ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Psychological Science in the Public Interest
2020, Vol. 21(2) 55–97

DOI: 10.1177/1529100620915848
www.psychologicalscience.org/PSPI
SAGE

Persistence and Fade-Out of Educational-Intervention Effects: Mechanisms and Potential Solutions

Drew H. Bailey[1], Greg J. Duncan[1], Flávio Cunha[2], Barbara R. Foorman[3], and David S. Yeager[4]
[1]School of Education, University of California, Irvine; [2]Department of Economics, Rice University; [3]Florida Center for Reading Research, Florida State University; and [4]Department of Psychology, University of Texas at Austin

Abstract
Some environmental influences, including intentional interventions, have shown persistent effects on psychological characteristics and other socially important outcomes years and even decades later. At the same time, it is common to find that the effects of life events or interventions diminish and even disappear completely, a phenomenon known as *fade-out*. We review the evidence for persistence and fade-out, drawing primarily on evidence from educational interventions. We conclude that (a) fade-out is widespread and often coexists with persistence; (b) fade-out is a substantive phenomenon, not merely a measurement artifact; and (c) persistence depends on the types of skills targeted, the institutional constraints and opportunities within the social context, and complementarities between interventions and subsequent environmental affordances. We discuss the implications of these conclusions for research and policy.

Keywords
fade-out and persistence, interventions, education

The possibility that time-limited experiences can alter a person's life trajectory has interested psychological researchers and the public for generations. Prominent examples include Freud's (1924/1961) writings on the hypothesized role of adverse childhood experiences in the etiology of hysteria; Bowlby's (1973) research on the consequences of separation and attachment; adoption studies that compared adult siblings who had been reared in different homes (Bouchard, Lykken, McGue, Segal, & Tellegen, 1990); studies of the influence of early-childhood education (ECE) experiences on adult-labor-market outcomes (Heckman, 2006); and research on the contribution of childhood home and contextual experiences to the intergenerational transmission of inequality (Chetty & Hendren, 2018; Felitti et al., 1998).

Despite this long-standing interest, psychology lacks a strong general scientific framework for understanding the circumstances under which the effects of a childhood experience *persist* (with children continuing to do better or worse than they would have done had they not had the experience) or *fade out* (with children's long-run outcomes not depending on whether they had the experience). This is evident from seemingly contradictory ideas in the field. For example, some argue that our default expectation should be that the effects of treatments (e.g., higher quality schooling) that change psychological characteristics are transitory because the underlying traits (Costa & McCrae, 1980; Jensen, 1998) and environmental experiences (D. T. Campbell & Frey, 1970) that produce individual differences in the first place remain unchanged. Others argue for persistence as the default expectation because earlier levels of psychological characteristics tend to be strongly correlated with, and likely affect, later levels of psychological characteristics (Duncan et al., 2007).

The temporal stability of many psychological characteristics has been taken both as evidence that

Corresponding Author:
Drew H. Bailey, School of Education, 2072 Education, University of California, Irvine, Irvine, CA 92697
E-mail: dhbailey@uci.edu

potential treatments that change those characteristics will have persistent effects (Roisman & Fraley, 2013) and as evidence that treatment effects will likely fade out (Bailey, Watts, Littlefield, & Geary, 2014). Some have argued that the benefits from childhood interventions will fade out if, after the intervention ends, children enter environments that fail to support the skills developed by the intervention (Currie & Thomas, 2000), but others have argued that the most learning-conducive environments will produce rapid fade-out of the effects of prior interventions because they give students who did not receive the intervention a chance to be "treated" (Bailey, Duncan, Odgers, & Yu, 2017; D. T. Campbell & Frey, 1970).

These contradictions reflect, in part, variation in definitions of persistence and fade-out, the nature of the interventions, the developmental periods in which they occur, the psychological domains studied by these researchers, and, of course, the differing training, prior beliefs, experiences, and biases of the researchers themselves.

The purpose of this review is to describe the current state of research on the persistence and fade-out of the effects generated by interventions. The primary focus of this review is on educational interventions, although we will also discuss some noneducational interventions when they help elucidate the potential mechanisms through which specific kinds of educational interventions might operate, or might help us understand why the effects of educational interventions persist or fade out. Our goal is to develop a set of concepts, terms, and hypotheses that can be applied across domains and tested empirically.

We begin by providing more precise definitions of fade-out and persistence and reviewing empirical research on their patterns. We conclude that some degree of fade-out is the norm in interventions targeting psychological outcomes, such as skills, interests, tendencies, capacities, and beliefs, but that the relatively small number of long-run follow-ups of initially effective interventions often show at least some degree of persistence.

We then review explanations for fade-out in interventions targeting skills, capacities, and opportunities of children and youth. Some researchers consider fade-out to be a mere measurement artifact; others attribute fade-out to substantive psychological and social processes. We conclude that although its precise magnitude and timing can be sensitive to measurement, fade-out is real and in need of explanation. However, how fade-out is shaped by substantive psychological and social processes is likely to differ across kinds of interventions, populations, and subsequent experiences.

Next, we discuss mechanisms for persistence, which include (a) intervening on the kinds of skills, capacities, and beliefs likely to generate persistent effects; (b) the removal of postintervention constraints and/or the realization of opportunities afforded by institutions; and (c) the features of postintervention environments most likely to support persistence.

Finally, we discuss implications of this literature for future research into persistence and fade-out, as well as implications for policy. Most notably, a strong model-based theory of persistence and fade-out would be useful for both scientific and applied purposes, but such a theory can be constructed only by increasing the number of intervention studies that include long-term follow-up assessments of outcomes and assessments of the mechanisms listed above.

Terminology

Intervention *impacts* or *effects* are typically defined as the differences in outcomes among the group of people who were offered (or actually received) some treatment relative to a group of people who were not offered that treatment. Researchers have used the term *fade-out* to refer to a variety of associated but distinct phenomena related to the time course of effects after the completion of an intervention. For some, fade-out is a dichotomous characterization of whether the long-term effects of initially effective interventions are essentially null across a range of long-term outcomes (Heckman, 2017). In other cases, and for the purposes of the current review, it refers to a pattern of diminishing intervention impacts after the end of treatment (Elango, García, Heckman, & Hojman, 2015). This latter definition allows for fade-out and persistence to coexist for the same intervention, for the same sample, and even across measurement occasions for the same outcome.

Coexistence occurs when fade-out-generating processes such as forgetting co-occur with persistence-generating processes such as remembering and transfer of learning (Kang et al., 2019). For example, the process of beginning to learn a new language often involves a sometimes frustrating combination of learning, forgetting recently learned words, using previously acquired knowledge to make inferences about new knowledge (e.g., conjugating new verbs and spelling new words), and, one hopes, some remembering. Despite the setbacks, and despite the concern that students will forget what they were taught, educators hope students will perform better if they take a class than if they do not.

This co-occurrence-based conceptualization of persistence and fade-out also has implications for the distinction between two important concepts in the fade-out literature—*emergence* and *reemergence*. We restrict our use of reemergence to cases in which intervention impacts are initially favorable, then fade out, but subsequently reemerge within the same domain. A version of reemergence for which there is strong evidence is *savings*, the phenomenon in memory research whereby fewer trials are required to learn a list of items to mastery after the list has already been mastered a first time (for review, see Murre & Dros, 2015).

We argue that in the case of intensive multifaceted interventions, it is more common to find emergence; in other words, some degree of fade-out within one set of domains (e.g., academic and social skills) is followed by the emergence of impacts in related but distinct domains (e.g., high school graduation, adult earnings, or physical and mental health). Emergence may occur for a variety of reasons, including persistent impacts on key, often unmeasured, constructs directly affected by the intervention and indirect effects that arise because, although constructs targeted by an intervention and tracked in its evaluation fade out completely, the intervention affected other important skills that were not directly targeted by the intervention.

Although impacts are typically defined as the differences in outcomes between groups of people who did or did not receive a treatment, experts in causal inference propose a different, individual-specific definition. In their view, the effect of an intervention is the difference between a treated person's outcomes relative to an imaginary (i.e., counterfactual) world in which that same person did not receive the treatment (or received a different kind of treatment). Because we cannot observe imaginary worlds, we generally use the outcomes of a comparison group as a stand-in for a person's own counterfactual outcomes (Rubin, 2004).

It is useful to think about fade-out in the context of a randomized experiment. Although most intervention research reports provide careful descriptions of the treatment, they often fail to provide an equally thorough description of counterfactual conditions (i.e., what *would have happened* to individuals if they had not received the intervention, often approximated by the measured experiences of the control groups). This can be a problem because the nature of counterfactual conditions has important implications for understanding the magnitude and time path of intervention impacts in the real world. When conditions in the absence of the intervention benefit control-group members, then impact estimates (i.e., differences between the treatment and control groups) are smaller than would otherwise be the case. Example of factors benefiting control-group members include maturation (e.g., typically developing children in the control group for an intervention to increase self-regulation would have developed similar levels of self-regulation skills in any case), the availability of alternative interventions (e.g., if children who are eligible for a reading tutoring program but are not chosen for that program receive other additional services from the school), or, in the case of an active control group, the members of that group benefiting from the key ingredients responsible for the program's success.

Fade-out may even occur when treatment-group individuals enjoy rapid improvement on an outcome of interest after the end of an intervention, as long as members of the control group improve more rapidly. As we will see in the next section, this "catchup" pattern of fade-out stemming from more rapid posttreatment improvement for children in the control group, as opposed to those in the treatment group, is probably the norm in the case of ECE interventions targeting academic skills.

How Widespread Is Fade-Out?

Fade-out and persistence are observable at several timescales: minutes in the case of memory (Ebbinghaus, 1885); months and years in the case of many academic interventions; and even generations, as with the introduction of the food-stamp program in the United States in the 1960s and 1970s (Hoynes, Schanzenbach, & Almond, 2016) and famous ECE programs such as Perry Preschool. We divide our review into interventions focused on specific skills, capacities, and beliefs and then on more general education and other environmental interventions. Basic descriptive knowledge on the temporal pattern of impacts with these kinds of interventions is hindered by the fact that relatively few evaluations measure outcomes beyond the end of their treatments.

We begin our look at skill-targeted interventions with data from an evaluation (TRIAD, or Technology-enhanced, Research-based, Instruction, Assessment, and professional Development) of one implementation of an unusually well-developed prekindergarten (pre-K) mathematics curriculum called Building Blocks (Clements, Sarama, Spitler, Lange, & Wolfe, 2011; Clements, Sarama, Wolfe, & Spitler, 2013). One of the evaluations involved random assignment of the curriculum at the school level to 834 students in 30 elementary schools serving low-income neighborhoods in either Buffalo, New York, or Boston, Massachusetts.[1] Students' mathematics achievement was measured at the beginning and end of the preschool year; at the end of kindergarten, first grade, and fifth grade; and during the fourth-grade year. Drawn from data in Clements et al. (2013) and Bailey et al. (2016), Figure 1a shows that treatment/control differences at the end of the pre-K year amounted to 0.63 *SD*—a large impact.

But Figure 1a also shows that this large initial treatment effect at the end of preschool quickly faded—to about 40% of its initial value by the end of first grade and almost completely by fourth grade. Figure 1b recasts TRIAD data through age 7 by showing vertically scaled math scores separately for treatment and control groups. The more rapid growth in math skills across the pre-K year for the treatment group relative to the

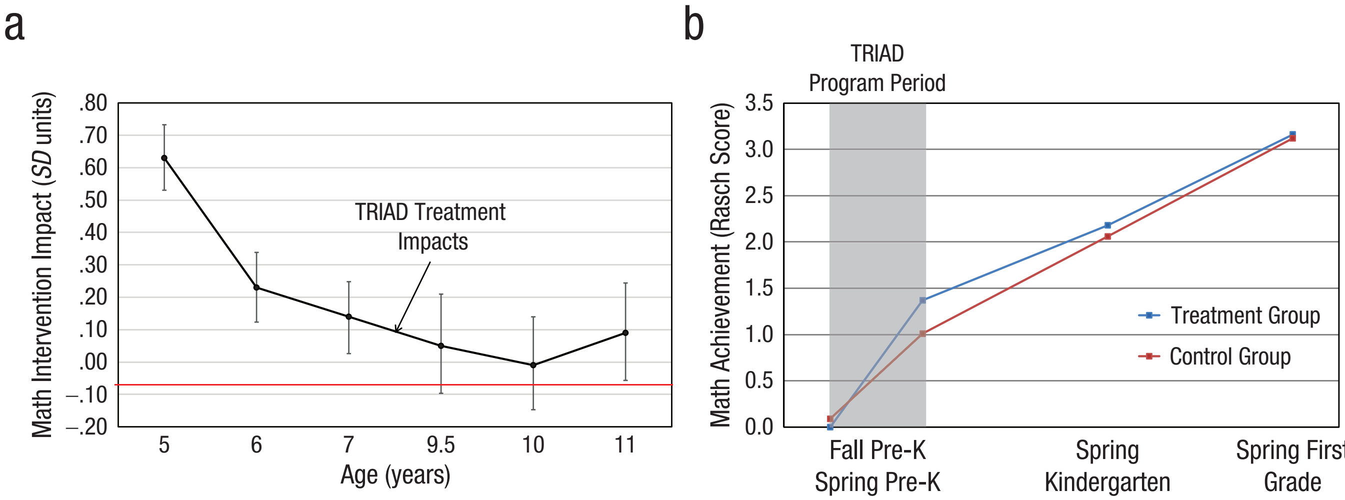

Fig. 1. Treatment- and control-group differences and trajectories in the TRIAD intervention. In (a), math intervention impact is graphed as a function of age. Error bars indicate 95% confidence intervals. The red line indicates a treatment effect of 0. In (b), math achievement is graphed as a function of grade and season, separately for the treatment and control groups. Rausch score *SD* = 1.

control group produces the large end-of-treatment effect shown in Figure 1a. Between the end of treatment and the end of first grade, math scores grow for both groups, but they grow faster, on average, for children in the control group than for those in the treatment group. Thus, the fade-out observed in Figure 1a appears to be the result of control-group catchup.

By and large, meta-analyses of interventions targeting specific skills also show that end-of-treatment impacts begin to fade out quickly and, in some cases, disappear completely. The case of educational and environmental interventions is more difficult to summarize succinctly, in part because differential fade-out is observed from one domain to the next (e.g., academic compared with socioemotional skills) and in part because long-run outcomes such as completed schooling, labor market success and health in adulthood lack closely associated counterpart domains in childhood.

Widespread fade-out in skill-focused interventions?

A pattern of declining treatment effects has been observed in many experiments testing the effects of interventions designed to boost children's academic skills:

- In a meta-analysis of 51 such studies, Hattie, Biggs, and Purdie (1996) estimated an average end-of-treatment impact of 0.45 *SD*. However, in studies that followed up with students after the end of treatment (with follow-up intervals averaging between 3 and 4 months after the end of treatment), the average impact was 0.10 *SD*.
- Of the 36 evaluations assessing impacts on phonological awareness included in Bus and van IJzendoorn's (1999) study of phonological-awareness training programs, only seven assessed long-term effects, which, on average, occurred only 8 months after the end of the programs. Effect sizes fell by about half of their end-of-treatment values in these follow-up assessments. In the case of reading outcomes, the measurement period was longer (averaging 18 months) but so too was the amount of fade-out (nearly 80%).
- Protzko (2015) conducted a meta-analysis of 23 different evaluations of interventions targeting IQ. End-of-treatment effect sizes averaged 0.37 *SD*, but these effects declined by about 0.10 *SD* per year. Limiting the sample to evaluations of interventions that could reasonably be expected to raise IQ increases the end-of-treatment impact estimate to 0.52 *SD*, but also the amount of fade-out (0.13 *SD* per year). Protzko finds that IQ fade-out is caused mainly by declines in the IQs of treatment-group children across the follow-up period. However, because IQ scores are age normed, it is not clear whether treatment groups experienced a net loss of skill, on average, following the interventions included in this meta-analysis.
- Takacs and Kassai (2019) summarize the large literature on interventions designed to promote children's executive-function skills. As with the other literature reviews, only a small fraction of studies (15 of 90) included follow-up assessments, and these were conducted, on average, 22 weeks after the end of treatment. In contrast to the substantial end-of-treatment effect sizes, there was no convincing evidence that impacts persisted on these follow-up assessments.

All of these meta-analyses of a diverse set of skill-targeted interventions suggest that some degree of fade-out is the norm. Although it is difficult to generalize to the universe of skill-based interventions, we know of no meta-analyses of interventions targeting specific skills in children, with substantial variation in follow-up period, in which full persistence prevails.

Persistence in some intensive environmental interventions

In contrast to the ubiquity of fade-out in interventions that target specific skills, large bodies of research document substantial persistent effects of some interventions that explicitly target or unintendedly affect a broader set of skills and capacities. A dramatic example is that children who are adopted into relatively more advantaged families tend to exhibit improvements across a number of domains, including cognitive skills (Kendler, Turkheimer, Ohlsson, Sundquist, & Sundquist, 2015), health-related behavior, educational attainment, and earnings (Sacerdote, 2007). It should be noted that, relative to typical education interventions, adoption is remarkably intensive, in terms of both duration and ubiquity.

Less intensive environmental interventions can also generate persistent socially significant effects. Positive early-life treatments, such as iodine fortification (Adhvaryu & Nyshadham, 2016), and negative early-life shocks, such as pneumonia exposure in infancy (Bhalotra & Venkataramani, 2015; for review, see Almond, Currie, & Duque, 2018), are associated with changes in adult-labor-market outcomes in studies based on strong quasiexperimental designs—a finding that has appeared in various forms at least since Elder's (1974) classic study of children affected by the Great Depression. Differences between identical twins on a variety of psychological characteristics show stability across several years (Bailey & Littlefield, 2017; Tucker-Drob & Briley, 2014; von Stumm & Plomin, 2018), which is consistent with the notion that changes induced by environments that differ between children raised together can persist across time.

Some economic interventions appear to generate long-run benefits as well. Hoynes et al. (2016) take advantage of the fact that the U.S. food-stamp program was rolled out on a county-by-county basis between the early 1960s and the mid 1970s to show that children conceived or born into counties with the program already in place reported lower levels of cardiovascular symptoms in their 30s and 40s than children living in counties that adopted the food-stamp program later. In Bouguen, Huang, Kremer, and Miguel's (2019) review of cash transfers and health interventions in developing countries, several conditional cash transfer programs were identified as producing long-term impacts on cognitive, educational, and labor-market outcomes. Moreover, some unconditional cash-transfer programs generated long-term impacts only when they were combined with intensive training and support.

A set of clinical interventions presents a notable potential exception to the regular pattern of fade-out described above. These interventions include preventive and therapeutic clinical interventions for antisocial behavior (Farrington & Welsh, 2003; Sawyer, Borduin, & Dopp, 2015) and social-emotional learning interventions based in schools (R. D. Taylor, Oberle, Durlak, & Weissberg, 2017). Often involving both skill building and environmental changes, they have sometimes generated impacts persisting for months or years, and many impacts have remained stable in magnitude from the end of treatment to long-run follow-ups. Similar patterns of strong persistence have been found in adulthood for cognitive behavioral therapy (van Dis et al., 2019) and for (largely clinical) interventions that have estimated impacts on components of adult personality (B. W. Roberts et al., 2017).

Persistence in education interventions

Some interventions designed to improve cognition and behavior appear to generate persistent benefits. For example, evaluations have shown that some ECE programs promote educational attainment in early adulthood (for reviews, see Elango et al., 2015; McCoy et al., 2017), whereas additional years of education at the end of schooling appear to improve cognitive test scores (Ritchie & Tucker-Drob, 2018) and earnings (Card, 2001) in adulthood.

