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

Author note: 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. Bailey, by the Institute of Education Sciences, U.S. Department of Education, through Grants R305K05157 and R305A120813 to D. 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 following 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 fadeout by emphasizing differences in scores near typical levels of first-graders while treating differences elsewhere as unimportant. 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 review, see Bailey, Duncan, Odgers, & Yu, 2017). 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 tests are scaled across age. By scaling we mean assigning numbers to increasing levels of performance on a test. 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.

Suppose preschool mathematics curriculum has a persistent effect on children's mathematics learning after the intervention ends. Children who receive the curriculum learn more mathematics skills than children in the control group at the end of the one-year intervention, and they have the same amount of more learning than children in the control group in each of the subsequent years. When the same amount of learning is worth fewer points on later achievement tests relative to learning in earlier years, it will create the illusion of fadeout. This will happen if the variance of knowledge increases over time but the test is always scaled to have the same variance at each time point. However, under a different scale, the same amount of learning might later be worth the same number or even more points on the achievement test than at the end of the intervention. As a result, we will see no fadeout or even a reversal of fadeout.

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.

The current study tests the robustness of the fadeout effect to scaling choice. It draws on 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 theoretically possible levels of fadeout and persistence, analogously to Bond & Lang (2013). 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, Sarama, Wolfe, & Spitler, 2013). We examine how various scale transformations, which all preserve the order of the original scale, affect the evolution of the mathematics achievement test-score gap between children in the control and treatment groups over this period.

We find that fadeout is robust when we constrain scale variance to be constant across waves, and under the transformation that maximizes the ability of early scores to predict later

scores. We can, however, eliminate fadeout using scales optimized to do so. These scales largely shrink scale differences for 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. This causes the variance in math achievement to increase dramatically with age. While 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), adopting the fadeout-eliminating scale would have substantial implications for interpreting the effects of educational interventions.

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, Fuchs, Gilbert, Geary, & Fuchs, 2018; Clarke et al., 2016; Clements et al., 2013; Hassler Hallstedt, Klingberg, & Ghaderi, 2018; Smith, Cobb, Farran, Cordray, & Munter, 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 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, Duncan, Clements, Sarama, & Bailey, in press). 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, Claessens, & Finch, 2013; McCormick, Hsueh, Weiland, & Bangser, 2017).[1]

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 for Head Start on educational attainment and other employment relevant outcomes (Deming, 2009; Garces, Thomas, & Currie, 2002; Gibbs, Ludwig, & Miller, 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.[2] Another possibility is that the interventions influenced cognitive 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 a methodological artefact. Perhaps impact of early interventions do not really fade over time. Instead, fadeout merely reflects defects in how researchers measure learning across development. Specifically, fadeout

---

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

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

may be partially a statistical artefact of the way achievement tests are scaled. If we measure 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. 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 obtain 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, Bloom, Black, & Lipsey, 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 used, and the CNLSY 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 gap growth 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.

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 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, Lefgren, & Sims, 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 gap growth estimates 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.[3]

III. Data and Methods

We test the robustness of the fadeout effect to a variety of different scaling decisions that determine changes in variance across time and the intervals between scores across the score distribution. We use data from an RCT evaluating an early mathematics curriculum: the Technology-enhanced, Research-based, Instruction, Assessment, and professional Development (TRIAD) evaluation study (Clements, Sarama, Spitler, Lange, & Wolfe, 2011; Clements et al., 2013; Sarama, Clements, Wolfe, & Spitler, 2012). Forty-two low-income schools in Buffalo, NY, and Boston, MA, were randomly assigned to one of three conditions: 1) control (n = 378; school N = 16), 2) *Building Blocks* preschool mathematics curriculum treatment in preschool (n = 456; school N = 14), or 3) *Building Blocks* treatment in preschool with follow-through (n = 471; school N = 12). In both *Building Blocks* treatment conditions, teachers received pedagogical

[3] 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. However, the current study does not consider such scales.

development and implemented the *Building Blocks* mathematics curriculum during preschool. For schools in the *Building Blocks* 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).

The *Building Blocks* curriculum was designed 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 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 et al., 2011).

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 15% of the control and 20% of the *Building Blocks* group, resulting in a final sample of 720 observations. See Figure 1 for the cumulative distribution functions and Figure S1 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 logistic regression models with clustered standard errors at the school level (for sex and ethnicity). The table shows that the original randomization, after attrition, matched the control groups well in 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, Sarama, & Liu, 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 allows for accurate comparisons of scores between groups and across ages (Clements et al., 2011). 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 gap is .401 on the Rasch scale, but falls to .134 a year later and to .040 two years later, at which point it ceases to be statistically significant. We define the test score gap as the standardized treatment effect: the difference between the mean mathematics test scores of the *Building Blocks* (BB) group and control group

divided by the overall standard deviation of test scores of the full sample in that grade. We note, however, that if as discussed in section II, the true variance is increasing over time, this will lead to us finding fadeout.

