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

Drew H. Bailey, Greg J. Duncan, and Tyler Watts
University of California, Irvine

Doug H. Clements and Julie Sarama
University of Denver

Developmental theories often posit that changes in children’s early psychological characteristics will affect their much later psychological, social, and economic outcomes. However, tests of these theories frequently yield results that are consistent with plausible alternative theories that posit a much smaller causal role for earlier levels of these psychological characteristics. Our article explores this issue with empirical tests of skill-building theories, which predict that early boosts to simpler skills (e.g., numeracy or literacy) or behaviors (e.g., antisocial behavior or executive functions) 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 these skill-building processes. Using the example of skill-building in mathematics, we argue that longitudinal correlations, even if adjusted for an extensive set of baseline covariates, constitute an insufficiently risky test of skill-building theories. We first show that experimental manipulation of early math skills generates much smaller effects on later math achievement than the nonexperimental literature has suggested. 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 can reproduce the pattern of experimental impacts on children’s mathematics achievement. Implications for developmental theories, methods, and practice are discussed.

Keywords: early childhood, interventions, skill-building, cognitive development, education

Supplemental materials: http://dx.doi.org/10.1037/amp0000146.supp
Robinson, 2013), a linkage that requires causal evidence. One cannot have it both ways.

We argue that regression-adjusted correlations often provide insufficiently “risky” tests of 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 can 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. Nonzero regression-adjusted correlations between children’s early psychological characteristics and their much later psychological, social, and economic outcomes do not constitute a risky test of a developmental theory because such correlations are consistent with a number of plausible competing theories, including those positing that a combination of differentially stable general cognitive abilities, personality, and environmental affordances is responsible for generating the correlational patterns. We explore these issues in the context of a well-trodden area in child development and education: the substantial correlations between children’s early and much later academic achievement.

**Correlational Tests of Skill-Building Theories**

Even after adjusting for a large set of controls, including baseline measures of other academic and socioemotional skills and capacities, domain-general cognitive abilities, and socioeconomic status, strong longitudinal correlations are often observed in studies of academic domains of school readiness, across many years (Aunola, Leskinen, Lerkkanen, & Nurmi, 2004; Bailey, Siegler, & Geary, 2014; Duncan et al., 2007; Geary, Hoard, Nugent, & Bailey, 2013; Jordan, Kaplan, Ramineni, & Locuniak, 2009; Siegler et al., 2012; Watts, Duncan, Siegler, & Davis-Kean, 2014). These longitudinal correlations constitute an important part of the empirical basis for skill-building theories.

Researchers, including authors of this article (Bailey, Siegler, et al., 2014; 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. 1,430) 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.” Focusing on the development of children’s mathematics achievement, this article uses experimental evidence, falsification tests, and alternative model structures to show that riskier tests suggest a much smaller causal role for skill-building processes than is commonly believed.

**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). Such findings might reasonably lead one to predict that the children with the most solid early foundations of math skills, in the context of K–12 instruction will 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 bidirectional 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 that it allows for simpler skills to support more sophisticated skills, 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” to produce 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, the “math to math,” “reading to reading,” and “anti-social to anti-social” interwave correlations with fall of kindergarten values, calculated from national data from the 1998–99 Early Childhood Longitudinal Study (ECLS-K), appear in Figure 1 (measure and sample details in the online supplemental material Appendix). These correlations 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).

Although clearly consistent with skill-building developmental theories, these correlations do not constitute a risky test of those theories because they are also consistent with alternative theories in which the causal effects of early skills on much later skills play a minor role in contributing to the stability of psychological characteristics during 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 captured sufficiently with regression controls or other techniques for reducing omitted-variables bias (OVB), then the apparent role of skill-building processes in generating cross-time correlations could be seriously overstated.