Some of the best-known instances of persistence of educational interventions show an interesting pattern of fade-out, followed by the emergence of long-term effects, although not always on the same kinds of developmental outcomes (e.g., F. Campbell et al., 2014; Heckman, 2006; Schweinhart et al., 2005). In the case of teacher effects, Jacob, Lefgren, and Sims (2010) concluded that teacher-induced (value-added) learning has low persistence: Three quarters or more of teacher effects on achievement-test scores fade out within 1 year. However, Chetty et al. (2011) and Chetty, Friedman, and Rockoff (2013) found long-run impacts on both attainment and behavior when the same children were tracked through adulthood via administrative records. A pattern of fade-out and reemergence in young adulthood has also been documented for early social-skills training. The Fast Track program provided a range of behavioral and academic services to a random subset of first-grade boys exhibiting conduct problems. Impacts

Table 1. Proposed Explanations of Fade-Out as a Methodological Artifact.

Explanation	Description	Why it is likely insufficient
	Artifactual explanations	
Misleading effect size reporting	Changes in standardized effect sizes over time can be misleading, particularly when variance on the underlying construct increases with age.	Fade-out has been observed on a variety of measures, scaling decisions, constructs, and age ranges. Effects sometimes reverse in sign.
Publication bias	If follow-up assessments are more likely to be conducted in the case of evaluations showing larger end-of-treatment impacts, then the end-of-treatment impacts in studies with follow-up assessments would be positively selected on sampling error and thus upwardly biased.	Publication bias can make fade-out look more or less severe. Fade-out is observable in quasiexperimental designs for which all outcome waves have been collected before analysis.
	Partly artifactual explanations	
Overalignment	Initial overalignment between treatments and outcomes creates a spuriously large estimate of end-of-treatment impacts.	Fade-out has been observed for combinations of broad treatments and measures (including measures other than cognitive tests), where a strong degree of alignment is unlikely.
Multidimensionality	Interventions may meaningfully affect some psychological attribute, but at follow-up, long-term impacts are misestimated because a different construct is measured.	Fade-out has been observed on outcome measures other than psychological attributes with straightforward interpretations, such as employment and earnings.

in elementary school were uniformly positive, producing improvements in the boys' prosocial behaviors and classroom social competence and reductions in their aggressive and oppositional behaviors (Conduct Problems Prevention Research Group, 1999a, 1999b). By middle or high school, most of these effects had disappeared for all but the highest-risk boys (Conduct Problems Prevention Research Group, 2011), although impacts on some of these outcomes reappeared when the participants were assessed in their mid 20s (Dodge et al., 2015).

All in all, it appears that some well-designed and well-implemented cognitive, social, and emotional interventions produce immediate positive impacts on child and adolescent outcomes. Sharp reductions in subsequent intervention effects are typically observed among the regrettably small fraction of interventions for which follow-up data are available. And, in a handful of some of the most rigorously implemented and evaluated early-childhood interventions, this pattern of rapid intervention-effect fade-out has been followed by the detection of impacts in other domains in adulthood, such as attainment, behavior, and sometimes health.

Methodological and Substantive Explanations of Fade-Out and Persistence

Explanations of the pattern of diminishing treatment effects after the end of an initially successful intervention fall along a continuum (Bailey, 2019). At one end is the idea that fade-out is an artifact of measurement and requires no substantive explanation. At the other end are purely substantive explanations (e.g., fade-out is a result of forgetting and might be mitigated by teaching children skills to mastery). In the middle are explanations of fade-out as either artifactual or "real" depending on how the explanation is framed. For example, some observed fade-out may result as an artifact of teaching to the test—impacts on children's test-specific knowledge might persist following the end of an effective intervention, but that increase in knowledge might not be useful in children's later academic pursuits. We begin our discussion by highlighting some of the methodological problems associated with tracking patterns of fade-out and persistence. Brief descriptions of these explanations, and why we conclude that each of them is insufficient to provide a full account of fade-out, appear in Table 1.

Fade-out as an artifact of misleading effect-size reporting

Because fade-out is defined as a temporal pattern of diminishing effect sizes following the end of an effective intervention, a first question is whether fade-out is robust to the types of measures used to calculate effect sizes. For example, for various reasons, children progress far more rapidly on vertically scaled standardized achievement tests in the early grades than in the later grades when progress is measured in standard deviation units. Using national norming data on seven major standardized tests in reading and six tests in math, Hill, Bloom, Black, and Lipsey (2008) calculated annual growth rates of more than a full standard deviation per year in kindergarten but less than one third of a standard deviation per year in middle school. Thus, when expressed as months of schooling, a treatment effect of *x* standard deviations in kindergarten would be the same as a treatment effect of *x*/3 standard deviations in middle school. Moreover, test-score variance also increases with age on some tests, which means that a standard deviation group difference represents fewer raw points on a test in a given year than in the following year. Thus, treatment impacts reported in standard deviation units could in principle fade over time, even as months of schooling or unstandardized group differences on some policy-relevant scale (e.g., predicted future earnings, rank in the distribution of effect sizes of interventions at some later age) remain the same or increase.

Changes in effect sizes across time are notoriously difficult to study. Bond and Lang (2013, 2018) show that apparent growth in gaps between Black and White students' achievement-test scores in the early school years are sensitive to how the tests are scaled and may in large measure be an artifact of measurement error in test scores in the early grades. Although a pattern of decreasing measurement error with age would bias end-of-treatment impacts downward, making fade-out look less dramatic, the more general idea that changes in group differences can be highly sensitive to scaling decisions should be considered in interpreting the literature on fade-out and persistence.

Although the influence of test-score construction and reporting on fade-out and persistence is not fully understood, there are good reasons to think that these explanations are insufficient to account for fade-out. First, test-score fade-out expressed in standard deviation units tends to be far more rapid, on average, than year-to-year fluctuations in growth rates, falling by 50% or more in the first year after the end of treatment in reviews of educational interventions targeting children in early childhood and the early school years (Bailey, Duncan, Watts, Clements, & Sarama, 2018; Li et al., 2017). Second, within the set of interventions measured with vertically scaled achievement tests, changes in test score variance over time are too small to fully explain fade-out (Cascio & Staiger, 2012). Third, fade-out can be readily observed on measures that are not vulnerable to the problems of vertically scaled achievement tests, such as tasks designed to measure cognitive processes (Bailey, Fuchs, Gilbert, Geary, & Fuchs, 2020a; G. Roberts et al., 2016), life-satisfaction surveys (Lucas, 2007), and economic consumption (Bouguen et al., 2019). Finally, a recent evaluation of a preschool program reported estimated effects on achievement-test scores that flipped from positive at the end of pre-K to negative by second grade (Lipsey, Farran, & Durkin, 2018), a pattern difficult to reconcile with the scaling-based explanations for fade-out above. Still, the possibilities that fade-out may be sensitive to the use of different scales and that different scales may be more useful for researchers with different goals (e.g., researchers interested in the development of skill-building for its own sake vs. researchers interested in long-term impacts on socially important outcomes influenced by the accumulation of human capital) are important reasons to consider a variety of scales when comparing earlier and later impacts.

Fade-out as an artifact of publication bias

Protzko (2015) noted that the temporal pattern of impacts following the end of an intervention may be sensitive to publication practices. In particular, publication bias will reduce the apparent magnitude of fade-out if positive impacts are more likely than null impacts to be published following the end of the evaluation of some intervention. On the other hand, other kinds of publication selection could lead to an overestimation of fade-out. For example, follow-up studies showing that an initially promising intervention faded out over time could be surprising and more likely to be viewed as novel and publishable. Furthermore, if follow-up assessments are more likely to be conducted in the case of evaluations showing larger end-of-treatment impacts (e.g., if funding agencies support follow-ups only after knowing that a treatment worked initially), then the end-of-treatment impacts in studies with follow-up assessments would be positively selected on sampling error and thus upwardly biased. Subsequent effect sizes could regress to smaller magnitudes, creating the illusion of fade-out, even if the true impacts of interventions remained stable following the end of the intervention.

Alternatively, if researchers selectively report positive relative to null follow-up impacts, this would upwardly bias long-term impact estimates and could lead to an underestimate of fade-out. This may be a concern in the clinical literature described above, for which persistence is commonly observed, but long-run outcome measures are quite heterogeneous. For example, in long-term evaluations of studies of cognitive-behavioral therapy in adults, the combination of (a) high rates of uncertainty about whether outcomes are fully reported and (b) the regularity with which intervention developers coauthor intervention evaluations is potentially worrisome (van Dis et al., 2019).

Nor can we discount the possibility of publication bias in meta-analytic estimates of fade-out and persistence. This is particularly true for interventions with small impacts, imprecisely measured relative to their magnitudes, in which the difference between complete fade-out and complete persistence is very small, and both of these outcomes can be obtained within a range of plausible model specifications. Various methods are also available to probe whether the average role of publication bias in any particular literature is sufficient to create illusory fade-out or persistence (e.g., McShane, Böckenholt, & Hansen, 2016).

That said, our confidence that fade-out effects are not fully artifacts of publication bias is bolstered by the fact that observed instances of fade-out come from designs with properties that make publication bias unlikely. For example, some field randomized controlled trials (RCTs) of educational programs contain a single, prespecified outcome measure at any given wave (e.g., Clements et al., 2013). Others have measured the same set of outcomes several times following the end of an intervention, finding corresponding diminishing impacts of treatment on all of them (e.g., Bailey et al., 2020a).

Also inconsistent with a publication-bias explanation is the fact that fade-out has been observed in many retrospective quasiexperimental studies estimating the causal effect of some quasirandomly assigned treatment on a set of outcomes all measured by the time the analysis is conducted. In these cases, positive selection of reported long-term effect sizes is plausibly at least as severe as positive selection of reported end-of-treatment effect sizes. A small subset of educational interventions that have been evaluated years after the end of treatment and shown to produce fade-out on achievement-test scores over time include Head Start attendance (Deming, 2009), high-quality teachers (Jacob et al., 2010), and taking two periods of sixth-grade math (E. Taylor, 2014).

Baseline or posttreatment imbalance

Baseline imbalance between groups in RCTs (i.e., "unhappy randomization") or quasiexperimental designs, or imbalance on posttreatment experiences not caused by the treatment, may lead groups to diverge or converge following the end of treatment in ways that will deviate from the population-level long-run effects of the treatment. However, because these factors are likely to vary randomly and not systematically (at least in the absence of the other artifactual explanations discussed in greater detail above), they are just as likely to lead to inflated estimates of persistence as to inflated estimates of fade-out. Attrition may systematically bias findings toward fade-out or persistence, depending on the processes through which individuals select out of the study. If attrition is greater among control-group members who are more disadvantaged, this may create the illusion of fade-out. R. D. Taylor and colleagues (2017) reported findings that are perhaps consistent with this pattern, estimating that, across studies of the impacts of social-emotional learning interventions, higher levels of attrition were associated with smaller long-term impacts. However, if disadvantaged treatment-group members drop out, this may create the illusion of persistence. For example, Ou, Arteaga, and Reynolds (2019) estimated larger long-term impacts of the Chicago Child-Parent Center program for children who spent more time in the program, but these children were also less disadvantaged, on average. Finally, being in the treatment group may *cause* individuals to have subsequently different environmental experiences, but these would be mediators through which fade-out and persistence would be observed, rather than measurement artifacts (we discuss this possibility as a substantive explanation for impact persistence in the Mechanisms for Persistence section).

All of the fade-out-as-artifact explanations discussed above deserve careful consideration in individual studies and across specific bodies of intervention research. Note that different practices related to multiple testing and decisions to collect follow-up data make different predictions about patterns of publication bias, and we encourage researchers to test these systematically in future meta-analytic work. However, we doubt that these explanations can fully account for the fade-out effect, because the conditions under which they apply (namely, increasing test-score variance over time, the use of vertically scaled tests on which participants are rapidly improving, and the sort of publication bias in which inflated initial effects are reported and more realistic follow-up effects are reported) do not characterize all of the cases in which fade-out has been observed.

One notable body of research to which none of these explanations apply is the study of the forgetting curve (Ebbinghaus, 1885; Murre & Dros, 2015; Pashler, Rohrer, Cepeda, & Carpenter, 2007), in which participants are taught information they would never otherwise learn and repeatedly tested on the same items. This work is important because it points to an important substantive explanation for some instances of fade-out: forgetting, which we discuss below.

Fade-out as substantively meaningful but misleading

We turn now to explanations positing that fade-out results from a lack of correspondence between constructs influenced by or measured at the end of interventions and constructs measured during the follow-up period. Depending on the nature of these constructs, these explanations might be viewed as characterizing fade-out as an artifact or as a substantive phenomenon.

Overalignment between treatments and outcomes. One important explanation for fade-out is that initial overalignment between treatments and outcomes creates a spuriously large estimate of end-of-treatment impacts. Overalignment occurs when an intervention is evaluated with measures closely tied to the treatment itself. An example might be in the evaluation of an early math curriculum that stresses geometry using an assessment containing a preponderance of items testing for understanding of shapes and sizes. Using terminology from experimental psychology, overalignment might be conceptualized as using a manipulation-check measure to estimate program impacts.

For tests measuring skills that accumulate with training, overalignment can result from what is sometimes called *teaching to the test*. And when outcomes are measured by self-reported behavior rather than performance-based assessment, participants may know or guess what the experimenters' hypotheses are and respond or behave in a way that is consistent with those hypotheses.

There are some reasons to believe that teaching to the test contributes to some of the observed patterns of fade-out. First, everyday instruction appears to involve some amount of teaching to tests. Linn, Graue, and Sanders (1990) and Linn (2000) documented instances in which a test used for accountability purposes was replaced within a state or district. Standardized scores for the presumably otherwise comparable cohorts in the following year dropped substantially on the new test, and then slowly rebounded over time as teachers aligned their instruction with the new test, resulting in what the researchers call a "sawtooth" pattern.

Second, a large meta-analysis of educational interventions found that impacts estimated with experimenter-designed measures were roughly twice as large as interventions using independent measures (Cheung & Slavin, 2013). Perhaps apparent fade-out would be observed if impacts on all tests remained the same following an intervention, but a researcher-created outcome was measured at the end of treatment, and grades or standardized test scores were measured at a follow-up assessment.

Teaching to the test is a common explanation for fade-out following cognitive-training interventions. Jensen (1998) referred to gains following cognitive training as "hollow." However, the precise links between teaching to the test and fade-out are rarely articulated and may not be as direct as many psychologists assume (Protzko, 2015, 2016). We discuss some of the relevant ideas using the case of cognitive-training interventions, for which the goal is often to elicit transfer of learning to some general causes of individual differences in cognitive ability.

In one sense, fade-out following overly aligned treatments and posttest measures as well as teaching to the test is real—performance or self-reported psychological characteristics did indeed change relative to some counterfactual condition, and this difference was no longer detectable at some later point. In another sense, fade-out on overly aligned outcome measures is artifactual if the targeted underlying psychological characteristics were never changed at all. Completely artifactual cases are better conceptualized as instances in which there was no initial impact rather than cases in which an initial impact diminished after the end of treatment.

It is noteworthy that explanations based on overalignment are not likely to apply to the same degree across all interventions that raise children's test scores and show some degree of fade-out after the end of treatment. Although it is sensible that short-term, task-specific coaching interventions will lead to shallow knowledge that may be forgotten, many educational interventions are designed to build children's conceptual understanding within some underlying domain (e.g., Clements et al., 2011; Gonzalez et al., 2011) and may look quite different—in duration of treatment, amount of contact with an instructor, and depth of content—from experiments in which participants receive repeated test practice (e.g., Estrada, Ferrer, Abad, Román, & Colom, 2015; we discuss ways in which the breadth of training impacts might relate to fade-out and persistence in Box 1). Still other interventions change children's contexts intensively for a substantial amount of time, as when children attend preschool for several hours a day rather than staying home. In such cases, the idea that participant benefits are fully captured by a narrow assessment is unlikely. Yet some degree of fade-out on cognitive test scores is the norm in both kinds of interventions.

Multidimensionality. A broader set of possible explanations of fade-out, of which overalignment may be considered a subset, is that the appearance of fade-out is a predictable consequence when an intervention *meaningfully* affects some psychological attribute, but at follow-up, long-term impacts are misestimated because a different construct is measured. For example, B. Clarke and colleagues (2016) found substantial impacts of a kindergarten mathematics intervention at the end of kindergarten on a variety of tests focused primarily on children's number sense. At the end of first grade, researchers observed no impact on a novel broad math achievement test. In this case, one cannot rule out the possibility that, although impacts never emerged on that broad math achievement test, impacts may have persisted on number sense. If number sense continues to influence performance or learning during the follow-up period, it would

Box 1. Case Study: Multidimensionality, Fade-Out, and Persistence in Cognitive Skill Effects.

Some psychologists are tempted by the idea that a necessary condition for generating persistent impacts on cognitive skills is to improve an underlying broad construct of set of latent constructs, such as general cognitive ability. Consistent with this possibility, within the cognitive training literature, both lack of broad transfer (e.g., Colom et al., 2013; Sala & Gobet, 2018) and fade-out (e.g., Melby-Lervåg, Redick, & Hulme, 2016; Takacs & Kassai, 2019) are commonly observed. But two kinds of counterexamples show that this co-occurrence does not prove that boosting general cognitive ability will always lead to persistence and that changing specific skills will always lead to fade-out. First, impacts on specific kinds of skills may persist in ways that manifest themselves on broad cognitive tests. Indeed, persistent effects on test scores have been identified as a result of intensive interventions, such as adoption (Kendler, Turkheimer, Ohlsson, Sundquist, & Sundquist, 2015) or a year of extra schooling (Ritchie & Tucker-Drob, 2018), despite some evidence that these gains are in specific skills and not solely in a single latent general cognitive ability (Ritchie, Bates, & Deary, 2015; te Nijenhuis, Jongeneel-Grimen, & Armstrong, 2015). Second, inconsistent with the idea that changes to higher level cognitive abilities will always persist, gains induced by sustained contextual interventions that occur across a broad set of cognitive measures, as would be expected if the intervention affected higher level cognitive constructs, can fade out (Protzko, 2016).

be misleading to interpret the null follow-up impact on the broad math achievement test as being indicative of complete fade-out.

Multidimensionality of assessment content across waves is most clearly a concern in skill-building domains, in which different tests are administered at different points during development or in which the same test includes different kinds of items for individuals who are at different points in their development. In such cases, impacts on skills targeted by an intervention might diminish on subsequent assessments in part because those assessments are no longer measuring the content targeted during the initial intervention.

For example, children's reading skills follow a predictable skill-building sequence: Throughout development, correlations between decoding and reading comprehension decrease as correlations between linguistic comprehension and reading comprehension increase (Hoover & Gough, 1990). Foorman, Petscher, and Herrera (2018) examined the prediction of reading comprehension by decoding and linguistic comprehension in a large sample of students in Grades 1 through 10 and found that their contributions to reading achievement were dynamic. Figure 2a shows that decoding explained more unique variance than linguistic comprehension in Grade 1 reading achievement; the majority of the explained variance was shared between the two predictors. Figure 2b shows that this pattern changed dramatically by Grade 6, such that more than half of the total variance in reading achievement was explained by the unique contribution of linguistic comprehension and nearly all the remaining variance was explained by the shared variance between decoding and linguistic comprehension. By Grade 10 (Fig. 2c), unique variance due to linguistic comprehension explained an even larger proportion of the variance, and the shared variance between decoding and linguistic comprehension continued to account for almost all of the remaining variance. By high school, reading comprehension and written language understanding at the word and discourse level were nearly psychometrically indistinguishable.

Thus, consider an intervention generating a persistent impact on decoding. If the heightened decoding skills did not transfer to linguistic comprehension, then the intervention would appear to have a diminishing impact on a reading comprehension test from one grade to the next. If reading comprehension scores are considered to be a good index of the intervention's long-term success, then one might consider this to be a clear case of fade-out. Indeed, many literacy researchers have noted that the skills required for success vary dynamically according to what individuals need to learn or do to be successful (e.g., Ackerman, 2017; Cunha & Heckman, 2007).

Alternatively, one might consider fade-out in this hypothetical case to be a measurement artifact, given that treatment impacts on the trained cognitive skill (decoding) do not diminish after the end of treatment. Our view is that, in such cases, both persistence on the decoding measure and fade-out on the reading comprehension measure are real, and their relative importance will depend on the importance of 10th-grade decoding and 10th-grade reading comprehension for children's later-life outcomes.