We take two broad approaches to testing whether fadeout is a scaling artefact. We use a theory-driven approach, in which we start with a reason 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 assuming constant test reliability across grades and rescaling the test to the time specific standard deviation. Another example is from Bond and Lang (2013), who transform earlier and later test scores to maximize the correlation between the two. If we were worried that there is a difference in test reliability, or some nonlinearity in the relation between cognitive skills and earlier or later test scores, including range restriction, then the correlation-maximizing transformation could make fadeout smaller or even go away. In this paper, we use both variance-equating and correlation-maximization to test for the robustness of fadeout to possible variations in test score variance and test reliability across time.

We also use a data-driven approach, similar to Bond and Lang (2013), to test the robustness of fadeout: we look for the transformations that maximize or fully eliminate 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 correlation and minimizing the gap growth 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 test score gap growth is an artefact of higher measurement error in the early grades.

Our data-driven approach draws on Bond and Lang (2013). We look for the bounds of the transformations, the one that: i) maximizes the gap growth (and thus minimizes fadeout) from the end of treatment in preschool to the spring of grade 1; ii) minimizes gap growth (and maximizes fadeout). 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, β_0 - β_6 and k are constants. We use an optimization function in Stata/SE 14.0 to search for the values of β_0 - β_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 gap growth from preschool to grade 1. Similarly, we maximize the objective function given by D_{max}. Since the sixth-degree polynomial does not require monotonicity, our algorithm checks for it and rejects parameters that violate the condition.[4] Figures 2 and 3 show the densities of the scores of different scales, while Figure 4

[4] 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.

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 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 (maximize) the objective function given by $D_{\min}$ ($D_{\max}$). The histograms of the transformation scores are presented in Figure S2 and Figure S3, and the relation among them is displayed in Figure 5.

Following Cascio and Staiger (2012), our theory-driven approach scales down initial treatment impact by multiplying the baseline treatment impact by the ratio of the earlier to the later standard deviation of test scores. We also use the sixth-degree polynomial discussed above to find the transformation that maximizes the correlation between end of treatment and grade 1 scores in the control group. As discussed, this addresses our concerns about possible differences in test reliability and the test range restriction.

IV. Results

Figure 1 shows the cumulative distribution functions of scores, re-normed to range from 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 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, we observe first-order stochastic dominance (FOSD) in the initial period so that 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 .17. 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 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 shows differences among results close to one 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 and above .9. 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 column shows the baseline pattern. The treatment effect decreases from .561 SD in preschool to .059 SD by 1st grade. We observe a similar pattern under the gap-growth-minimizing (and 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 gap-growth-maximizing (and fadeout-minimizing) polynomial transformation reduces the gap at the spring of preschool to .239 SD, but the gap in first grade remains close to 0. However, if we allow for a discrete scale, fadeout becomes amplification, although the growth in the gap is not statistically significant.[5]

Following Bond and Lang (2013), an alternative way of expressing gaps under different transformations is to calculate what the test score gap would be in each grade, under each transformation, if the control group had all the lowest scores and treatment group all the highest.

---

[5] 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.

This is the maximum possible test gap; we express the observed gap for each time-transformation combination as a proportion of this maximum. 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 5% 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, 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 scales, the maximum possible gap drops to .37 SD, approximately 1/4 of the baseline maximum possible gap. Notably, under this transformation, fadeout is actually *larger*, expressed as a difference in proportions of the maximum possible gap, than it is under other transformations. As with the polynomial transformations, the gaps in preschool are comparable across discrete transformations and again actually largest under the fadeout-minimizing discrete transformation, wherein the maximum possible gap is reduced dramatically. The first grade gaps, expressed as a proportion of the maximum possible gap, are similar across discrete transformations. The variance in the fadeout-minimizing discrete transformation is again greatly reduced, such that the maximum possible test gap increases by a factor of greater than 10, from .162 SD at the spring of preschool to 1.820 SD in first grade.

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. 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.

V. 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 gap growth is at least partially a measurement artefact, we found that almost all rescaling choices show nontrivial fadeout.

We can eliminate 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 makes the variance of math achievement much higher in the later period.

Given the lack of converging evidence for fadeout's sensitivity to scaling decisions across

other approaches, should we seriously consider this scale? More specifically, 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.

In conclusion, the results seem to reconcile the measurement-based explanation of fadeout with the substantive theoretical explanations of fadeout whereby fadeout and persistence may happen at the same time. In other words, it may imply that a large impact and its fadeout is happening on preschool mathematics skills while a small and persistent impact is happening on first grade mathematics skills. It will be both theoretically important and potentially practically useful to test whether this is the only way to make fadeout disappear, or whether there are interventions where fadeout can be eliminated without compressing the variance of scores at the early period. Notably, both interpretations of the discrete fadeout-minimizing transformation results make the prediction that teaching more advanced knowledge will yield larger longer-term treatment impacts. Educational interventions may be improved if economists and other researchers contributed additional effort to further developing and testing the long list of substantive explanations of fadeout, along with their educational implications.

