designed to measure the mathematics understanding of children between age 3 and 8 (Clements et al., 2008). It was administered through two one-on-one interviews, which were taped and coded, and students were rated on both their correctness and strategy use. Test items were ordered by difficulty. Testing ceased after children answered four consecutive questions incorrectly. The exam covered counting, number recognition, addition and subtraction, patterning, measurement, and shape recognition. This measure defined mathematics achievement as a latent trait using the Rasch model, a one parameter IRT model**,** which ideally would allow for accurate comparisons of scores between groups and across ages.[8] The measure has been found to have high internal consistency (Cronbach's $\alpha = .94$) and to correlate highly ($r = .74$) with the Woodcock-Johnson Applied Problems subtest (Clements, Sarama, & Wolfe, 2011).

Table 1 provides the basic evidence for fadeout. At the end of the intervention (spring of preschool) when the children were, on average, five years old, the score difference between treatment group and control group is .414 on the Rasch scale, but falls to .184 a year later and to .099 two years later, at which point it ceases to be statistically significant. We note, however,

[8] Assuming the model is correctly specified, an IRT scale ensures that a score of *t* in preschool reflects the same level of mathematics as the same score in another grade. However, we do not know whether the assumptions of the Rasch model are correct.

that if as discussed in section II, the true variance is increasing over time, this will lead to us finding fadeout.

IV. Methods

We take two broad approaches to testing whether fadeout is a scaling artefact. We refer to the first as a theory-driven approach, because we start with reasons why fadeout might be a scaling artefact, and then attempt to correct for these potential explanations and test whether we still observe fadeout. For example, Casio and Staiger (2012) start with the intuitive hypothesis that the variance of knowledge is increasing across grades and examined whether it could account for fadeout of intervention effects in test scores. They attempt to adjust for this potential statistical artefact by modeling variance of measures based on test reliability across grades and rescaling the test to the time specific standard deviation.

In addition, we adopt a strategy in Bond and Lang (2013) which transforms earlier and later test scores to maximize the correlation between the two. Implicitly, this maximizes the objective of making a linear function of the initial test score the best linear prediction of the second test score and vice versa. Thus, it chooses the scale that maximizes the reliability of the tests. If the underlying latent variable we are trying to capture is ability to do well on tests of this form, this appears to be a sensible rescaling, but there is no guarantee that it will improve the validity of the scale with respect to some other interpretation of the underlying latent variable such as mathematics achievement. In this paper, we use both variance-equating and correlation-maximization to test for the robustness of fadeout.

We also use a data-driven approach, similar to Bond and Lang (2013): we search for

monotonic transformations of the published scale which maximize or minimize fadeout, and then assess what the new scales imply about whether fadeout is plausibly a scaling artefact.

Although both approaches may provide useful evidence about whether fadeout is a substantive phenomenon or a measurement artefact, conclusions can be strongest when these different methods yield converging results. For example, in one analysis, Bond and Lang (2013) find that both maximizing the cross-test correlation and minimizing the growth of black-white test score gap lead to reducing the gap in later grades while keeping the gap at entry grade similar to that of the original scale. In other words, the black-white gap does not widen if it is measured based on predicted future outcomes or based on the growth-minimizing transformation. Both the theory-driven and data-driven transformations in Bond and Lang (2013) provide support for the idea that the growth of the test score gap is an artefact of higher measurement error in the early grades.

We use the same method as in Bond and Lang (2013) when finding our data-driven transformations. We bound the amount of fadeout by searching for two transformations: i) one that maximizes the growth of the treatment-control test score gap (and thus minimizes fadeout) from the end of treatment in preschool to the spring of grade 1; ii) one that minimizes the growth of the gap (and maximizes fadeout). We define the test score gap as the standardized treatment effect: the difference between the mean mathematics test scores of the treatment group and control group divided by the overall standard deviation of test scores of the full sample in that grade. To impose smoothness, we use a sixth-degree polynomial given by $T(u) = \beta_0 + \beta_1(u - k) + \beta_2(u - k)^2 + \beta_3(u - k)^3 + \beta_4(u - k)^4 + \beta_5(u - k)^5 + \beta_6(u - k)^6$,

where T is the transformed score, u represents untransformed score, $\beta_0 - \beta_6$ and k are constants. Given its flexibility, the sixth-degree polynomial can approximate a wide array of continuous functions.

We use an optimization function in Stata/SE 14.0 to search for the values of $\beta_0 - \beta_6$ and k that minimize the objective function given by $D_{min} = \min (G_1 - G_p)$, where G_1 is the test score gap in grade 1, G_p is the test gap in preschool, and D is the growth of the treatment-control test score gap from preschool to grade 1. Similarly, we maximize the objective function given by D_{max}. Since the sixth-degree polynomial does not require monotonicity, a property making direction of the change in transformed score stay the same as the change in original score, our algorithm checks for it and rejects parameters that violate the condition.[9] Figures 2 and 3 show the densities of the scores of different scales, while Figure 4 shows the relation among them.

In a complementary set of analyses designed to test the limits of how much fadeout can be manipulated, we relax smoothness and only assume monotonicity of our transformations, making the approach even more data-driven. We discretize the scale by obtaining percentile ranks associated with each test score across preschool and first grade in the data. We impose the transformation, $T(u+1) = T(u) + a_{u+1}^2$, where T represents the transformed score, u represents

[9] We adopt the algorithm used in Bond and Lang (2013). In the algorithm, if monotonicity fails at any score within the range of observed scores, the objective function is penalized one unit. This will lead to a discontinuity in the objective function and create many local minima or maxima. Therefore, we tried several different starting values and picked the best one. Chang (2019) recently developed a Stata routine to apply the Bond-Lang algorithm which uses 1,000 different starting values. According to Chang (private communication), his routine, when applied to Bond and Lang (2013), yields the same optimal values in nearly every case and no substantive differences.

score in the discrete scale, and a_{u+1} is a real number. We again use an optimization function in Stata/SE 14.0 to search for the values of $a_2 - a_{100}$ that minimize (or maximize) the objective function given by $D_{\min}$ (or $D_{\max}$). The histograms of the transformation scores are presented in Figure S2 and Figure S3 in the Appendix, and the relation among them is displayed in Figure 5.

For our theory-driven approach, we try both the transformation that equates the variance of scores over time and the transformation that maximizes the correlation between preschool and 1st grade scores. Following Cascio and Staiger (2012), the variance-equating transformation scales down the baseline-scale treatment impacts by multiplying them and the ratio of the earlier to the later standard deviation of test scores. [10] Thus, $T(u_1) = u_1 * SD_p/SD_1$, where T represents the transformed score, u_1 represents untransformed score in grade 1, SD_p and SD_1 are the standard deviations of test scores in preschool and grade 1 respectively. For the correlation maximization transformation, we use a sixth-degree polynomial given by $T(u) = \beta_0 + \beta_1(u - k) + \beta_2(u - k)^2 + \beta_3(u - k)^3 + \beta_4(u - k)^4 + \beta_5(u - k)^5 + \beta_6(u - k)^6$, where T is the transformed score, u represents untransformed score, $\beta_0 - \beta_6$ and k are constants. We use an optimization function in Stata/SE 14.0 to search for the values of $\beta_0 - \beta_6$ and k that maximize the correlation between end of treatment and grade 1 scores in the control group.

V. Results

Figure 1 shows the cumulative distribution functions of scores, re-normed to range from

[10] Since the reliability for the REMA is similar across age groups in our study, we do not adjust for it like what Cascio and Staiger did in their original study.

0 (the lowest score in the spring of preschool) to 1 (the highest score in the spring of first grade), for both groups (see Figure S1 in the Appendix for the density plots). It is evident that in the early period no student scores above .7 while in the later period no student scores below .4. Moreover, in the initial period, we observe first-order stochastic dominance (FOSD), a necessary and sufficient condition for ensuring that all scales rank average performance in the same way. Thus, any scale that distinguishes among scores in the 0 to .7 range will show a positive effect of the intervention right after its completion. In particular, no treated student scores below about .15. We do not observe stochastic dominance two years following treatment, but we do observe a higher density of scores among the treated group almost everywhere above about .7. It follows that we can get a very large fadeout effect if we treat the differences between scores below about .4 as very large and those above .7 as unimportant. In that case, we will observe a very big score gap between treatment group and control group in preschool but a nearly zero gap in grade 1. In contrast, if differences in scores below .7 are minimal, there is almost no immediate treatment effect, but we can choose values of the remaining scores that produce a large long-term effect. The remainder of our analysis largely formalizes these intuitions.

The relation between the original and polynomial transformed scales is shown in Figure 4. As suggested by Figure 1, the fadeout-maximizing scale treats differences in the baseline scale scores between roughly .58 and .85 as essentially unimportant. In contrast, the transformation that minimizes the fadeout effect does its best to eliminate all meaningful differences in scores at the end of preschool while emphasizing the importance of differences in the scores that no student obtains in preschool.