The importance of the multidimensionality problem for understanding fade-out depends on the answers to two questions: (a) How important are variations in basic

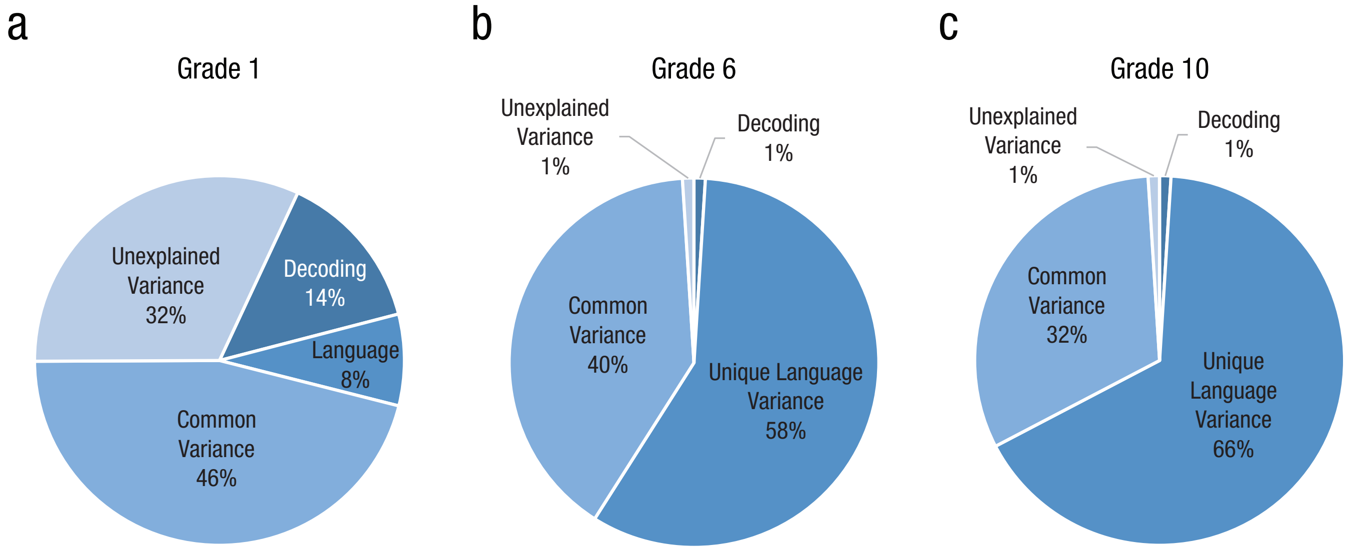

Fig. 2. Total percentage of variance explained in reading comprehension decomposed into unique and common effects of language, decoding, and unexplained variance for Grade 1, Grade 6, and Grade 10.

skills, above and beyond some limited amount of transfer to intermediate measured skills, for advanced skill development and other socially important outcomes? (b) If variations in basic skills continue to causally influence these long-term outcomes, do impacts on basic skills persist in cases in which basic skills are manipulated and more advanced skills are measured at follow-up?

The answer to the first question likely varies substantially across interventions, populations, and outcomes: For example, increasing children's decoding skills may influence their educational success primarily because they transfer to early reading skills. In contrast, if educational interventions improve children's socioemotional skills, these skills may improve academic achievement in the short term and directly reduce high school dropout in the long-term by improving socioemotional behaviors. However, as we argue in the Mechanisms for Persistence section, skills that are no longer measured after the end of the posttest may often develop very quickly under counterfactual conditions, leading to catchup by the control group, and thus our hypothetical example of persistence on decoding may not be realistic in many cases. For example, by 2 years after a number-knowledge tutoring intervention that gave first graders a substantial amount of practice with speeded arithmetic, children from the control group fully caught up to those in the treatment group on the measure of speeded arithmetic given at the end of treatment (Bailey et al., 2020a).

To summarize our discussion of multidimensionality, understanding changes in measures and constructs over time may be useful for developing a process-based understanding of fade-out in some cases. Further, a variant of the multidimensionality problem (i.e., treatment impacts on some constructs are substantial and persistent but unmeasured) may be an important route to the emergence of long-term effects of interventions.

Fade-out as a predictable consequence of psychological processes

Some explanations for fade-out invoke psychological and/or social processes that occur following the end of interventions. Each explanation makes predictions about which factors will promote persistence.

Forgetting. As noted above, research on forgetting has provided some of the clearest and most robust evidence for fade-out. Moreover, the shape of fade-out curves and the shape of forgetting curves are often very similar. Might forgetting play an important role in fade-out from skill-building interventions? Forgetting is likely under conditions in which individuals practice learned information less and new information more (for review, see Wixted, 2004). Both of these conditions are likely to prevail when children receive a one-time skill-building intervention and then return to an environment with children who did not.

A simplistic interpretation of the hypothesis that forgetting accounts for fade-out—that children forget everything they learned immediately following the end of an effective educational intervention—can be ruled out: As described above, the most common pattern of fade-out observed in ECE interventions is one of control-group catchup rather than absolute declines in the skill of treatment-group children. However, control-group catchup does not rule out a role for forgetting: D. T. Campbell and Frey (1970) presented a very simple model of simultaneous learning and forgetting to account for fade-out following effective ECE intervention. In brief, for some subset of related skills and equal environmental inputs in the period following an effective intervention, if forgetting is steeper for individuals who have acquired more skill, and learning is steeper for individuals with less skill, then skill levels of the more skilled treatment group and the less skilled control group will converge.

Evidence for the role of forgetting in fade-out comes from a study of the TRIAD pre-K mathematics intervention (Kang et al., 2019). One year after the end of treatment, children in both the treatment group and the control group were given a math achievement test. Children in the treatment group were more likely than children in the control group to incorrectly answer an item that they had incorrectly answered on the end-of-treatment assessment. This difference, which does not capture instances in which children in the treatment group forgot and then relearned items, was approximately one quarter the size of the total fade-out effect during this period.

Modest transfer. Although transfer of learning (i.e., when learning one skill influences the performance or the learning of another) is likely to play an enormous role in children's academic development, the striking degree of fade-out following initially effective skill-building interventions suggests that transfer of learning may play a smaller role than is often assumed. We argue that both of these statements can be true, particularly when (a) the set of possible skills on which one might intervene is very large, (b) development under counterfactual conditions is rapid, and (c) one-time skill-building interventions are constrained by time and resources to focus on a narrow subset of these skills.

The degree of fade-out observed in studies of academic skill-building casts doubt on the hypothesis that transfer of learning within a single domain (e.g., literacy or math) primarily accounts for the large correlations between children's early and much later academic achievement. A common and influential finding in developmental psychology has been that children's early academic achievement strongly predicts their much later academic achievement, after controlling statistically for a comprehensive set of covariates (e.g., Duncan et al., 2007). When interpreted causally, these findings imply that approximately 40% of the initial magnitude of an early math intervention's treatment effect will persist approximately 5 years after the end of treatment—a degree of persistence obviously at odds with the kind of data presented in Figure 1.

Moreover, the kinds of partial correlations presented in Duncan et al. (2007) show some odd patterns that are difficult to reconcile with cognitive and educational psychological theory. For example, if interpreted causally, they imply that early math interventions will have persistent effects on reading skills many years later and that these effects will be comparable in size to the effects of early reading interventions on later reading achievement (Bailey et al., 2018). In contrast, patterns of fade-out following the end of an effective early math intervention imply that treatment effects decay by approximately 60% each year following the end of an intervention (whether they approach an asymptote at 0 or some small positive value is not easily determined although it is perhaps of theoretical importance). These results imply that transfer of learning within the math domain and across domains accounts for some of the longitudinal stability in children's early and later academic skills—but far from all of it and perhaps for only a trivial part of the stability across multiyear time lags and across domains.[2]

Mechanisms for Persistence

Bailey et al. (2017) proposed three distinct processes that might sustain the benefits of interventions for children and adolescents: skill building, sustaining environments, and institutional opportunities and constraints.[3]

Key to the *skill-building* process is the idea that simpler skills support the learning of more sophisticated ones and, in some economic models, that skills acquired before a given skill-building or capacity-building intervention increase the productivity of that investment. Skill-building processes can be readily seen in both math and literacy learning. In math, counting serves as a cognitive basis for children's early addition-problem solving, and addition is often employed as a subroutine of children's multiplication-problem solving (Baroody, 1987; Lemaire & Siegler, 1995). In reading, 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 (LaBerge & Samuels, 1974). It is important to note that "skills" in skill-building models are conceived to be much broader than conventional academic skills such as literacy and math. Indeed, they encompass any skill, behavior, capacity, or psychological resource that helps individuals attain successful outcomes.

In the case of skill-building interventions, Bailey et al. (2017) point to the potential importance of targeting what they call "trifecta" skills—ones that are malleable, fundamental, and would not have developed in the absence of the intervention. In this framework, malleability is not an absolute property of any skill but is defined relative to what can be changed across the range of currently available interventions. Fundamentality is the extent to which a skill affects positive life outcomes in adulthood or affects other intermediate skills that reliably affect such outcomes. Bailey and colleagues argue that all three conditions are needed to generate long-term effects and that the third "trifecta" condition—eventual skill development in counterfactual conditions—is particularly problematic for long-term impacts of interventions focused on building early literacy and math skills because most children are likely to eventually acquire at least minimal levels of these skills soon after entering school. This kind of "catchup-driven" explanation may explain the widespread fade-out reported in our above review of interventions targeting specific skills.

A second approach to understanding fade-out is termed the *sustaining-environments* perspective by Bailey et al. (2017). It recognizes the importance of interventions that build important skills and capacities but views the quality of environments subsequent to the completion of the intervention as crucial for maintaining initial skill advantages.

An example of *un*sustaining environmental processes is the *constraining-content hypothesis*, which is based on the idea that fade-out is a consequence of high-achieving students' limited opportunities to build on learning from an effective educational intervention after it ends. This might be the case for intervention gains among the higher achieving students if subsequent instruction is aimed at lower achieving students. Support for this idea comes from studies showing fade-out following Head Start attendance, which has been hypothesized to result from Head Start children attending lower quality elementary schools (Currie & Thomas, 2000) and U.S. kindergarten teachers having been observed to dedicate little instructional time, on average, to advanced academic skills that might be most likely to build on skills learned in high-quality, skill-building preschools (Engel, Claessens, & Finch, 2013).

Despite this kind of supportive evidence, more general tests of the constraining-content hypothesis have yielded mixed results. One important prediction is that fade-out will be largest for children with the highest levels of achievement and thus the least opportunity to be exposed to content beyond their level of mastery. However, similar degrees of fade-out for high- and low-achieving students are often found after early educational interventions (Bailey et al., 2016; Bitler, Hoynes, & Domina, 2014). Also relevant to the constraining-content hypothesis are studies that have estimated statistical interactions between early and later educational quality. We review this literature in the Sustaining Environments section below.

As explained in the "gateway" sections below, interactions between children's developmental trajectories and schools is key to the perspective on institutional opportunities and constraints. In this case, successful interventions equip a child with the right skills or capacities at the right time to avoid imminent risks (e.g., grade failure, teen drinking, or teen childbearing) or to seize emerging opportunities (e.g., entry into honors classes, SAT prep) associated with the organization of schools or other social structures. The skill or capacity boosts need not be permanent, as with SAT prep that boosts chances of acceptance into a college with greater resources, a key step in a positive cascade that might influence human-capital and labor-market outcomes. For SAT prep, it is the enriched college resources, rather than any lingering test-prep knowledge, that might lead to a higher paying job.

A formal model of skill building

Our discussion of fade-out and persistence thus far has been inductive and presented in narrative form rather than as formal theory. Formal models such as the one presented below can be very helpful in clarifying concepts and making explicit assumptions that are often hidden and sometimes unintended (Lave & March, 1993). Formal models are best developed in a cyclical process. First, the theorist studies data from sound empirical studies to inform model assumptions then proceeds by deriving model predictions that can inform empirical studies to reject such predictions (or not). Then, given rejections, the new data are used to update model assumptions, thus resuming the cycle in which formal models are created.

In this section, we extend the kinds of skill-building models developed by Cunha and Heckman (2007) by showing the conditions under which these models can explain the patterns of fade-out or persistence of intervention effects (details of the model are provided in the Appendix). We also describe conditions under which skill-building models can produce the emergence of long-term effects in adult outcomes reported in some studies in the literature.

Skill-building models developed by economists view skill building by children as a sequential process of combining investments of time and money (specific to these skills) with the stock of children's skills developed at earlier stages of the child's life cycle. Economists refer to the process by which more complex skills are built on simpler skills as *self-productivity*. A second key component of skill-building models focuses on the nature of the interactions between the level of incoming foundational skills developed before the intervention (period $t - 1$) and the levels of time and money investments undertaken in the current stage (period t). A key issue is whether higher foundational skill levels increase or decrease the productivity of investments made during period t. Sustaining environments are possible when the productivity of investments undertaken in the time period following the end of an intervention (period $t + 1$) is higher for individuals whose skills have been boosted by the period t intervention than for otherwise similar individuals not exposed to the period t intervention. Cases in which these investments yield greater benefits to individuals with lower as opposed to higher levels of incoming skills will lead to faster fade-out.

The nature of these interactions can be framed in the following fashion: Consider two children, Child A and Child B, who differ in their levels of counting knowledge; Child A has higher foundational skills than Child B. If both receive the same amount of teaching time and effort to learn addition and subtraction, which child will profit the most from the instruction? If Child A will profit most, we say that the skill-building model features dynamic *complementarity*—the teaching investment complements a child's incoming level of foundational skills and produces a Matthew effect, where the rich get richer. On the other hand, if teaching investments are more productive for Child B, then we say that the model features dynamic *substitutability*—the teaching investment is compensatory by raising skills already mastered by Child A but not Child B.

In the case of sequential interventions, different skills may exhibit dynamic complementarity or dynamic substitutability with different interventions. Dynamic substitutability is more likely to occur when two skill investments are redundant in their content. An example that is consistent with dynamic substitutability is Clements et al.'s (2013) TRIAD evaluation a pre-K math intervention (see Figs. 1a and 1b). The results of the evaluation show that although the math skills of children in the treatment group grew faster (relative to the math skills of children in the control group) during the intervention period, the math skills of the treatment group actually grew more slowly than the skills of children in the control group in the years after the end of the intervention. Key to dynamic substitutability in this case is that what determines the importance of math skills targeted by the pre-K TRIAD intervention are not the investments in the preschool years but the total sequence of investments across a number of periods. Dynamic substitutability is thus closely related to the notion of the rate of development under counterfactual conditions as proposed by Bailey et al. (2017). We return to this issue below.

Dynamic substitutability also has implications for the composition of the target population for early-childhood interventions. Such interventions (e.g., Head Start at age 4) usually target children who are at risk for having experienced relatively low levels of investment (e.g., by parents or programs such as Early Head Start) period to age 4. But if Early Head Start and Head Start investments are dynamically substitutable, then the highest returns to Head Start would be for children who had not received services from Early Head Start. An example of dynamic substitution is Rossin-Slater and Wüst's (2016) study of a Danish preschool program and a nurse home-visiting program, which were rolled out across communities in a haphazard way during the same period. Although each produced long-term effects on educational attainment, they were almost completely dynamic substitutes for one another: Receiving both programs yielded no larger effects than receiving just one of them.

Dynamic complementarity presents difficulties for policymakers concerned with educational equity. It implies that skill-building interventions in, say, adolescence may be less productive if early investment levels are low (Cunha & Heckman, 2007). This argument is used for advocates of early-childhood investments who identify low levels of early investments as a significant bottleneck for the development of human capital once children growing up in disadvantaged circumstances reach school age. If, say, educational investments in the K–12 schooling years dynamically complements kindergarten-entry skill levels, then preschool investments in children's school readiness will pay learning dividends across all of the years of formal schooling. The opposite would be the case if the two sets of investments are dynamic substitutes.

There is less discussion of a second public-policy implication of dynamic complementarity: If the skill-formation process features dynamic complementarity, then the rate of return on early investments is expected to be greater when levels of late investments are higher. Therefore, when dynamic complementarity is in place, interventions that boost early skill formation will generate larger long-term impacts when accompanied by higher levels of later investments. These ideas support a direct linkage between dynamic complementarity and the concept of "sustaining environments," as described in Bailey et al. (2017) and elaborated on below.

Another important result from skill-building models is that the effects of early interventions will fade out if the sequence of postintervention investments is the same for the control and treatment groups. The intuition is as follows: Although it is true that skills produced in the early stages of the life cycle are an input into the production of skills produced at later stages, early skills are not the only input. Indeed, in skill-building models that feature malleability, skills require investments as well. Regardless of the degree of dynamic complementarity or substitutability, if later investments are equated across groups—and below the initial boost in the treatment condition—fade-out will generally happen regardless of whether the model features dynamic complementarity or substitutability.[4]

As highlighted in the previous paragraph, skill-building models will generate fade-out if there are no differences between control and treatment groups regarding the amount of investment in postintervention periods. Unfortunately, very few studies have systematically investigated the differences in postintervention investments. Two studies that have done so focused on the Head Start program. First, Currie and Thomas (2000) used data from the Children of the National Longitudinal Study 1979 to compare the quality of postintervention environments between children who attended the Head Start program and children who did not. They found that Black Head Start children go on to attend lower quality schools than other Black children do. This finding suggests that treated children experienced lower levels of investment during postintervention years than control children. Gelber and Isen (2013) used data from the Head Start Impact Study and found a small but positive difference in postintervention home investments between treated children and control children, between 0.03 and 0.08 *SD*.

Given these empirical findings, the puzzle, from the perspective of economic models of skill formation, is not why fade-out or persistence exists, but why convergence or divergence of investments exist in postintervention periods between control and treatment groups. An exogenous increase in the levels of investment in period $t-1$ induces two forces in investments in period t that operate in opposite directions in the presence of dynamic complementarity and in the same direction in the presence of dynamic substitutability. An exogenous increase in investments in period $t-1$ *raises* the returns on investments in period t in the presence of dynamic complementarity (because of its synergistic nature) but *lowers* the returns on investments in period t in the presence of dynamic substitutability

On the other hand, an exogenous increase in investments in period $t-1$ raises skill levels in period $t-1$ and, via self-productivity, also indirectly raises skill levels in period t (everything else constant). Thus, this second force reduces the incentives for higher levels of investment in period t. If there is dynamic substitutability, both forces predict that an exogenous increase in investments in period $t-1$ would reduce investments in postintervention years. If there is dynamic complementarity, the forces work in opposite direction, and determining which force prevails depends on the relative strength of dynamic complementarity and self-productivity. The stronger the dynamic complementarity and the weaker the self-productivity of skills, the stronger the incentives for investments to be higher in period t.

We now return to the discussion concerning the concept of skills that would not develop well under counterfactual conditions (Bailey et al., 2017). Skill-building models predict that proficiency in skills that have sensitive periods of development is more closely determined by the total amount of investment during sensitive periods than by how the investments are distributed within those sensitive periods. This finding automatically implies fade-out of intervention effects. To see why, consider a skill for which the sensitive period of development lasts 2 years, which we denote by year t and year $t+1$. Assume that we randomly allocate individuals to an intervention or control group. In the control group, the individuals receive two units of investment in each period. In the treatment group, the individuals receive three units of investment in the first period and one unit of investment in the second period. Both groups receive four units of investments, but these investment units are distributed differently across groups. If there is equal malleability and a high degree of dynamic substitutability, skill-building models predict the treatment group will surge during year y, but that the control group will eventually catch up by the end of year $t+1$. Thomas and colleagues (2019) call the frustrating implication of malleability for fade-out the "double-edged sword of plasticity" (p. 718). When sensitive periods of development are long, final skill proficiency is determined by total amount of investment within sensitive periods and not how investments are allocated.

