Title page with All Author Information

Risky Business: Correlation and Causation in Longitudinal Studies of Skill Development

Drew Bailey[1]
Greg J. Duncan[1]
Tyler Watts[1]
Doug Clements[2]
Julie Sarama[2]

Acknowledgment
We are grateful to the Eunice Kennedy Shriver National Institute of Child Health & Human Development of the National Institutes of Health under award number P01-HD065704. This research was also supported by the Institute of Education Sciences, U.S. Department of Education through Grants R305K05157 and R305A120813. The opinions expressed are those of the authors and do not represent views of the U.S. Department of Education. We would also like to thank Peg Burchinal, Paul Hanselman, Fred Oswald, David Purpura and Deborah Stipek for helpful comments on a prior draft and Ken T.H. Lee for his help with analyses. Finally, we would like to express appreciation to the school districts, teachers, and students who participated in the TRIAD research.

[1] School of Education, University of California, Irvine
[2] Morgridge College of Education, University of Denver

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

Abstract

To what extent do early boosts to simpler skills or behaviors support the long-term development of more sophisticated skills or behaviors? Substantial longitudinal associations between academic or socioemotional skills measured early and then later in childhood or adolescence are often taken as support of theories of skill-building processes. We argue that longitudinal correlations, even if adjusted for an extensive set of baseline covariates, constitute an insufficiently "risky" test of skill-building theories because they are consistent with alternative theories as well. Using the example of skill-building in mathematics, we consider a set of riskier tests of skill-building processes. We first show that experimental manipulation of early math skills generates much smaller "effects" on later math achievement than both bivariate and longitudinal non-experimental models with extensive baseline controls. We then conduct falsification tests that show puzzlingly high cross-domain associations between early math and later literacy achievement. Finally, we show that a skill-building model positing a combination of unmeasured stable factors and skill-building processes is able to reproduce the pattern of experimental impacts on children's mathematics achievement.

Keywords:

early childhood; interventions; skill-building; cognitive development; education

Strong longitudinal correlations are often observed in studies of academic and socioemotional domains of school readiness, across many years. Such longitudinal correlations constitute an important part of the empirical basis for various developmental and educational theories, including skill-building (e.g., Baroody, 1987; Stanovich, 1986; Cunha & Heckman, 2007), life-course development of psychopathology (e.g., Moffitt, 1993), and theories that posit reciprocal effects between skills and environments (e.g., Scarr & McCartney, 1983). Many of these correlations persist after adjusting for a large set of controls, including baseline measures of other academic and socio-emotional skills and capacities, domain-general cognitive abilities, and socioeconomic status (Aunola, Leskinen, Lerkkanen, & Nurmi, 2004; Bailey, Siegler, & Geary, 2014a; Duncan et al., 2007; Geary, Hoard, Nugent, & Bailey, 2013; Jordan, Kaplan, Ramineni, & Locuniak, 2009; Siegler et al., 2012; Watts, Duncan, Siegler, & Davis-Kean, 2014).

Researchers, including authors of this paper (Bailey et al., 2014a; Duncan et al., 2007; Watts et al., 2014), attribute to these robust correlations varying degrees of causality. For example, the Duncan et al. (2007; p. 1430) study of school readiness states: "…we implement rigorous analytic methods that attempt to isolate the effects of school-entry academic, attention, and socioemotional skills by controlling for an extensive set of prior child, family, and contextual influences that may be related to children's achievement." A cautious, yet superficial, alternative approach adopted by many authors writing about these kinds of correlations is to assert that because only random-assignment designs can prove causation, correlational evidence should not be interpreted as evidence of causality. But many of these same authors then go on to discuss the policy implications of their evidence (Reinhart et al., 2013), a linkage that requires causal evidence. One cannot have it both ways.

We argue that longitudinal correlations, even when adjusted for an extensive set of baseline controls, constitute an insufficiently "risky" test of skill-building and other popular developmental theories. We borrow from Meehl's (1978, 1990) insight that when diverse theories make the same predictions, it is important to conduct "risky" tests that have the ability to distinguish among them. A prediction is considered risky if the probability of such a prediction being true, *assuming that the theory is false*, is low. We argue that the correlational patterns documented in Duncan et al. (2007) do not constitute a risky test of skill-building theories, because they are likely to be found under a number of plausible competing theories, including ones positing that a combination of differentially stable general cognitive abilities, personality, and environmental affordances is responsible for generating the observed within-domain temporal correlations.

Focusing on the development of children's mathematics achievement, we use experimental evidence, falsification tests, and alternative model structures to show that riskier tests suggest a much smaller role for skill-building processes than commonly believed. We close with some recommendations for future studies.

Causal Mechanisms and Correlational Patterns

What are the causal mechanisms through which boosts in early school readiness skills and behaviors promote the development of much later academic and socioemotional skills? Skill-building models provide one clear answer: for math and literacy, early academic skills are the foundations upon which later skills are built. In the case of math, counting serves as a basis for children's early addition problem solving (Baroody, 1987), and addition is often employed as a subroutine of children's multiplication problem solving (Lemaire & Siegler, 1995). An implication of these models is that the children with the most solid early foundations of math

skills will, in the context of K-12 instruction, tend to maintain higher levels of math skills throughout childhood and adolescence.

In the development of reading skills, children's ability to match letters to sounds supports their learning to recognize written words, which in turn supports their vocabulary learning, which then supports their reading comprehension. Causal relations among these literacy skills are likely bi-directional with, for example, increases in reading comprehension facilitating more reading, which increases vocabulary (Stanovich, 1986).

The skill-building model of Cunha and Heckman (2007) is more comprehensive in allowing for these kinds of processes, but also posits a kind of multiplier effect in which early skills and capacities can increase the productivity of subsequent schooling and other investments. Moreover, it assumes that the list of "inputs" for the production of any particular skill or behavior may include a wide array of past skills and behaviors.

Substantial longitudinal correlations within domains of academic achievement and socioemotional behaviors are predicted by these skill-building causal models and are generally found in studies that estimate zero-order and regression-adjusted correlations within many domains of achievement and socioemotional skills across time (e.g., Duncan et al., 2007). For example, in national data from the 1998-99 Early Childhood Longitudinal Study (ECLS-K; Figure 1; measure and sample details in the appendix) the "math to math," "reading to reading," and "anti-social to anti-social" inter-wave correlations with fall of kindergarten values decay the most across the kindergarten year, but then flatten out to a moderate (about .45 to .65) magnitude by fifth grade. These kinds of patterns have been well documented in longitudinal correlational studies of children's cognitive abilities (Bayley, 1949; Tucker-Drob & Briley, 2014) and in the development of personality (Anusic & Schimmack, 2015).

[Figure 1 here]

Substantial cross-time correlations within domains of achievement and behavior such as these are clearly consistent with skill-building developmental theories. However, they do not constitute a risky test of such theories, because there are several plausible reasons one might expect to see them even if skill-building played a negligible role in cognitive development. We focus on one set of competing theories: those that posit an important role for some combination of foundational and relatively stable psychological characteristics and persistent environmental characteristics such as family functioning or neighborhood poverty. If these influences are not sufficiently well-captured with regression controls or other techniques for reducing omitted-variable bias, then the apparent role of skill-building processes in generating cross-time correlations could be seriously overstated.

In short, although skill-building theories of development are consistent with an overwhelming body of correlational evidence, so too are competing theories. When it comes to children's academic and social development, an inability to distinguish among such theories hampers the development of both theory and practice. Can riskier tests help us do better?

Riskier Tests

We discuss three promising approaches to estimate the importance of skill-building processes. First, and most important, is an experimental manipulation of children's early skills or behaviors, which provides the riskiest (i.e., most at risk of being refuted) test of theories of children's skill development. Second, we detail how risky falsification tests can be applied more widely to correlational data. And third, we show that longitudinal correlations and experimental impact patterns can be modeled in ways that make more precise (and thus riskier) predictions about the effects of prior skills or behaviors on later skills or behavior. Our empirical evidence

on these approaches is taken exclusively from the domain of math achievement. As we argue below, our analysis has implications for a much broader set of developmentally important skills and behaviors.

Experimental evidence. Randomly assigned boosts in school readiness skills and behaviors provide a very risky test of the theory that early skills are a powerful cause of the learning of new content in a manner that allows students with early skill advantages to maintain this advantage throughout school. If, as Duncan et al. (2007) imply, controls for an extensive set of prior child, family, and contextual influences enable an analyst to use nonexperimental data to compare otherwise similar groups of children who differ only in one particular school-entry skill or behavior, then the multivariate regression approach of Duncan et al. (2007) and others could identify the causal impact of a particular skill or behavior on later school success. We would then expect the patterns of "impacts" from these regression models to match those generated by a genuine random-assignment experiment.

To investigate whether well controlled correlational models of long-run achievement patterns reliably generate causal estimates, we draw data from the TRIAD (Technology-enhanced, Research-based, Instruction, Assessment, and professional Development) study. This study evaluated the long-run effectiveness of a learning intervention model, called TRIAD, which featured a preschool mathematics curriculum, called *Building Blocks*, as its key component (see Clements & Sarama, 2008). As explained in the appendix, the TRIAD study randomly assigned 42 schools with state-funded preschool programs in Massachusetts and New York either to a treatment condition in which the *Building Blocks* curriculum was implemented in preschool classes, or to a control condition in which preschool math was taught as usual.[1] In treatment schools, the curriculum was administered over the course of the preschool year. Math

achievement was measured in the fall and spring of the pre-kindergarten year, in the spring of the kindergarten, first, fourth and fifth-grade years, as well as in the fall of the fourth-grade year. Random assignment checks showed that treatment and control groups were balanced (Clements et al., 2011).

We first ignored experimental variation in the TRIAD data and used the study's control group to generate cross-time correlations between math achievement in the spring of the pre-kindergarten year and math achievement measured in all of the study's follow-ups. In contrast to Figure 1, we adjusted these correlations for the baseline achievement and demographic measures described in the appendix and also show 95 percent confidence intervals associated with each of the estimates. These confidence intervals were derived from standard errors that were adjusted for school-level clustering. The correlations shown in the "TRIAD regression-adjusted correlations" line of Figure 2 display the same kind of asymptotic pattern found with the unadjusted ECLS-K-based "math-to-math" correlations shown in Figure 1.[2]

Regression adjustments drop the estimated effect by around .20 SD-units, although these estimates still exceed .40 SD in fifth grade. Duncan and colleagues (2007) reported an average regression-adjusted math-to-math estimated "impact" of a remarkably similar .42 SD when comparing an early measure of children's mathematics achievement with later measures across six data sets. If the baseline covariates included in these regression-adjusted TRIAD estimates eliminate bias due to omitted variables, then we would expect to see a similar pattern in the experimental data.