**Riskier Tests**

We discuss three promising approaches to understanding 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 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**

Random assignment to programs that boost 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 predicted 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 a test of the
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. 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 prekindergarten year; in the spring of the kindergarten year; in the spring of the kindergart ten, 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, Sarama, Spitler, Lange, & Wolfe, 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 prekindergarten 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% confidence intervals associated with each of the estimates. These confidence intervals were derived from SEs 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 1.1

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 predicted 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 OVB, then we would expect to see a similar pattern in the experimental data.

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 predicted 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/.63.2 Rescaled impact estimates fall to about .46 SD within a year and drop to statistically nonsignificant .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.3 Overall, the correlation-based estimate of the treatment effect is very close to the observed treatment effect 1 year after the end of treatment, but then much higher than the observed treatment effects at all subsequent waves. In highlighting the discrepancy between correlational estimates and experimental impacts, we do not intend to discourage the use of early intervention (we discuss possible implications for research and practice below). Rather, our goal is to highlight the inaccuracy of estimated theoretically important causal effects when confounds are assumed to be largely or fully controlled in a regression model.

TRIAD’s ability to shed light on math skill-building processes is a function of the comprehensiveness of its initial impacts. End-of-preschool mathematics knowledge was assessed using the Research-Based Early Maths Assessment (REMA; Clements, Sarama, & Liu, 2008), described more completely in the supplemental materials, which are available online). The REMA assessed children’s conceptual and procedural knowledge, as well as problem-solving and strategic competencies in the domain of early mathematics, and has been shown to strongly correlate with

1 Full correlation matrices for measures of children’s mathematics achievement administered at several waves across development are available in Bailey, Watts, Littlefield, and Geary (2014).

2 This means, for example, that TRIAD’s .63 SD impact at the end of pre-kindergarten is shown as 1.0 and its .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.

3 Using a 2-level random intercept logistic regression model, Clements and colleagues (2017) found a statistically significant treatment impact of the TRIAD intervention at the spring of fifth grade.
other measures of early learning (e.g., *Applied Problems, Child Math Assessment*). Furthermore, it has been shown to strongly predict later mathematics achievement measured through Grade 5 (Watts, Duncan, Clements, & Sarama, 2017). Thus, the REMA should provide a strong measure of the early mathematical competencies needed to build later skills in mathematics.

As explained in the Appendix, the end-of-Pre-K REMA test showed considerable variation in four 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 ($\beta = 0.45$, $SE = 0.06$), patterning ($\beta = 0.36$, $SE = 0.06$), geometry ($\beta = 0.67$, $SE = 0.06$), and measurement ($\beta = 0.20$, $SE = 0.06$), all of which have been shown to predict later mathematics knowledge (Nguyen et al., 2016). Because the intervention boosted a wide variety of preschool math skills, treated children should have had a much stronger base of math competencies from which to build further math skills when compared with children in the control group. In other words, these robust causal impacts at the end of the pre-K year suggest that the TRIAD 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 skill building processes at work in 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, Cobb, Farran, Cordray, & Munter, 2013; for
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 appears that experimentally induced changes in early skills may have temporary effects on children’s subsequent learning. Yet, 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 antisocial behavior but decidedly not for math-to-reading correlations, which are virtually indistinguishable from math-to-math correlations beyond first 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 data sets (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 metaanalytic estimates in Duncan et al., 2007) are much higher than correlations between early reading and later math (.10). Is it plausible that boosting children’s early mathematics skills would affect later reading skills 1 to 2 years later to the same extent that it affected children’s mathematics skills? On one hand, the TRIAD early mathematics intervention did show effects on some measures of early oral language skills (Sarama, Lange, Clements, & Wolfe, 2012). On the other hand, the effects did not generalize to other tests of early reading skills, and the statistically significant effects were much smaller than they were for children’s mathematics skills. Learning mathematics may have a non-zero effect on children’s early reading achievement, but we doubt that the effect would be almost identical to the effect on children’s achievement in the same domain. Still, we consider additional falsification tests below.