The fact that the fadeout-maximizing scale emphasizes differences in scores close to 1 while the fadeout-minimizing scale does so at very low scores may reflect the requirement of continuity and the restriction on curvature imposed by the polynomial. To address this directly, Figure 5 shows the test scores using a discretized scale. As we surmised based on Figure 1, the fadeout-maximizing scale treats differences among scores below roughly .2 as inconsequential while magnifying differences among scores between roughly .25 and .65. In contrast, the fadeout-minimizing scale eliminates all differences among the scores received by preschoolers and magnifies the differences among most scores in the upper range.

Finally, the transformation that maximizes the correlation between end of treatment and first grade scores in the control group produces results almost identical to those obtained with the untransformed scores. So does the transformation that constrains variance to be constant across grades. Table 2 shows the mathematics test score gap from the spring of preschool through the spring of first grade under the original scale and different transformed scales. The Baseline scale column shows the pre-transformation pattern. The treatment effect decreases from .578 SD in preschool to .146 SD by 1st grade. We observe a similar pattern under the fadeout-maximizing polynomial transformed scale, but the effect in first grade is approximately 0. The change is even more extreme when we allow for discrete jumps in the scale (column 6); the gap at the end of 1st grade is reversed although it is not statistically significant. The fadeout-minimizing polynomial transformation reduces the gap at the spring of preschool to .304 SD, but the gap in first grade remains close to the one obtained by the baseline scale. However, if we allow for a discrete scale, the gap becomes amplification over time, although the growth in the gap is not statistically

significant.[11] We test the robustness of the results by using the sample including only preschool and kindergarten scores (see Tables S3 and S4 in the Appendix), and by searching for the transformation that either maximizes the preschool treatment effect or maximizes the first-grade treatment effect (see Table S5 in the Appendix). We found that patterns of fadeout were robust under these conditions. [12]

As mentioned in Bond and Lang (2013), the choice of different transformations may limit the potential magnitude of between-group differences. For example, consider an intervention conducted on 100 students equally divided between treatment and control, and then assessed on a two-point scale (proficient and non-proficient). The treatment effect will always be largest whenever all 50 treated students outperform the best performing control student, but the measurement of this outperformance will depend on how the scale defines "proficient." If "proficient" were defined as the top 50% of students, then the maximum observable treatment effect would be two standard deviations. If instead "proficient" were defined as the top 75% of students, the maximum treatment effect would be approximately 1.1 standard deviations. We assess the extent to which our scales can mechanically create larger or smaller gaps by performing a similar exercise. We calculate for each grade, on each scale, what the size of the

---

[11] While the density plot shows no students in preschool scoring above approximately .7, this is not quite accurate. The discrete scale thus does produce a small difference between the mean scores of the two groups and a small standard deviation rather than producing something undefined.

[12] We also adjusted for the crossover of schools by using the original assignment as an instrument for the treatment (see Table S6 in the Appendix). Patterns of fadeout were robust under these specifications across all transformations we considered.

test gap would be if the highest performing observed scores all came from the treatment group. We then calculate the proportion of this maximum our gaps are in practice.

Table 3 shows the results of this exercise. Again, results are similar across most transformations: the treatment effect is about 1/3 of the maximum possible gap in preschool and less than 10% of the maximum possible gap in first grade. Under the baseline and correlation maximizing polynomial transformed scales, the maximum possible gap is nearly identical at both waves (about 1.59-1.60 SD in preschool and about 1.61-1.64 SD in first grade; see columns 1 and 4), suggesting that restriction of range is not causing gaps at one wave to be underestimated compared to gaps at another wave.

In the fadeout-minimizing polynomial transformed scale, the maximum possible gap in preschool drops to .50 SD, approximately 1/3 of the baseline maximum possible gap (column 3). Notably, under this transformation, fadeout is actually *larger*, expressed as a difference in proportions of the maximum possible gap (i.e., from 60.6% in preschool to 9.1% in first grade), than it is under other transformations. As with the polynomial transformations, the fadeout is actually larger under the fadeout-minimizing discrete transformation, wherein the maximum possible gap is reduced dramatically (from 74.2% in preschool to 10.9% in first grade; see column 6). The variance in the fadeout-minimizing discrete transformation is greatly increased (i.e., by 1016.9%), such that the maximum possible test gap increases by a factor of greater than 9, from .20 SD at the spring of preschool to 1.82 SD in first grade. Although we cannot rule out the possibility that the variance of performance is approximately an order of magnitude higher at the end of first grade than at the end of pre-K, this exercise certainly raises the concern that the

changes in the treatment-control gap produced by these transformations reflect changes in scale sensitivity instead of changes in the real achievement gap.

Histograms of test scores at the end of treatment in preschool and spring of first grade for each discrete values-transformed scale are displayed in Figures S2 and S3 in the Appendix. The relation between the original and discrete values transformed scales appears in Figure 5. The discrete fadeout-minimizing transformation is a more severe version of the polynomial fadeout-minimizing transformation: it almost fully compresses the part of the distribution where preschool scores fall, and increases the test's sensitivity in the area of the distribution in which the treatment group continued to outperform the control group in first grade in Figure 1.

VI. Discussion and Conclusions

We show that the fadeout of the effect of a preschool mathematics intervention is preserved across most of the monotonic transformations we considered. Using the discrete fadeout-minimizing transformation, fadeout was eliminated because variance during the preschool year was nearly eliminated.

In some respects, our findings resemble Bond and Lang's (2013) investigation of the robustness of growth of the black-white test score gap across different scales. In both studies, scales that limited the extent of the maximum possible gap in the early years produced smaller early gaps and more positive (in Bond & Lang) or less negative (our study) gap growth. Bond and Lang's gap growth-maximizing transformation and our fadeout-minimization transformation best exemplified this pattern. However, the evidence for scaling artefacts differs across these two studies. While Bond and Lang found converging evidence across these theory-driven and data-

driven transformations that the growth of black-white score gap is at least partially a measurement artefact, we found that almost all rescaling choices show nontrivial fadeout of intervention effects.

We can eliminate or even reverse fadeout by using a scale that assigns minimal importance to variation in the range of achievement we observe in preschool regardless of whether students are in preschool or 1st grade but which emphasizes the importance of variation near typical levels of achievement for 1st graders. In effect, this lowers treatment effects at preschool age and raises them in first grade, relative to the original scale.

Given the lack of converging evidence for fadeout's sensitivity to scaling decisions across other approaches, should we seriously consider this scale? Under the fadeout minimizing scale, the lower range of scores among 1st graders is not supposed to capture much variance relevant to later math learning, which means, variation within low 1st grade scores within the control group might not predict later math scores. We test this by examining the relation between preschool, 1st grade, and 5th grade test scores in the control group (see Figure S4 in the Appendix). We find that earlier scores are predictive of 5th grade scores across the distribution of early scores, providing no support for the interpretation that fadeout is an artefact of 1st grade test scores only showing meaningful variance above some threshold.

On the other hand, why might the true score variance of math achievement on a vertically scaled test increase dramatically in a two-year period? One possibility is that first-grade mathematics knowledge is substantially more cognitively complex than preschool mathematics knowledge. The idea has some face validity: for example, a first-grader might be asked to solve

the problem “8 + _ = 11”. Variation in item responses to this question could depend on variation in a variety of underlying knowledge states, such as knowing the meaning of the equal sign, the ability to visualize the problem, the ability to break 8 + 3 into the easier two problems “8 + 2” and “10 + 1”, and/or the ability to symbolically translate the problem to “11 – 8 = _”. This problem shares demands with a problem that might be asked of a kindergartener, “Which is larger: 8 or 3?”, in that both problems may require students to know the meaning of the symbols “8” and “3”, but the former problem requires additional cognitive processing. Indeed, there is some evidence that vertically scaled achievement tests administered to older children inadequately account for increases in item complexity, underestimating growth in math achievement across years (Bolt, Deng, & Lee, 2014).

Assuming for the sake of argument that the discrete fadeout-minimizing scale realistically expresses individual differences in math achievement and that this carries over to the tests that have been used in other studies, this has important implications for the study of individual differences and educational interventions. Almost by definition if differences in math achievement at levels associated with preschool are minimal, there cannot be gaps associated with race or class at this age. Only when differences become meaningful can gaps emerge. Thus gaps (e.g., by class and race), when measured in standard deviations, would be substantially overestimated in earlier years relative to later years relative to the true gap in achievement on some absolute scale.