[Figure 2 here]

The "TRIAD Treatment Impacts" line in Figure 2 shows that this is not at all the case. Treatment and control differences at the end of the pre-K year amounted to .63 SD – a large

impact. To establish comparability between this .63 SD impact and the 1.0 SD "impact" implicit in the regression-adjusted estimate shown in Figure 2's top line, we rescale this and all other experimental impact estimates by multiplying by 1/.63.[3] Rescaled impact estimates fall to about .46 SD within a year, and drop to statistically non-significant .08 SD and -.02 SD values for the two fourth-grade tests. The partial recovery of impacts in fifth-grade (.14 SD) is intriguing, but statistically indistinguishable from zero in this analysis.[4] Overall, the correlation-based estimate of the treatment effect is very close to the observed treatment effect one year after the end of treatment, but then much higher than the observed treatment effects at all subsequent waves.

TRIAD's ability to shed light on math skill-building processes is a function of the comprehensiveness of its initial impacts. As explained in the appendix, the end-of-Pre-K REMA test showed considerable variation in 4 subdomains of preschool mathematics knowledge: counting, patterning, measurement, and geometry. Bearing in mind that the psychometric properties of the overall REMA test, but not its subscales, have been established (Clements, Sarama, & Liu, 2008; Weiland et al., 2012), we grouped REMA items into each of these subdomains and created four measures defined as the proportion of correct responses on the items included in each category. We then tested the impact of the treatment on each standardized subdomain score. The intervention generated statistically significant impacts on all four subdomains: counting (β= 0.45, *SE*=0.06), patterning (β= 0.36, *SE*=0.06), geometry (β= 0.67, *SE*=0.06), and measurement (β= 0.20, *SE*=0.06). Because the intervention boosted a wide variety of preschool math skills, treated as opposed to control-group children should have had a considerably broad base of math competencies from which to build further math skills. In other words, these robust causal impacts suggest that the Building Blocks intervention provides an excellent foundation for tests of subsequent skill-building processes.

Returning to the patterns of experimental impacts in Figure 1, it is noteworthy that TRIAD treatment effects do not disappear completely immediately following the conclusion of treatment, which is indeed consistent with the impacts of skill building on children's academic development. However, skill-building processes following the conclusion of the intervention do not appear to sustain a substantial treatment effect much beyond first grade. This pattern of declining treatment effects is consistent with the patterns observed in many randomized controlled trials testing the effects of interventions designed to boost children's early academic skills (Bus & van IJzendoorn, 1999; Puma et al., 2012; Smith et al., 2013; for review, see Bailey et al., 2015, and Protzko, 2015).

The divergent lines in Figure 2 pose a profound challenge to the large correlational literature (including our own work) that has relied on longitudinal trajectories based on nonexperimental data to infer developmental processes. It would appear that experimentally induced changes in early skills may have temporary effects on children's subsequent learning, but, in the longer run, and in the context of the elementary schools most of these children attended, children's skills converge to trajectories governed by other processes.

A falsification test based on cross-domain correlations. Even in the absence of experimental data, it is possible to subject skill-building models to riskier tests using correlational data. Falsification tests provide one example. One set is based on the argument that if *within*-domain skill building processes were generating the strong pattern of longitudinal correlations shown in Figure 1, then *cross*-domain correlational patterns should be much weaker. Specifically, in the case of mathematics and reading, correlations between early and later math achievement scores should be persistently higher than correlations between early math and later reading. Figure 1 shows this is the case for math and anti-social behavior but decidedly not for

math-to-reading correlations, which are virtually indistinguishable from math-to-math correlations beyond 1st grade. That school-entry mathematics achievement is a robust predictor of children's long-term reading outcomes was also observed by Duncan and colleagues (2007) in their analysis of six longitudinal datasets (including the ECLS-K).

Skill-building models of children's academic development have a difficult time explaining why early mathematics achievement would exert a strong causal impact on later reading achievement. To be sure, skills such as language comprehension are common to both mathematics and reading achievement. But other evidence shows that correlations between early math and later reading scores (.26 in meta-analytic estimates in Duncan et al., 2007) are much higher than correlations between early reading and later math (.10).

Models consistent with temporal patterns of within-domain correlations. Consider the asymptotic rather than complete decline in the within-domain correlations show in Figure 1 and the regression-based correlations in Figure 2. What kind of skill-building processes would cause the later impacts of early achievement to become constant throughout development? A simple skill-building model could explain the shape of these lines if learning a basic skill earlier than one's peers persistently enabled a child to learn more advanced skills before his or her peers. For example, if learning to count before one's peers resulted in a high probability of learning to add before one's peers, which in turn resulted in a high probability of learning to multiply before one's peers, the correlation between counting skills and multiplication fluency could be high.

However, probabilities of learning later skills conditional on learning an early skill are the *product* of these interim probabilities. This model is shown with solid lines in Figure 3, where the paths MS_1 and MS_2 represent the impacts of a previous math skill on the immediately

following math skill. As long as these probabilities are less than one, we should observe some kind of exponential decay in early-to-late correlations as skills become more advanced.

[Figure 3 here]

The correlational estimates in Figures 1 and the top line of Figure 2 show a different pattern. They decay a bit over time, which is predicted by the hypothesis that skill building plays a role in the stability of individual differences in children's early academic achievement. However, they soon show a great deal of stability, especially after the first year of the study, which suggests that other factors or processes are at work.

Skill-building models may account for an asymptote in the effects of early math achievement on much later math achievement if individual differences in early achievement skills provide a basis for learning *across* development. This theory, illustrated by the dashed line in Figure 3, is appealing, given that skills acquired early in development clearly provide a basis for children's subsequent academic development. However, given that the most basic skills are quickly mastered by the vast majority of children (Engel, Claessens, Watts, & Farkas, 2016; Paris, 2005), individual differences in such skills are unlikely to account for robust longitudinal associations between earlier and later academic skills. For example, if almost all fifth graders can count to 10, it is difficult to imagine how children's ability to count to 10 would underlie individual differences in fifth graders' learning, despite the obvious importance of being able to count to 10 for learning mathematics throughout development. It is an open question whether broader foundational proficiencies (as described by the National Research Council, 2001) which are not quickly or easily mastered (e.g., OECD, 2014) could be developed early and result in more persistent effects on children's mathematics achievement.

An Alternative Developmental Model

What might account for the lingering discrepancy between experimental and correlational estimates of the effects of changes in early academic skills on academic skills several years later? The discrepant correlational and experimental patterns shown in Figure 2 and the patterns of correlations across time within and between academic domains estimated all suggest that omitted variables may be imparting a substantial upward bias to the correlational estimates. Directly and precisely measuring all of the important variables omitted from the regressions producing the top line of Figure 2, and then controlling for them in a regression, is a Herculean task.

An instructive alternative approach is to partition variance in children's academic and behavioral development into factors that exert a stable influence on children's development and – to continue with the example of mathematics achievement – children's mathematics knowledge assessed in the immediately preceding wave of data collection. A simple version of one such model is depicted in Figure 4. Developed by Steyer (1987) and implemented by Bailey and colleagues (2014b), this so-called latent state-trait model considers children's mathematics achievement at any time to be caused by two sets of factors:

1) Children's mathematics skill at the immediately preceding measurement occasion. The impacts of children's immediately preceding mathematics achievement on their subsequent mathematics achievement, indicated in Figure 4 by MS_1, MS_2, …, MS_{k-1}, could occur through content overlap and two aspects of skill-building – transfer of learning and indirect effects of mathematics achievement via increased motivation, teacher placement, or other mediators.
2) Characteristics with a stable influence on children's learning throughout development, represented by the loading of children's mathematics achievement at each occasion on a latent variable, labeled in Figure 4 as "Unmeasured persistent factor." This factor is likely comprised of environmental and personal factors that differ between individuals in a similar

manner across development and could include a much broader set of stable influences than is usually implied by the term "trait."

The model depicted in Figure 4 requires at least three waves of data to estimate, but has several advantages over traditional regression and SEM-based approaches for estimating the effects of children's prior and subsequent skills and behaviors. First, in the case of math, it simultaneously estimates effects of children's prior achievement on their later achievement during several inter-wave periods (the *MS* paths), thereby allowing for simultaneous tests of several theoretically important predictions (e.g., that a treatment effect on children's time 1 mathematics achievement will be reduced to the treatment effect times MS_1 at the second wave, to the original treatment effect times MS_1 times MS_2 by the third wave, etc.).

Second, the model can account for the apparent discrepancy between correlational and experimental estimates of the effects of children's prior mathematics achievement on their later mathematics achievement. If one assumes relatively large effects of stable environmental and personal factors on children's mathematics learning, then the model predicts an approximately exponential decay of treatment effects as the time between measurement occasions increases – a pattern consistent with experimental estimates – accompanied by high correlational stability.

Third, owing to the accumulating effects of stable factors on children's achievement across development, the model generates a testable prediction of increasing inter-year stability as children get older (Cole, Martin, & Steiger, 2005). This pattern is well established in the development of children's general cognitive ability, both at the phenotypic and genetic levels (Bayley, 1949; Tucker-Drob & Briley, 2014), and in the development of personality (Anusic & Schimmack, 2015). It is also evident in the ECLS-K dataset, where the correlations between mathematics achievement scores across waves increase, despite growing inter-wave intervals.

Children's mathematics achievement in the spring of kindergarten correlates .77 (*SE* = .01) with their mathematics achievement one year later in the spring of first grade, which correlates .80 (*SE* = .01) with their mathematics achievement two years later than that in the spring of third grade, which correlates .89 (*SE* = .01) with mathematics achievement two years later than that in the spring of fifth grade.

Fourth, the model can be easily adapted into an experimental design, in which the first wave of post-treatment achievement and the stable latent variable are simultaneously regressed on treatment status. In the TRIAD data used above, Watts and colleagues (2016) found a substantial impact of the intervention on children's math skills, but no effect on the latent variable representing stable factors that influence children's achievement across development. However, this finding warrants replication under conditions in which persistence may be most likely, including for subgroups, treatments, and populations for which skills affected by the intervention are least likely to develop under counterfactual conditions.

To be sure, the model depicted in Figure 4 leaves much to be desired because it merely assigns a key role to unmeasured persistent factors but does not identify them. As discussed above, the ideal test of what constitutes a persistent factor, and indeed, of whether such a factor truly exists, is to regress the unmeasured persistent factor on a source of exogenous variation, such as a randomly assigned intervention. In correlational datasets, the unmeasured persistent factor can be regressed on hypothesized sources of persistent variation (Bailey et al., 2014b), but this analysis is vulnerable to problems associated with all cross-sectional regression analyses, such as omitted-variable bias. Furthermore, the model is only one of many possible explanations for how early and later mathematics achievement are related. We hope other models will be

compared to the one we use here, both on the basis of fit to correlational datasets and their predictions about experimentally induced effects across time.