Ideally, we would be able to use these implications of skill-building models to inform public policy. From the point of view of skill-building models, the design should start by answering this question: "What is the long-term outcome of interest that the intervention is supposed to improve?" Once the answer is determined, then "What are the most fundamental skills that determine this outcome?" The answer to this question is very important because not all outcomes are equally affected by different skills (e.g., Heckman, Stixrud, & Urzua, 2006). The third question: "At what stage of the life cycle are these skills most malleable, and is the sensitive period of development long or short?" The answer to this question determines the target age range for the intervention and whether the timing of investment is crucial or not. The fourth question: "In practice, what constitutes investment in the formation of these fundamental skills during the sensitive periods of development?" And, finally, "Who is at risk for low levels of investment during these fundamental skills' sensitive development period?"

In summary, skill-building models describe a rich process of human-capital formation. There are many families of skills, each one has multiple members, and each member has different degrees of fundamentality, distinct levels of malleability, and diverse durations of sensitive periods of development.

There are two ways that skill-building models can explain why impacts fade out over time. First, impacts will fade out because of malleability. Each member in a family of skills requires investments. Thus, interventions that take place in only certain segments of an individual's life may affect the skill-family members that are malleable during those segments, but this advantage can be sustained over time only if later skills also receive higher levels of investment from sustaining environments or if self- or cross-productivity is reliably large. Alternatively, skill-building models predict that impact will fade out over time if interventions just anticipate investments to early stages of sensitive periods of development.

There are also two ways in which skill-building models predict that interventions will not lead to long-term impacts. First, interventions will not affect long-term outcomes if the interventions do not focus on fundamental skills. Second, interventions that simply anticipate investments to early stages of sensitive periods will also not lead to long-term impacts because long-term outcomes are affected by final performance in fundamental skills and not by performance at each point in time.

What skills to target? Bailey and colleagues (2017) highlighted the potential trade-offs among the criteria for trifecta skills. Malleable and fundamental skills (e.g., basic math and reading skills) may develop quickly under counterfactual conditions because educators and other interventionists know they are malleable and foundational and have designed early-grade curricula accordingly. Analogously, psychological factors with known limited sensitive periods are often those for which the required inputs will be predictably present and are the subject of pediatrician screening. Examples include basic visual input and exposure to the sounds that appear in one's first language; for a review on mechanisms underlying sensitive periods in cognitive development, see Knudsen (2004). Under such conditions, increased investment in early skills that are both malleable and fundamental would be dynamic substitutes for schooling and other counterfactual environments, resulting in fade-out. Exceptions would include cases in which these inputs are predictably absent, most notably cases of extreme deprivation, or in educational settings, perhaps the case of second-language acquisition. Skills that are foundational and do not develop quickly under counterfactual conditions may be difficult to change (this may apply most to broad domains that are influenced by a large number of inputs, such as working memory capacity or conscientiousness). Skills that are malleable but do not develop under normal conditions may not be taught because educators have decided they are not foundational (e.g., learning Aramaic).

Because of these trade-offs, Bailey and colleagues (2017) argued that the list of potential trifecta skills may be short; their list included advanced academic and vocational skills and social cognitive factors, such as children's implicit theories of intelligence. It also included additional potential targets, such as nutrition, toxic stress, parenting, and basic problem-solving skills, for children in particularly adverse environments. We think the list is plausible, but also that the binary category of trifecta skill probably does not reflect the underlying distribution of promise of all of the possible skills that could yield long-term benefits. Malleability, fundamentality, and development under counterfactual conditions are all difficult to measure quantitatively, are continuously distributed, and vary across populations as well as across time within individuals.

Measures of skills omitted from this list because of a hypothesized lack of malleability, such as personality traits and general cognitive ability, appear to change in response to environmental inputs (Bleidorn et al., 2019; Ritchie & Tucker-Drob, 2018) in ways that are plausibly consequential for long-term outcomes. We think that more formal models of skill development, along with increased use of cost-benefit analysis (both of which tend to rely on long-term follow-up studies from causally informative evaluations of interventions, when available) are better sources of evidence for deciding which skills to target, and we discuss these later in the Implications for Research section. However, in Box 2, we describe an example of how reasoning about these criteria might be used to make judgments about what kinds of skills to target.

Sustaining environmental processes

In our review of skill-building models, we introduced the concepts of dynamic complementarity and dynamic substitutability. Both are likely to be important for educational interventions. Bailey and colleagues (2017) discussed mechanisms through which dynamic substitutability would lead to fade-out and through which dynamic complementarity would lead to persistence. As described above, teaching children skills they are likely to learn quickly under counterfactual conditions

Box 2. A Case Study From Effects of Promising and Unpromising Targets of Reading Intervention.

The results from the systematic reviews and meta-analyses of well-designed studies show that alphabetic skills, word reading, and spelling are proven intervention targets for students who struggle to learn to read (Gersten, Newman-Gonchar, Haymond, & Dimino, 2017; Wanzek et al., 2018). These are malleable, foundational skills that are necessary but not sufficient for reading success. The fact that they are not universally mastered by all students with or without intervention is clear from National Assessment of Educational Progress (NAEP) results. There may indeed be phonics programs that perseverate on constrained skills that are transient and fail to generalize to other reading skills (Paris, 2005). Examples are continuing to teach phonological awareness skills once students in kindergarten or Grade 1 can already decode words or measuring words correct per minute as a substitute for reading comprehension. Yet well-designed phonics programs that provide a scope and sequence of alphabetic instruction and practice in decodable text, with sufficient opportunities to differentiate instruction so that all students can progress to text that can be read independently with accuracy and comprehension, are a valuable instructional support (Chingos & Whitehurst, 2012; Foorman, Francis, Davidson, Harm, & Griffin, 2004) and can promote achievement gains if well implemented (Foorman et al., 2003; Foorman, Francis, Fletcher, Schatschneider, & Mehta, 1998; Kim et al., 2016).

Building skill in oral language would seem to be a critical target for reading interventions if students are to comprehend written text. Indeed, the large amount of variance that language and decoding share in predicting reading comprehension in the primary grades and the gradual amalgamation of language and reading comprehension into one dimension during the secondary grades (Foorman et al., 2018) suggest that instruction may require integrating language and reading skills (Foorman, Petscher, Stanley, & Herrera, 2017). If decoding skills are to lead to efficient word identification and lexical access, they must be connected not only to a word's pronunciation but also to its multiple meanings, orthographic representation, and morphological structure (Perfetti, 2007). This lexical knowledge must be integrated with world knowledge and inferencing skills through comprehension processes that accurately represent the propositions in text (Perfetti & Stafura, 2014).

Thus, in addition to developing oral language skills, reading interventions must build word knowledge by targeting efficient retrieval of word meanings through accurate orthographic representations and their integration with mental models of the text (Cain & Oakhill, 2007; Perfetti & Stafura, 2014). A challenge for successful reading interventions is to accomplish this word-to-text integration within readers' limited attentional and memory resources and to teach oral language skills when they have not developed during normal circumstances (however, for promising follow-up effects of a language intervention with students in the upper elementary grades, see P. Clarke, Snowling, Truelove, & Hulme, 2010).

may lead to rapid fade-out. This is an example of dynamic *substitutability*: An ECE intervention and subsequent schooling are substitutes in the task of building children's academic skills. Truly sustaining environments require something different—some sort of mechanism whereby high-quality postintervention contexts sustain the effects of early interventions. Sustaining environments rely on dynamic *complementarity*: later high-quality investments will pay off more for children who received an effective earlier intervention than for children who did not.

Intensive interventions that affect children's contexts throughout development may be most likely to yield persistent benefits. However, such interventions are very expensive, making them difficult to implement at scale. From a policy perspective, identifying reliable sustaining environments would be quite useful, because this would allow for a combination of intensive targeted early intervention for children at risk for persistently poor outcomes, and continued investment in later environmental quality would benefit everyone.

The sustaining-environments hypothesis has received mixed support. The strongest supporting example is Johnson and Jackson's (2019) analysis of variation in changes in Head Start and K–12 funding in the United States in the 1960s through 1980s. The estimated effects of Head Start attendance on educational attainment and adult wages/earnings were stronger for children who also experienced a court-ordered increase in school spending when they were in elementary school. Although this analysis provides compelling evidence for dynamic complementarity, the exact mechanisms (be they cognitive or noncognitive skill-building, institutional gateways, both, or something else altogether) through which these changes influenced adult outcomes are not clear. In the Rossin-Slater and Wüst (2017) study of a Danish preschool program and nurse home-visitation program mentioned earlier, receiving both programs yielded no larger effects than receiving just one of them.

In a recent meta-analysis on the sustaining-environments hypothesis, Bailey, Jenkins, and Alvarez-Vargas (2020b) found that estimated interactions between the effects of early educational interventions and measures of later educational quality were approximately 0. Further, in the studies analyzed, it was rare for either early or late educational quality to be randomly assigned (for an exception, see Jenkins et al., 2018). Some main effects of preschool programs were estimated to be negative (likely because of selection bias), making it difficult to test the theoretical predictions of dynamic complementarity.

Perhaps more causally informative are two-shock designs—or "lightning strikes twice" designs—in which two random or quasirandom sources of variation influence individuals during development, and the interaction between these two shocks can be estimated. For example, in Bangladesh, a randomized controlled trial of vitamin A supplementation was conducted on infants, many of whose mothers were affected by a tornado during pregnancy (Gunnsteinsson et al., 2019). The vitamin A supplementation was found to mitigate the effects of tornado exposure 6 months later. A recent review of this small but growing literature by Almond, Currie, and Duque (2018) suggests that when interactions were present, interventions *compensated* for early disadvantage, consistent with dynamic substitutability.

Although the sustaining-environments hypothesis has received only limited support, designing targeted interventions that are likely to complement children's likely subsequent investments may be a worthy pursuit. Important questions remain unanswered. For example, might it be the case that in the early grades, school content is sufficiently redundant that observed variation in quality or advanced content is unlikely to moderate the persistence of early intervention impacts on later learning? Still, better alignment of curricula across preschool and the early school years may benefit children who receive effective instruction in preschool (Phillips et al., 2017). Perhaps most importantly, if the key mediators of the long-term effects of schooling are not reflected in test scores, then more direct tests of the sustaining-environments hypothesis will rely on better measures of these factors.

Further, the potential for complementarities across domains, contexts, and treatments is large, and exploration may benefit from stronger theories of when they might be expected to operate. Walton and Yeager (2020) hypothesized that interventions targeting only beliefs (but not teaching skills) would generally be more effective following the end of treatment in contexts that reinforce the idea that the belief is valid and actionable. Supportive evidence comes from the National Study of Learning Mindsets (Yeager, Hanselman, Walton, et al., 2019), which evaluated a short, online "growth mindset" intervention designed to motivate ninth-grade students by teaching them that abilities were malleable. The intervention showed effects on grades at the end of the school year only when the peer climate was supportive of the growth-mindset message.

Institutional constraint and opportunity explanations of fade-out and persistence

> Early transitions can have enduring consequences by affecting subsequent transitions, even after many years and decades have passed. They do so, in part, through behavioral consequences that set in motion cumulating advantages and disadvantages. (Elder, 1998, p. 7)

One way of thinking about sustaining environments is that the persistence of intervention effects requires environments that apply a positive force that is equal and opposite to the forces that would return a person's outcomes to the original, undesirable levels. This assumption leads to a model in which interventions cannot generate persistent effects in the absence of sustained and strong support of the intervention target. Although this may be a useful way to conceptualize fade-out in some instances, it fails to appreciate the ways in which individuals are not isolated agents but instead are embedded within and surrounded by institutions. We argue that an important kind of pathway through which interventions may lead to persistent impacts is one in which certain kinds of interactions with institutions at critical moments in the life course put constraints on individuals' outcomes far into the future. We call this the *institutional-gateway mechanism* for intervention effects.

As detailed below, examples of gateway mechanisms in adolescence are policies or interventions generating small changes in access to college-admissions tests or scores (e.g., Hurwitz, Mbekeani, Nipson, & Page, 2017; Hyman, 2017) or to financial aid (e.g., Bettinger, Gurantz, Kawano, Sacerdote, & Stevens, 2019) that in turn help qualified high school students make the transition to college. These interventions do not teach a fundamental skill, but they nevertheless could improve long-term outcomes because they help students enroll in competitive colleges that have more resources to support students on the path to graduation.

Agency and structure. Our line of reasoning regarding institutional constraints and opportunities is most directly inspired by a classic debate in sociology about the tension between agency and structure (e.g., Coleman, 1990). Throughout its history, sociology has described the ways in which institutions ("structures") constrain the possible effects of human choices ("agency"), in particular across group lines (e.g., between rich and poor). Glen Elder (1998) summarized well the view of the field:

> Some individuals are able to select the paths they follow, a phenomenon known as human agency, but these choices are not made in a social vacuum. All life choices are contingent on the opportunities and constraints of social structure and culture. (p. 2)

The contextual factors that are well described in this literature have important implications for theories of persistence and fade-out. Life-course theories in sociology have argued that human agency can matter a great deal during life transitions (Crosnoe, 2011; Elder, 1998) because they are moments of flexibility in a person's engagement with institutional structures. Examples include the transition from one school to the next, from youth to financial independence, from single to married life, and so on.

The institutional-gateway mechanism can be illustrated with the metaphor of a train leaving a station (Yeager, Hanselman, Muller, & Crosnoe, 2019). A small intervention, such as a sprint across the platform or the right advice about how to navigate the train station, can enable a person to catch the train. Once aboard, the train will take a person to a destination, even without additional choices on the person's behalf. That is, the same burst of agency or advice at a later moment will not cause the train to change tracks. It might cause people to run from the back to the front of the train or lead them to choose a different seat, but the effect of those changes are dwarfed by the much larger impact of the final destination determined by the track (also see Goyer et al., 2017).

This metaphor has three implications for how to think about the institutional-gateway mechanism. The first is that small interventions can have large and lasting effects if they are timed well—when the train is in the station and its doors are still open. Second, somewhat counterintuitively, it may not be essential to retain the immediate target of the intervention in order for someone to continue to benefit from intervention receipt. An example might be a ninth-grade program to help struggling students learn enough math to stay on track for graduation. The math skills themselves might not transfer well to higher order math skills, but passing ninth-grade algebra keeps the student moving toward graduation. The factors that helped people board the train (e.g., knowing more algebra) are not the same ones that carry them forward to their long-term destination (e.g., meeting graduation requirements).

The third implication is that institutional gateways are an important class of mechanisms for understanding heterogeneity in fade-out across or within studies. Although, conceptually speaking, institutional gateways are mediators, they can be studied as moderators, as when experiments are conducted in settings in which it was possible to board the train or not.

The present framing of institutional gateways draws on a rich intellectual history. Ross and Nisbett (1991) describe "channel factors"—institutional factors that make change easier—growing out of Kurt Lewin's (1952) pioneering research that bridged psychology and sociology. In recent years, Cohen and colleagues built on the Lewinian framework in their research on "fast-flowing" channels (which carry people forward to positive outcomes) and "slow-flowing" channels (which do not; Goyer et al., 2017). In behavioral economics, Rogers and colleagues (e.g., Frey & Rogers, 2014) called the institutional-gateways mechanism a "rip current," in reference to the fact that stepping into the ocean from one beach can pull a person to another (or out to sea) without additional effort (other than treading water). In developmental psychology, the institutional-gateway mechanism is an example of a "developmental cascade," in which the ultimate effects of previous steps in a process are thought to be fully mediated by consequences at later steps (Dodge et al., 2009; Obradović, Burt, & Masten, 2010).

Gateway to and through college examples of institutional constraints and opportunities.

Financial aid. A classic example of an institutional-gateway mechanism for persistent treatment effects comes from a recent evaluation of the effect of college financial aid on educational-attainment and labor-market outcomes (Bettinger et al., 2019). Between the years 1998 and 2000, students in California were eligible for financial aid if their high school GPAs were above an arbitrary threshold. Crucial for the analysis was that this threshold fluctuated from year to year and was not communicated to families. This created an opportunity for researchers to compare the outcomes for students whose grades were just below the threshold (and who were *not* eligible for financial aid) with students whose grades were just above the threshold (and who *were* eligible for financial aid). Assuming that very small differences in grades are essentially random and do not indicate substantial unmeasured differences in academic skills or motivation, then any long-term difference between the groups can be attributed to the availability of financial aid. This study and many of the examples of evidence for institutional gateways use regression-discontinuity designs for making causal inferences from nonexperimental data (Thistlethwaite & Campbell, 1960).

Bettinger and colleagues showed that the availability of financial aid for those just across the GPA threshold increased, by 5.6 percentage points, the rate of enrollment in more selective 4-year colleges that tend to have higher graduation rates. Once in those high-graduation-rate colleges, students who were eligible for aid were more likely to graduate with a 4-year college degree, and some subgroups showed higher earnings 14 years after high school (although some estimates were imprecise; Bettinger et al., 2019).

Studies of random-assignment interventions have produced similar effects. In a canonical example of a "nudge" intervention (Thaler & Sunstein, 2008), Bettinger and colleagues (Bettinger, Long, Oreopoulos, & Sanbonmatsu, 2012) contracted tax professionals to simplify the financial-aid-form process for students applying to college (i.e., the Free Application for Federal Student Aid [FAFSA]). This increased high school students' likelihood of successfully applying for and receiving financial aid, which in turn resulted in higher rates of full-time enrollment for the first 2 years of college (which was the full follow-up period).

These financial-aid studies illustrate the institutional-gateway mechanism in several ways. The FAFSA intervention does not teach students essential skills; all it does is help students across a threshold to be admitted to colleges from which, by virtue of their enrollment, they will be more likely to graduate. Therefore, students can continue to benefit from the intervention quite independently of whether they continue to receive assistance with financial-aid paperwork. Stated differently, one does not need to assume a "sustaining environment" of financial-aid support to explain persistent treatment effects.

College entrance exams. Students are not eligible for most 4-year colleges without college-entrance-exam (e.g., SAT and ACT) scores. These examinations, of course, measure prior skills and predict success in college, but they also come with administrative and personal costs that could keep otherwise-qualified students from making it through the college-enrollment gateway (see a discussion in Tough, 2019). Like completing financial-aid forms, taking a college entrance exam can be expensive and logistically challenging, which can pose a barrier to college access that is independent from students' prior preparation. Not only that, but students may attempt to improve their performance with test preparation, which involves learning test-taking strategies, among other things. Therefore students must find and use study-prep materials and find time to engage with them. In addition, as in the case of financial-aid forms, many of the steps students need to take to get through the entrance-exam barrier are unlikely to influence future college success. Because the tests themselves are unlikely to provide benefits later in the life course, interventions that promote test taking are good candidates for tests of the institutional-gateway mechanism.

When the state of Michigan implemented universal ACT testing in high school, it raised the rate of college enrollment and graduation among students who were generally qualified for college but were on the fence about whether they would go (i.e., those with a medium-to-low prior probability of going to college; Hyman, 2017). In a related vein, interventions that attempt to improve SAT preparation by enhancing self-regulation (e.g., mental contrasting and implementation intentions) could in principle have persistent effects if they caused students to earn high enough scores that they change their likelihood of college enrollment (Duckworth, Grant, Loew, Oettingen, & Gollwitzer, 2011).

Gateways in middle and high school: examples of institutional constraints and opportunities psychological barriers to institutional gateways. Cultural belief systems—such as the stereotypes about your group's ability or belonging, or the belief that the difficulty one feels while taking on a challenge means that one lacks talent (Murphy & Walton, 2013; Steele, 1997)—can be pervasive and impede educational progress. Yet social-psychological interventions targeting those barriers can increase a student's motivation to learn and cause greater engagement and better academic outcomes over time (Harackiewicz & Priniski, 2018; Walton & Wilson, 2018; Yeager & Walton, 2011); this can happen in part because of the institutional-gateway mechanism.