An alternative interpretation of this result is that fadeout may be reconceptualized as different items measuring different knowledge states. It seems that the discrete fadeout-

minimizing transformation turns the test into a measure of first grade math achievement instead of a measure of math achievement across years. Fadeout may be conceptualized as a consequence of small effects of marginal changes to the skills comprising earlier math achievement on the skills comprising later math achievement.

Like Bond and Lang (2013), we cannot know what the true effect sizes are. But, our analysis of fadeout under different scales points to more fruitful way of looking at the data than trying to determine "the one true measure of fadeout." As we note in the introduction, in some sense each scale could be correct. However, taken together, they make it clear that the TRIAD intervention did not have durable effects at or below test scores typical of first graders. On the other hand, it appears to have had some lasting effects on test scores reflecting high levels of performance. Determining a single measure of fadeout requires taking a stand on the appropriate weighting of these two findings, which seems more scientifically challenging than simply acknowledging both.

We can imagine numerous objectives a policymaker might be pursuing by adopting TRIAD. If it is ensuring that "no child is left behind," then the current implementation must be adjusted to address fadeout. If it is to induce meaningful improvement for first graders who would otherwise perform near average, fadeout seems to be less of a concern. If it is to improve preparation for primary school math, it appears to have had mixed success. If it is to increase some latent achievement in mathematics, then we believe the answer is unknowable given our findings.

But if the goal is to understand whether the difference in average scores changes, except

for the fadeout-minimizing scale that effectively disregards differences in preschool scores, our analysis strongly points to the gap declining (i.e., fadeout is real) in this particular study. As to whether fadeout over the full range of scores is over- or under-estimated by the baseline scale, our results do not strongly support either possibility.

Table 1

Descriptive statistics

Variables	Treatment group	Control group	Group differences	*p* value for group differences
Spring of preschool math	-1.846	-2.260	.414	.002
	(.651)	(.729)		
Spring of kindergarten math	-1.025	-1.209	.184	.083
	(.651)	(.675)		
Spring of 1st grade math	-.064	-.163	.099	.346
	(.677)	(.673)		
Male	.512	.500	.012	.721
Ethnicity				
Black	.535	.494	.041	.749
Hispanic	.174	.261	-.087	.372
White	.256	.170	.086	.409
Ethnicity- Other	.035	.075	-.040	.190
Age (years) fall preschool	4.347	4.377	-.030	.643
Observations	402	318		

Note. Standard deviations are in parentheses for variables. *p*-values indicate the extent to which treatment participants were different from controls on each variable. They are calculated using simple regression models (for math scores and age) and logit (for sex and ethnicity) with clustered standard errors at the school level (n = 30 schools). The math scores were standardized such that a score of "0" approximates the achievement level of a student in first grade.

Table 2

Evolution of the Treatment-Control test gap under various polynomial transformations and discrete transformations of math scores

Variables	Baseline scale (1)	Fadeout effect maximization polynomial (2)	Fadeout effect minimization polynomial (3)	Correlation maximization polynomial (4)	Baseline (constant variance) (5)	Fadeout effect maximization discrete (6)	Fadeout effect minimization discrete (7)
Spring of preschool math	.578**	.570**	.304**	.578**	.578**	.613***	.147
	(.165)	(.167)	(.104)	(.163)	(.165)	(.156)	(.076)
Spring of kindergarten math	.277	.285	.235	.277	.297	-	-
	(.154)	(.142)	(.150)	(.154)	(.165)		
Spring of 1st grade math	.146	.034	.154	.147	.155	-.038	.198
	(.152)	(.118)	(.150)	(.151)	(.162)	(.106)	(.145)
Fadeout effect	.432***	.536***	.150	.431***	.423***	.651***	-.051

Note. Test gaps are measured in standard deviations, and standard errors are in parentheses. Fadeout effect = gap in preschool math – gap in 1st grade math. $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$.

Table 3

The Treatment-Control test gap as a percentage of boundary under various polynomial transformations and discrete transformations of math scores

Variables	Baseline scale (1)	Fadeout effect maximization polynomial (2)	Fadeout effect minimization polynomial (3)	Correlation maximization polynomial (4)	Fadeout effect maximization discrete (5)	Fadeout effect minimization discrete (6)
Spring of preschool T-C test gap	.578	.570	.304	.578	.613	.147
Spring of preschool maximum test gap	1.589	1.619	.502	1.595	1.622	.198
Spring of preschool % of maximum gap	36.4%	35.2%	60.6%	36.2%	37.8%	74.2%
Spring of 1st grade T-C test gap	.146	.034	.154	.147	-.038	.198
Spring of 1st grade maximum test gap	1.605	.759	1.688	1.637	1.046	1.819
Spring of 1st grade % of maximum gap	9.1%	4.5%	9.1%	9.0%	3.6%	10.9%
% of 1st grade SD to preschool SD	94.4%	44.0%	59.8%	107.4%	67.5%	1016.9%

Note. T-C = Treatment-Control, SD = Standard Deviation. Test gaps are measured in standard deviations. Maximum test gap is the test gap that would be observed if all the lowest scores belonged to control group and all the highest scores belonged to treatment group.

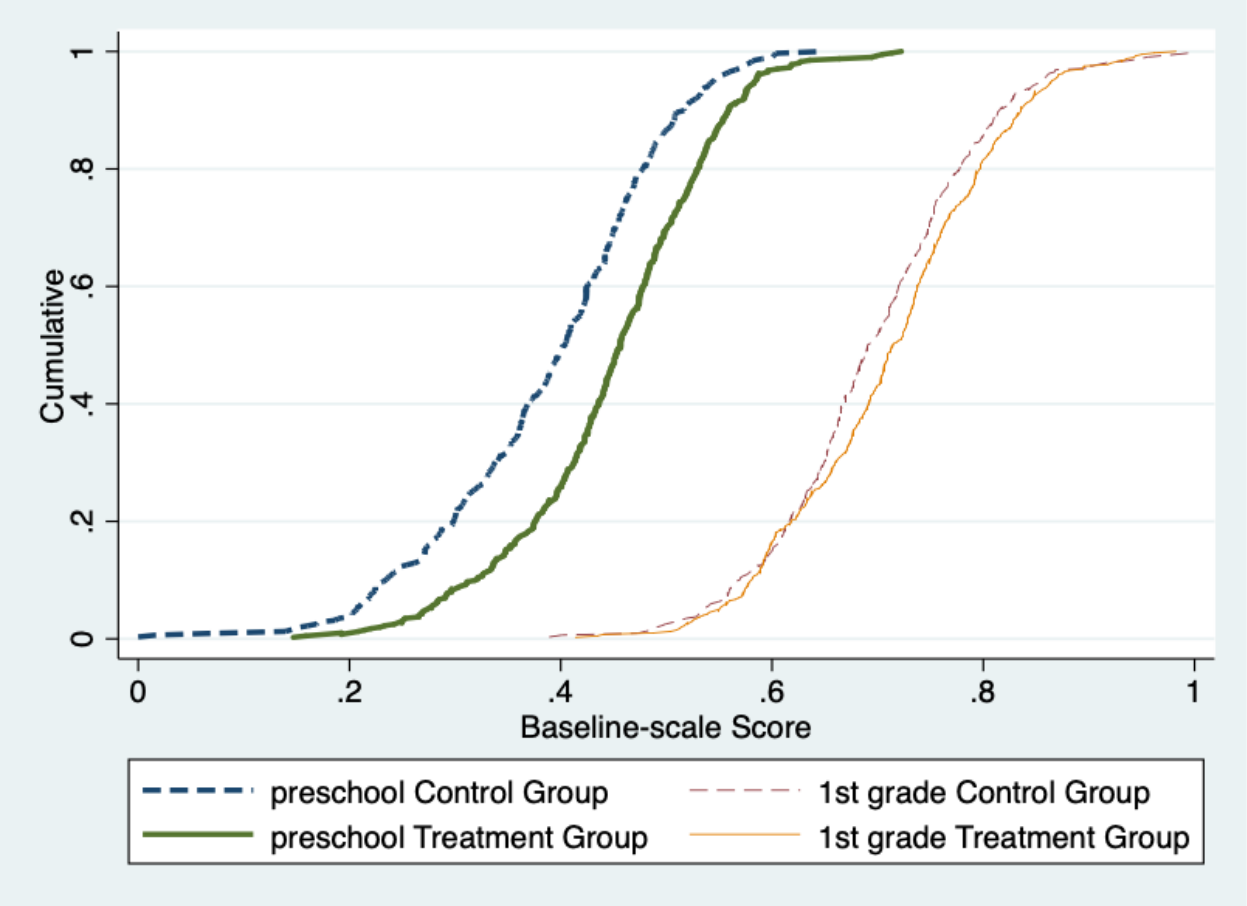

Figure 1. Cumulative distribution functions of baseline-scale scores

Note: The scores have been normalized to range from 0 to 1.