Table 1

Descriptive statistics

Variables	Building Blocks	Control group	Group differences	p value for group differences
Spring of preschool math	-1.842	-2.243	.401	.002
	(.655)	(.724)		
Spring of kindergarten math	-1.044	-1.178	.134	.217
	(.646)	(.686)		
Spring of 1st grade math	-.089	-.129	.040	.712
	(.674)	(.679)		
Male	.506	.507	-.001	.978
Ethnicity				
Black	.530	.501	.029	.826
Hispanic	.182	.248	-.066	.496
White	.255	.176	.079	.464
Ethnicity- Other	.034	.075	-.041	.182
Age (years) fall preschool	4.334	4.391	-.057	.376
Observations	385	335		

Note. Standard deviations are in parentheses for variables. P-values are got from regressions and indicate the extent to which treatment participants different from controls on each variable . In each regression, standard errors were adjusted for clustering at the school level (n = 30 schools).

Table 2

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

Variables	Baseline (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	.561**	.548**	.239*	.561**	.561**	.595**	.090
	(.166)	(.167)	(.090)	(.166)	(.166)	(.154)	(.060)
Spring of kindergarten math	.200	.226	.140	.204	.215	-	-
	(.159)	(.146)	(.147)	(.159)	(.171)		
Spring of 1st grade math	.059	-.004	.033	.061	.063	-.075	.142
	(.158)	(.118)	(.144)	(.156)	(.167)	(.113)	(.151)
Fadeout effect (preschool math minus 1st grade math)	.502***	.552***	.206	.500***	.498***	.670***	-.052

Note. Test gaps are measured in standard deviations, and standard errors are in parentheses. $^{*}p<.05$, $^{**}p<.01$, $^{***}p<.001$.

Table 3

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

Variables	Baseline	Fadeout effect maximization polynomial	Fadeout effect minimization polynomial	Correlation maximization polynomial	Fadeout effect maximization discrete	Fadeout effect minimization discrete
Spring of preschool *BB-Control test gap*	.561	.548	.239	.561	.595	.090
Spring of preschool *maximum test gap*	1.557	1.547	.368	1.571	1.558	.168
Spring of preschool *% of maximum gap*	36.0%	35.4%	64.9%	35.7%	38.2%	53.6%
Spring of 1st grade *BB-Control test gap*	.059	-.004	.033	.061	-.075	.142
Spring of 1st grade *maximum test gap*	1.595	.790	1.198	1.610	2.506	1.713
Spring of 1st grade *% of maximum gap*	3.7%	0.5%	2.8%	3.8%	3.0%	8.3%
% of 1st grade SD to preschool SD	94.4%	41.5%	90.5%	83.3%	70.6%	859.6%

Note. 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 BB group.