Primary and secondary effects. Sociologists note that cultural belief systems (and therefore interventions that target these belief systems) can have *primary effects* or *secondary effects* on human capital (Crosnoe & Muller, 2014). Primary effects comprise improved qualifications for a given institutional track. For instance, students who are stereotyped as less able may be less motivated to learn, and, knowing less, may perform more poorly on tests and may not be eligible for mainstream or advanced curricular tracks (Goyer et al., 2017). Because such students may not make it through key institutional gateways, they may accumulate additional qualifications at a slower rate, resulting in differences in human capital in the long run. But because the mechanism of primary effects involves accumulation of skills that will continue to be useful after crossing the institutional gateway, primary effects are not a clear example of the gateway mechanism and are perhaps better conceptualized as skill building when followed by a sustaining environment.

But cultural belief systems can also have *secondary effects*—behaviors or relationships that are consequential for moving through institutional gateways but are not necessarily indicative of individuals' underlying qualifications. These are more directly illustrative of the institutional-gateway mechanism. For example, students who are stereotyped or who have fewer family advantages might be recommended less frequently for advanced coursework or might be put in remedial tracks more frequently than other students, even with equal prior qualifications. For instance, students from Mexican American immigrant families are less likely to be recommended for advanced coursework in high school after controlling for prior ability (Crosnoe, Lopez-Gonzalez, & Muller, 2004). Secondary effects represent a clearer example of the institutional-gateway mechanism than primary effects because the skills or behaviors or resources that cause a person to cross a threshold at a given moment in time may not be the ones that they will continue to use on the other side of the gateway.

Sociologists argue that one reason for secondary effects is that institutional gateways are guarded by *gatekeepers*. Gatekeepers often have imperfect information about individuals' abilities, which means that they have to rely on peripheral information (such as stereotypes or character judgments) to assess which students get access to which resources. In some high schools, a single guidance counselor or teacher can have sole authority over which students are recommended for advanced coursework or not, and therefore who gets on a "fast flowing" track or not (see Goyer et al., 2017). In middle school, an assistant principal might have considerable latitude to decide the severity of the punishment a student receives, for instance by sending students to suspension or expelling them. A student who acts out and has been suspended at a critical time—such as the end of seventh grade—might be put on a path of extreme discipline that involves expulsion and subsequent interaction with the juvenile justice system, which has a strong negative association with a person's eventual likelihood of on-time high school graduation (Yeager, Purdie-Vaughns, Hooper, & Cohen, 2017).

Secondary effects can also occur because contexts might influence individuals differently depending on the developmental period in which a child is exposed to that context. For example, in a large RCT, receiving a voucher to move to a lower poverty neighborhood had beneficial impacts on educational attainment and earnings for individuals who moved before they were 13 years old but had little effect on individuals who moved later (Chetty, Hendren, & Katz, 2016).

Values-affirmation interventions. Psychological interventions that address the effects (or prevalence) of cultural belief systems during a key institutional transition can have persistent effects on human capital by preventing the impact of secondary effects. For instance, G. L. Cohen, Garcia, Purdie-Vaughns, Apfel, and Brzustoski (2009) used a values-affirmation intervention to help African-American seventh-grade students feel that their teachers cared about their core values—a technique hypothesized to buffer racial and ethnic minority students from some of the effects of negative group-based stereotypes. Treated students earned higher grades and were less likely to be retained in seventh grade (G. L. Cohen et al., 2009); they were then more likely to be enrolled in a selective 4-year college 6 years later (Goyer et al., 2017). Additional information about institutional gateways for these effects came from a parallel experiment with Latinx seventh-grade students who received the same values-affirmation intervention (or control). Treated minority students became less likely to be placed in remedial clinics, more likely to be sent to a mainstream (versus alternative) high school, and more likely to take advanced coursework.

The gateways identified by Goyer and colleagues (2017) may play a role in eventual on-time high school graduation as well as eventual qualification for college. However, they may do so quite apart from the initial buffering from negative stereotypes induced by the intervention. Once students are held back from promotion to the next grade or placed in an alternative high school, it will be difficult for them to get back on a "college track," even if they were later provided with a stereotype-threat-buffering intervention. In addition, it is noteworthy that in all of the cases reviewed in this section, skill building (in elementary or secondary school, college, or even in some occupation) may be a pathway through which institutional gateways affect later outcomes.

Comprehensive school-based interventions. Examples of potential institutional gateways from early childhood are interventions that reduce a student's chances of being retained in grade or placed in a special-education classroom (McCoy et al., 2017). These actions can serve as gateway mechanisms if they change a student's eventual educational attainment. However, evidence on whether this is in fact the case is mixed (Chesmore, Ou, & Reynolds, 2016; Martorell & Mariano, 2018). Another institutional pathway through which some ECE programs may operate is via their influence on the subsequent schools children attend.

Three recent studies of the medium-run effects of ECE programs provide evidence for this pathway:

1. Students who attended Tulsa's pre-K program were more likely to remain in Tulsa's public school system and have access to Tulsa's magnet middle and high schools than matched control

students (Gormley, Phillips, & Anderson, 2018; Kitchens, Gormley, & Anderson, 2020).
2. Children who won a lottery to attend a pre-K housed in a Boston elementary school with higher Grade 3 test scores scored higher on tests in Grade 3 than children who lost the lottery (Unterman & Weiland, 2019).
3. Children in the Chicago School Readiness Project, a program that attempted to improve the quality of Head Start classrooms in low-income neighborhoods, raised the likelihood that students would attend higher performing elementary schools (Watts et al., 2020).

Furthermore, comprehensive school-based interventions that focus on social-emotional factors rather than academics can boost long-term academic outcomes—such as completing high school or college—via the institutional-gateway mechanism. For instance, teen-volunteering interventions that prevent teenage pregnancy might also prevent high school dropout and affect long-term educational attainment via continued engagement with the institution of high school, even if there is no continued encouragement to volunteer in the community (Allen, Philliber, Herrling, & Kuperminc, 1997). Likewise, interventions that temporarily prevent the first onset of major depressive disorder in high school or college might prevent school dropout and therefore benefit long-term labor market outcomes, even if individuals still go on to experience major depressive disorder once they are on the other side of the institutional gateway (see Thapar, Collishaw, Pine, & Thapar, 2012). Even for interventions designed to target children's achievement-test scores, institutional gateways may serve as important mediators for long-term impacts. For one program that placed children in late elementary school into an accelerated curriculum, long-term impacts on attainment were no larger at sites that raised treated children's achievement-test scores than at sites that did not; rather, long-term impacts were larger at the sites for which treated children were more likely to stay on track in meeting institutional milestones, such as lack of retention and taking the SAT (Cohodes, 2020).

Implications of institutional-gateway mechanisms. The gateway mechanism is optimistic because it means that well-timed and psychologically informed treatments can generate sustained effects (we describe some hypothesized examples of time-sensitive institutional gateways in Box 3). Moreover, it suggests that fruitful educational interventions are possible long after whatever sensitive periods early childhood may hold. Taken as a whole, psychological interventions may offer a cost-effective and scalable means for addressing the underdevelopment of human capital (Benartzi et al., 2017; Yeager, Hanselman, Walton, et al., 2019).

But the institutional-gateway mechanism is also a source of pessimism because it reveals something profound about the ways in which institutional interactions structure and reinforce access to resources. If random or nearly random structural advantages can substantially influence socially important outcomes for vulnerable populations, this means that, absent the intervention, there was potential for individuals in those populations to have better outcomes than those observed. That is, a "promising intervention effect" can be seen as an indictment of systemic structures that could keep willing and skilled people out of the higher tiers of educational progress. This implies that perhaps ideal long-term solutions are substantially different from short-term solutions. Stated differently, the conclusion that society's structures are optimally suited for interventions just because we can get interventions to work on the margins may be misguided.

A concern related to relying on institutional gateways for persistence is the possibility that these gateways, arbitrary as they may be, serve some zero-sum or even positive-sum sorting function. In such cases, eliminating these gateways could lead to their quick reemergence or to negative unintended consequences. An example from education comes from California's movement to enroll all eighth-grade students into algebra. The well-meaning policy reflects the idea that students who take advanced coursework will fare better relative to their similarly achieving peers who do not have the opportunity to enroll in an advanced course (Domina, 2014). However, placing all children into algebra, although it removed a somewhat arbitrary institutional gateway, also changed the nature of the "treatment" at scale. When this policy was implemented, children who took algebra had different kinds of peers and coursework, on average, than when only more advanced students were enrolled in algebra. Penner, Domina, Penner, and Conley (2015) estimated that the policy negatively influenced the achievement of higher achieving students and did not improve the achievement of lower achieving students, labeling this a "collective effects problem." They encourage policymakers to consider not just the individual-level effects but also the effects for the population of individuals affected by a policy change. Policies such as selective enrollment, tracking, and various school-accountability systems may also have collective effects that are not the same as the average treatment effect for individuals most affected by these policy changes, although, of course, this does not mean that these are zero- or negative-sum policies. This is particularly relevant for policies intended to change institutional gateways.

Box 3. Hypothesized Examples of Time-Sensitive Institutional Gateways.

When interventions influence outcomes primarily via institutional gateways, rather than human capital immediately influenced by the intervention, local structural factors may moderate the link between the intervention and long-run impacts. Thus, the institutional-gateway mechanism has direct implications for understanding heterogeneity of the effects of an intervention within a particular evaluation and about kinds of interventions that will or will not produce reliably replicable effects.

Deworming interventions and schooling. An intriguing example comes from experiments in development economics. In 1998, more than 32,000 primary school students in Kenya were randomly assigned to a universal "deworming" treatment, which reduces infections of intestinal worms that are spread through soil and water. Ten years later, boys were more likely to have higher paying jobs than boys who were not treated and girls were more likely than girls who were not treated to have attended secondary school (Baird, Hicks, Kremer, & Miguel, 2016). How was this possible? At first glance, intestinal worms seem to have little to do with educational attainment; after all, deworming medications do not teach math or science or literature. The institutional-gateway mechanism could play an important role. Children infected with intestinal worms become lethargic and anemic and have weaker immune systems; being tired and sick, they have a hard time attending school. Deworming therefore increases children's ability to cross the threshold of health that allows them to attend school and learn. Once children learn more, they are more competitive applicants for the next transition in life: high-status jobs or competitive entrance examinations for secondary school.

Deworming interventions are a useful example because deworming is not directly an educational intervention. Rather, it is an intervention that allows children to benefit from educational institutions' resources, quite independently from the initial intervention effects (the reduced prevalence of hookworms). Although it is possible that some of the educational benefits from deworming come from cognitive development outside of school settings, evidence for the institutional-gateway mechanism comes in part from a moderation analysis that assessed the time sensitivity of the intervention and found that the deworming treatments had no effects on educational outcomes when they were delivered to older adolescents who had already passed the age at which health could plausibly affect their likelihood of enrollment in secondary school (Croke & Atun, 2019).

Double-dose math-course taking and educational attainment. Another type of intervention that may influence long-run outcomes primarily via institutional gateways is one that requires students to take two periods of math courses instead of one. In one instance when the intervention was given to high school freshmen, it yielded impressive effects on a variety of outcomes, including test scores and educational attainment (Cortes & Goodman, 2014); in another instance, when the intervention was given to sixth graders, impacts on test scores faded out by high school, and no impacts were found on educational attainment (E. Taylor, 2014). What might account for the difference in long-run impacts? One possibility is skill building: Perhaps algebra is more fundamental to later math learning than sixth-grade math. Another possibility is that failing a high school algebra class is a more consequential institutional gateway for later difficulty than failing a sixth-grade math course. Bailey and colleagues (2017) argued that the institutional-gateway mechanism was a likely candidate for the positive high school impacts reported by Cortes and Goodman because the intervention raised a variety of outcomes—including a larger estimated impact on ACT verbal scores than on ACT math scores. This finding is difficult to reconcile with the hypothesis that algebra, per se, was the sole active ingredient in the intervention's success.

Mechanisms for Emergence of Impacts Following Fade-Out

In the case of ECE programs, consider the patterns of impacts generated by the two most famous early-childhood interventions—the Perry Preschool Program and the Abecedarian Project. Both provided classroom-based educational curricula for children as well as other services (although these services differed between the two programs). In both cases, treatment/control differences extended well into adulthood and generated economic benefits far in excess of program costs (F. Campbell et al., 2014; García, Heckman, Leaf, & Prados, 2016; Heckman, Moon, Pinto, Savelyev, & Yavitz, 2010; Schweinhart et al., 2005). However:

- *Patterns of long-run impacts differed.* Although both programs appeared to boost adult earnings, treatment/control differences in criminal behavior were large enough to dominate the dollar value of benefits generated by the Perry program. Treatment- and control-group children in the Abecedarian Project did not differ in terms of their criminal behavior.
- *IQ impact trajectories differed.* Although the achievement-test scores of children attending both the Perry and Abecedarian programs were persistently higher than comparison-group children across childhood and adolescence, IQ-score impacts in Perry had disappeared by age 8, whereas IQ differences in Abecedarian persisted throughout childhood and adolescence.
- *Impacts on other hypothesized mediators differed.* Children attending Perry persistently scored higher on an assortment of socioemotional behaviors, whereas children attending Abecedarian, if anything, exhibited worse socioemotional behaviors, especially in the early grades. Moreover, Abecedarian children were considerably less likely to be placed in special-education classes or to be retained in grade across their years of school. These patterns are consistent with more positive early trajectories that may have contributed to Abecedarian's long-run successes. In contrast, these kinds of mediational impacts were not apparent for children attending Perry Preschool.
- *Mediational processes are not well understood.* Mediation analyses in both studies on selected long-term outcome measures are often able to explain less than half of the total effect (Elango et al., 2015), suggesting that some mediators may not be well-measured in these studies and may include poorly measured institutional gateways, socioemotional skills, and improved health (Abenavoli, 2019; Bailey et al., 2017; Gibbs, Ludwig, & Miller, 2011; Heckman & Karapakula, 2019; Ludwig & Miller, 2007). In these and other studies of long-term impacts, forecasts of long-run labor-market outcomes based on medium-run achievement impacts are underestimated relative to forecasts based on short-run achievement impacts (Bartik, 2014; Chetty et al., 2011). It is unclear, however, whether early achievement is a causal mediator in these cases or effects on other, unmeasured mediators persist. However, three sources of evidence support the latter explanation. First, as described above, some evidence suggests that therapeutic interventions that target skills other than academic achievement appear to produce more persistent impacts than interventions that focus on cognitive skills. Second, later interventions that have changed children's test scores sometimes do not change their long-term educational or labor-market outcomes (Dobbie & Fryer, 2016; E. Taylor, 2014). Third, new evidence from studies of the effects of teachers suggest that impacts on students' socioemotional skills and behaviors, although difficult to measure, may be more consequential for persistent effects of teachers on student outcomes (Jackson, 2018; Kraft, 2019; Liu & Loeb, 2019).

The overarching lesson from Perry and Abecedarian is that both are *existence proofs* of long-term impacts from ECE interventions, although how these long-run impacts came about is not clearly understood and may have differed between the two programs.

Strong quasiexperimental follow-up studies suggest that larger programs, such as Head Start, have also influenced long-term educational and labor-market outcomes (Deming, 2009; Thompson, 2018). But differences in impacts on possible mediational processes and outcome domains in adulthood preclude consensus conclusions about why the programs were successful. Even more troublesome are the policy-related issues of external validity—would we expect that a Perry- or Abecedarian-type program begun today would generate the same kind of long-term benefits as the originals?

Perry and Abecedarian were designed and run by researchers, whereas most of the ECE programs that are the focus of today's evaluations are being operated at scale in states, school districts, or communities. Moreover, the counterfactual conditions facing children *not* assigned to these programs have improved dramatically since the 1960s and 1970s: Low-income mothers finish several more years of formal schooling, family sizes have fallen dramatically, and the availability of alternative forms of center-based care has increased substantially (Duncan & Magnuson, 2013). This raises the bar for demonstrating effectiveness in the case of modern-day preschool programs.

Although the children in recent rigorous evaluations of the Head Start and state pre-K programs have not yet reached adulthood, impacts on middle-term mediators are not encouraging. The Head Start Impact Study randomly assigned a national sample of children on Head Start waiting lists in 2002 either to attend Head Start classrooms or not (Puma et al., 2012). For 4-year olds entering Head Start for the first time, end-of-treatment impacts were null for both math and behavioral outcomes, but on six of the eight language and literacy measures, impacts were positive and statistically significant at $p < .10$ and had an average effect size of 0.18

SD across all eight measures. Follow-ups in elementary school were uniformly disappointing. Math and behavioral differences failed to emerge, and differences in only one of nine literacy measures attained $p < .10$ at the end of first grade. Absent large impacts on other kinds of mediators, it is hard to imagine that a follow-up conducted in adolescence or adulthood would show notable group differences. On the other hand, given the uncertainty around the processes mediating the effects of ECE programs on adult outcomes and the large range of plausible mediators discussed above, perhaps a substantial policy change before obtaining long-term outcome data might well do more harm than good (Gibbs et al., 2011).

Implications for Research

The complicated and at times contradictory nature of theory and evidence on short- and long-term patterns of intervention impacts calls out for more theorizing and empirical research. If nothing else, empirical researchers in the habit of gathering outcome data only at the conclusion of their interventions should be sobered by the evidence indicating that partial or complete fade-out is the norm rather than the exception. So should editors and reviewers who judge the contributions of such studies, as should practitioners and policymakers who seek interventions that alter developmental trajectories. In this section, we outline some productive ways in which research efforts might be directed. All of the implications for research that we discuss are rooted in the idea that optimally designed interventions (or sequences of interventions) require a strong developmental theory linking the treatment to the outcome.[5]

Include longer follow-up periods

Our review of meta-analyses found that only a small fraction of intervention studies assessed impacts beyond the end of the interventions themselves. Given how pervasive fade-out appears to be, particularly in skill-targeted interventions, studies without follow-ups teach us very little about the nature of their impacts. An obvious implication is to set funding and publication standards in ways that value follow-ups.

But how to allocate scarce funding and researcher energy between the tasks of inventing new interventions and conducting follow-up studies on existing ones? Education interventions teach us the hazards of choosing follow-ups according to the size of early impacts. IQ impacts had disappeared by age 8 in the Perry Preschool intervention, and we would never have known about the impressive collection of treatment/control group differences that emerged in adulthood had the IQ fade-out led to the study's termination. Further, the extreme focus on short-term impacts for educational interventions may incentivize practices (e.g., overalignment and publication bias) that inflate such impacts but are unlikely to translate into long-term benefits.

Watts, Bailey, and Li (2019) offer several ideas about how to improve the incentives for testing for long-term impacts (and, ideally, generating them). Primarily, they argue for the importance of requiring researchers applying for funding for intervention RCTs to specify in the proposal whether they expect any long-term impacts and why. The currently common practice is to cite the small experimental literature on long-term impacts or the larger correlational literature with only superficial attention to the mechanisms through which the proposed intervention would improve children's long-term outcomes. This practice allows researchers to imply that their work might improve children's long-term outcomes without making these assertions a key component of the funding discussion. Watts and colleagues propose that funders could promote longer follow-up periods in several ways, but they favor a mechanism that would randomly select among funded studies that meet requirements for the quality of implementation, randomization, and lack of attrition after the end of treatment. Selecting at least some studies for follow-up on the basis of their methodological strength, rather than the size of short-term impacts, has several benefits, including limiting additional incentives for methodological decisions that inflate short-term impacts and allowing for stronger tests of theories of persistence by creating a larger pool of follow-up impacts for meta-analysis and more cross-study variation in short-term impacts.

However, perhaps the most useful consequences of such a policy would be to increase the level of attention given by interventionists and reviewers to the mechanisms through which interventions might produce long-term impacts. This funding policy, for example, might incentivize more thinking about interventions that *complement*, rather than *substitute*, for children's subsequent educational experiences, as well as incentivize collaborative projects that might align children's educational experiences over a multiyear period (Stipek, Franke, Clements, Farran, & Coburn, 2017).