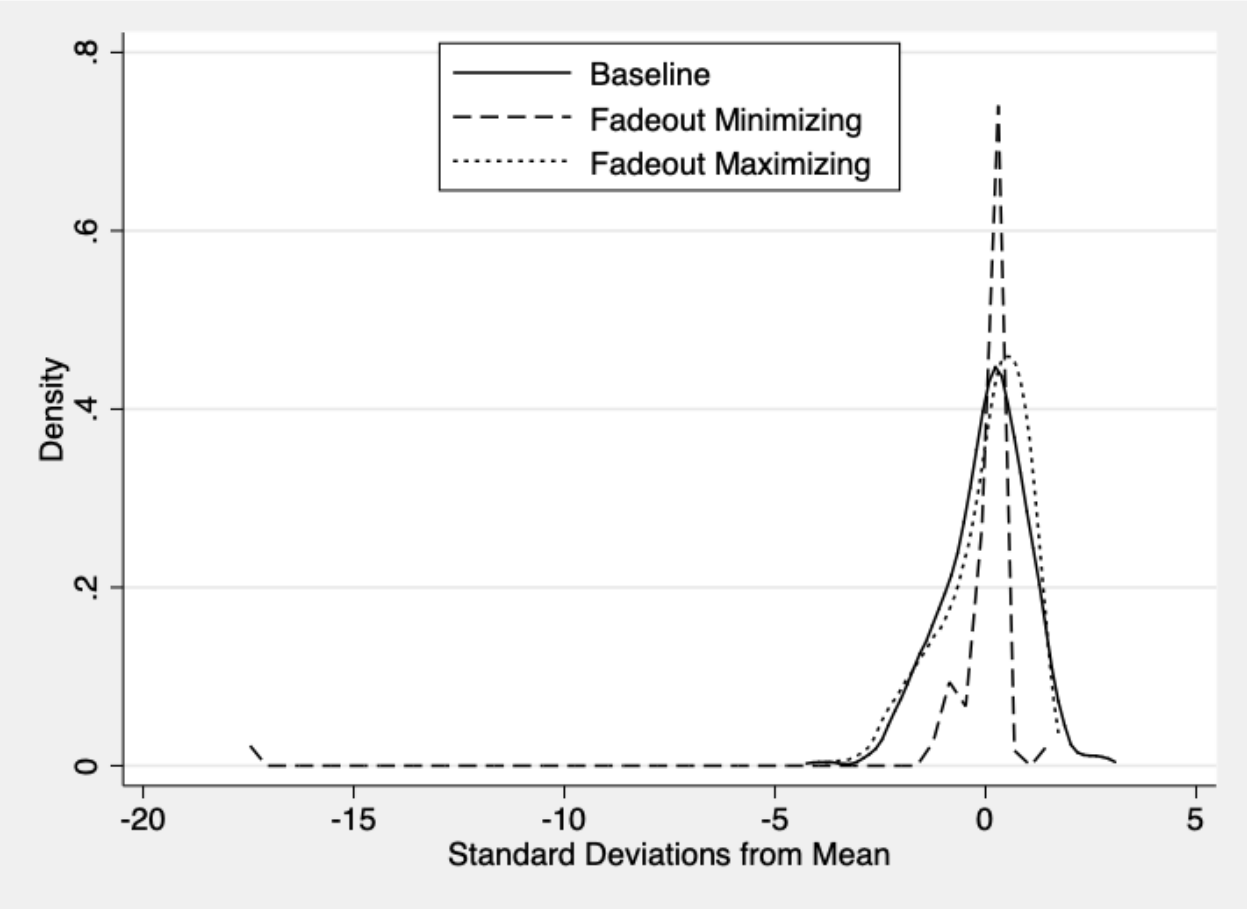

Figure 2. Spring of preschool densities under polynomial transformations

Note: The figure displays densities of transformed test scores in spring of preschool under polynomial transformations that minimize and maximize the fadeout effect.

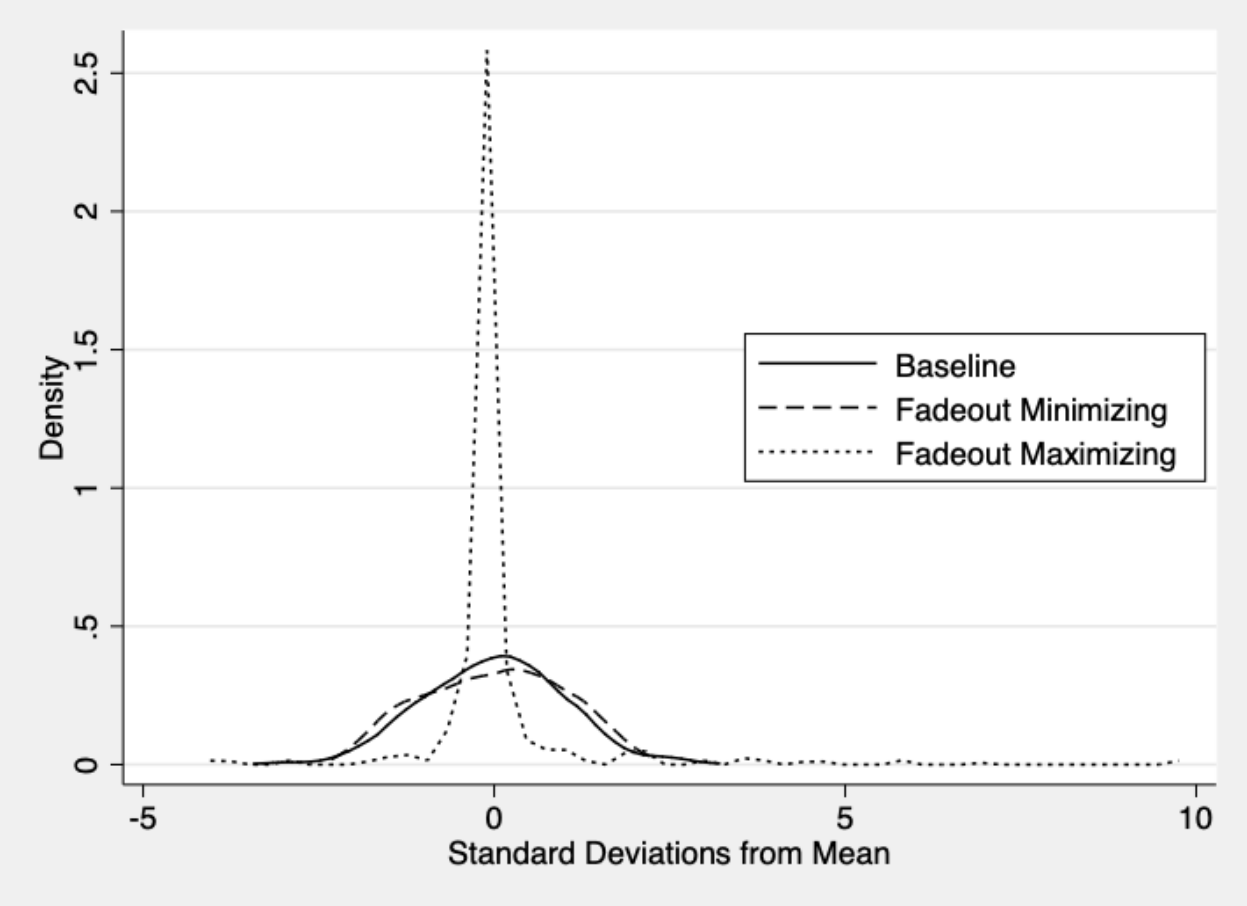

Figure 3. Spring of 1st grade densities under polynomial transformations

Note: The figure displays densities of transformed test scores in spring of 1st grade under polynomial transformations that minimize and maximize the fadeout effect.