Comparing model with experimental estimates. Assuming no effects of early mathematics interventions on the unmeasured persistent factor (an assumption that deserves continued scrutiny), the key model parameters for estimating the pattern of treatment effects over time are the *MS* paths in Figure 4. Table 1 shows estimates of the MS paths from all of the correlational studies of the development of children's math achievement across pre-kindergarten and the elementary school years to which the state-trait model has been applied, to our knowledge. The average 1-year lagged *MS* paths from the three datasets are all modest, ranging from .29 to .35, depending on whether effect estimates are weighted by their underlying sample sizes. Put another way, the model depicted in Figure 4 predicts that the treatment effect should decay to approximately one-third of its previous magnitude each year.

How well do these estimates track patterns observed in the TRIAD experimental study? The bottom half of Table 1 shows estimates of *MS* paths implied by experimental impacts from three early mathematics interventions that followed children for at least a year following the conclusion of the given intervention included in the bottom half of Table 1. One-year lagged *MS* paths can be calculated by dividing experimental impacts at the end of a one-year period by the impacts at the beginning of that period. The average 1-year lagged *MS* path from the three studies ranged from .39 to .44, depending on how effects were weighted across studies. In other words, *MS* paths inferred from patterns of experimental impacts across time follow a pattern of decay that is only slightly less steep than that predicted by estimates derived from models based on a persistent latent factor. Patterns of impacts across time predicted by estimated and inferred weighted study average *MS* paths appear in Figure 5. These are calculated using the formula MS^t,

where *MS* is the estimated (.35) or inferred (.44) weighted study average *MS* path and *t* is the number of years since the end of the treatment. They are similar to each other and to the pattern of impacts in the TRIAD study (this is unsurprising for the inferred paths, given that these were based in large part on the TRIAD impact estimates). In fact, the average *MS* path estimated from the state-trait model falls within the confidence interval of every observed impact in the TRIAD study, while this is true of only 1 out of 5 regression-based estimates.

[Figure 5 here]

This pattern may generalize well to children's academic achievement outcomes. In a review of experimental estimates from 67 high-quality studies of early childhood education programs, Li and colleagues (under review) reported an average end-of-treatment effect size of .23, with estimates 0-1 years after treatment averaging approximately .10. Impact estimates from subsequent waves were smaller, but their precise values were sensitive to inclusion criteria.

As mentioned above in the discussion of Figure 2, correlational data track experimental impact estimates from TRIAD much more closely at the end of kindergarten than in later grades. In light of the estimates of the *MS* paths shown in Table 1, the model depicted in Figure 4 appears to track experimental estimates more closely in later grades than at the end of kindergarten. An obvious possible explanation for the larger 1-year lagged *MS* paths inferred from experimental estimates is some misspecification or bias in the Figure 4 model. Another possible explanation for the difference is that early mathematics interventions generate transitory impacts on a broader set of children's capacities (e.g., literacy skills [Sarama et al., 2012] or motivation), which independently boost children's later mathematics achievement. Although in the latter case, the state-trait estimates would actually provide more accurate estimates of *MS*

paths than those inferred from experimental impacts, the experimental impacts are more policy-relevant than the state-trait estimates in either case.

In summary, a model that allows for persistent unmeasured factors produces estimates consistent with the exponential decay in treatment effects observed in the most relevant set of experimental studies on children's early mathematics achievement, whereas traditional methods fail to do so after the first year or so. Notably, the correlational estimates most in line with experimental data were produced by the state-trait model from the largest sample to which it has been applied, which also spanned the longest time interval. Whether this particular model is the ideal specification (see Cole et al., 2005, for a discussion), whether intervention effects are likely to be confined to occasion-specific variation, rather than stable environmental and personal factors, and more generally, what comprises variation in these stable environmental and personal factors influencing children's mathematics learning throughout development remain open questions.

What stable factors are missing from our regression models? As noted above, we think that the "unmeasured persistent factors" in Figure 4 that influence children's academic and social development are in all likelihood a set of stable environmental and personal factors. Probable influences on child achievement throughout development include domain-general cognitive abilities, personality, and environmental affordances. Intelligence and working memory have been strongly implicated as key drivers of children's academic development in correlational studies (Deary, Strand, Smith, & Fernandes, 2007; Geary, Hoard, Nugent, & Bailey, 2012; Szücs, Devine, Soltesz, Nobes, & Gabriel, 2014), so much so that a general factor extracted from various cognitive tests was found to correlate .83 with a general factor extracted from academic achievement tests (Kaufman et al., 2012). Personality may also play a role, with

conscientiousness being the strongest Big Five personality correlate of academic achievement (Rimfield, Kovas, Dale, & Plomin, 2016).

Both personality and domain-general cognitive abilities are substantially influenced by differences in both genes and environments (Bouchard & McGue, 2003). Aside from latent environmental effects inferred from imperfect correlations between identical twins, strong designs have also identified effects of measured environments, such as adoption (van Ijzendoorn, Juffer, & Poelhuis, 2005; Kendler et al., 2015), maternal nutrition during prenatal development (Almond & Mazumder, 2011), or a very intensive early childhood education program (Campbell et al., 2001), on cognitive abilities many years later.

Implications for Design and Analysis in Developmental Research

We have argued that commonly used approaches to inferring skill-building processes from longitudinal correlation are based on insufficiently "risky" tests. In particular, we have shown that longitudinal correlations imply a much stronger skill-building process than does more direct evidence from experimental studies; that the similarity of within- and cross-domain correlations over time constitutes a falsification test that a simple math skill-building model does not pass; and that at least one alternative developmental model, which accounts for unmeasured factors, better reproduces the declining pattern of impacts generated by a large random-assignment evaluation of an intensive math skills intervention.

Although our review has been confined to mathematics learning, we suspect that we would find similar patterns in data on literacy, given the parallel nature of skill-building in those two domains of learning and the similar patterns of within-domain longitudinal correlations shown in Figure 1. Whether our conclusions about math generalize to other domains of interest in developmental research, such as anti-social behavior or executive functions, is less obvious.

The differing heights of patterns of math-to-reading and math-to-antisocial behavior correlational lines shown in Figure 1 suggest that whatever latent factor may underlie math and reading trajectories does not substantially impact anti-social behavior across development. However, an analogous developmental story may apply: Correlations between kindergarten anti-social behavior and anti-social behavior at subsequent waves follow a pattern similar to those for children's academic skills (Figure 1), suggesting that other factors may generate stability in children's anti-social behavior across time. If these factors are not well measured, the auto-regressive effects of anti-social behavior will be exaggerated. Consistent with this possibility, Anusic and Schimmack (2015) observed substantial stability in the inter-wave correlations of personality, affect, self-esteem, and life satisfaction, with the highest stability in personality.[5] It is important to emphasize that the existence of effects of stable environmental and personal factors on children's academic development does not preclude skill-building processes, nor does the existence of empirically stable environmental and personal factors imply that such factors are immutable. One possible reason for fadeout in early mathematics learning interventions such as Building Blocks is that it is unrealistic to expect impact persistence if children who receive high-quality interventions remain in low-resource communities, including low-quality schools (Brooks-Gunn, 2003; Duncan & Magnuson, 2013). In this case, proficiencies developed by interventions are not evanescent, but rather result in enhanced knowledge states that are subsequently left fallow by environments that do little or nothing to build upon these early learnings[6].

This fadeout problem has not been lost on intervention designers. For example, Clements and colleagues (2013) included a follow-through condition in the TRIAD study. In this alternative treatment arm, kindergarten and first grade teachers received additional professional

development designed to help them build upon the gains students made during preschool. Although the original treatment impacts still faded for this group, they faded more slowly than for the students that only received the preschool treatment without follow-through.

Following a demonstration of mismatched correlational and experimental findings, it is common to call for more randomized controlled trials. We certainly endorse RCTs because they can provide the strongest evidence on skill-building processes, but we recognize that many (including most lab-based experiments) have limited external validity, sometimes target a bundle of constructs that may benefit children but render implications for developmental processes unclear, and rarely track longer-term persistence. We also endorse the pursuit of data from sibling, neighborhood and school fixed effects models and from so-called "natural experiments" such as school policy changes that provide significantly more intensive academic training to a subset of students who can be compared with a very similar group of "untreated" students. An example is the "double dose" algebra training introduced to low-achieving ninth-graders in the Chicago Public Schools as defined by an eighth-grade math test score cutoff (Cortes and Goodman, 2014). Natural experiments (including the double-dose program) are not unproblematic (they often face the same problems with external and construct validity), but they can help to adjudicate competing theories of learning.

Should we give up on the idea of using correlational data analysis to make theoretical or policy-relevant inferences about children's academic development? We are not so pessimistic. Indeed, we believe that when exposed to riskier tests and informed by prior experimental work, correlational data analyses can also help triangulate to the most useful theories of children's academic development: theories that can accurately predict when the effects of academic interventions will fade out or persist.

References

Almond, D., & Mazumder, B. (2011). Health capital and the prenatal environment: the effect of Ramadan observance during pregnancy. *American Economic Journal: Applied Economics*, 56-85.

Anusic, I., & Schimmack, U. Stability and change of personality traits, self-esteem, and well-being: Introducing the meta-analytic stability and change model of retest correlations. *Journal of Personality and Social Psychology, 110*, 766-781.

Aunola, K., Leskinen, E., Lerkkanen, M.-L., & Nurmi, J.-E. (2004). Developmental dynamics of math performance from pre-school to Grade 2. *Journal of Educational Psychology, 96*, 699–713.

Bailey, D. H., Duncan, G., Odgers, C., & Yu, W. (2015). Persistence and fadeout in the impacts of child and adolescent interventions (Working Paper No. 2015-27). Retrieved from the Life Course Centre Working Paper Series Website: http://www.lifecoursecentre.org.au/working-papers/persistence-and-fadeout-in-the-impacts-of-child-and-adolescent-interventions

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 pre-existing differences? *Developmental Psychology*.

Bailey, D. H., Siegler, R. S., & Geary, D. C. (2014). Early predictors of middle school fraction knowledge. *Developmental Science, 17*, 775-785.

Bailey, D. H., Watts, T. W., Littlefield, A. K., & Geary, D. C. (2014b). State and trait effects on individual differences in children's mathematical development. *Psychological Science, 25*, 2017-2026. doi:10.1177/0956797614547539

Baroody, A. J. (1987). The development of counting strategies for single-digit addition. *Journal for Research in Mathematics Education*, 141-157.

Bayley, N. (1949). Consistency and variability in the growth of intelligence from birth to eighteen years. *The Pedagogical Seminary and Journal of Genetic Psychology, 75*, 165–196. doi:10.1080/08856559.1949 .10533516

Bouchard, T. J., & McGue, M. (2003). Genetic and environmental influences on human psychological differences. *Journal of Neurobiology, 54*, 4-45.