Models Consistent With Temporal Patterns of Within-Domain Correlations

Consider the asymptotic rather than complete decline in the within-domain correlations shown 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 $MS_2$ (i.e., “math skills”) represent the impacts of a previous math skill on the immediately following math skill. As long as these probabilities are less than 1, we should observe some kind of exponential decay in early-to-late correlations as skills become more advanced. Research on transfer of learning, wherein knowledge or skills learned in one domain or setting are applied to other situations, provides a strong theoretical basis for this
In particular, this work suggests that as the features of domains and settings diverge from those in which the initial learning takes place, transfer becomes decreasingly likely (Perkins & Salomon, 1988). As children progress through school, the settings and content to which they are exposed grow increasingly dissimilar, on average, to the settings and content to which they are exposed at school entry.

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, Claeysens, 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 Kilpatrick, Swafford, & Findell, 2001), which are not quickly or easily mastered (e.g., OECD, 2016) 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 estimated patterns of correlations across time within and between academic domains all suggest that omitted variables influencing development throughout the observed period 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 the causes of children’s academic development into factors that exert a stable influence on children’s academic skills throughout development and—to continue with the example of mathematics achievement—children’s mathematics knowledge assessed in the immediately preceding wave of data collection. This approach provides an additional risky test of the hypothesis that skill-building effects can be recovered in longitudinal data sets. If confounds are sufficiently controlled in standard regression models, then removing variance attributable to stable factors influencing children’s learning across development should not impact estimates of the effects of early achievement on later achievement. If, however, persistent confounds are responsible for the discrepancy between experimental and regression derived estimates, then controlling for these confounds across development should enable us to reproduce experimental estimates using correlational data.

A simple version of one such model is depicted in Figure 4. Developed by Steyer (1987), and implemented in structural equation modeling software (for discussion, see Cole, Martin, & Steiger, 2005; Steyer & Schmitt, 1994; for an
accessible introduction and code, see Prenoveau, 2016), 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 $MS_1$, $MS_2$, ..., $MS_{n-1}$, 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 some other 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 interwave 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 $MS_2$ times $MS_3$ by the third wave, etc.).

Second, the model may plausibly 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. In the presence of stable confounding, more common alternative approaches such as the cross-lagged panel model might yield upwardly biased predicted effects across development (Hamaker, Kuiper, & Grasman, 2015).

Third, because of the accumulating effects of stable factors on children’s achievement across development, the model generates a testable prediction of increasing interyear stability as children get older (Cole et al., 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 interwave intervals. Children’s mathematics achievement in the spring of kindergarten correlates .77 ($SE = .01$) with their mathematics achievement 1 year later in the spring of first grade, which correlates .80 ($SE = .01$) with their mathematics achievement 2 years later than that in the spring of third grade, which correlates .89 ($SE = .01$) with mathematics achievement 2 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 posttreatment achievement and the stable latent variable are simultaneously regressed on treatment status. Relying on the TRIAD data used previously, 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 previously, 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 data sets, the unmeasured persistent factor can be regressed on hypothesized sources of persistent variation (Bailey, Watts, Littlefield, & Geary, 2014), but this analysis is vulnerable to problems associated with all cross-sectional regression analyses, such as OVB. 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 with the one we use here, both on the basis of fit to correlational data sets and their predictions about experimentally induced effects across time.

**Comparing the Persistent Factor 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 develop-
ment of children’s math achievement across prekindergarten 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 data sets 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 1-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 $MS_t = MS_0 \times e^{-t}$, where $MS_0$ 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 one out of five regression-based estimates.