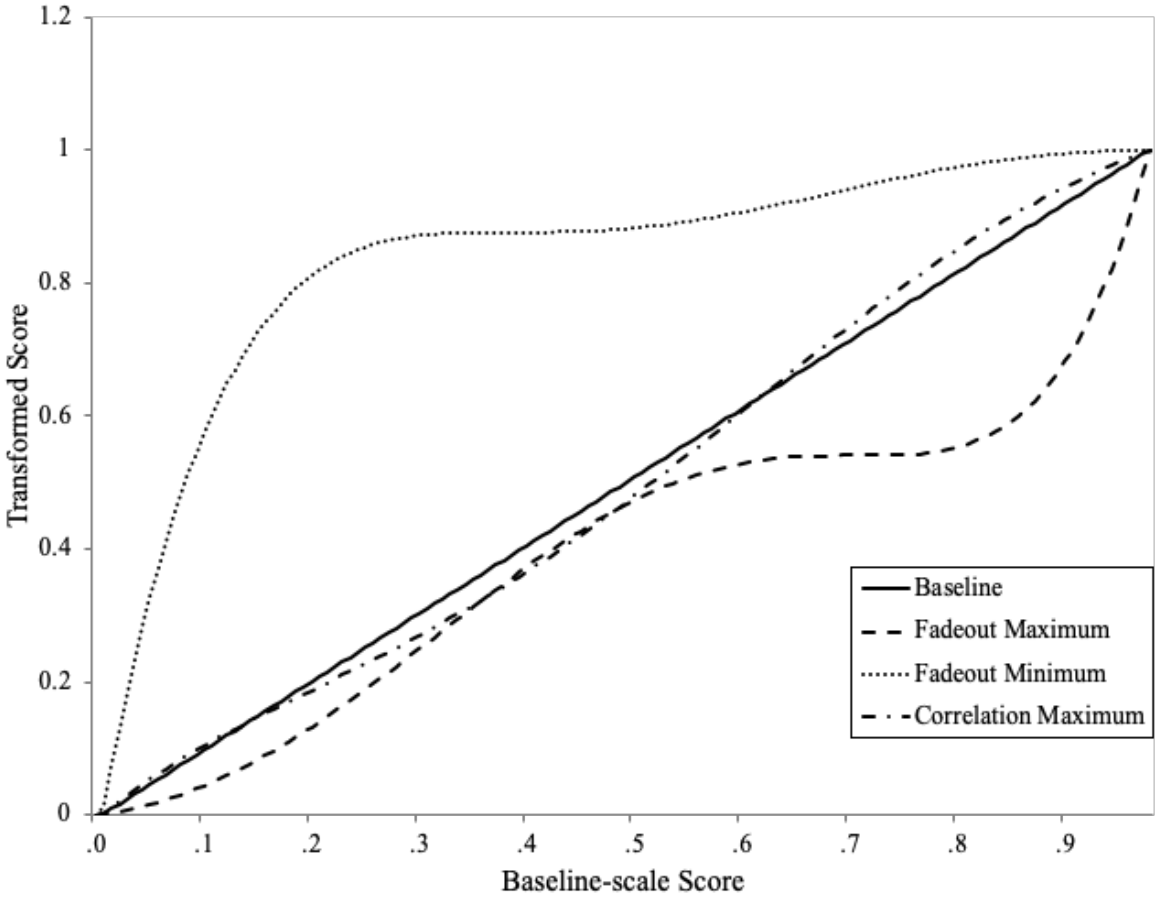

Figure 4. Polynomial transformation functions

Note: The figure displays the relation between the original scale and the transformed scales: polynomial transformation functions that minimize and maximize the fadeout effect. Transformations have been normalized to be over the same range (from 0 to 1) as the original scales.

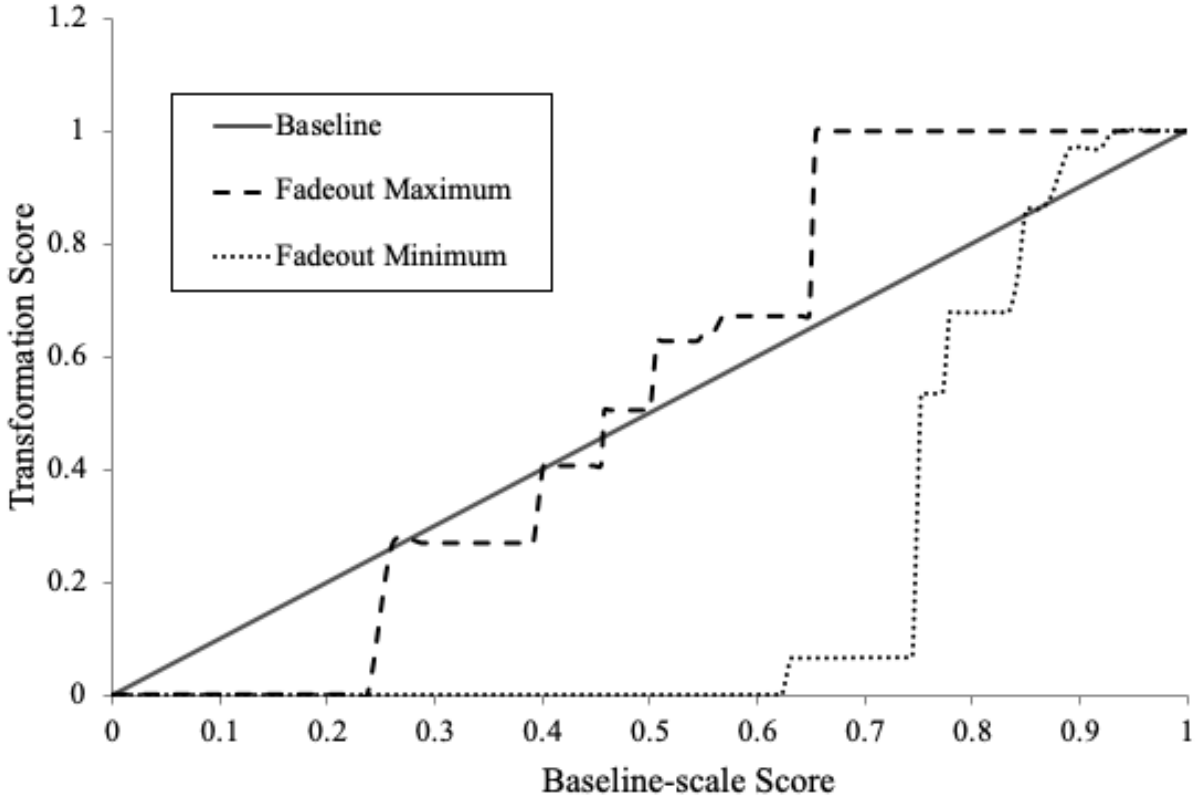

Figure 5. Discrete transformation functions

Note: The figure displays the relation between the original scale and the transformed scales: discrete transformation functions that minimize and maximize the fadeout effect. Transformations have been normalized to be over the same range (from 0 to 1) as the original scales.

References

Bailey, D. H., Duncan, G. J., Cunha, F., Foorman, B. R., & Yeager, D. S. (2020). Persistence and Fadeout of Educational Intervention Effects: Mechanisms and Potential Solutions. *Psychological Science in the Public Interest.*

Bailey, D. H., Fuchs, L. S., Gilbert, J. K., Geary, D. C., & Fuchs, D. (2018). Prevention: Necessary but Insufficient? A Two-Year Follow-Up of Effective First-Grade Mathematics Intervention. *Child Development.*

Bailey, D. H., Nguyen, T., Jenkins, J. M., Domina, T., Clements, D. H., & Sarama, J. S. (2016). Fadeout in an early mathematics intervention: Constraining content or preexisting differences?. *Developmental Psychology, 52*(9), 1457.

Bitler, M. P., Hoynes, H. W., & Domina, T. (2014). *Experimental evidence on distributional effects of Head Start* (No. w20434). National Bureau of Economic Research.

Bolt, D. M., Deng, S., & Lee, S. (2014). IRT model misspecification and measurement of growth in vertical scaling. *Journal of Educational Measurement, 51*, 141-162.

Bond, T. N., & Lang, K. (2013). The evolution of the Black-White test score gap in Grades K–3: The fragility of results. *Review of Economics and Statistics*, *95*(5), 1468-1479.

Bond, T. N., & Lang, K. (2018). The Black–White Education Scaled Test-Score Gap in Grades K-7. *Journal of Human Resources*, *53*(4), 891-917.

Campbell, F., Conti, G., Heckman, J. J., Moon, S. H., Pinto, R., Pungello, E., & Pan, Y. (2014). Early childhood investments substantially boost adult health. *Science*, 343(6178), 1478–1485. doi:10.1126/science.1248429

Campbell, D. T., & Frey, P. W. (1970). The implications of learning theory for the fade-out of gains from compensatory education. *Compensatory education: A national debate*, *3*, 455-463.

Cascio, E. U., & Staiger, D. O. (2012). *Knowledge, tests, and fadeout in educational interventions* (NBER Working Paper No. 18038). Cambridge, MA: National Bureau of Economic Research.

Chang, A. Y. (2019). Test score gap robustness to scaling: The scale transformation command. *World Bank Policy Research Working Paper*, (8986).

Clarke, B., Doabler, C., Smolkowski, K., Kurtz Nelson, E., Fien, H., Baker, S. K., & Kosty, D. (2016). Testing the immediate and long-term efficacy of a Tier 2 kindergarten mathematics intervention. *Journal of Research on Educational Effectiveness*, *9*(4), 607-634. doi:10.1080/19345747.2015.1116034

Clements, D. H., & Sarama, J. (2008). Experimental evaluation of the effects of a research-based preschool mathematics curriculum. *American Educational Research Journal*, 45(2), 443-494.