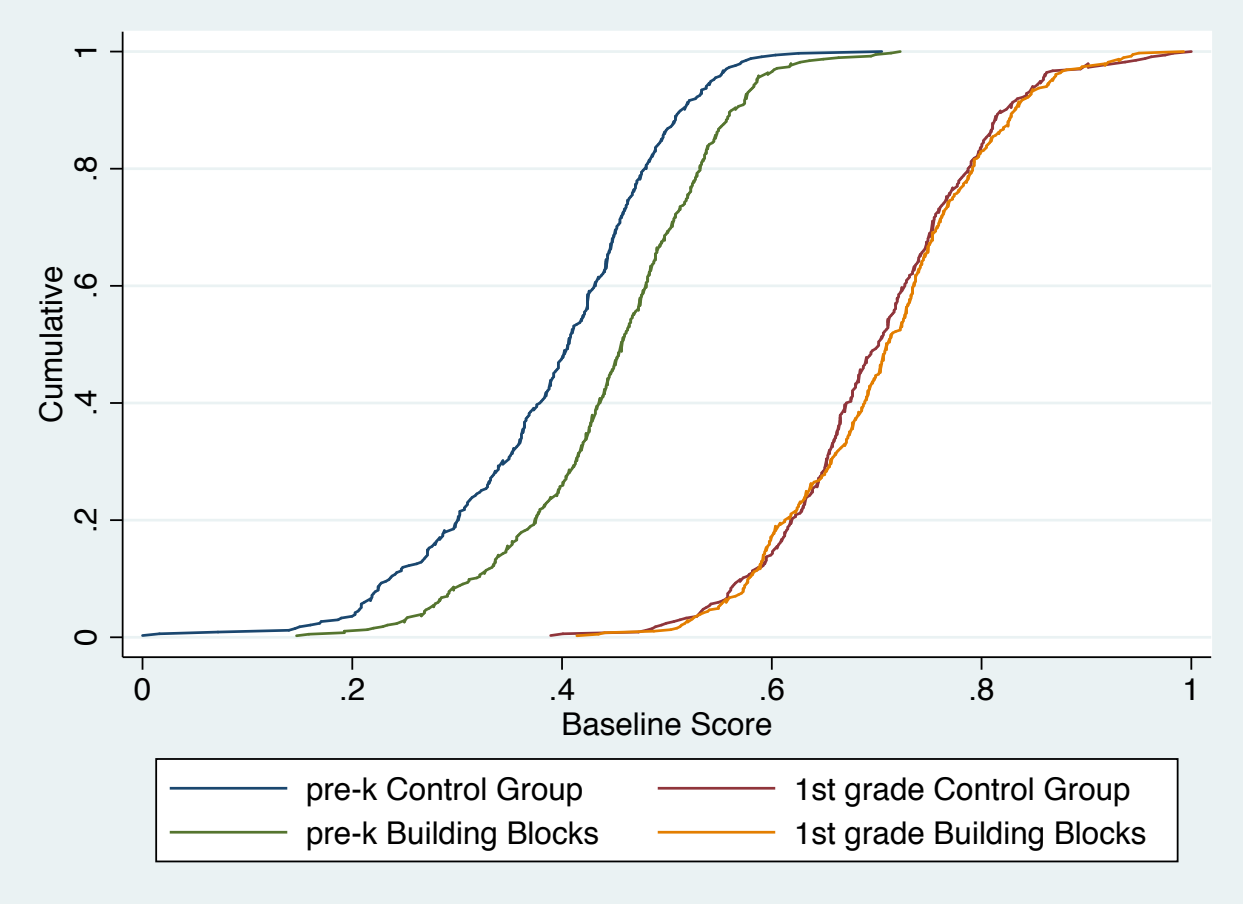

Figure 1. Cumulative distribution functions of baseline scores

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

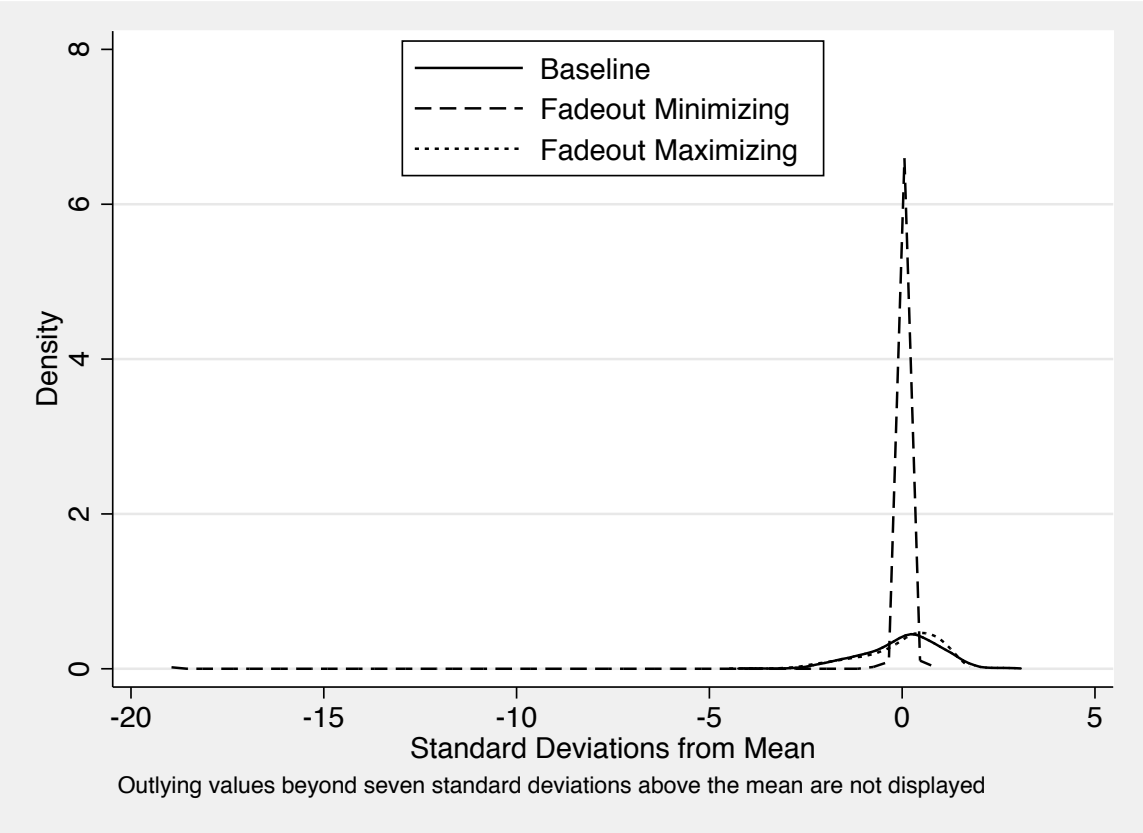

Figure 2. Spring of preschool densities under polynomial transformations