Bus, A. G., & van IJzendoorn, M. H. (1999). Phonological awareness and early reading: A meta-analysis of experimental training studies. *Journal of Educational Psychology, 91*(3), 403.

Campbell, F. A., Pungello, E. P., Miller-Johnson, S., Burchinal, M., & Ramey, C. T. (2001). The development of cognitive and academic abilities: Growth curves from an early childhood educational experiment. *Developmental Psychology, 37*(2), 231-242. doi:http://dx.doi.org/10.1037/0012-1649.37.2.231

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

Clements, D. H., & Sarama, J. (2013). *Building Blocks, Volumes 1 and 2*. Columbus, OH: McGraw-Hill Education.

Clements, D. H., Sarama, J., Khasanova, E., & Van Dine, D. W. (2012). *TEAM 3-5—Tools for elementary assessment in mathematics*. Denver, CO: University of Denver.

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*(4), 457-482. doi: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., & 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. doi: 10.3102/0002831212469270

Cortes, K. E., & Goodman, J. S. (2014). Ability-tracking, instructional time, and better pedagogy: The effect of Double-Dose Algebra on student achievement. *The American Economic Review*, 104(5), 400-405. doi:http://dx.doi.org/10.1257/aer.104.5.400

Cunha, F., & Heckman, J. J. (2007). The technology of skill formation. *American Economic Review, 97*(2): 31–47. doi:10.3386/w12840

Deary, I. J., Strand, S., Smith, P., & Fernandes, C. (2007). Intelligence and educational achievement. *Intelligence*, *35*, 13-21.

Duncan, G. J., Dowsett, C. J., Claessens, A., Magnuson, K., Huston, A. C., Klebanov, P., ... & Japel, C. (2007). School readiness and later achievement. *Developmental Psychology, 43*(6), 1428-1446. doi:http://dx.doi.org/10.1037/0012-1649.43.6.1428.supp

Engel M., Claessens A., Watts, T. W., Farkas, G. (2016). Mathematics content coverage and student learning in kindergarten. *Educational Researcher, 45*(5), 293-300.

Geary, D. C. (2010). Missouri longitudinal study of mathematical development and disability. *British Journal of Educational Psychology Monograph Series II, 7,* 31–49.

Geary, D. C., Hoard, M. K., Nugent, L., & Bailey, D. H. (2012). Mathematical cognition deficits in children with learning disabilities and persistent low achievement: A five-year prospective study. *Journal of Educational Psychology, 104*, 206-223.

Geary, D. C., Hoard, M. K., Nugent, L., & Bailey, D. H. (2013). Adolescents' functional numeracy is predicted by their school entry number system knowledge. *PLoS One, 8*(1), e54651.

Gresham, F., and Elliot, S. (1990). *Social skills rating system*. Circle Pines, MN: American Guidance Services, Inc.

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*, 172-177. doi:10.1111/j.1750-8606.2008.00061.x

Jordan, N. C., Kaplan, D., Ramineni, C., & Locuniak, M. N. (2009). Early math matters: Kindergarten number competence and later mathematics outcomes. *Developmental Psychology*, *45*(3), 850–867. doi:10.1037/a0014939

Kaufman, S. B., Reynolds, M. R., Liu, X., Kaufman, A. S., & McGrew, K. S. (2012). Are cognitive g and academic achievement g one and the same g? An exploration on the Woodcock–Johnson and Kaufman tests. *Intelligence, 40*, 123-138.

Kendler, K. S., Turkheimer, E., Ohlsson, H., Sundquist, J., & Sundquist, K. (2015). Family environment and the malleability of cognitive ability: A Swedish national home-reared and adopted-away cosibling control study. *Proceedings of the National Academy of Sciences, 112*, 4612-4617.

Lemaire, P., & Siegler, R. S. (1995). Four aspects of strategic change: contributions to children's learning of multiplication. *Journal of Experimental Psychology: General, 124*(1), 83-97.

Massachusetts Department of Elementary and Secondary Education, (2011). *Massachusetts curriculum framework for mathematics.* Malden: Massachusetts Department of Elementary and Secondary Education.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology, 46*, 806-834.

Meehl, P. E. (1990). Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports, 66*, 195-244.

Moffitt, T. E. (1993). Adolescence-limited and life-course-persistent antisocial behavior: a developmental taxonomy. *Psychological review, 100*, 674-701.

National Council of Teachers of Mathematics, (2000). *Principles and standards for school mathematics (Vol. 1).* Reston, VA: National Council of Teachers of Mathematics.

NICHD Early Child Care Research Network. (2002). Early child care and children's development prior to school entry: Results from the NICHD Study of Early Child Care. *American Educational Research Journal, 39*, 133–164.

Nguyen, T., Watts, T. W., Duncan, G. J., Clements, D. H., Sarama, J. S., Wolfe, C., & Spitler, M. E. (2016). Which preschool mathematics competencies are most predictive of fifth grade achievement? *Early Childhood Research Quarterly, 36*, 550-560.

Paris, S. G. (2005). Reinterpreting the development of reading skills. *Reading Research Quarterly, 40*, 184-202. doi: 10.1598/RRQ.40.2.3

Protzko, J. (2015). The environment in raising early intelligence: A meta-analysis of the fadeout effect. *Intelligence, 53*, 202-210.

Pollack, J. M., Rock, D. A., Weiss, M. J., Burnett, S. A., Tourangeau, K., West, J., & Hausken, E. G. 2005a. "Early Childhood Longitudinal Study-Kindergarten Class of 1998-99 (ECLS-K), psychometric report for the fifth grade." U.S. Department of Education, National Center for Education Statistics, Washington, DC.

Pollack, J. M., Rock, D. A., Weiss, M. J., Burnett, S. A., Tourangeau, K., West, J., & Hausken, E. G. 2005b. "Early Childhood Longitudinal Study- Kindergarten Class of 1998-99 (ECLS-K), psychometric report for the third grade." U.S. Department of Education, National Center for Education Statistics, Washington, DC.

Rock, D. A. & Judith M. P. 2002. "Early Childhood Longitudinal Study-Kindergarten Class of 1998-99 (ECLS-K), psychometric report for kindergarten through first grade." U.S. Department of Education, National Center for Education Statistics, Washington, DC.

Puma, M., Bell, S., Cook, R., Heid, C., Broene, P., Jenkins, F., Mashburn, A., & Downer, J. (2012). *Third Grade Follow-up to the Head Start Impact Study Final Report, OPRE Report # 2012-45*. Washington, DC: Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.

Reinhart, A. L., Haring, S. H., Levin, J. R., Patall, E. A., & Robinson, D. H. (2013). Models of not-so-good behavior: Yet another way to squeeze causality and recommendations for practice out of correlational data. *Journal of Educational Psychology*, *105*(1), 241-247.

Rimfeld, K., Kovas, Y., Dale, P. S., & Plomin, R. (2016, February 11). True grit and genetics: Predicting academic achievement from personality. *Journal of Personality and Social Psychology.* Advance online publication. http://dx.doi.org/10.1037/pspp0000089

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.

Scarr, S., & McCartney, K. (1983). How people make their own environments: A theory of genotype→ environment effects. *Child Development*, 424-435.

Schenke, K., Lam, A. C., Rutherford, T., & Bailey, D. H. (2016). Construct confounding among predictors of mathematics achievement. *AERA Open*, *2*, 2332858416648930.

Siegler, R. S., Duncan, G. J., Davis-Kean, P. E., Duckworth, K., Claessens, A., Engel, M., et al. (2012). Early Predictors of High School Mathematics Achievement. *Psychological Science*, *23*, 691-697.

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. doi:10.3102/0002831212469045

Stanovich, K. E. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Reading Research Quarterly*, 360-407.

Steyer, R. (1987). Konsistenz und Spezifitaet: Definition zweier zentraler Begriffe der Differentiellen Psychologie und ein einfaches Modell zu ihrer Identifikation [Consistency and specificity: Definition of two central concepts of differen- tial psychology and a simple model for their identification]. *Zeitschrift für Differentielle und Diagnostische Psychologie*, *8*, 245–258.

Szűcs, D., Devine, A., Soltesz, F., Nobes, A., & Gabriel, F. (2014). Cognitive components of a mathematical processing network in 9- year- old children. *Developmental Science*, *17*(4), 506-524.

Tucker-Drob, E. M., & Briley, D. A. (2014). Continuity of genetic and environmental influences on cognition across the life span: A meta-analysis of longitudinal twin and adoption studies. *Psychological bulletin*, *140*, 949-979.

Tourangeau, K., Nord, C., Lê, T., Sorongon, A. G., Najarian, M., & Hausken, E. G. (2009). Early childhood longitudinal study, kindergarten class of 1998-99 (ECLS-K) combined user's manual for the ECLS-K Eighth-Grade and K-8 full sample data files and electronic codebook. U.S. Department of Education, National Center for Education Statistics, Washington, DC.

Van Ijzendoorn, M. H., Juffer, F., & Poelhuis, C. W. K. (2005). Adoption and cognitive development: a meta-analytic comparison of adopted and nonadopted children's IQ and school performance. *Psychological Bulletin*, *131*(2), 301-316.

Watts, T. W., Clements, D. H., Sarama, J., Wolfe, C. B., Spitler, M. E., & Bailey, D. H. (2016). Does early mathematics intervention change the processing underlying children's mathematics achievement? *Journal of Research on Educational Effectiveness.*

Watts. T.W., Duncan, G. J., Clements, D. H., Sarama, J., Wolfe, C. B., & Spitler, M. E. (under review). *Early mathematical competencies and long-run achievement: Which kindergarten domains predict later achievement?*

Watts, T. W., Duncan, G. J., Siegler, R. S., & Davis-Kean, P. E. (2014). What's past is prologue: Relations between early mathematics knowledge and high school achievement. *Educational Researcher*, *43*(7), 352-360. doi: 10.3102/0013189X14553660

Weiland, C., Wolfe, C. B., Hurwitz, M. D., Clements, D. H., Sarama, J. H., & Yoshikawa, H. (2012). Early mathematics assessment: Validation of the short form of a prekindergarten and kindergarten mathematics measure. *Educational Psychology*, *32*(3), 311-333.