Clements, D. H., Sarama, J. H., & Liu, X. H. (2008). Development of a measure of early mathematics achievement using the Rasch model: The Research-Based Early Maths Assessment. *Educational Psychology*, 28, 457–482. http://dx.doi.org/10.1080/01443410701777272

Clements, D. H., Sarama, J., Spitler, M. E., Lange, A. A., & Wolfe, C. B. (2011). Mathematics learned by young children in an intervention based on learning trajectories: A large-scale cluster randomized trial. *Journal for Research in Mathematics Education*, 42, 127–166.

Clements, D. H., Sarama, J., & Wolfe, C. B. (2011). TEAM—Tools for early assessment in mathematics. *Columbus, OH: McGraw-Hill Education*.

Clements, D. H., Sarama, J., Wolfe, C. B., & Spitler, M. E. (2013). Longitudinal evaluation of a scale-up model for teaching mathematics with trajectories and technologies: Persistence of effects in the third year. *American Educational Research Journal*, *50*(4), 812-850.

Cunha, F., & Heckman, J. (2007). The technology of skill formation. *American Economic Review*, *97*(2), 31-47.

Cunha, F., Heckman, J. J., Lochner, L., & Masterov, D. V. (2006). Interpreting the evidence on life cycle skill formation. *Handbook of the Economics of Education*, *1*, 697-812.

Cunha, F., Heckman, J. J., & Schennach, S. M. (2010). Estimating the technology of cognitive and noncognitive skill formation. *Econometrica*, *78*(3), 883-931.

Currie, J., & Thomas, D. (1998). *School quality and the longer-term effects of Head Start* (No. w6362). National Bureau of Economic Research.

Deming, D. (2009). Early childhood intervention and life-cycle skill development: Evidence from Head Start. *American Economic Journal: Applied Economics*, *1*(3), 111-34.

Engel, M., Claessens, A., & Finch, M. A. (2013). Teaching students what they already know? The (mis) alignment between mathematics instructional content and student knowledge in kindergarten. *Educational Evaluation and Policy Analysis, 35,* 157–178.

Feller, A., Grindal, T., Miratrix, L., & Page, L. C. (2016). Compared to what? Variation in the impacts of early childhood education by alternative care type. *The Annals of Applied Statistics, 10*(3), 1245-1285.

Fryer Jr, R. G., & Levitt, S. D. (2006). The black-white test score gap through third grade. *American Law and Economics Review*, *8*(2), 249-281.

Garces, E., Thomas, D., & Currie, J. (2002). Longer-term effects of Head Start. *American economic review*, *92*(4), 999-1012.

Gibbs, C., Ludwig, J., & Miller, D. L. (2011). *Does Head Start do any lasting good?* (No. w17452). National Bureau of Economic Research.

Gibbs, C., Ludwig, J., & Miller, D. L. (2013). Head Start origins and impacts. *Legacies of the War on Poverty*, 39-65.

Hassler Hallstedt, M., Klingberg, T., & Ghaderi, A. (2018). Short and long-term effects of a mathematics tablet intervention for low performing second graders. *Journal of Educational Psychology*, *110*(8), 1127.

Heckman, J. J. (2006). Skill formation and the economics of investing in disadvantaged children. *Science*, *312*(5782), 1900-1902.

Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives*, *2*(3), 172-177.

Jacob, B. A., Lefgren, L., & Sims, D. P. (2010). The persistence of teacher-induced learning. *Journal of Human resources*, *45*(4), 915-943.

Johnson, R. C., & Jackson, C. K. (2017). *Reducing inequality through dynamic complementarity: Evidence from Head Start and public school spending* (No. w23489). National Bureau of Economic Research.

Kang, C. Y., Duncan, G. J., Clements, D. H., Sarama, J. S., & Bailey, D. H. (2019). The roles of transfer of learning and forgetting in the persistence and fadeout of early childhood mathematics interventions. *Journal of Educational Psychology.*

Kline, P., & Walters, C. R. (2016). Evaluating public programs with close substitutes: The case of Head Start. *The Quarterly Journal of Economics, 131*(4), 1795-1848.

Lang, K. (2010). Measurement matters: Perspectives on education policy from an economist and school board member. *Journal of Economic Perspectives*, *24*(3), 167-82.

Li, W., Duncan, G. J., Magnuson, K., Schindler, H. S., Yoshikawa, H., & Leak, J. (2017). *Timing in early childhood education: How cognitive and achievement program impacts vary by starting age, program duration, and time since the end of the program* (UCI SoE Working Paper). Irvine, CA: Graduate School of Education, University of California, Irvine.

McCormick, M., Hsueh, J., Weiland, C., & Bangser, M. (2017). The challenge of sustaining preschool impacts. *Expanding Children's Early Learning Network*, 1-11.

Puma, M., Bell, S., Cook, R., Heid, C., Broene, P., Jenkins, F.,…Downer, J. (2012). *Third grade follow-up to the Head Start Impact Study: Final report* (OPRE Report No. 2012-45). Retrieved from http://eric.ed.gov/?idDED539264

Sarama, J., Lange, A. A., Clements, D. H., & Wolfe, C. B. (2012). The impacts of an early mathematics curriculum on oral language and literacy. *Early Childhood Research Quarterly*, *27*(3), 489-502.

Schweinhart, L. J., Montie, J., Xiang, Z., Barnett, W. S., Belfield, C. R., & Nores. M. (2005). *Lifetime effects: The High/Scope Perry Preschool study through age 40*. Ypsilanti, MI: High/Scope Press.

Smith, T. M., Cobb, P., Farran, D. C., Cordray, D. S., & Munter, C. (2013). Evaluating math recovery: Assessing the causal impact of a diagnostic tutoring program on student achievement. *American Educational Research Journal*, *50*(2), 397-428.

Taylor, E. (2014). Spending more of the school day in math class: Evidence from a regression discontinuity in middle school. *Journal of Public Economics*, *117*, 162-181.

Appendix

Table S1

Comparing intervention impacts of TRIAD with the Head Start Impact Study.

	TRIAD	Head Start Impact Study
Initial impact	.64 (.18)	.21 (.04)
Later impact	.16 (.17)	.08 (.06)
Sample size	720	1601

Note. Intervention impacts are measured in standard deviations, and standard errors are in parentheses. For TRIAD, we include the sample in the current study from preschool to grade 1. The values used here are instrumental variables (IV) estimates. For the Head Start Impact Study, we include the entering 4-year-olds cohort from end-of-Head-Start to grade 1 with the average of Peabody Picture and Vocabulary Test (PPVT) and Woodcock Johnson III (WJIII) scores as the outcome. The values for the Head Start Impact Study are IV estimates reported by Kline and Walters (2016).

Table S2

Regression Estimates of Attrition on Spring of Preschool Math Achievement

	Spring of preschool math
Treatment Group	.414** (.118)
Attrited	-.075 (.110)
Interaction: Treatment X Attrited	-.068 (.129)
Constant	-2.260*** (.093)
R-squared	.077**
Observations	834

Note. Robust standard errors were adjusted for clustering at the school level (n = 30 schools), and are displayed in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table S3
Sample size of different analyses

	Total N	Treatment N	Control N
Original TRIAD sample	834	473	361
Manuscript analysis sample	720	402	318
Only require preschool and kindergarten scores	779	435	344

Note: As we mentioned in the manuscript, we use only two groups (treatment only in preschool group and control group) of the original TRIAD study.

Table S4

Evolution of the Treatment-Control test gap under various polynomial transformations and discrete transformations of math scores from preschool to kindergarten

Variables	Baseline scale	Fadeout effect maximization Polynomial	Fadeout effect minimization polynomial	Correlation maximization polynomial	Baseline (constant variance)	Fadeout effect maximization discrete	Fadeout effect minimization discrete
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Spring of preschool math	.584**	.533**	.325**	.571**	.584**	.559**	.198
	(.160)	(.151)	(.100)	(.156)	(.160)	(.154)	(.114)
Spring of kindergarten math	.255	.122	.204	.257	.276	.051	.275*
	(.159)	(.118)	(.150)	(.159)	(.172)	(.091)	(.130)
Fadeout effect	.329***	.411**	.121	.314***	.308***	.508***	-.077

Note. Test gaps are measured in standard deviations, and standard errors are in parentheses. Fadeout effect = gap in preschool math – gap in 1st grade math. $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$.

Table S5

Evolution of the Treatment-Control test gap under transformation that maximizes either the preschool treatment effect or the 1st grade treatment effect

Variables	Baseline scale (1)	Preschool effect maximization (2)	1st grade effect maximization (3)
Spring of preschool math	.578**	.579**	.552**
	(.165)	(.165)	(.154)
Spring of 1st grade math	.146	.124	.155
	(.152)	(.150)	(.153)
Fadeout effect	.432***	.455***	.397***

Note. Test gaps are measured in standard deviations, and standard errors are in parentheses. Fadeout effect = gap in preschool math – gap in 1st grade math.