**Table 1**

*Estimates of Math Skills (MS) Paths From Observational and Experimental Data*

<table>
<thead>
<tr>
<th>Source</th>
<th>Sample size</th>
<th>Period</th>
<th>Implied 1-year MS estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>State-trait estimates</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bailey, Watts, et al., 2014: Missouri Math Study</td>
<td>292</td>
<td>Grade 1–Grade 2</td>
<td>.26</td>
</tr>
<tr>
<td>Bailey, Watts, et al., 2014: Missouri Math Study</td>
<td>292</td>
<td>Grade 2–Grade 3</td>
<td>.18</td>
</tr>
<tr>
<td>Bailey, Watts, et al., 2014: Missouri Math Study</td>
<td>292</td>
<td>Grade 3–Grade 4</td>
<td>.20</td>
</tr>
<tr>
<td>Bailey, Watts, et al., 2014: SECCYD</td>
<td>1124</td>
<td>Grade 1–Grade 3</td>
<td>.58</td>
</tr>
<tr>
<td>Bailey, Watts, et al., 2014: SECCYD</td>
<td>1124</td>
<td>Grade 3–Grade 5</td>
<td>.30</td>
</tr>
<tr>
<td>Watts et al., 2016: TRIAD</td>
<td>834</td>
<td>PreK–K</td>
<td>.25</td>
</tr>
<tr>
<td>Watts et al., 2016: TRIAD</td>
<td>834</td>
<td>K–Grade 1</td>
<td>.04</td>
</tr>
<tr>
<td>Watts et al., 2016: TRIAD</td>
<td>834</td>
<td>Grade 1–Grade 4</td>
<td>.51</td>
</tr>
<tr>
<td>Simple average</td>
<td></td>
<td></td>
<td>.29</td>
</tr>
<tr>
<td>Unweighted study average</td>
<td></td>
<td></td>
<td>.31</td>
</tr>
<tr>
<td>Weighted study average</td>
<td></td>
<td></td>
<td>.35</td>
</tr>
<tr>
<td>Experimental estimates</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Current article: TRIAD</td>
<td>834</td>
<td>PreK–K</td>
<td>.46</td>
</tr>
<tr>
<td>Current article: TRIAD</td>
<td>834</td>
<td>K–Grade 1</td>
<td>.48</td>
</tr>
<tr>
<td>Current article: TRIAD</td>
<td>834</td>
<td>Grade 1–Grade 4</td>
<td>.48</td>
</tr>
<tr>
<td>Current article: TRIAD</td>
<td>834</td>
<td>Grade 4–Grade 5</td>
<td>N/A</td>
</tr>
<tr>
<td>Hofer et al., 2013: TRIAD</td>
<td>1192</td>
<td>Pre-K–K</td>
<td>.28</td>
</tr>
<tr>
<td>Hofer et al., 2013: TRIAD</td>
<td>1129</td>
<td>K–Grade 1</td>
<td>.67</td>
</tr>
<tr>
<td>Smith et al., 2013</td>
<td>320</td>
<td>Grade 1–Grade 2</td>
<td>.22</td>
</tr>
<tr>
<td>Simple average</td>
<td></td>
<td></td>
<td>.43</td>
</tr>
<tr>
<td>Unweighted study average</td>
<td></td>
<td></td>
<td>.39</td>
</tr>
<tr>
<td>Weighted study average</td>
<td></td>
<td></td>
<td>.44</td>
</tr>
</tbody>
</table>

*Note: SECCYD = Study of Early Childcare and Youth Development (see NICHD Early Child Care Research Network, 2002); TRIAD = Technology-Enhanced, Research-Based, Instruction, Assessment, and Professional Development. 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. All state-trait models corrected correlations among math tests for measurement error by setting the path from the latent state factor to the mathematics measure equal to the square root of the reliability of the test. For information regarding the Missouri Math Study, see Geary, 2010. Information regarding the TRIAD study is presented in the Appendix and in Clements et al., 2013. ^ Fourth grade is the average of two fourth-grade scores. Average interval is 2.75 years. ** Fifth-grade estimate is higher than fourth-grade estimate; neither is statistically distinguishable from 0. *** Treatment effects were calculated as the average of the three standardized mathematics tests administered at both waves.*
These patterns may generalize well to other kinds of studies of children’s academic achievement outcomes. In a review of experimental estimates from 67 high-quality studies of early childhood education programs, Li and colleagues (2016) reported an average end-of-treatment effect size of .23, with estimates 0–1 year after treatment averaging approximately .10. Impact estimates from subsequent waves were smaller, but their precise values were sensitive to inclusion criteria.