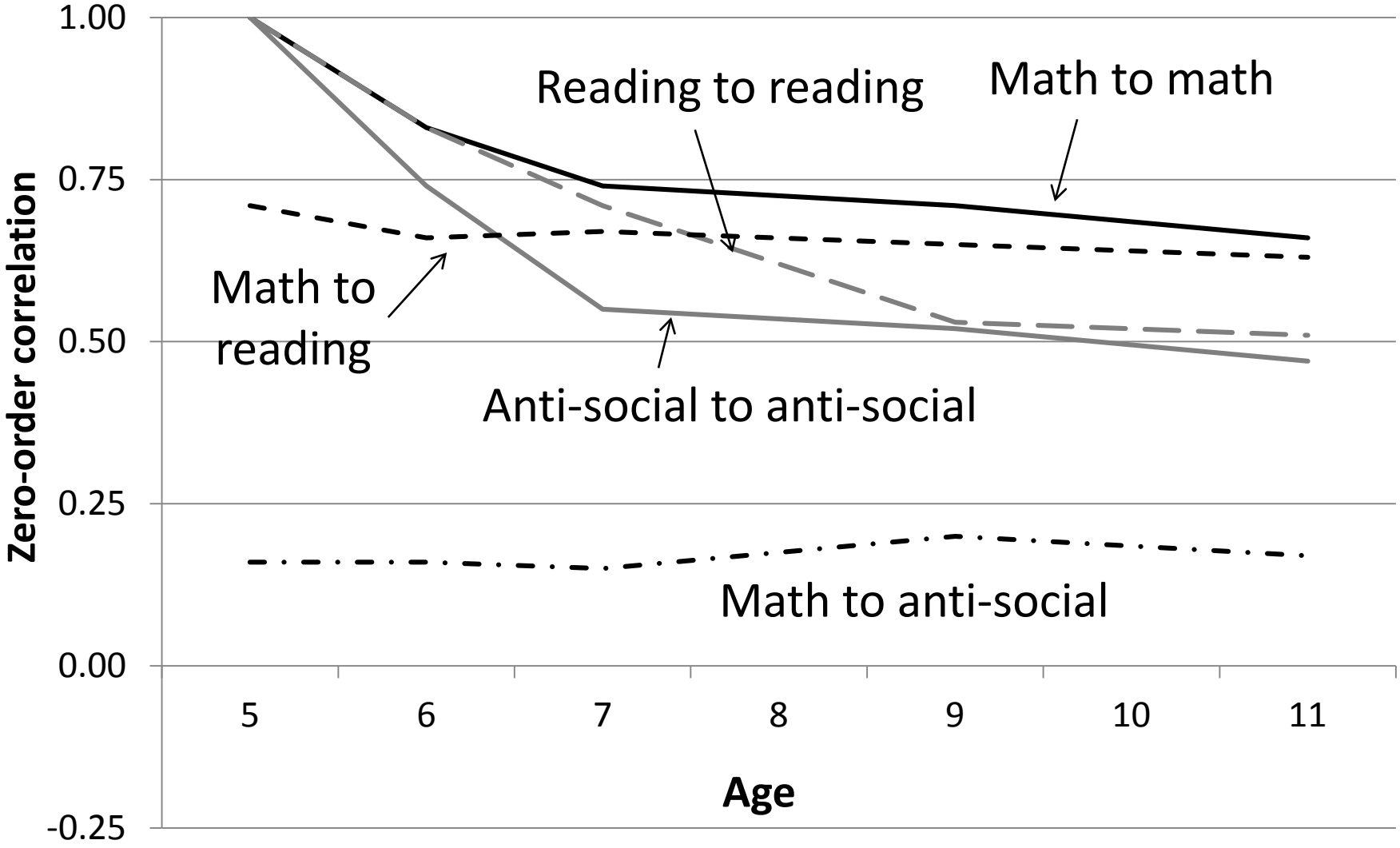
Wicherts, J. M. (2016). The importance of measurement invariance in neurocognitive ability testing. *The Clinical Neuropsychologist*, 1-11.

Table 1
Estimates of MS (Math Skills) paths from observational and experimental data

	Source	Sample size	Period	Implied 1-year *MS* estimate
State-trait estimates	Bailey et al., 2014b: Missouri Math Study	292	Grade 1-Grade 2	0.26
	Bailey et al., 2014b: Missouri Math Study	292	Grade 2-Grade 3	0.18
	Bailey et al., 2014b: Missouri Math Study	292	Grade 3-Grade 4	0.20
	Bailey et al., 2014b: SECCYD	1124	Grade 1-Grade 3	0.58
	Bailey et al., 2014b: SECCYD	1124	Grade 3-Grade 5	0.30
	Watts et al., 2016: TRIAD	834	PreK-K	0.25
	Watts et al., 2016: TRIAD	834	K-Grade 1	0.04
	Watts et al., 2016: TRIAD	834	Grade1-Grade 4	0.51
	Simple average			**0.29**
	Unweighted study average			**0.31**
	Weighted study average			**0.35**
Experimental estimates	Current paper: TRIAD	834	PreK-K	0.46
	Current paper: TRIAD	834	K-Grade 1	0.48
	Current paper: TRIAD	834	Grade 1-Grade 4	0.48*
	Current paper: TRIAD	834	Grade 4-Grade 5	N/A**
	Hofer et al., 2013: TRIAD	1192	Pre-K-K	0.28
	Hofer et al., 2013: TRIAD	1129	K-Grade 1	0.67
	Smith et al., 2013	320	Grade 1-Grade 2	0.22***
	Simple average			**0.43**
	Unweighted study average			**0.39**
	Weighted study average			**0.44**

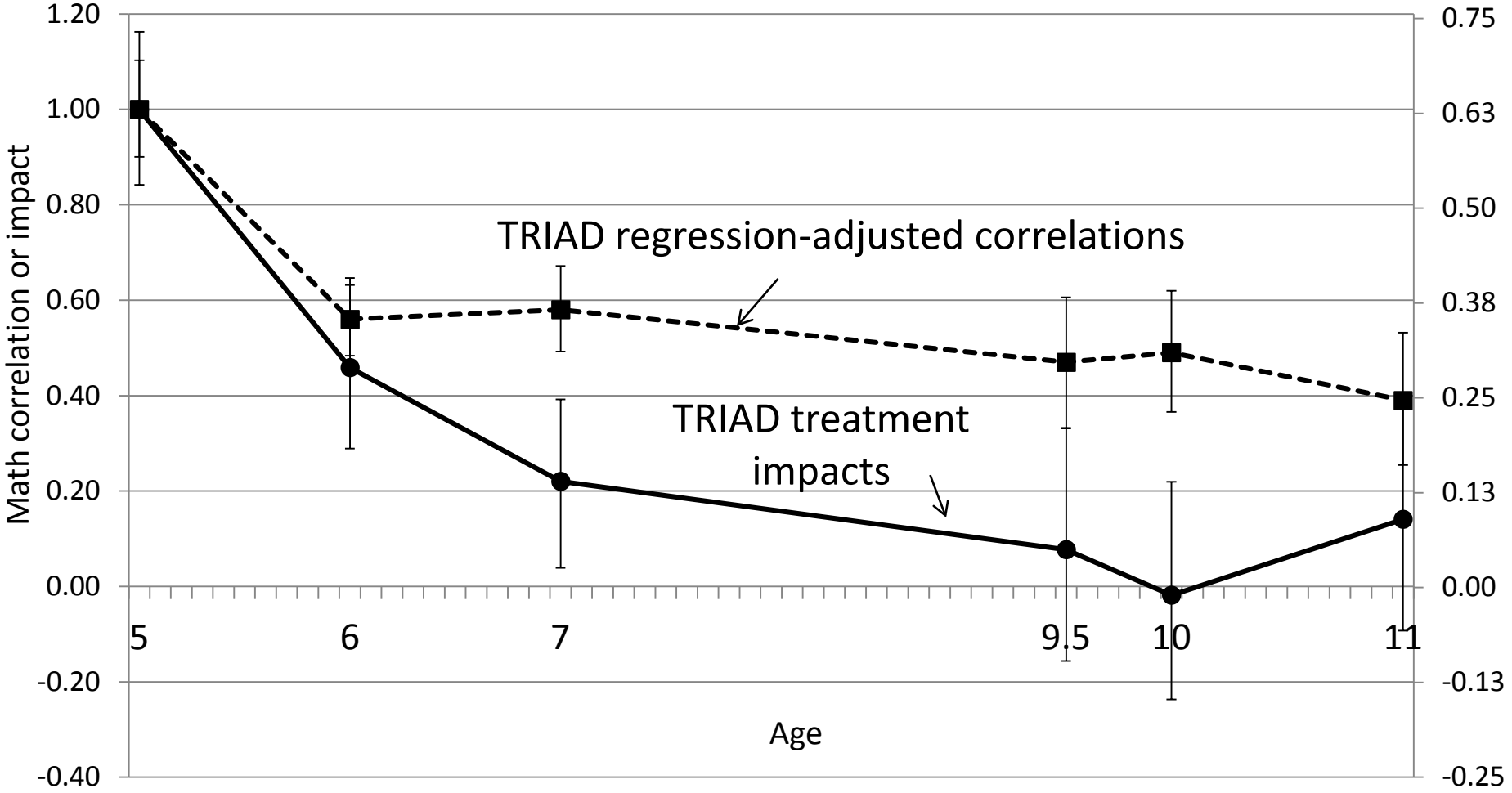
Note. One-year MS estimates from intervals with multiple lags are calculated by raising the reported estimate to the power of 1/t, where t is the number of years in the given interval. MS estimates from experimental studies are calculated by dividing a treatment effect by a prior treatment effect, and are corrected using the same exponential transformation when intervals between measurements vary by an amount different from 1 year. For information regarding the Missouri Math Study, see Geary, 2010. SECCYD stands for the Study of Early Childcare and Youth Development (see NICHD Early Child Care Research Network, 2002). Information regarding the TRIAD (Technology-enhanced, Research-based, Instruction, Assessment, and professional Development) study is presented in the Appendix and in Clements et al., 2013.
* 4th grade is the average of 2 4th grade scores. Average interval is 2.75 years
** 5th grade estimate is higher than 4th grade estimate; neither is statistically distinguishable from 0
*** Treatment effects were calculated as the average of the 3 standardized mathematics tests administered at both waves.

Figure 1: Bivariate correlations with Fall of Kindergarten measures



Source: ECLS-K 1998-1999 cohort. All correlations are p<.05

Figure 2: Regression-adjusted correlations and experimental impacts in TRIAD



Note: All 4th and 5th grade impacts are p>.05. All correlations and other impacts are p<.05. Impacts are rescaled to be 1.0 in the spring of pre-K, Right scale shows non-rescaled impacts. Vertical lines depict 95% confidence intervals.