*$p < .05$, **$p < .01$, ***$p < .001$.

Table S6

Robustness check of the evolution of the Treatment-Control test gap under various transformations of math scores

Variables	Baseline scale	Fadeout effect maximization polynomial	Fadeout effect minimization polynomial	Correlation maximization polynomial	Baseline (constant variance)	Fadeout effect maximization discrete	Fadeout effect minimization discrete
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Spring of preschool math (ITT)	.578**	.570**	.304**	.578**	.578**	.613***	.147
	(.165)	(.167)	(.104)	(.163)	(.165)	(.156)	(.076)
Spring of preschool math (IV)	.638**	.628**	.336**	.637**	.638**	.676**	.162
	(.183)	(.185)	(.115)	(.180)	(.183)	(.174)	(.086)
Spring of kindergarten math (ITT)	.277	.285	.235	.277	.297	-	-
	(.154)	(.142)	(.150)	(.154)	(.165)		
Spring of kindergarten math (IV)	.305	.314	.259	.305	.327	-	-
	(.172)	(.159)	(.167)	(.172)	(.185)		
Spring of 1st grade math (ITT)	.146	.034	.154	.147	.155	-.038	.198
	(.152)	(.118)	(.150)	(.151)	(.162)	(.106)	(.145)
Spring of 1st grade math (IV)	.161	.037	.169	.163	.171	-.041	.218
	(.169)	(.130)	(.167)	(.168)	(.180)	(.117)	(.162)
Fadeout effect (ITT)	.432***	.536***	.150	.431***	.423***	.651***	-.051
Fadeout effect (IV)	.477***	.591***	.167	.474***	.467***	.717***	-.056

Note. Test gaps are measured in standard deviations, and standard errors are in parentheses. $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$. ITT = intention-to-treat estimates, IV = Instrumental variables estimates. Fadeout effect = gap in preschool math – gap in 1st grade math.

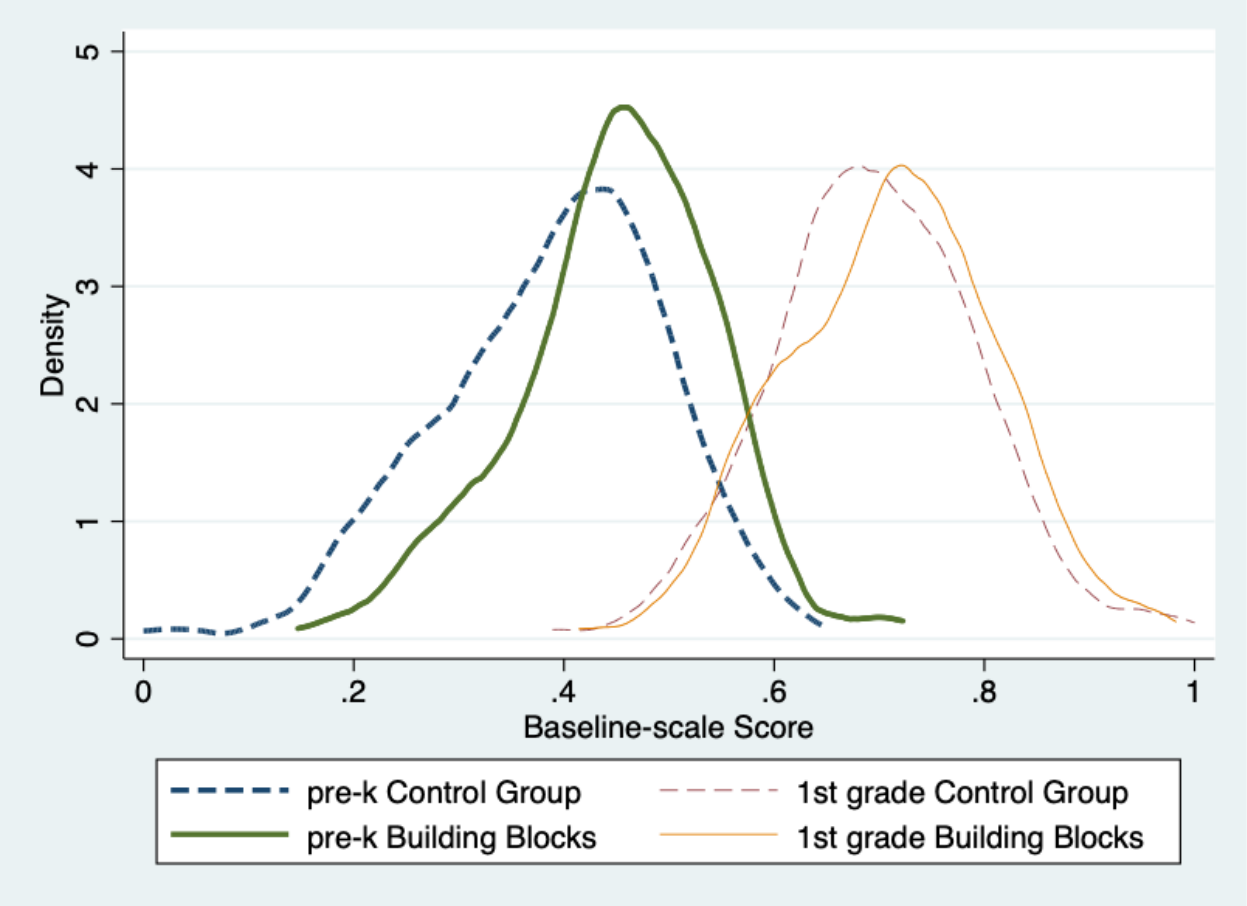

Figure S1. Probability density functions of baseline-scale scores

Note: The scores have been normalized to range from 0 to 1.

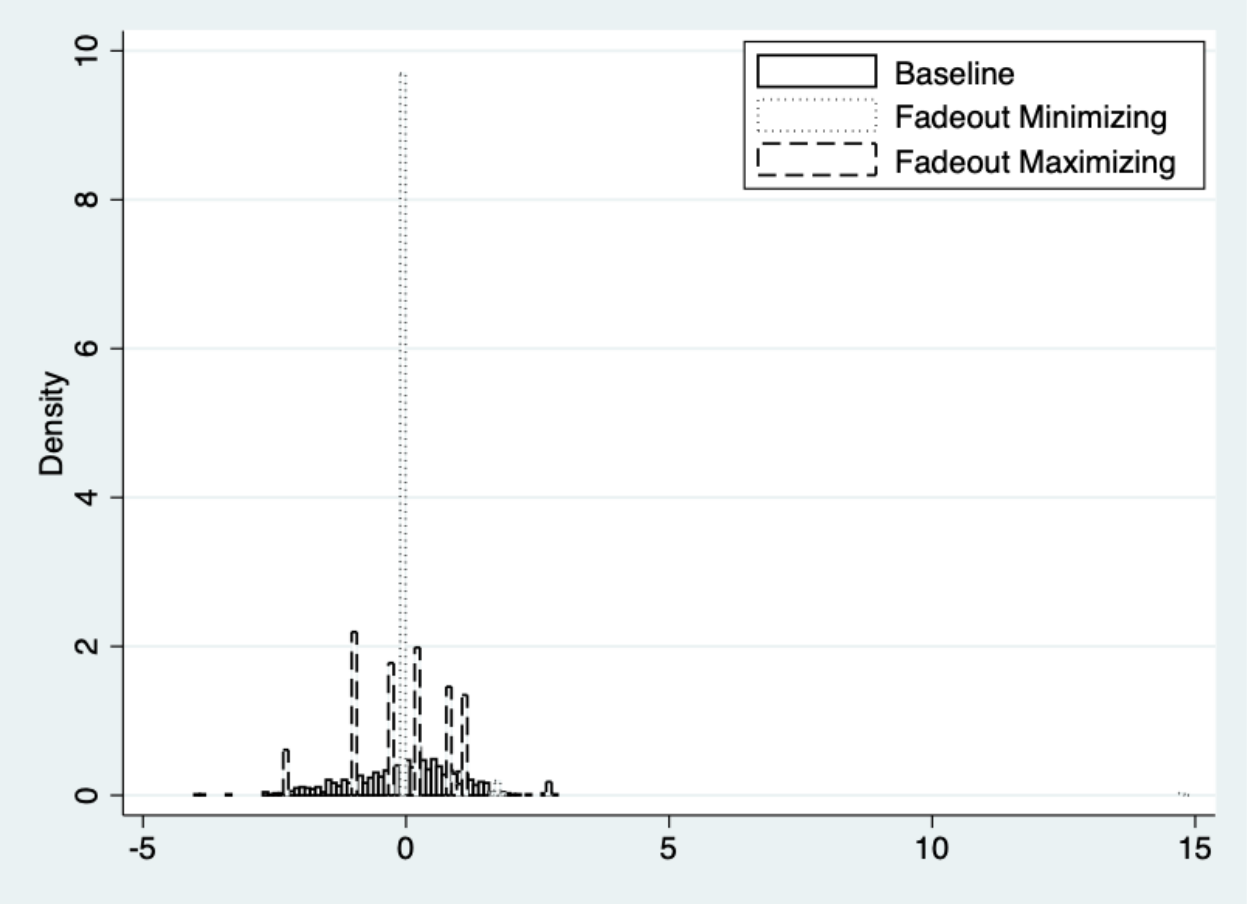

Figure S2. Spring of preschool histograms under discrete transformations

Note: The figure displays histograms of transformed test scores in spring of preschool under discrete transformations that minimize and maximize the fadeout effect.