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

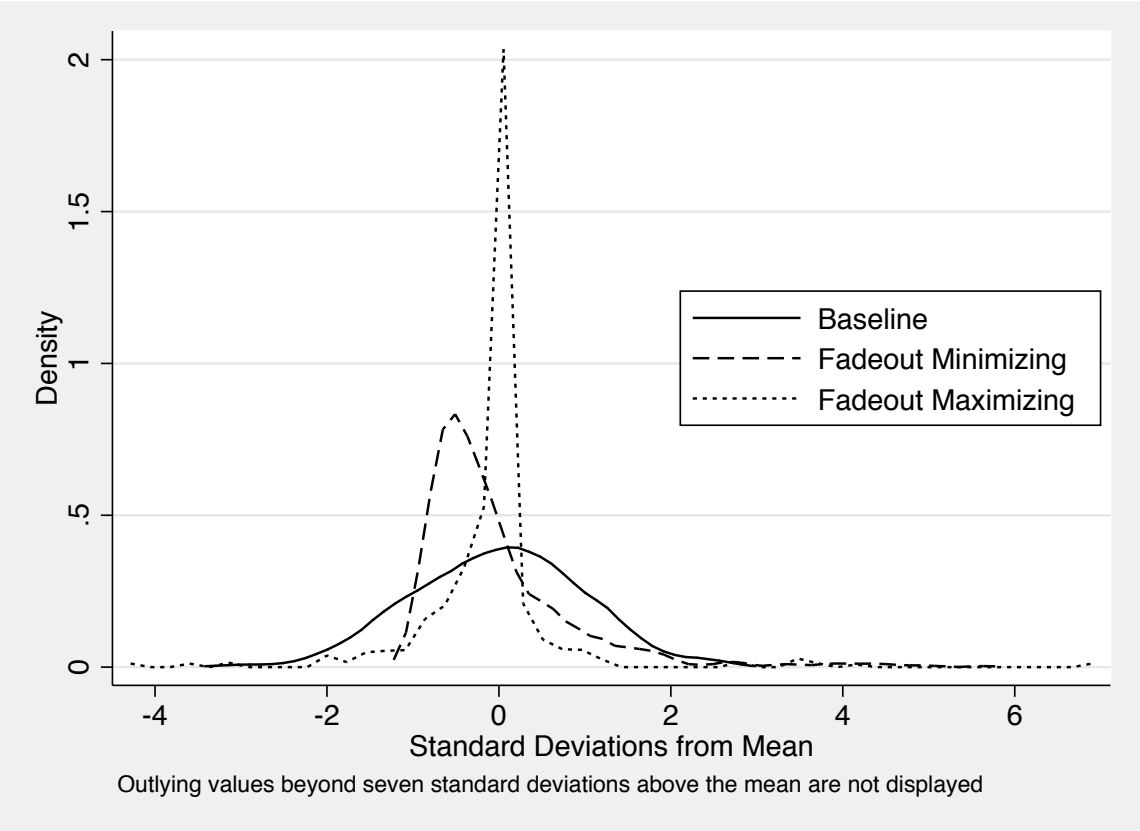

Figure 3. Spring of 1st grade densities under polynomial transformations

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

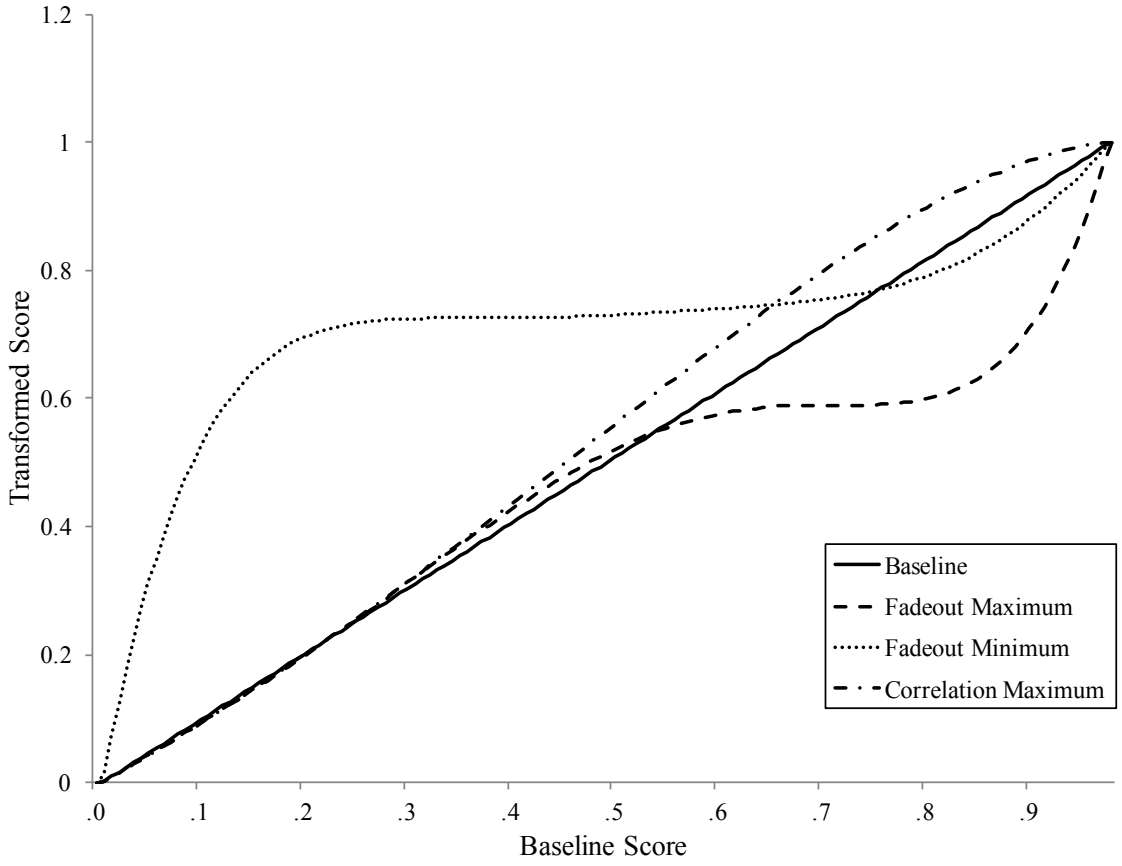

Figure 4. Polynomial transformation functions