Testing for and reducing overalignment

Here we consider two issues that may lead to upwardly biased estimates of short-term program impacts. The first relates to the fact that assessments may not

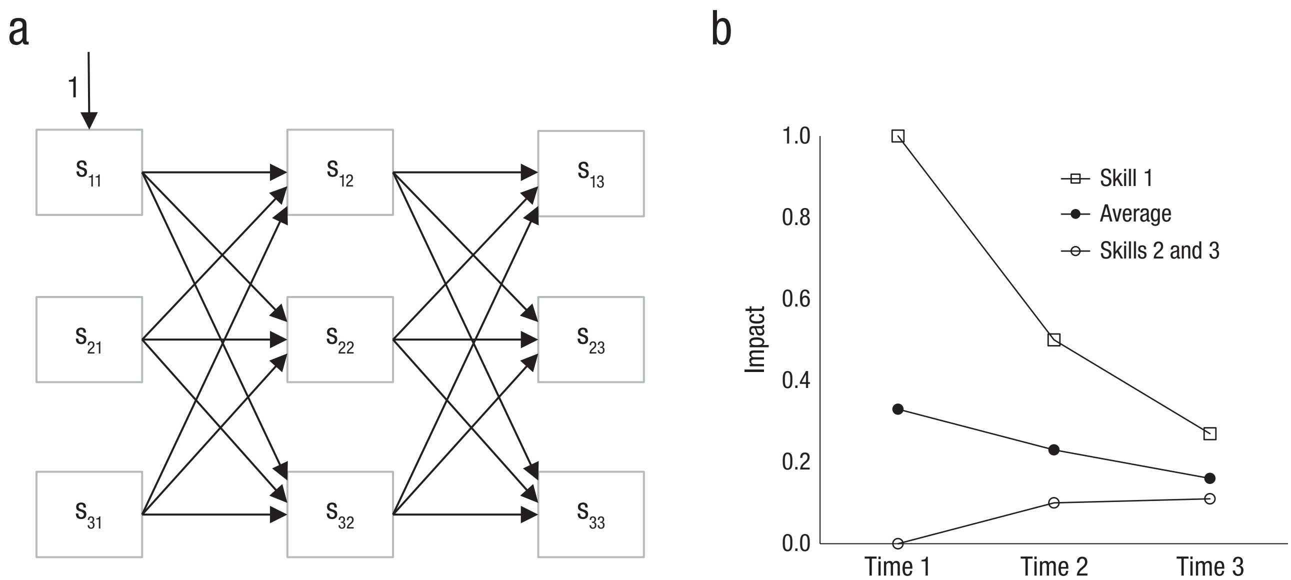

Fig. 3. Hypothetical skill-building model and projected impacts from hypothetical intervention. The diagram in (a) shows how a hypothetical intervention influences s_{11} by 1 unit. s_{ij} = skill *i* at year *j*. The graph in (b) shows projected impacts on Skill 1 and Skills 2 and 3 at each wave, assuming additive effects, no measurement error, all auto-regressive paths = 0.5, and all cross-lagged paths = 0.1. The average impact at year *j* across Skills 1 to 3 is shown by the line labeled "average."

measure fundamental skills. One way of testing for and mitigating this problem is to anchor assessments to long-term outcomes so that the estimation of short-term impacts is less influenced by how an intervention affects nonfundamental skills that receive high weights in the construction of unanchored developmental scales.

The second issue arises when the intervention is designed to close skill gaps measured in a particular instrument of child development. This problem is similar to the teaching-to-the-test phenomenon encountered in schools. Using a narrowly tailored set of outcome measures can yield a biased estimate of the magnitude of an intervention's effect on developmental outcomes more broadly. Consider a hypothetical case in which three broad skill groups influence each other across time (Fig. 3). After a hypothetical intervention that influences Skill 1 in Year 1 but not Skills 2 and 3, a measure of Skill 1 will yield an incomplete and optimistic measure of a child's overall skill level, whereas measures of Skills 2 and 3 will yield incomplete and pessimistic measures of a child's overall skill level. Across time, these impacts will converge, such that the impact on Skill 1 will look less optimistic and the impacts on Skills 2 and 3 will become less pessimistic. Including a broad set of measures at the end of treatment would yield a more nuanced view of the impact of the intervention and allow for a better forecast of its long-term effects.

Both of these approaches suggest that interventions should address the gaps in the environments that foster fundamental and malleable skills rather than the gaps in skills measured by developed scales. Both also suggest that a better understanding of the processes underlying skill development will yield insights into the most efficient levels at which to intervene. Two additional technical solutions, *measurement-invariance testing* (e.g., Wicherts, 2016) and *anchoring* (e.g., Bond & Lang, 2018; Cunha, Heckman, & Schennach, 2010), may also help diagnose and fix problems related to differences in the kinds of skills reflected on tests within the control group and those caused by specific interventions.

Building sustaining environments into evaluation studies: designing research to study Intervention × Context interactions

At least since Cronbach's (1957) APA presidential address, psychologists have known that any intervention intended to alter trajectories of human development will depend on the context in which it is administered, and therefore variability in contexts should be studied directly. If nothing else, our review of fade-out across different kinds of interventions highlights the potential importance of environmental conditions after an intervention has ended. How could we design research studies from the ground up to produce replicable and generalizable insights about theory-informed Intervention × Context interactions? There is a large and growing body of evidence in the field of educational evaluation and statistics that provides guidance (see a review in Tipton, Yeager, Iachan, & Schneider, 2019; also see Stuart, Bell, Ebnesajjad, Olsen, & Orr, 2017; Tipton & Hedges, 2017; Yeager, Hanselman, Walton, et al., 2019).

A first step is to specify plausible contextual gateways in advance. If the study is conducted around only one gateway (e.g., proximity to school transitions such as the transition to high school or college, in the case of research on some social cognitive interventions hypothesized to operate via institutional gateways), the authors should explain this in the Method section or a preregistration plan so that others replicating the study can understand the theory. And even in gateway periods, there may be contexts in which there are more open or less open gateways. Therefore, there should be a plan for measurement of those gateways. Successful measurement and conceptualization of context mechanisms may require psychologists to collaborate with researchers with different perspectives on contextual influences (e.g., sociologists and public policy researchers) and practitioners with knowledge of the local context.

The second step is to construct a sample in which the context mechanism could plausibly be detected. This is hard to do because it means sampling sites in which the context is closed and open, with sufficient power to detect an interaction effect. The latter is challenging because researchers commonly handpick sites for which they expect strong main effects, and it is exceedingly rare to create samples that include many sites for which one expects weak or null effects. However, understanding contextual mechanisms requires varying contextual factors, so inclusion of sufficient sites with weaker effects is essential. One way to do so is to draw a random sample of sites (e.g., Puma et al., 2012; Yeager, Hanselman, Walton, et al., 2019).

Moderation analysis also plays an important role in tests of sustaining environments (Abenavoli, 2019; Bailey et al., 2017, 2020a). Whether children with a TRIAD-induced boost of math skills or a Head Start-induced boost of literacy skills in their pre-K classrooms have more positive achievement trajectories presumably depends, in part, on the structure and content of instruction they receive once they enter kindergarten. An obvious research implication in this case is to design two-stage interventions with random assignment to preschool conditions, followed by a second-stage random assignment into early-grade conditions thought best for sustaining impacts.

Because statistical interactions are imprecisely estimated relative to main effects, and a reasonable prior is that sustaining-environments interactions will be smaller than their corresponding main effects (Bailey et al., 2020a), it is likely to be most productive to concentrate sustaining-environments treatment in cases in which initial interventions generate substantial end-of-treatment impacts. In addition, despite their higher costs, four-way assignment to initial and subsequent treatments are more useful than an assignment scheme in which only the first-stage intervention group receives a second-stage intervention. This is because first-stage control-group children may benefit just as much from a second-stage enrichment as first-stage treatment-group children, but we will never know that in the absence of four-way random assignment.

A final area of need for future research is to consider context as a dynamic factor that can be influenced by interventions. Returning to the example of interventions that move large numbers of disadvantaged students into more advanced math classes, how does this change the classroom dynamic? How does this affect decision-makers' allocation of resources to the higher level math courses—especially if decision-makers harbor negative stereotypes about disadvantaged students?

Further complicating matters is the possibility that marginal effects of systemic opportunities offered by an intervention may not generalize when they are afforded to everyone. Such was the result when research on the marginal benefit of taking Algebra I by eighth grade was scaled-up by creating universal access to Algebra I for all students in the state, as described above. In contrast, systemic advantages may accrue to untreated individuals when some programs are implemented at scale (for review, see Greenberg & Abenavoli, 2017). For example, substantial benefits to untreated peers (Dodge, Bai, Ladd, & Muschkin, 2017; List, Momeni, & Zenou, 2019) or siblings (Heckman & Karapakula, 2019) have been estimated in studies that have identified random or quasirandom variation in the implementation of preschool programs. Note that precise estimates of peer spillover effects require studying higher level units (e.g., classrooms, schools, or districts, rather than children). In RCTs, this means using much larger samples. In quasiexperimental designs, it introduces the possibility of a greater number of potential confounds. But given the importance of contextual constraints and opportunities in fade-out and persistence, we think conducting RCTs with higher levels of randomization may yield important insights into the optimal conditions for generating persistent benefits.

Currently, researchers know very little about how our conclusions about Intervention × Context interactions are sensitive to general equilibrium effects. Learning more about these possibilities might allow intervention designers to think more carefully about the conditions under which early demonstration studies might be scaled up effectively to be used as universal treatments at a population level. Although conceptualizing and studying the context with an assumption that it is dynamic will pose a major challenge for research design, it has the potential to be fruitful for theory and policy.

Develop models to forecast long-term impacts

Current methods for identifying promising intervention targets may not be smoothly aligned with the goal of promoting persistent effects of educational interventions. In developmental psychology, partial correlations between children's early skills or environments and their much later educational or labor market outcomes are frequently cited as evidence for the importance of intervention on these early targets. However, validation of the causal informativeness of these partial correlations is rare. This is understandable, both because these correlations are often presented with ambiguously causal interpretations (thus making it difficult to understand what kind of causal estimates, if any, to which they could be compared) and because such experiments would require a rare combination of a somewhat narrowly focused randomly assigned intervention and a long-term follow-up assessment.

In the case of children's mathematics achievement, Bailey and colleagues (2018) present evidence that influential correlational estimates of the effects of children's early math skills on their achievement several years later are upwardly biased. They argue for developing models that can accurately forecast the long-term impacts of interventions, conditional on the short-term impacts. Doing so would require (a) interpreting parameter estimates in longitudinal models as causal estimates and (b) checking them repeatedly against causally informative information. Models that can reliably forecast long-term intervention impacts conditional on short-term impacts would be very useful in identifying the most promising short-term targets to aim for.

Different substantive explanations of fade-out make very different predictions about what kinds of interventions will lead to the most persistent effects. If forgetting is a major contributor to fade-out across a variety of interventions, this has important implications for what kinds of skill-building interventions will produce more persistent effects. For example, factors found to promote retention of learned information, such as spaced practice, frequent testing, and interleaving different kinds of problems into instructional materials (Rohrer & Pashler, 2010), may promote persistence. On the other hand, especially for early interventions, if children are likely to continue practicing the information they learned during the intervention in their regular educational contexts, intervention time may be better spent teaching different kinds of skills (e.g., skills that are advanced or otherwise likely to be ignored during subsequent instruction; Paris, 2005). Because these predictions are different and may even conflict in many instances, this is an important area for future research on fade-out and persistence.

Thus far, somewhat informal attempts at forecasting have been made, for example by multiplying treatment impacts on end of treatment or intermediate outcomes by regression-estimated effects of these intermediate outcomes on labor-market outcomes (e.g., Deming, 2009; Kline & Walters, 2016) or within a meta-analytic framework, using end-of-treatment impacts on different kinds of outcomes to predict long-term impacts across studies (R. D. Taylor et al., 2017). A group of economists has started a project attempting to develop a database of "surrogate indices" that can be used to forecast the long-term impacts of a wide variety of interventions, conditional on a theory-informed index of short-term impacts and has shown some success in doing so with a job-training intervention (Athey, Chetty, Imbens, & Kang, 2019). Using a model designed to account for stable confounds during development, Bailey and colleagues (2018) showed some success at forecasting the impacts of a math intervention on math achievement several years into the future.

Implications for Policy

Many of the interventions featured in this review were designed to inform policies that would enhance the development of children and youth. How should we think about their successes in doing so? Many skill-building interventions appeared to generate substantial improvements in the short run but evaluations showed that only a handful have provided evidence of long-term impacts. Should long-term success be required before policy implications are considered? Impacts on adult well-being from interventions such as the Abecedarian Project were impressive, but so too was its cost—more than $80,000 per child in today's dollars. Were the benefits worth the cost? Impacts from many social-psychological interventions on socially important outcomes are likely to be much smaller, but at low per-child costs (often less than $100), do they constitute wise social-policy investments?

In this final section we offer some thoughts on the answers to these questions. As might be expected from the disciplinary orientation of two of the authors of this article, we argue that the policy value of intervention evaluations is enhanced by providing even very rough estimates of their likely costs and benefits (Duncan & Magnuson, 2007). The logic follows that of business executives who want to know how an investment would affect their company's bottom line: Policymakers should find it useful to ask not only whether government expenditures have the intended effects but also whether investing in child and adolescent programs provides "profits" to the children themselves, to taxpayers, and to society as a whole (Gramlich, 1990; Levin, 1988). Below we note a number of important considerations in thinking about the benefits and costs of interventions.

Standardized effect sizes are a poor guide to good policy

J. Cohen's (1988) widely cited guidelines categorize effect sizes as large (≥ 0.80 *SD* change in the dependent variable), medium (0.30–0.80 *SD*), or small (< 0.30 *SD*). Although it is tempting to infer that interventions with "large" effect sizes make for better policy than interventions with "small" effects, that is not the case. Although effect sizes can help standardize impact estimates and ensure that statistical significance will not be the sole arbiter of meaningful effects, they provide incomplete and at times misleading guidance to policymakers. A cost–benefit approach is more useful because evidence-based policy decisions must compare the value of a program's effects with the costs incurred in achieving them. An inexpensive program that produces small but economically valuable outcomes may make for good policy (Greenberg & Abenavoli, 2017), whereas a very expensive program that produces larger—but not proportionately larger—effects may not (for an updated discussion of effect sizes, costs, and benefits in the context of educational interventions, see Kraft, 2020).

Not all mediators are created equal

Society invests a lot of money in its public schools—in the United States, it amounted to about $11,763 per pupil in 2016 (U.S. Census Bureau, 2016).

Children retained in grade before graduation spend an additional year in a school system and, all other things being equal, cost that system thousands of extra dollars per occurrence. Likewise, special-education classes, with their high staff-to-student ratios, can cost a school system and its tax payers more than twice as much as classes for conventional students. Accordingly, interventions that can reduce costly short-run outcomes, such as placement in special-education classes or grade retention, can provide important social benefits even if long-term follow-ups show no differences between children who did and did not take part in the intervention program.

That indeed seems to be the case for at least some early education programs. McCoy et al.'s (2017) meta-analysis of 22 experimental and quasiexperimental studies conducted between 1960 and 2016 found that participation in ECE leads to statistically significant reductions in special education placement ($d = 0.33$ *SD*, 8.1 percentage points) and retention in grade ($d = 0.26$ *SD*, 8.3 percentage points).

Another use of mediators in the absence of long-term outcomes is to use them to forecast what those long-term outcomes might be. For example, because we have a number of fairly strong studies estimating the costs of long-term consequences of dropping out of high school, intervention impacts on dropout can be valued by these estimates. Levin, Belfield, Muennig, and Rouse (2007) estimate that each new high school graduate yields a public benefit of $209,000 in higher government revenues and $82,000 lower government spending on public health, crime and justice, and welfare.

However, these forecasts may be unreliable and even predictably biased under some circumstances. For example, as noted above, across studies of ECE interventions, test-score impacts during the middle of elementary school appear to underforecast impacts on adult-labor-market outcomes, perhaps indicating a prominent role of noncognitive and/or structural factors in mediating these impacts. In contrast, a program in Louisiana that provided private-school vouchers to disadvantaged students showed large negative impacts on student test scores (Abdulkadiroğlu, Pathak, & Walters, 2018), whereas an analysis of a different cohort of treated students indicated a small, nonsignificant positive impact on college enrollment (Holmes Erickson, Mills, & Wolf, 2019), perhaps indicating some combination of beneficial unmeasured mediators of private-school attendance and test overalignment within the public schools. In both kinds of cases, a better understanding of the processes underlying fade-out and persistence may improve our ability to project costs and benefits for a particular type of intervention.

A point we have encountered in discussing fade-out and persistence with interventionists is the idea that interventions' benefits may exceed their costs even when impacts fade out completely within the few years after the end of the intervention. A frequent analogy is to medical interventions that alleviate painful symptoms, which can be worthwhile in many instances. Calculating the benefits of temporarily effective educational interventions requires a judgment of the value of a short-term advantage on educational skills, and we think that in some cases, such effects can be nontrivial (e.g., if improving students' academic skills for a limited period of time relieves stress on the child and family experience for that limited period). However, we do not see this as a strong argument for funding any intervention at scale that finds short-term positive impacts on participant satisfaction, noting that an alternative intervention that produces long-term impacts on socially important outcomes may produce similar or larger short-term benefits as well.

Rough estimates of costs

Essential for policy discussion is an order-of-magnitude estimate of the likely costs of recommended policy changes. Are costs per child or family likely to amount to $100, $1,000, or $10,000? If the policy needs to run for several years to produce its effects, then per-year costs should be multiplied accordingly. Does the recommended policy provide one-on-one or group services? What level of professional training is required of the service providers? Detailed cost estimates are best, but the policy relevance of research findings can be greatly enhanced even with crude information about likely program costs.

Direct service program costs are often dominated by the salary costs of staff, so the single most valuable piece of cost information is the number of hours of professional time spent per child or per family. Again, order-of-magnitude estimates are useful: Does the number of hours of professional time per participating child or family amount to 10, 100, or 1,000?

Kraft (2020) offers a set of policy-relevant benchmarks for evaluating effect sizes (less than 0.05 *SD* on an achievement test is small, 0.05 to 0.20 is medium, and 0.20 or above is large), costs (less than $500 per pupil is low, $500 to $4,000 is moderate, and $4,000 or above is high), and scalability (a qualitative judgment of whether an intervention would be easy to scale, reasonable to scale, or hard to scale). We agree with Kraft's warning that such benchmarks should be used in conjunction with contextual information, but we find them generally reasonable. On the basis of the present review, we would add that such effects on achievement tests are far more impressive because they are measured years after the end of treatment (see A Formal Model of Skill Building above) and because the intervention does not principally target the knowledge assessed by the standardized tests (see the Testing for and Reducing Overalignment section and Fig. 3).

Conclusion

We offer several general conclusions. First, fade-out is widespread and often coexists with persistence, such that even as impacts diminish after the end of treatment, impacts on other outcomes—and even the same outcome—may be consistently positive. Second, fade-out is a substantive phenomenon, not merely a measurement artifact. And third, persistence may depend on the types of skills targeted, the institutional constraints and opportunities within the social context, and complementarities between interventions and subsequent environmental affordances. In this article, we present a formal model of skill building that might be used to make predictions about persistence and fade-out. However, data remain an important limiting factor: Although persistence is an explicit goal of many psychological interventions, long-term follow-up assessments are rare. We included considerations for research design and analysis that can be used to further knowledge within this important area. All of these conclusions are tentative, and we hope that tests of our falsifiable predictions will help the field make important progress in improving theories of human development, along with educational practice and policy.

Appendix

A formal skill-building model

In this appendix, we formalize ideas presented in the body of the chapter on the nature of predictions from skill-building models. In particular, these models provide a complementary way of understanding such concepts as malleability, fundamentality, and counterfactual conditions. Remarkably, these models suggest that skill fade-out following the end of interventions is inevitable under most reasonable conditions but also that short-term fade-out does not preclude the emergence of long-term impacts.