Figure 3: Direct and indirect paths in a math skill-building model

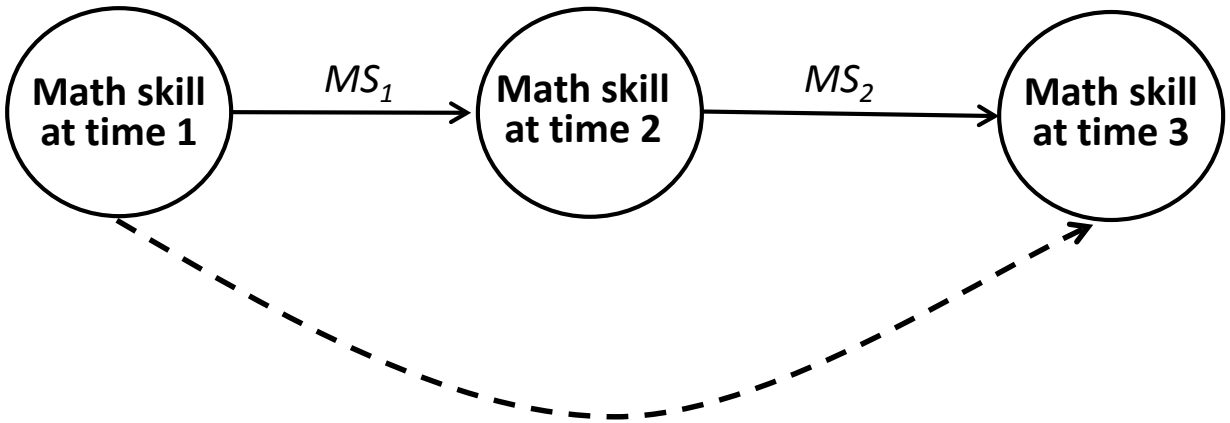
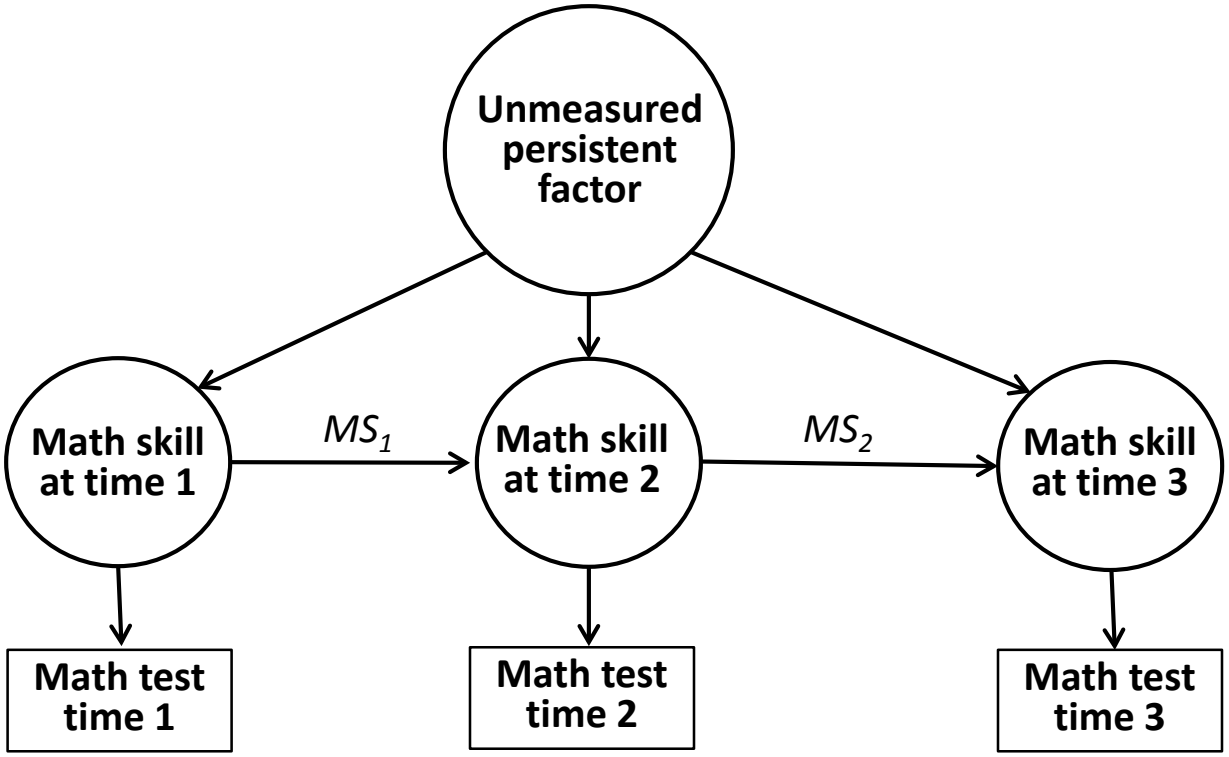


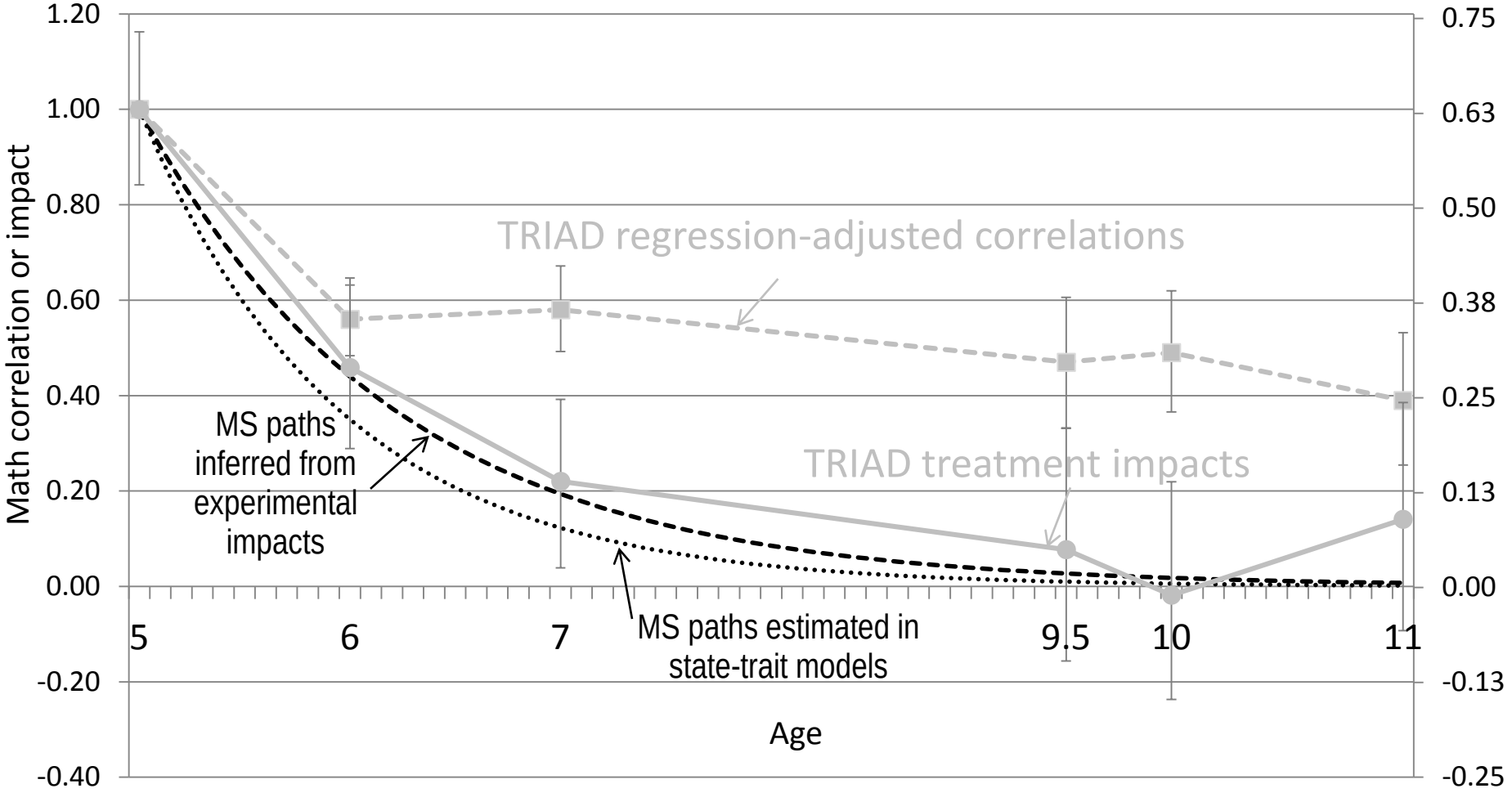
Figure 4: An alternative math skill-building model with unmeasured persistent influences



Predicted standardized treatment effects on math skill following 1 SD boost in Math skill at time 1:

Math skill at time 1: 1
Math skill at time 2: MS_1
Math skill at time 3: $MS_1 * MS_2$

Figure 5: Correlations inferred from *MS* path estimates in Table 1



Note: All 4th and 5th grade impacts are p>.05. All correlations and other impacts are p<.05. Impacts are rescaled to be 1.0 in the spring of pre-K, Right scale shows non-rescaled impacts. Vertical lines depict 95% confidence intervals.

END NOTES

1. A second treatment condition, which we ignore, assigned students to receive the pre-K *Building Blocks* curriculum coupled with "follow-through" math support to kindergarten and first-grade teachers
2. Full correlation matrices for measures of children's mathematics achievement administered at several waves across development are available in Bailey et al. (2014b).
3. This means, for example, that TRIAD's .63 SD impact at the end of pre-kindergarten is shown as 1.0 and its the .29 SD experimental impact at the end of the kindergarten is shown as .29 SD (= .18 /.63). The scale for the correlations and rescaled impact estimates are shown on the left y-axis, while the non-rescaled estimates appear on the right y-axis.
4. Using a 2-level random intercept logistic regression model, Clements and colleagues (under review) found a statistically significant treatment impact of the TRIAD intervention at the spring of 5^{th} grade.
5. Anusic and Schimmack (2015) reported values of "1-year stability of the change component" comparable to the 1-year *MS* paths reported in this article, Table 1. For personality, affect, self-esteem, and life satisfaction, the authors reported values of .25, .88, .79, and .78, respectively. Thus, the .35 estimate in Table 1 indicates that inter-individual stability across time in children's mathematics achievement more closely resembles the pattern observed for personality than for affect, self-esteem, or life satisfaction.
6. A closely related measurement issue that deserves attention is the need for more nuanced assessments of children's mathematics knowledge. Measures in extant studies are based on single scores rather than more cognitively-complex and diagnostic assessments (cf. Tatsuoka et al., in press). Thus, children assigned the same score are assumed to have the same knowledge state, such as children who raised their scores via participation in an intervention group matched to controls who achieved that score without the intervention (Bailey et al., 2016). However, the control children will likely have a far longer, far more extensive, set of experiences that led to the same score; for example, building parallel distributed process networks of broad reach across the brain, which, because they have been reinforced for years, have established retrieval paths that are myriad, strong, and stable. Thus, future measures of the two groups of children may yield different scores even if subsequent experiences are the same. Future measurement studies, including those that investigate measurement invariance across subgroups (Wicherts, 2016), are warranted to investigate such possibilities.