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.

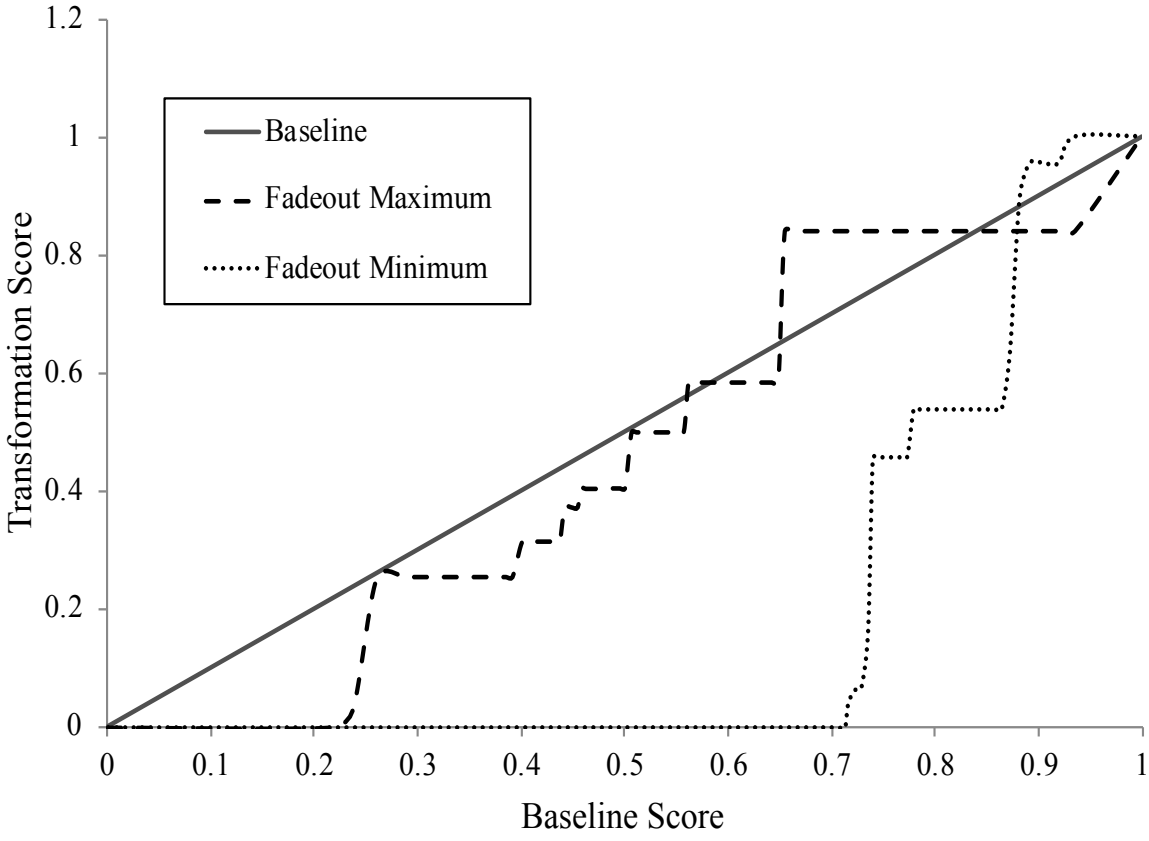

Figure 5. Discrete transformation functions

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., 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.*

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.