To keep things simple, we focus on the process of developing a single family of skills. An example might be a family of math skills: counting knowledge, addition and subtraction, multiplication and division, exponents and logarithms, real analysis, functional analysis, etc. Each one of these skills can be understood to represent a distinct member of the math-skill family. Thus, in principle, we could represent math skills as a vector (and, more generally, human capital as a vector of vectors).

Individual differences in math skills arise because some individuals acquire more elements of this vector than others (e.g., some individuals know functional analysis and others do not), and also because of proficiency differences in a given skill (e.g., among individuals who know functional analysis, some know it better than others). Although generalizing the model to include multiple families of skills (e.g., executive functioning skills, socioemotional skills) is straightforward, we do not consider models featuring multiple families of skills because their additional insights are not fundamental for the purpose of this appendix: to understand the conditions under which skill-building processes do and do not produce persistent impacts if skills are augmented by early-life interventions.

A skill-building model. Let θ_t denote the skills a child has acquired by the time he or she is t years old and x_t denote the amount of additional investment in the acquisition of skill θ_t that takes place during period t. Skill-building models such as those developed by Cunha and Heckman (2007) draw from economic models of production and focus on the key concepts of *self-productivity* and *dynamic complementarity or substitutability* by using the following functional form:

$$\theta_t = \left[\gamma_t(\theta_{t-1})^{\phi_t} + \rho_t(x_t)^{\phi_t}\right]^{\frac{1}{\phi_t}},$$

where γ_t, $\rho_t \geq 0$ and $\phi_t \leq 1$.[6] Because it relates current (period t) to past (period $t-1$) levels of θ, the parameter γ_t characterizes the degree of self-productivity γ of skills. The higher the value of γ_t, the higher the self-productivity of θ_t. In the extreme, if $\gamma_t = 1$ and $\rho_t = 0$, then the development of θ_t is automatic and requires no investment (that is, $\theta_t = \theta_{t-1}$).

To use a math example, if counting knowledge led automatically (i.e., without teaching) to an understanding of addition and subtraction, this would be a situation in which $\gamma_t = 1$ and $\rho_t = 0$. The parameter ρ_t captures the degree of malleability of skill θ_t with respect to investment x_t. For example, if $\rho_t = 1$ and $\gamma_t = 0$, then the skill θ_t is fully malleable with respect to investment x_t. Again, drawing from a math example, if counting knowledge depended only on teaching the child how to count, regardless of the child's prior math knowledge contained in θ_{t-1}, this would be a situation in which $\rho_t = 1$ and $\gamma_t = 0$.

This model formulation is also convenient because it allows us to characterize the degree of synergy between a student's past skill level and current math instruction with the coefficient ϕ_t. Economists term conditions with negative values of ϕ_t as having *dynamic complementarity* (i.e., highest skill students learn the most from x_t) and conditions with positive values of ϕ_t as having *dynamic substitutability* (i.e., the lowest skill students learn the most).[7]

Linking child and adolescent skills to adult well-being. A noteworthy feature of our formal skill-building model is that individuals are assumed to form one kind of math skill at each point in time. So, for example, θ_1 might describe how well the individual counts, θ_2 depicts the individual's mastery of addition and subtraction skills, and so on. By the end of period t, math skills can be described by the vector $\Theta^t = (\theta_1, \ldots, \theta_t)$.

Assume that a child becomes an adult person at period $T + 1$ and let Θ^T denote the vector of all skills that the child has developed across all periods before adulthood (i.e., $\Theta^T = (\theta_1, \ldots, \theta_T)$). The next step in skill-building models is to map the vector Θ^T to adult human capital. Let H denote the stock of human capital once the child becomes an adult. In skill-building models, H determines educational attainment, labor-force attachment, labor earnings, and other positive life outcomes. We assume that H is defined by an individual's accumulated skills according to the formula:

$$H = g(\Theta^T)$$

This definition of adult human capital is general and is consistent with a number of possible skill-building processes. For example, it may be the case that adult human capital H depends only on the skills (e.g., functional analysis) the child developed in the last period, θ_T. Alternatively, and more realistically, it is also consistent with adult human capital as a function of many (and possibly all) skills developed across childhood and adolescence.

Malleability and fundamentality. The model is useful in thinking about the Bailey et al. (2017) idea of malleability. Theirs is a practical rather than conceptual definition—a skill is malleable if it can be affected by a range of available interventions under realistic counterfactual conditions. In terms of the above skill-building model, this means that skills are enhanced by investment x_t. In the context of skill-building models, malleable skills are skills for which $\rho_t > 0$. The higher the level of ρ_t, the higher the malleability of θ_t.

Bailey et al. (2017) introduce the concept of *fundamental skills* as

> those upon which later skills are built, and that influence positive life outcomes, such as attainment or labor market success. These long-term impacts may be direct effects of persistent skill gains (e.g., trained vocational skills may be rewarded in the labor market) or indirect effects of early skill gains via their effects on other skills (e.g., an early boost in English language proficiency may allow some children to learn more science, which may be rewarded in the labor market). (p. 17)

We can represent fundamentality, as defined by Bailey et al. (2017), in skill-building models in two cases. The first case is to say that skill θ_t is fundamental if it has a direct effect on H.[8] As we show below, this first concept of fundamentality is crucial for understanding why there might be long-term impacts of early interventions even when there is short-term fade-out (as measured by, say, standardized IQ or achievement tests). This kind of fundamentality might be understood as a kind of "cumulative" fundamentality (Funder & Ozer, 2019), whereby individual differences in skill θ_t continue to affect human capital across time.

The second case of fundamentality rests on an indirect channel. Consider two members of the family of math skills: counting and addition–subtraction. Suppose that counting does not affect H directly but that addition–subtraction does. Suppose, further, that the acquisition of addition–subtraction skills requires counting skills plus instructional investments that use counting skills to help build an understanding of simple arithmetic. In this case, counting would be considered a fundamental skill because it promotes the acquisition of a subsequent skill (i.e., addition–subtraction) that directly affects H.[9] It is possible to show that this indirect channel of fundamentality implies some degree of self-productivity (which can be arbitrarily small) and rules out extreme cases of dynamic substitutability.

In sum, the concepts of malleability and fundamentality impose restrictions on skill-building models and the types of skills that should be targeted by interventions at different stages of the life cycle. Malleability implies that $\rho_t > 0$, so interventions should target skills that are acquired from instruction and other environmental factors. Fundamentality implies that intervention should target skills that have direct impact on adult outcomes of interest or that help individuals acquire more sophisticated skills that themselves have direct impacts.

Skill-building models, counterfactual conditions, and sensitive periods. How do our skill-building models relate to the Bailey et al. (2017) idea that interventions should target skills that do not develop well in counterfactual conditions? Skill-building models can capture this idea by allowing each member (e.g., counting) of a family of skills to be strongly malleable for many periods and in a way that investments in these different periods are dynamically substitutable. For example, consider an intervention that targets the formation of certain math skills when the children are t years old. Further suppose that these skills are malleable between ages t and $t + 1$ and that investments in the formation of these skills that take place between years t and $t + 1$ are dynamically substitutable. This implies that the distribution of investments across years t and $t + 1$ does not matter. This situation is illustrated with the following hypothetical intervention that is implemented in experimental fashion: treatment children receive three units of investment in period t and one unit of investment in period $t + 1$. In contrast, control children receive two units of investment in period t, and an additional two units of investment in period $t + 1$. If there is perfect dynamic substitutability, then at the end of period $t + 1$, both the treatment and control children will have, on average, the same levels of human capital.

This shows that skill-building models have an important message for the design of interventions to foster human-capital formation in the case of skills that are malleable over many years and when investments are dynamically substitutable over those years of malleability. First, identify the periods in which skills are malleable. Second, identify the population at risk for having low levels of investment in all periods of malleability. Third, design an intervention that increases the sum of investments for this target population across all periods of malleability. To see why this is so potent, consider a case of extreme dynamic substitutability for the case in which the skill is equally malleable over 2 years: $\theta_{t+1} = \gamma\theta_{t-1} + \rho x_t + \rho x_{t-1}$. Then, note that $\theta_{t+1}\ \gamma\theta_{t-1} + \rho(x_t + x_{t+1})$ and what matters for θ_{t+1} is the total sum of $x_t + x_{t+1}$, not how it is distributed over time.

Likewise, suppose that the skill is malleable over years t and $t + 1$, but now investments are dynamically complementary. In this case, the distribution of investments over t and $t + 1$ matter. The intervention should not try to increase investments in some years but not others. Rather, it should increase investments in a proportional way in both periods. To see why, consider the case of extreme dynamic complementarity: $\theta_{t+1} = \min\{\theta_{t-1}, x_t, x_{t+1}\}$. In this case, the distribution of investments over time matter. So, if x_t is high, but x_{t+1} is low (and vice-versa), then θ_{t+1} will be low.

In skill-building models, we say that a skill has sensitive periods of development if we can identify segments in childhood or adolescence in which the skill has strong malleability. For some skills, these periods may last for days or weeks, whereas for other skills these periods may span many years. Below we describe how interventions that target skills that have sensitive periods of development may produce fade-out (without emergence of long-term impacts). First, however, we show the more general implications of skill-building models for fade-out and persistence.

Why fade-out is inevitable in most skill-building models. We are now in a position to describe what skill-building models that include the restrictions imposed by malleability and fundamentality imply for the fade-out or persistence of impacts of interventions. Our thought experiment consists of an intervention in which a large number of children are randomly assigned to control or treatment conditions. Our group-based formulation helps because it increases the likelihood that the distribution of initial conditions θ_0 is identical across control and treatment groups.

Suppose that control-group children receive a constant amount of investment every year across childhood and adolescence, so that $x_t = \bar{x}$ for $t = 1, \ldots, T$. In contrast, suppose that the investment in the treatment-group children is larger than the investment in control-group children only in period 1: $x_t = \alpha\bar{x}$, $\alpha > 1$, for $t = 1$, but $x_t = \bar{x}$ for $t = 2, \ldots, T$.[10] What implications does our skill-building model generate for the evolution of skill differences between children in the treatment and control groups over time? The key question is this: Under what conditions are end-of-treatment differences between treatment and control children maintained (impact persistence) or not (impact fade-out)?

Let θ_t^T and θ_t^C denote, respectively, the average stock of skills in treatment and control children in any given period t. Assume that the parameters that govern self-productivity, malleability, and dynamic complementarity are constant over time, so that $\gamma_t = \gamma$, $\rho_t = \rho$, and $\phi_t = \phi$ for all $t = 1, \ldots, T$.

According to the skill-building model, at the end of the intervention period ($t = 1$), the skills of treatment and control children are

$$\theta_1^T = \left[\gamma(\theta_0^T)^\phi + \rho(\alpha\bar{x})^\phi\right]^{\frac{1}{\phi}}$$

and

$$\theta_1^C = \left[\gamma(\theta_0^C)^\phi + \rho(\bar{x})^\phi\right]^{\frac{1}{\phi}}.$$

This difference involves nonlinear terms, but it can be manipulated to arrive at the following equality:

$$(\theta_1^T)^\phi - (\theta_1^C)^\phi = \rho\left[(\alpha\bar{x})^\phi - (\bar{x})^\phi\right].$$

This equation supports the reasonable expectation that a one-period intervention can be expected to boost the skills of treatment-group children relative to control-group children at the end of treatment as long as $\alpha > 1$ (i.e., as long as the investment in the treatment-group children is larger than the investment in control-group children).

Now impose our assumption that after the intervention is finished, both control-group and treatment-group children receive the same amount of investment $x_t = \bar{x}$ for $t = 2, \ldots, T$. Then, according to the model, skills 1 year after the end of the intervention (i.e., at $t = 2$) for the two groups are

$$\theta_2^T = \left[\gamma(\theta_1^T)^\phi + \rho(\bar{x})^\phi\right]^{\frac{1}{\phi}}$$

and

$$\theta_2^C = \left[\gamma(\theta_1^C)^\phi + \rho(\bar{x})^\phi\right]^{\frac{1}{\phi}}.$$

If $0 < \gamma\rho < 1$, the skill-building model predicts that fade-out emerges at the end of period $t = 2$ because the difference in skills between treatment-group and control-group children is smaller than it was at the end of period $t = 1$:

$$(\theta_2^T)^\phi - (\theta_2^C)^\phi = \gamma\left[(\theta_1^T)^\phi - (\theta_1^C)^\phi\right] = \gamma\rho\left[(\alpha\,\bar{x})^\phi - (\bar{x})^\phi\right].$$

This is the case because malleability and fundamentality imply that $\gamma \in (0,1)$, and, as long as γ is less that one, skill differences at the end of period $t = 2$ have to be smaller than skill differences at the end of period $t = 1$:

$$\frac{(\theta_2^T)^\phi - (\theta_2^C)^\phi}{(\theta_1^T)^\phi - (\theta_1^C)^\phi} = \gamma < 1.$$

This is true for any value of ϕ (i.e., in the cases of both dynamic complementarity and dynamic substitutability) as long as $\gamma < 1$ and $\gamma\rho < 1$. Furthermore, for any future period, it follows that

$$\frac{(\theta_t^T)^\phi - (\theta_t^C)^\phi}{(\theta_1^T)^\phi - (\theta_1^C)^\phi} = \gamma^{t-1}.$$

Under the conditions described above, the inevitability of fade-out from this model is both remarkable and intuitive. It is remarkable because the skill-building model it is based on is a very general one and allows for a wide range of linkages between past and current skills, as well as a very wide range of relationships between skill-building interventions and the levels of skills that children bring to them. As long as skill-building is not in complete lockstep from one period to the next, then fade-out is inevitable. But the intuition for this is also strong: Once skill building depends on continuing investment in those skills, and those investments are assumed to be similar for treatment- and control-group children, then end-of-treatment differences between treatment- and control-group children will erode.

The rate of erosion depends on the parameter ϕ. When there is dynamic complementarity, the rate of erosion is faster because the timing of investments matters. In contrast, when there is dynamic substitutability, the timing of investment is not important as long as the average level of investment is high across all periods.

Development under counterfactual conditions. We can now return to the discussion of how fade-out is linked to the concept of skills that would not develop well under counterfactual conditions. For skills that have sensitive periods of development and investment and are dynamically substitutable, skill-building models predict that proficiency is more strongly determined by the total amount of investment during sensitive periods than by how investments are distributed within sensitive periods of development. This property of skill-building models implies fade-out of intervention effects in which control groups catch up to treatment groups.

To see why, consider a skill for which the sensitive period of development lasts 2 years, which we denote by year t and year $t + 1$. Assume that we randomly allocate children to an intervention or control group. In the control group, the children receive two units of investment in each period. In the treatment group, the children receive three units of investment in the first period and one unit of investment in the second period. Both groups receive four total units of investment, but their temporal distribution differs across groups. If there is equal malleability and a high degree of dynamic substitutability, skill-building models predict that the skills of treatment-group children will surge during year t but that the skills of control-group children will eventually catch up by the end of year $t + 1$. When sensitive periods of development are long and investments are dynamically substitutable, final proficiency of skills are determined by total amount of investment within years of sensitive periods and not how investments are allocated across these years.

Skill-building models also predict fade-out when investments are dynamically complementary, but the way fade-out occurs is different. It is not that the control group catches up, but rather that the treatment group "converges down" to the control group. This occurs because under dynamic complementarity, the distribution of investments over time matters for eventual skill acquisition. As this discussion illustrates, fade-out is consistent with dynamic complementarity or dynamic substitutability, but each one implies different forms of fade-out.

Emerging impacts later in life. Conditions for emergence can also be derived from the formal skill-building model presented earlier. Skill-building models can generate effects on positive life outcomes even while predicting fade-out of early interventions: Consider the situation in which the parameter that governs dynamic complementarity/substitutability is equal to zero ($\phi = 0$ for all t), and self-productivity parameters are constant over time (so that $\gamma_t = \gamma$ for all t) and $\rho = 1 - \gamma$:

$$\ln \theta_t = \psi + \gamma \ln \theta_{t-1} + \rho \ln x_t, \text{ for } t = 1, \ldots, T.$$

Finally, we assume that adult human capital is the weighted sum of skills acquired by the end of each period of childhood and adolescence:

$$H = \sum_{t=1}^{T} \delta_t \ln \theta_t.$$

The parameter δ_t captures the fundamentality (in the direct sense) of skill θ_t. Specifically, skill θ_t is fundamental if $\delta_t > 0$ (i.e., if more of the skill increases the child's eventual stock of human capital). Under these parameterizations, it is possible to show that the impact of early investments on adult outcomes is the cumulative sum of its indirect impacts on all skills that are fundamental.[11]

Now, consider the example intervention in which control children have a constant level of investments equal to $\bar{x}$ for all t and the treatment children have a higher level of investment in the first period, $x_1 = \alpha \bar{x}$, but $x_t = \bar{x}$ for $t = 2, \ldots, T$. The difference in adult human capital between treatment and control children is

$$H^T - H^C = \alpha \delta (1 - \gamma^T).$$

Therefore, skill-building models may predict long-term impacts of interventions even in situations in which there is short-term fade-out of intervention effects. This implication in skill-building models is possible because of fundamentality and malleability: Interventions that target highly fundamental (according to the first case) and highly malleable skills will produce long-term impacts even if the intervention is not sufficient to change the trajectory of skills developed in postintervention periods. The higher the fundamentality and the higher the malleability, the stronger the impact on long-term outcomes.

This prediction of skill-building models may explain puzzling findings from the Perry Preschool program through two different mechanisms. As is widely known, the Perry intervention had short-term impacts that faded out by age 10 (e.g., see Heckman et al., 2010). However, long-term follow-up shows that the intervention had sizeable impacts on important outcomes of interest (such as the propensity to commit felonies or violent crimes).

An example of the first mechanism, in which short-lived impacts on fundamental skills are followed by the emergence of impacts in adulthood, follows: IQ at age 5 is a set of fundamental skills, IQ at age 10 is malleable for several years between ages 5 and 10, and the investments are dynamically complementary. Fundamentality of IQ at age 5 explains the long-term impacts documented in the literature. Dynamic complementarity and malleability explain fade-out of IQ by age 10.

An example of the second mechanism, in which fundamental skills are malleable during a limited period during which the intervention takes place, leading to the emergence of impacts in adulthood, would be that the long-term impacts are due in part to the impact of the Perry Preschool program on executive-function skills (see Heckman and Karapakula, 2019). In this case, skill-building models would predict that these executive-function skills are fundamental and malleable over a limited number of years (during the periods in which the Perry intervention took place).

Skill-building models, therefore, separate the issue of fade-out from impact on long-term outcomes. Interventions that do not generate fade-out are not guaranteed to produce long-term outcomes. This situation may arise if the interventions target skills that are malleable but not fundamental. Analogously, skill-building models can accommodate situations in which there is fade-out and emergence of impacts on long-term outcomes. This situation arises when interventions take place in specific periods of individuals' lives and target fundamental skills. Fade-out will occur if the levels of postintervention investments are more or less similar between control and treatment groups, but if the skill is malleable and fundamental, there should be impacts on long-term outcomes despite fade-out. The situation in which there is no fade-out and there is long-term impact require, generally speaking, interventions that are long-lived (or sustained environments in the definition of Bailey et al., 2017). Finally, skill-building models predict that interventions will have fade-out and no long-term impact if the intervention only shifts investments within years of sensitive periods of development.

Transparency

Editor: Nora Newcombe

Declaration of Conflicting Interests

The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

Funding

D. H. Bailey is supported by a Jacobs Foundation Fellowship. D. S. Yeager's research reported in this publication was supported by *Eunice Kennedy Shriver* National Institute of Child Health and Human Development (NICHD) Grant R01HD084772 and by William T. Grant Foundation Award 182921. This research was also supported by NICHD Grant P2C-HD042849, awarded to the Population Research Center at The University of Texas at Austin. An in-person meeting among the authors was funded by the Association for Psychological Science. Some of the ideas presented in this manuscript were influenced by discussions during a meeting of the Consortium on Early Childhood Intervention Impact (CECII), which was made possible by NICHD Grant 1R01-HD095930-01A1. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health and other funders.