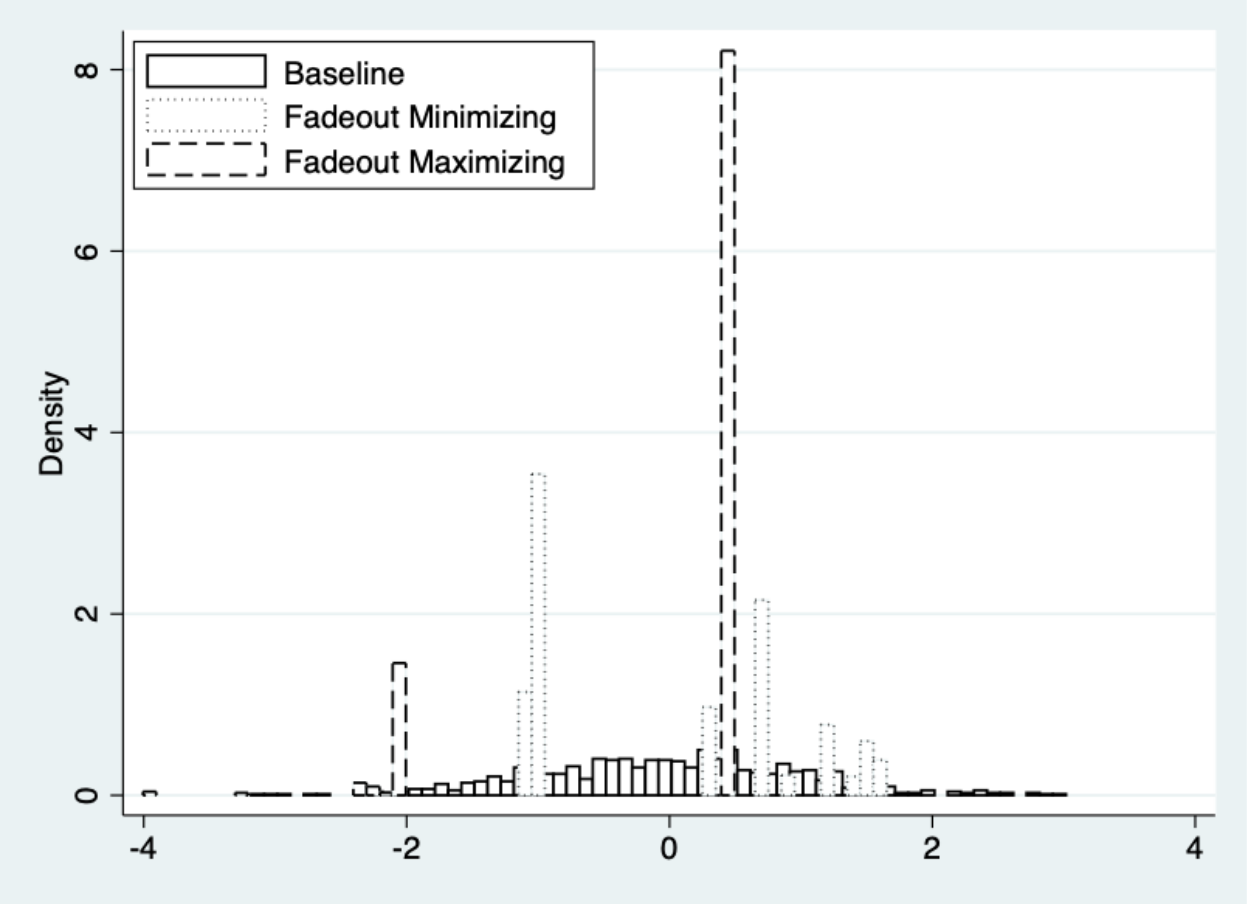

Figure S3. Spring of 1st grade histograms under discrete transformations

Note: The figure displays histograms of transformed test scores in spring of 1st grade under discrete transformations that minimize and maximize the fadeout effect.

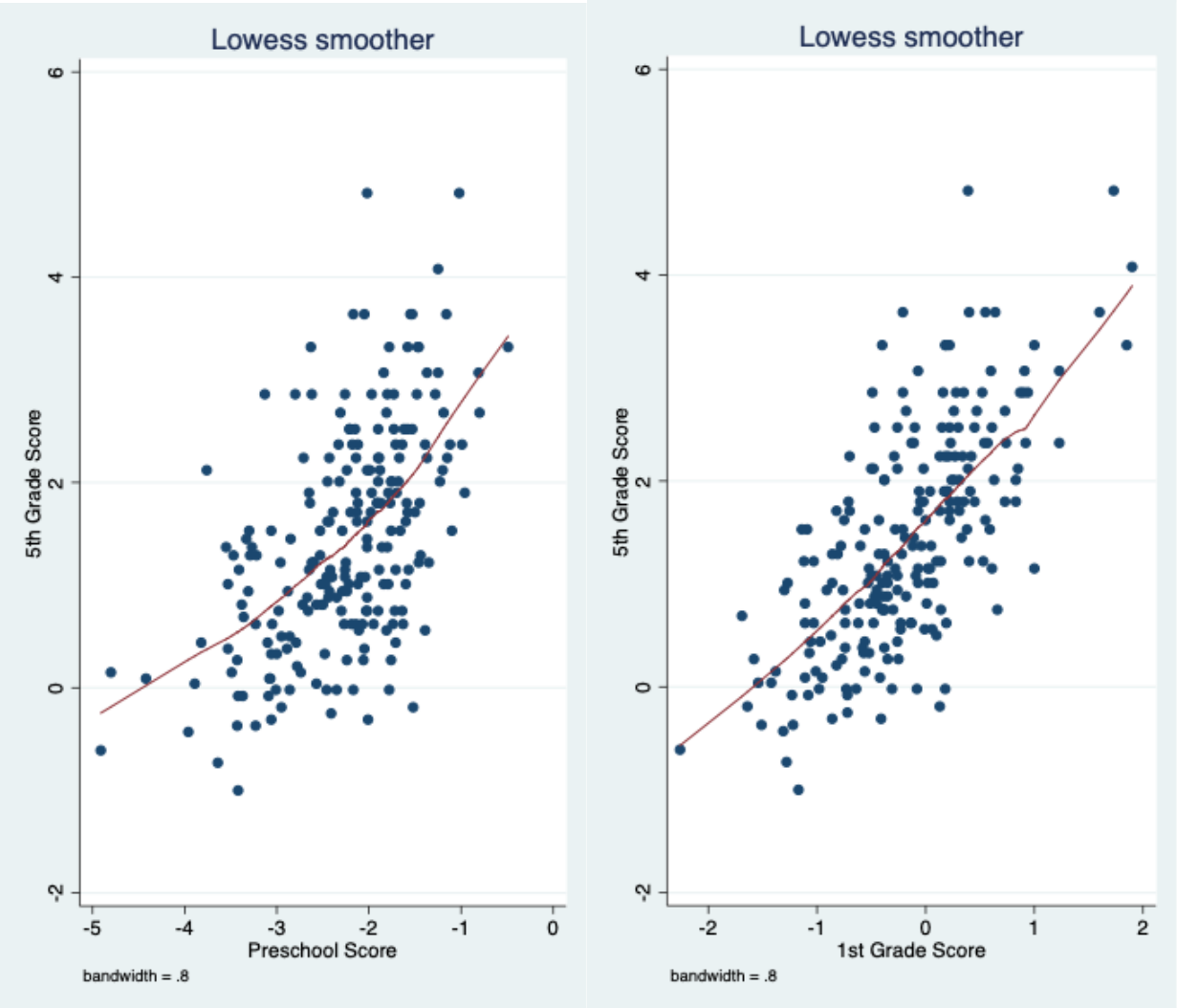

Figure S4. Scatterplots of preschool and 1st grade scores on 5th grade scores in control group