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.

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. (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. (in press). The roles of transfer of learning and forgetting in the persistence and fadeout of early childhood mathematics interventions. *Journal of Educational Psychology.*

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

Reardon, S. F. (2011). The widening academic achievement gap between the rich and the poor: New evidence and possible explanations. *Whither opportunity*, 91-116.

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.

Smit, E., Verdurmen, J., Monshouwer, K, & Smit, F. (2008). Family interventions and their effect on adolescent alcohol use in general populations: A meta-analysis of randomized controlled trials. *Drug and Alcohol Dependence*, 97(3), 195–206. doi:10.1016/j.drugalcdep.2008.03.032

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

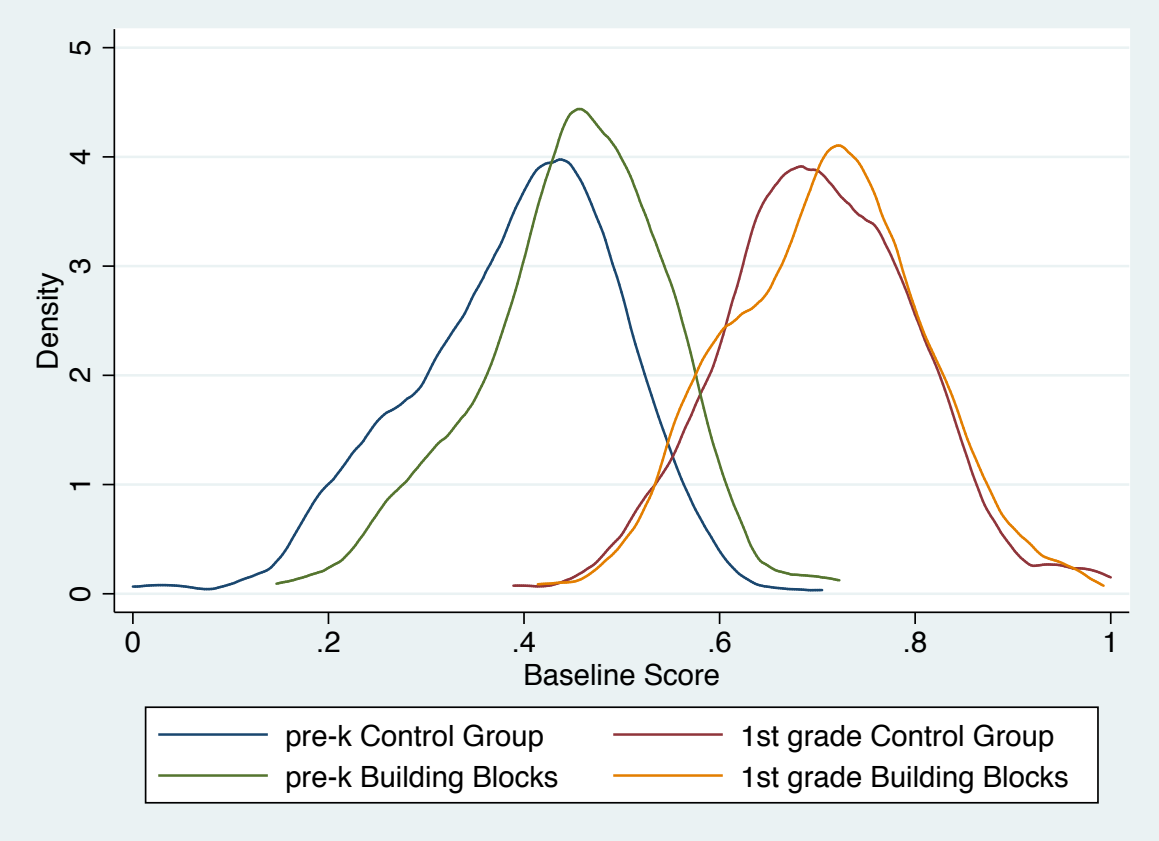

Figure S1. Probability density functions of baseline scores

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

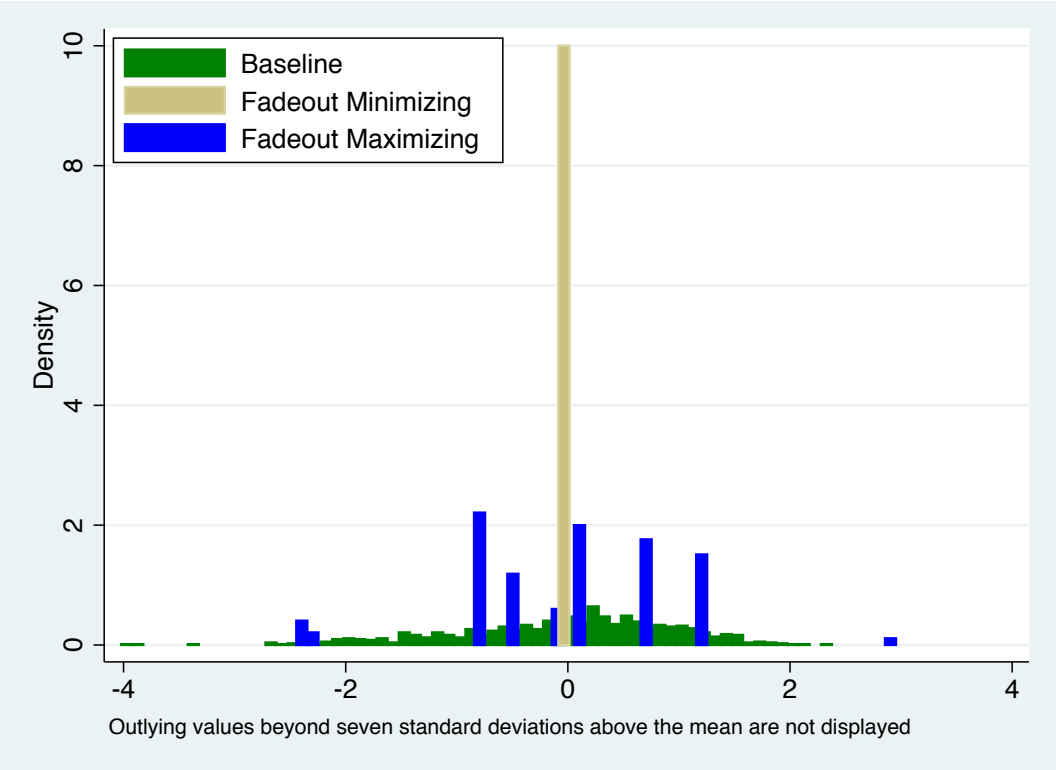

Figure S2. Spring of preschool histograms under discrete transformations

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

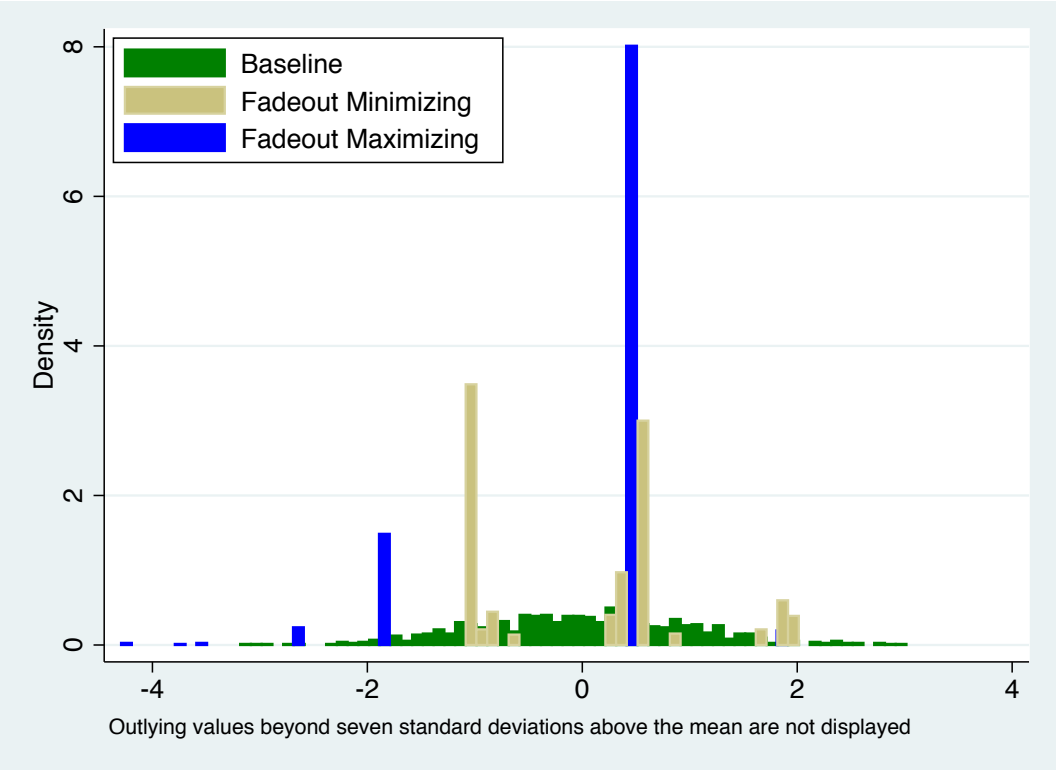

Figure S3. Spring of 1st grade histograms under discrete transformations

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