Online Supplementary Material for "Risky Business: Correlation and Causation in Longitudinal Studies of Skill Development"

Data

ECLS-K. Data for Figure 1 were drawn from the Early Childhood Longitudinal Study-Kindergarten Cohort (ECLS-K). In the fall of 1998, The National Center for Educational Statistics (NCES), began data collection for the ECLS-K, and a nationally representative sample of kindergarteners were drawn from approximately 1000 public and private schools across the country. The children were then followed longitudinally through the end of eighth grade, and data was collected from students, parents, teachers and administrators in kindergarten, first grade, third grade, fifth grade, and eighth grade. For the analyses presented in the current paper, we relied on data collected in the fall and spring of kindergarten, and in the spring of the first, third, and fifth grade years. NCES designed sampling weights for each wave, and for our analyses, we used panel weights designed for longitudinal analyses of child-level data from kindergarten through the end of fifth grade (see Table S2 note). The ECLS-K has been widely used to study K-12 education and child development, and further details regarding the study can be found in the NCES user manual (Tourangeau et al., 2009).

TRIAD. Data for Figure 2 were drawn from the TRIAD (Technology-enhanced, Research-based, Instruction, Assessment, and professional Development) scale-up evaluation, a multi-site, longitudinal, study designed to assess the long-run impacts of the *Building Blocks* preschool mathematics curriculum (see Clements & Sarama, 2013). The TRIAD study randomly assigned 42 elementary schools serving low-income neighborhoods in either Buffalo, New York or Boston, Massachusetts to one of three conditions: 1) *Building Blocks* preschool curriculum; 2) *Building Blocks* preschool curriculum with follow-through; 3) control (business as usual).

Schools were randomly-assigned through a blocking procedure, in which schools were grouped into 8 blocks based on similarity of state-achievement test scores, and random assignment then occurred within each blocking group. Schools assigned to either of the *Building Blocks* treatment conditions implemented the *Building Blocks* curriculum during preschool, and preschool teachers attended approximately 13 pedagogical development (PD) sessions designed to help them improve the mathematics taught in their classrooms. Control schools kept their pre-existing preschool mathematics programs. The "*Building Blocks* with follow-through" condition included additional PD for kindergarten and first grade teachers in which teachers were encouraged to build upon the mathematics the students learned during preschool.

The *Building Blocks* curriculum (Clements & Sarama, 2013) was based on developmental and cognitive theory, and the curriculum was designed to help children mathematize their everyday activities. The curriculum was organized into highly-sequenced learning trajectories, in which students developed conceptual understanding, procedural skill, and problem solving competencies in various foundational areas of mathematics (e.g., counting, geometry). The curriculum was also paired with the *Building Blocks* software, which further helped teachers personalize instruction to each child's unique needs. In total, the curriculum was designed to take approximately 15 to 30 minutes each day. Initial analyses showed that the curriculum was implemented with strong fidelity during preschool (Clements et al., 2011). For full details regarding the curriculum development and fidelity of implementation, see Clements and Sarama, 2013 and Clements et al., 2011.

TRIAD evaluators drew a random sample of 1375 students from the 42 participating schools, and students' mathematics achievement was measured during preschool, kindergarten, first grade, fourth grade, and fifth grade. Clements and colleagues (2011) reported a large initial

treatment effect at the end of preschool for students in either of the *Building Blocks* conditions and this effect faded by approximately 60% by the end of first grade for the treated students with no follow-through, and by 30% for the students that did receive follow-through (Clements et al., 2013).

In the current analyses, we only use data collected form children in either the *Building Blocks* treatment without follow-through condition or the control group (n= 834; school n= 30). Table S1 presents baseline descriptive characteristics for treatment and control students, and we found no significant differences between the groups on any background measure observed. Reflecting the low-income status of the sample, 85% of students qualified for free or reduced price lunch, 51% identified as Black, and 23% as Hispanic.

Measures

ECLS-K Achievement. We draw on measures of student achievement in mathematics and reading collected during the fall and spring of kindergarten, and during the spring of the first, third, and fifth grade years. The mathematics assessment was designed to assess conceptual and procedural knowledge in mathematics, and topics spanned basic counting in kindergarten to fractions and pre-algebra in fifth grade. The reading assessment was designed to assess vocabulary knowledge and reading comprehension, and topics spanned letter knowledge in kindergarten to understanding words in context and drawing inferences based on textual clues in fifth grade. The assessments were individually administered, and the assessments were adaptive such that the items students received were based on their performance on previous items. Assessment items in math and reading were drawn from the National Assessment of Educational Progress (NAEP), as well as the National Longitudinal Study of 1988 and the Educational Longitudinal Study of 2002. Items were ordered on a hierarchical scale, and proficiency scores

were calculated for each assessment in each wave. We rely on the IRT scores generated by NCES for each respective assessment at each respective wave. For more information regarding the ECLS-K achievement measures, see Pollack et al., 2005 and Tourangeau et al., 2009.

ECLS-K Behavior. We rely on teacher ratings of anti-social behavior (i.e., externalizing behavioral problems) measured during the fall and spring of kindergarten, as well as during the spring semesters of first, third, and fifth grade. The anti-social behavior scale was drawn from the Social Skills Rating System (Gresham & Elliott, 1990). Teachers used a frequency scale to indicate how often a student exhibited a certain behavior (items ranged from "1" = "never" to "4" = "very often"). NCES used factor analyses to confirm the items used in the scales, and items for the anti-social behavioral measure focused on whether the child disrupted class or if they acted aggressively toward other students. For the waves of data used here, reliability for the anti-social behavior problem scale ranged from 0.86 to 0.90 (see Rock & Pollack, 2002; Pollack et al., 2005a; 2005b).

TRIAD mathematics achievement. During the fall and spring of preschool, spring of kindergarten, and spring of first grade, mathematics achievement was assessed using the Research-based Early Math Assessment (REMA; Clements, Sarama, & Liu, 2008; Clements, Sarama, & Wolfe, 2011). The REMA was designed to measure the mathematics understanding of children between the ages of 3 and 8, and it was administered through two one-on-one interviews. The interviews were taped and coded, and students were rated on both their correctness and strategy use. Topics on the exam included counting, measurement, geometry and place value, and the REMA scores were converted to Rasch-IRT scales. The measure was validated across three different samples of young children, and the measure has been shown to have a 0.74 correlation with the Woodcock-Johnson Applied Problems subtest (Clements,

Sarama, & Wolfe, 2011). Further, the measure was found to have strong internal reliability (Cronbach's $\alpha = 0.94$; Clements, Sarama, & Wolfe, 2011).

In order to assess the degree to which the *Building Blocks* program impacted various aspects of preschool mathematics achievement, we also created sub-scores of the REMA. We replicated the measures used by Nguyen et al. 2016. In their study, they coded items of the REMA into various domains of preschool knowledge found across a wide variety of state-defined preschool mathematics standards documents (e.g., Massachusetts Department of Elementary and Secondary Education, 2011) and early childhood advisory panels and mathematics teaching organizations (e.g., National Council of Teachers of Mathematics, 2002). The final categories used were *Counting and Cardinality*, *Patterning*, *Geometry* and *Measurement and Data.* The *Counting and Cardinality* category included items that asked students to count objects and recognize numbers. *Patterning* items asked students to extend and duplicate patterns. The *Geometry* category was comprised of items that asked students to recognize shapes. Finally, the *Measurement and Data* category included items that asked students to identify the attributes of shapes by using measuring instruments.

In the fall and spring of fourth grade, and the spring of fifth grade, mathematics achievement was measured using the TEAM 3-5, a variant of the REMA (Clements, Sarama, Khasanova, & Van Dine, 2012). The TEAM 3-5 was a paper-and-pencil test that was aligned to the developmental progressions used in the REMA. However, simpler topics are retired (e.g., counting) and replaced with more advanced topics (e.g., fractions). Topics on the test included multiplication and division, measurement of area and volume, coordinate systems, and decimals, among other topics typical of late elementary school math curriculum. In the TRIAD sample, the TEAM 3-5 was found to have good internal reliability Cronbach's $\alpha = 0.91$), and it was also

found to have a strong correlation with state achievement tests in New York ($r(351)= 0.82$, $p < 0.001$) and Massachusetts ($r(110)= 0.76$, $p < 0.001$). As with the REMA, the TEAM 3-5 was transformed into a Rasch-IRT scale.

TRIAD baseline controls. In our TRIAD models, we control for a number of baseline demographic measures. Information regarding child gender, race, age at preschool entry, whether qualified for free or reduced price lunch, whether limited English proficient, and whether designated for special education was obtained from the study districts and schools. Mother's reported their highest level of education on a parent survey administered during the preschool year.

Data Analysis

Figure 1. For the estimates presented in Figure 1, we simply standardized each respective measure of math and reading achievement and teacher-rated anti-social behavior, and we then regressed each respective measure on the kindergarten measures of math, reading, and anti-social behavior. In order to make the correlations between math and anti-social behavior comparable to the correlations between math and reading, we reverse-scaled each anti-social behavior measure. Standard errors were adjusted for school-level clustering, and the regressions were weighted using the panel-weight for data spanning from kindergarten through fifth grade (see Table S2 note). Coefficients and standard errors from each of these regressions can be found in Table S2.

Figure 2. The red and blue lines displayed on Figure 2 were both generated from the TRIAD dataset. The red line displays correlational results generated from a series of regressions that modeled mathematics achievement measured at various timepoints between kindergarten and fifth grade as a function of end-of-preschool mathematics achievement and baseline controls,

site, and blocking group. All of these regression results were estimated only within the control group (n= 396), and full information maximum likelihood was used to account for missing data. Among the controls tested, 17% of students had missing data on "whether free or reduced price lunch," and 25% were missing on mother's education. Due to attrition, approximately 40% of the students were missing on follow-up tests during fourth and fifth grade. The coefficients and standard errors (adjusted for school-level clustering) estimated from these correlational models are displayed in Table S3.

The blue line on Figure 2 was generated from models in which we regressed end-of-treatment (i.e., end-of-preschool) and follow-up measures of mathematics achievement on treatment status and baseline controls (i.e., the same list of controls used for the correlational estimates shown in Table S3). In these models, the sample was restricted to students in only the treatment group without follow-through or the control group (n=880). We again used FIML to account for missing data, and we observed that 16% of students were missing on FRPL, and 22% were missing on mother's education. Further, we found that approximately 40% of students were missing on fourth and fifth grade math measures due to attrition. However, this pattern of attrition did not differ between treatment and control students, as we found only a 1% difference in the rate of attrition between the groups (p = .79). Coefficients and cluster-adjusted standard errors generated by these models are presented in in Table S4.