As mentioned previously 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., oral language [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. In our view, this model has advantages over standard ordinary least squares regression at approximating experimental impacts because it considers persistent unmeasured factors influencing children across development. However, the model does not resolve several important issues. First, it is unclear whether this particular model is the ideal specification (see Cole et al., 2005, for a discussion, and Hamaker et al., 2015, for alternatives), or whether intervention effects are likely to be confined to occasion-specific variation, rather than stable environmental and personal factors. Furthermore, more research should investigate what variables constitute the stable environmental and personal factors that influence children’s mathematics learning throughout development.

What Stable Factors are Missing From Our Regression Models?

As noted previously, 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, Reynolds, Liu, Kaufman, & McGrew, 2012). Personality also likely plays a significant and complex role in children’s academic development.

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 (Kendler, Turkheimer, Ohlsson, Sundquist, & Sundquist, 2015; van Ijzendoorn, Juffer, & Poelhuis, 2005), maternal nutrition during prenatal development (Almond & Mazumder, 2011), or a very intensive early childhood education program (Campbell, Pungello, Miller-Johnson, Burchinal, & Ramey, 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 antisocial 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 antisocial behavior across development. However, an analogous developmental story may apply: Correlations between kindergarten antisocial behavior and antisocial 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 antisocial behavior across time. If these factors are not well measured, the auto-regressive effects of antisocial behavior will be exaggerated. Consistent with this possibility, Anusic and Schimmack (2015) observed substantial stability in the interwave correlations of personality, affect, self-esteem, and life satisfaction, with the highest stability observed in personality. 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.

Better measurement of children’s skills enables riskier tests of developmental hypotheses. Measures in large longitudinal studies are often based on single scores rather than more cognitively complex and diagnostic assessments. 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). The intervention may have taught the former certain concepts and skills, raising their score. However, the latter, 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 are myriad, strong, and stable. The former children have none of these advantages. Therefore, future measures of the two groups of children may yield different scores even if subsequent experiences are the same. Future assessments and research designs, including those that investigate measurement invariance across subgroups (Wicherts, 2016), are needed to investigate such possibilities.

Some research has included much more specific measures of children’s knowledge in longitudinal studies. The advantage of this approach is that such studies can, in principal, provide riskier tests of skill building theories by relating gains in specific knowledge states to subsequent knowledge states. If researchers are testing a very specific theory of learning, this approach enables them to make very specific predictions about where nonzero expected test scores should appear, how big they should be, and perhaps most importantly, when they should vanish. This approach is what Shadish, Cook, and Campbell (2002) refer to as coherent pattern matching. The downsides to this approach are that (a) better measurement in one domain often comes at the cost of lower sample size and/or worse measurement in other domains, including family background characteristics.

---

4 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.
(b) some developmental theories are sufficiently open-ended that a very wide range of estimated effects can be argued to be plausible after the fact, and (c) in the context of early cognitive development, individual differences may be sufficiently nondifferentiated that differential prediction of later outcomes better reflects the extent to which a measure captures commonalities in knowledge states rather than uniqueness to a specific knowledge state (Paris, 2005; Purpura & Lonigan, 2013; Schenke, Lam, Rutherford, & Bailey, 2016). These shortcomings may be addressed by precision and completeness in measurement and theory and by formulating the riskiest tests possible. We see as particularly useful the practice of comparing correlational estimates to the most relevant experimental impacts whenever possible.

Following a demonstration of mismatched correlational and experimental findings, it is common to call for more randomized controlled trials. We certainly endorse randomized, controlled trials 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 & 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


Received August 22, 2016

Revision received December 21, 2016

Accepted January 22, 2017