Table S1
TRIAD Sample Characteristics

	Treatment	Control	p-values
Preschool Entry Math	-0.037	0.066	0.628
	(1.042)	(0.967)	
Site - New York	0.704	0.757	0.840
Male	0.502	0.497	0.570
Ethnicity			
African American	0.522	0.492	0.841
Hispanic	0.197	0.262	0.747
Ethnicity- Other	0.035	0.074	0.130
Age (years) at Baseline	4.331	4.392	0.810
	(0.353)	(0.348)	
Mother's Education			
Less than high school degree	0.142	0.139	0.900
High school degree	0.326	0.325	0.794
Free/Reduced Lunch	0.824	0.881	0.599
Limited English Prof.	0.132	0.230	0.499
Special Education	0.173	0.159	0.812
Observations	456	378	

Note. For each variable, mean values are displayed. Standard deviations are in parentheses. P-values 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 S2

Bivariate Regression-Adjusted Estimates of the Association Between Kindergarten Entry Competencies and Later Achievement and Behaviors

			Math		
	Fall K	Spring K	Spring 1st Gr	Spring 3rd Gr	Spring 5th Gr
K-Entry Math	-	0.83	0.74	0.71	0.66
	-	(0.01)	(0.01)	(0.02)	(0.02)
N	-	17703	14612	9522	9515

			Reading		
	Fall K	Spring K	Spring 1st Gr	Spring 3rd Gr	Spring 5th Gr
K-Entry Reading	-	0.83	0.71	0.53	0.51
	-	(0.01)	(0.02)	(0.04)	(0.04)
N	-	16749	13816	8902	8939
K-Entry Math	0.71	0.66	0.67	0.65	0.63
	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)
N	16747	17060	14376	9467	9508

			Anti-Social Behavior		
	Fall K	Spring K	Spring 1st Gr	Spring 3rd Gr	Spring 5th Gr
K-Entry Anti-Social Beh.	-	0.74	0.55	0.52	0.47
	-	(0.01)	(0.01)	(0.03)	(0.02)
N	-	16944	13126	8053	9012
K-Entry Math	0.16	0.16	0.15	0.20	0.17
	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)
N	17239	16925	13122	8078	9024

Note. All estimates were generated from separate regressions. Robust standard errors were adjusted for school-level clustering, and standard errors are displayed in parentheses. All coefficients were statistically significant ($p < 0.001$). We used the following panel weights for each respective set of regressions: kindergarten (BYCW0), first grade (C124CW0), third grade and fifth grade (C1_6FC0).

Table S3: *Correlational Estimates of the Association Between End-of-Preschool Mathematics Achievement and Later Mathematics Achievement, TRIAD*

	Spring of K (1)		Spring of 1st (2)		Fall of 4th (3)		Spring of 4th (4)		Spring of 5th (5)	
Math- End of Preschool	0.565	***	0.582	***	0.469	***	0.493	***	0.393	***
	(0.042)		(0.046)		(0.070)		(0.065)		(0.071)	
Controls										
Math- Preschool Entry	0.225	***	0.160	***	0.231	**	0.102		0.197	**
	(0.045)		(0.048)		(0.084)		(0.069)		(0.075)	
Male	0.099		0.197	**	0.031		0.064		0.015	
	(0.060)		(0.064)		(0.099)		(0.092)		(0.099)	
African American	-0.199	*	-0.125		0.054		-0.298	*	-0.138	
	(0.095)		(0.101)		(0.146)		(0.137)		(0.149)	
Hispanic	-0.070		-0.039		0.254		-0.177		-0.170	
	(0.107)		(0.114)		(0.172)		(0.164)		(0.180)	
Ethnicity- Other	0.376	**	0.439	**	0.661	**	0.387		0.766	*
	(0.142)		(0.152)		(0.220)		(0.211)		(0.226)	
Age (years) at Baseline	0.123		-0.165		0.026		-0.088		0.003	
	(0.099)		(0.106)		(0.153)		(0.149)		(0.164)	
Mom Ed.- No HS	-0.152		-0.342	**	-0.192		-0.145		-0.315	
	(0.114)		(0.121)		(0.169)		(0.164)		(0.176)	
Mom Ed.- HS	-0.189	*	-0.313	***	-0.423	***	-0.373	***	-0.387	**
	(0.077)		(0.082)		(0.114)		(0.107)		(0.117)	
Free/Reduced Lunch	0.179		0.060		0.443	*	0.486	**	0.510	**
	(0.109)		(0.117)		(0.200)		(0.176)		(0.189)	
Limited Eng Prof.	0.168		0.122		0.238		0.160		0.170	
	(0.096)		(0.103)		(0.155)		(0.152)		(0.171)	
Special Education	0.068		-0.122		0.128		-0.037		0.242	
	(0.083)		(0.089)		(0.141)		(0.126)		(0.139)	
Blocking Group										
2	-0.136		-0.225		-0.111		-0.095		0.118	
	(0.128)		(0.138)		(0.204)		(0.195)		(0.212)	
3	-0.274		-0.183		0.049		-0.246		-0.336	
	(0.154)		(0.164)		(0.235)		(0.227)		(0.236)	
4	0.062		0.184		-0.131		0.045		0.152	
	(0.113)		(0.124)		(0.179)		(0.169)		(0.187)	
5	0.106		-0.069		0.006		-0.205		-0.112	
	(0.109)		(0.121)		(0.184)		(0.175)		(0.188)	
6	0.060		-0.066		0.207		0.147		0.439	
	(0.127)		(0.141)		(0.220)		(0.204)		(0.228)	

7	-0.067	-0.207		-0.099	-0.239	-0.038
	(0.122)	(0.134)		(0.202)	(0.182)	(0.201)
8	-0.162	-0.305	*	-0.340	-0.533 **	-0.226
	(0.140)	(0.149)		(0.222)	(0.206)	(0.226)
Constant	-0.543	0.964		-0.341	0.546	-0.264
	(0.462)	(0.499)		(0.726)	(0.701)	(0.766)
Observations	378	378		378	378	378

Note. Models were estimated using the "SEM" commands in Stata 13.0, and full information maximum likelihood was used to account for missing data. Standard errors are presented in parentheses. Whites are the omitted reference group for race, and children from mother's with at least some college are the omitted reference group for mother's education. * $p < .05$ ** $p < .01$ *** $p < .001$

Table S4

Building Blocks Treatment Impact Estimates

	End of PreK (1)		Spring of K (2)		Spring of 1st (3)		Fall of 4th (4)		Spring of 4th (5)		Spring of 5th (6)	
Building Blocks Group	0.632	***	0.291	***	0.137	*	0.056		-0.004		0.094	
	(0.052)		(0.549)		(0.057)		(0.078)		(0.073)		(0.077)	
Controls												
Math- Prek Entry	0.537	***	0.489	***	0.466	***	0.413	***	0.394	***	0.367	***
	(0.028)		(0.029)		(0.031)		(0.044)		(0.039)		(0.041)	
Male	-0.056		-0.019		0.073		-0.016		-0.051		-0.061	
	(0.050)		(0.053)		(0.055)		(0.076)		(0.071)		(0.076)	
African American	-0.367	***	-0.440	***	-0.419	***	-0.255	*	-0.527	***	-0.442	***
	(0.073)		(0.077)		(0.080)		(0.111)		(0.103)		(0.107)	
Hispanic	-0.227	*	-0.172		-0.112		-0.006		-0.288	*	-0.290	*
	(0.088)		(0.095)		(0.099)		(0.138)		(0.133)		(0.143)	
Ethnicity- Other	-0.101		0.179		0.161		0.254		-0.022		0.320	
	(0.128)		(0.137)		(0.142)		(0.184)		(0.179)		(0.189)	
Age (years) at Baseline	0.484	***	0.359	***	0.095		0.151		-0.043		-0.007	
	(0.080)		(0.085)		(0.088)		(0.122)		(0.115)		(0.121)	
Mom Ed- No HS	-0.373	***	-0.439	***	-0.542	***	-0.451	***	-0.468	***	-0.509	***
	(0.085)		(0.087)		(0.091)		(0.119)		(0.113)		(0.123)	
Mom Ed- HS	-0.089		-0.145		-0.223	**	-0.249	**	-0.256	**	-0.199	*
	(0.066)		(0.068)		(0.071)		(0.093)		(0.087)		(0.091)	
Free/Reduced Lunch	0.117		0.145		0.007		0.073		0.016		-0.043	
	(0.080)		(0.088)		(0.091)		(0.133)		(0.125)		(0.132)	
Limited Eng Prof.	0.228	**	0.265	**	0.273	**	0.363	**	0.348	**	0.269	
	(0.084)		(0.091)		(0.095)		(0.136)		(0.132)		(0.140)	
Special Education	-0.111		-0.021		-0.095		-0.061		-0.162		-0.007	
	(0.068)		(0.073)		(0.077)		(0.106)		(0.099)		(0.105)	
Blocking Group												
2	0.474	***	0.143		0.155		0.113		0.204		0.293	
	(0.110)		(0.116)		(0.120)		(0.164)		(0.155)		(0.166)	
3	0.451	***	0.115		0.192		0.336	*	0.209		0.099	
	(0.108)		(0.116)		(0.121)		(0.169)		(0.162)		(0.168)	
4	0.391	***	0.219		0.412	***	0.178		0.326	*	0.294	*
	(0.094)		(0.100)		(0.103)		(0.142)		(0.135)		(0.143)	
5	0.310	**	0.186		0.306	**	0.271		0.103		0.362	*
	(0.092)		(0.097)		(0.101)		(0.140)		(0.135)		(0.142)	
6	0.426	***	0.397	***	0.458	***	0.552	**	0.490	**	0.562	**
	(0.105)		(0.113)		(0.120)		(0.167)		(0.159)		(0.175)	
7	0.487	***	0.149		0.253	*	0.215		0.056		0.132	

	(0.101)			(0.107)			(0.112)		(0.155)		(0.146)		(0.157)
8	0.423	***		0.002			0.017		-0.084		-0.042		0.056
	(0.121)			(0.127)			(0.132)		(0.186)		(0.166)		(0.177)
Constant	-2.673	***		-1.726	***		-0.485		-0.772		0.475		0.121
	(0.357)			(0.378)			(0.393)		(0.550)		(0.513)		(0.547)
Observations	834			834			834		834		834		834

Note. Models were estimated using the "SEM" commands in Stata 13.0, and full information maximum likelihood was used to account for missing data. Standard errors are presented in parentheses. Whites are the omitted reference group for race, and children from mothers with at least some college are the omitted reference group for mother's education. * p < .05 ** p < .01 *** p < .001