Acknowledgments

We thank Daniela Alvarez-Vargas, Paul Hanselman, Jade Jenkins, Tyler Watts, and three anonymous reviewers for useful comments on this project. We thank Rodney Atkins for helpful edits on a previous version.

Notes

1. The evaluation design involved two treatment arms and a control condition. We summarize results for children in classrooms assigned to the treatment arm consisting only of Building Blocks activities and teacher professional development during the pre-K year relative to a control condition in which no Building Blocks activities or related professional development took place during the pre-K year.
2. Cunha and colleagues (2010) report evidence that self-productivity, of which within-domain transfer can be thought of as an example, is higher at later stages in the life cycle.
3. Bailey et al. (2017) called this third process "foot-in-the-door."
4. Two special cases in which the skill-building model can generate persistent effects of early interventions are presented in the appendix.
5. We thank an anonymous reviewer for this point.
6. A more comprehensive parametric formulation of the skill-building model would be one in which the acquisition of skills at period t depended on the entire vector of skills acquired in all previous periods:

$$\theta_t = \left[\sum_{\tau=0}^{t-1} \gamma_{t,\tau} (\theta_\tau)^{\phi_t} + \left(1 - \sum_{\tau=0}^{t-1} \gamma_{t,\tau}\right) (x_t)^{\phi_t} \right]^{\frac{1}{\phi_t}}.$$

However, we choose not to pursue this specification because it adds much more notation and does not contribute any new insights about skill-building models.
7. The functional form of our skill-building model encompasses three special cases of interest involving the parameter ϕ_t For example, if $\phi_t = 1$, then θ_t is just the weighted average of levels of skills acquired in the prior period and current levels of investment: $\theta_t = \gamma_t \theta_{t-1} + \rho_t x_t$. This additive process represents an extreme form of dynamic substitutability: What ultimately determines performance is the average quality of teachers that children experience during their school years, regardless of how quality was distributed over time. To see why, replace θ_t recursively and note that $\theta_t = \gamma_t \gamma_{t-1} \theta_{t-2} + \gamma_t \rho_{t-1} x_{t-1} + \rho_t x_t$. Such an additive process is assumed, for example, in models that estimate teacher value added (e.g., Chetty et al., 2014). The teacher-value-added literature, therefore, assumes that teacher quality does not depend on the quality of teachers students experienced in previous years (for a discussion about this fact, see Rothstein, 2017). Another special case of interest arises when $\phi_t = 0$. Then, $\theta_t = \theta_{t-1}^{\gamma_t} x_t^{\rho_t}$. This multiplicative specification implies that there is synergy between the levels of prior skills and the current level of investment. For example, in such cases, the greater the quality of teachers in previous years, the greater the teacher value added. To see this fact, replace θ_t recursively and check that $\theta_t = \theta_{t-2}^{\gamma_t \gamma_{t-1}} x_{t-1}^{\gamma_t \rho_{t-1}} x_t^{\rho_t}$. Empirical studies that aim to estimate the equations that best describe skill-building processes report estimates of ϕ_t that are "close" to zero (see Attanasio et al., 2019; Cunha et al., 2010). Therefore, such studies indicate that this representation is empirically

relevant. Finally, there is the case in which acquisition of skills at period t is the minimum of the level of skills acquired in previous periods and the level of investment in the current period: $\theta_t = \min\{\theta_{t-1}, x_t\} = \min\{\theta_{t-2}, x_{t-1}, x_t\}$. This specification describes an extreme form of dynamic complementarity in that it implies that the productivity of current levels of investment is limited by the level of investment in previous periods.

8. Formally, according to this first case, skill θ_t is fundamental if $\partial H / \partial \theta_t > 0$.

9. Formally, consider two skills θ_t and θ_{t+k}, for any $k \in \{1, \ldots, T-t\}$,, and that $\partial H / \partial \theta_t = 0$ but $\partial H / \partial \theta_{t+k} > 0$. In words, θ_{t+k} is fundamental according to the first definition, but θ_t is not because increases in it do not increase human capital H. However, θ_t can be considered fundamental (via an indirect channel) if more of it boosts fundamental skill θ_{t+k} (i.e., if $\partial \theta_{t+k} / \partial \theta_t > 0$). If $k = 1$, this definition implies that $0 < \gamma_t \leq 1$. For a general value of k, the second definition implies that the product $\prod_{\tau=1}^{k} \gamma_{t+\tau} \neq 0$, which implies that $\gamma_{t+\tau} \neq 0$ for $\tau = 1, \ldots, k$.

10. We revisit this assumption later to discuss the predictions of the skill-building model when there are "sustaining environments."

11. Formally, this is represented by the following equation:

$$H = \left(\sum_{t=1}^{T} \delta_t \gamma^t\right) \ln \theta_0 + \rho\left(\sum_{t=1}^{T} \delta_t \gamma^{t-1}\right) \ln x_1 + \rho\left(\sum_{t=2}^{T} \delta_t \gamma^{t-1}\right) \ln x_2 + \ldots + \rho \delta_T \ln x_T.$$

References

Abdulkadiroğlu, A., Pathak, P. A., & Walters, C. R. (2018). Free to choose: Can school choice reduce student achievement? *American Economic Journal: Applied Economics*, *10*, 175–206.

Abenavoli, R. M. (2019). The mechanisms and moderators of "fade-out": Towards understanding why the skills of early childhood program participants converge over time with the skills of other children. *Psychological Bulletin*, *145*, 1103–1127.

Ackerman, P. L. (2017). Adult intelligence: The construct and the criterion problem. *Perspectives on Psychological Science*, *12*, 987–998.

Adhvaryu, A., & Nyshadham, A. (2016). Endowments at birth and parents' investments in children. *The Economic Journal*, *126*, 781–820.

Allen, J. P., Philliber, S., Herrling, S., & Kuperminc, G. P. (1997). Preventing teen pregnancy and academic failure: Experimental evaluation of a developmentally based approach. *Child Development*, *64*, 729–742. doi:10.1111/j.1467-8624.1997.tb04233.x

Almond, D., Currie, J., & Duque, V. (2018). Childhood circumstances and adult outcomes: Act II. *Journal of Economic Literature*, *56*, 1360–1446.

Athey, S., Chetty, R., Imbens, G. W., & Kang, H. (2019). *The Surrogate Index: Combining short-term proxies to estimate long-term treatment effects more rapidly and precisely* (NBER Working Paper No. 26463). Cambridge, MA: National Bureau of Economic Research.

Attanasio, O., Meghir, C., & Nix, E. (2019). *Human capital development and parental investment in India* (NBER Working Paper No. 21740). Cambridge, MA: National Bureau of Economic Research.

Bailey, D. H. (2019). Explanations and implications of diminishing intervention impacts across time. In D. C. Geary, D. B. Berch, & K. M. Koepke (Eds.), *Mathematical cognition and learning: Vol. 5. Cognitive foundations for improving mathematical learning* (pp. 321–346). San Diego, CA: Academic Press.

Bailey, D. H., Duncan, G. J., Odgers, C., & Yu, W. (2017). Persistence and fadeout in the impacts of child and adolescent interventions. *Journal of Research on Educational Effectiveness*, *10*, 7–39.

Bailey, D. H., Duncan, G. J., Watts, T., Clements, D., & Sarama, J. (2018). Risky business: Correlation and causation in longitudinal studies of skill development. *American Psychologist*, *73*, 81–94.

Bailey, D. H., Fuchs, L. S., Gilbert, J. K., Geary, D. C., & Fuchs, D. (2020a). Prevention: Necessary but insufficient? A 2-year follow-up of an effective first-grade mathematics intervention. *Child Development*, *91*, 382–400. doi:10.1111/cdev.13175

Bailey, D. H., Jenkins, M., & Alvarez-Vargas, D. (2020b). Complementarities between early educational intervention and later educational quality? A systematic review of the sustaining environments hypothesis. *Developmental Review*, *56*, Article 100910. doi: 10.1016/j.dr.2020.100910

Bailey, D. H., & Littlefield, A. K. (2017). Does reading cause later intelligence? Accounting for stability in models of change. *Child Development*, *88*, 1913–1921.

Bailey, D. H., Nguyen, T., Jenkins, J. M., Domina, T., Clements, D. H., & Sarama, J. S. (2016). Fadeout in an early mathematics intervention: Constraining content or preexisting differences? *Developmental Psychology*, *52*, 1457–1469.

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

Baird, S., Hicks, J. H., Kremer, M., & Miguel, E. (2016). Worms at work: Long-run impacts of a child health investment. *The Quarterly Journal of Economics*, *131*, 1637–1680. doi:10.1093/qje/qjw022

Baroody, A. J. (1987). The development of counting strategies for single-digit addition. *Journal for Research in Mathematics Education*, *18*, 141–157.

Bartik, T. J. (2014). *From preschool to prosperity: The economic payoff to early childhood education*. Kalamazoo, MI: W.E. Upjohn Institute for Employment Research. doi:10.17848/9780880994835

Benartzi, S., Beshears, J., Milkman, K. L., Sunstein, C. R., Thaler, R. H., Shankar, M., . . . Galing, S. (2017). Should governments invest more in nudging? *Psychological Science*, *28*, 1041–1055.

Bettinger, E. P., Gurantz, O., Kawano, L., Sacerdote, B., & Stevens, M. (2019). The long-run impacts of financial aid:

Evidence from California's Cal Grant. *American Economic Journal: Economic Policy*, *11*, 64–94.

Bettinger, E. P., Long, B. T., Oreopoulos, P., & Sanbonmatsu, L. (2012). The role of application assistance and information in college decisions: Results from the H&R Block FAFSA Experiment. *The Quarterly Journal of Economics*, *127*, 1205–1242. doi:10.1093/qje/qjs017

Bhalotra, S. R., & Venkataramani, A. (2015). *Shadows of the captain of the men of death: Early life health interventions, human capital investments, and institutions*. SSRN. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1940725

Bitler, M. P., Hoynes, H. W., & Domina, T. (2014). *Experimental evidence on distributional effects of Head Start* (NBER Working Paper No. 20434). Cambridge, MA: National Bureau of Economic Research. doi:10.3386/w20434

Bleidorn, W., Hill, P., Back, M., Denissen, J., Hennecke, M., Hopwood, C., . . . Orth, U. (2019). *The policy relevance of personality traits*. PsyArXive. doi: 10.31234/osf.io/a9rbn

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

Bond, T. N., & Lang, K. (2018). The black–white education scaled test-score gap in grades K-7. *Journal of Human Resources*, *53*, 891–917.

Bouchard, T. J., Lykken, D. T., McGue, M., Segal, N. L., & Tellegen, A. (1990). Sources of human psychological differences: The Minnesota study of twins reared apart. *Science*, *250*, 223–228.

Bouguen, A., Huang, Y., Kremer, M., & Miguel, E. (2019). Using RCTs to estimate long- run impacts in development economics. *Annual Review of Economics*, *11*, 523–561. doi:10.1146/annurev-economics-080218-030333

Bowlby, J. (1973). *Attachment and loss, Vol. II: Separation*. New York, NY: Basic Books.

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*, 403–414. doi:10.1037/0022-0663.91.3.403

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

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

Card, D. (2001). Estimating the return to schooling: Progress on some persistent econometric problems. *Econometrica*, *69*, 1127–1160.

Cain, K., & Oakhill, J. (2007). Reading comprehension difficulties: Correlates, causes, and consequences. In K. Cain & J. Oakhill (Eds.), *Challenges in language and literacy. Children's comprehension problems in oral and written language: A cognitive perspective* (pp. 41–75). New York, NY: Guilford Press.

Cascio, E. U., & Staiger, D. O. (2012). *Knowledge, tests, and fadeout in educational interventions* (NBER Working Paper No. 18038). Cambridge, MA: National Bureau of Economic Research. doi:10.3386/w18038

Chesmore, A. A., Ou, S. R., & Reynolds, A. J. (2016). Childhood placement in special education and adult well-being. *The Journal of Special Education*, *50*, 109–120.

Chetty, R., Friedman, J. N., Hilger, N., Saez, E., Schanzenbach, D. W., & Yagan, D. (2011). How does your kindergarten classroom affect your earnings? Evidence from Project Star. *Quarterly Journal of Economics*, *126*, 1593–1660.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014). Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood. *The American Economic Review*, *104*, 2633–2679.

Chetty, R., & Hendren, N. (2018). The impacts of neighborhoods on intergenerational mobility I: Childhood exposure effects. *The Quarterly Journal of Economics*, *133*, 1107–1162.

Chetty, R., Hendren, N., & Katz, L. F. (2016). The effects of exposure to better neighborhoods on children: New evidence from the Moving to Opportunity experiment. *American Economic Review*, *106*, 855–902.

Cheung, A. C. K., & Slavin, R. E. (2013). The effectiveness of educational technology applications for enhancing mathematics achievement in K-12 classrooms: A meta-analysis. *Educational Research Review*, *9*, 88–113.

Chingos, M. M., & Whitehurst, G. J. (2012). *Choosing blindly: Instructional materials, teacher effectiveness, and the common core*. Washington, DC: Brookings Institution.

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

Clarke, P., Snowling, M., Truelove, E., & Hulme, C. (2010). Ameliorating children's reading-comprehension difficulties: A randomized controlled trial. *Psychological Science*, *21*, 1106–1116. doi:10.1177/0956797610375449

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*, 812–850.

Cohen, G. L., Garcia, J., Purdie-Vaughns, V., Apfel, N., & Brzustoski, P. (2009). Recursive Processes in Self-Affirmation: Intervening to close the minority achievement gap. *Science*, *324*, 400–403. doi:10.1126/science.1170769

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. New York, NY: Routledge Academic.

Cohodes, S. R. (2020). The long-run impacts of specialized programming for high-achieving students. *American Economic Journal: Economic Policy*, *12*, 127–166.

Coleman, J. S. (1990). *Foundations of social theory*. Cambridge, MA: Harvard University Press.

Colom, R., Román, F. J., Abad, F. J., Shih, P. C., Privado, J., Froufe, M., . . . Karama, S. (2013). Adaptive n-back training does not improve fluid intelligence at the construct level: Gains on individual tests suggest that training may enhance visuospatial processing. *Intelligence*, *41*, 712–727.

Conduct Problems Prevention Research Group. (1999a). Initial impact of the Fast Track Prevention Trial for Conduct Problems: I. The high-risk sample. *Journal of Consulting and Clinical Psychology*, *67*, 631–647. doi:10.1037/0022-006X.67.5.631

Conduct Problems Prevention Research Group. (1999b). Initial impact of the Fast Track Prevention Trial for Conduct Problems: II. Classroom effects. *Journal of Consulting and Clinical Psychology*, *67*, 648–657.

Conduct Problems Prevention Research Group. (2011). The effects of the Fast Track preventive intervention on the development of conduct disorder across childhood. *Child Development*, *82*, 331–345. doi:10.1111/j.1467-8624.2010.01558.x

Cortes, K. E., & Goodman, J. S. (2014). Ability-tracking, instructional time, and better pedagogy: The effect of double-dose algebra on student achievement. *American Economic Review*, *104*, 400–405.

Costa, P. T., & McCrae, R. R. (1980). Influence of extraversion and neuroticism on subjective well-being: Happy and unhappy people. *Journal of Personality and Social Psychology*, *38*, 668–678.

Croke, K., & Atun, R. (2019). The long run impact of early childhood deworming on numeracy and literacy: Evidence from Uganda. *PLOS Neglected Tropical Diseases*, *13*(1), Article e0007085. doi:10.1371/journal.pntd.0007085

Cronbach, L. J. (1957). The two disciplines of scientific psychology. *American Psychologist*, *12*, 671–684.

Crosnoe, R. (2011). *Fitting in, standing out: Navigating the social challenges of high school to get an education*. New York, NY: Cambridge University Press.

Crosnoe, R., Lopez-Gonzalez, L., & Muller, C. (2004). Immigration from Mexico into the math/science pipeline in American education. *Social Science Quarterly*, *85*, 1208–1226. doi:10.1111/j.0038-4941.2004.00272.x

Crosnoe, R., & Muller, C. (2014). Family socioeconomic status, peers, and the path to college. *Social Problems*, *61*, 602–624. doi:10.1525/sp.2014.12255

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

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

Currie, J., & Thomas, D. (2000). School quality and the longer-term effects of Head Start. *Journal of Human Resources*, *35*, 755–774. doi:10.2307/146372

Deming, D. (2009). Early childhood intervention and life-cycle skill development: Evidence from Head Start. *American Economic Journal: Applied Economics*, *1*(3), 111–134.

Dobbie, W. S., & Fryer, R. G., Jr. (2016). *Charter schools and labor market outcomes* (NBER Working Paper No. 22502). Cambridge, MA: National Bureau of Economic Research. doi:10.3386/w22502

Dodge, K. A., Bai, Y., Ladd, H. F., & Muschkin, C. G. (2017). Impact of North Carolina's early childhood programs and policies on educational outcomes in elementary school. *Child Development*, *88*, 996–1014.

Dodge, K. A., Bierman, K. L., Coie, J. D., Greenberg, M. T., Lochman, J. E., McMahon, R. J., & Pinderhughes, E. E. (2015). Impact of early intervention on psychopathology, crime, and well-being at age 25. *American Journal of Psychiatry*, *172*, 59–70. doi:10.1176/appi.ajp.2014.13060786

Dodge, K. A., Malone, P. S., Lansford, J. E., Miller, S., Pettit, G. S., & Bates, J. E. (2009). A dynamic cascade model of the development of substance-use onset. *Monographs of the Society for Research in Child Development*, *74*(3), vii–119.

Domina, T. (2014). The link between middle school mathematics course placement and achievement. *Child Development*, *85*, 1948–1964.

Duckworth, A. L., Grant, H., Loew, B., Oettingen, G., & Gollwitzer, P. M. (2011). Self-regulation strategies improve self-discipline in adolescents: Benefits of mental contrasting and implementation intentions. *Educational Psychology*, *31*, 17–26. doi:10.1080/01443410.2010.506003

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*, 1428–1446. doi:10.1037/0012-1649.43.6.1428.supp

Duncan, G. J., & Magnuson, K. (2007). Penny wise and effect size foolish. *Child Development Perspectives*, *1*, 46–51.

Duncan, G. J., & Magnuson, K. (2013). Investing in preschool programs. *Journal of Economic Perspectives*, *27*, 109–132.

Ebbinghaus, H. (1885). *Über das Gedächtnis: Untersuchungen zur experimentellen Psychologie* [About memory: research on experimental psychology]. Leipzig, Germany: Duncker & Humblot.

Elango, S., García, J. L., Heckman, J. J., & Hojman, A. (2015). *Early childhood education* (NBER Working Paper No. 21766). Cambridge, MA: National Bureau of Economic Research.

Elder, G. H., Jr. (1974). *Children of the great depression*. Chicago, IL: University of Chicago Press.

Elder, G. H., Jr. (1998). The life course as developmental theory. *Child Development*, *69*, 1–12.

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

Estrada, E., Ferrer, E., Abad, F. J., Román, F. J., & Colom, R. (2015). A general factor of intelligence fails to account for changes in tests' scores after cognitive practice: A longitudinal multi-group latent-variable study. *Intelligence*, *50*, 93–99.

Farrington, D. P., & Welsh, B. C. (2003). Family-based prevention of offending: A meta-analysis. *Australian & New Zealand Journal of Criminology*, *36*, 127–151.

Felitti, V. J., Anda, R. F., Nordenberg, D., Williamson, D. F., Spitz, A. M., Edwards, V., . . . Marks, J. S. (1998). Relationship of childhood abuse and household dysfunction to many of the leading causes of death in adults: The Adverse Childhood Experiences (ACE) Study. *American Journal of Preventive Medicine*, *56*, 774–786.

Foorman, B. R., Chen, D.-T., Carlson, C., Moats, L., Francis, D., & Fletcher, J. (2003). The necessity of the alphabetic principle to phonemic awareness instruction. *Reading & Writing*, *16*, 289–324. doi:10.1023/A:1023671702188

Foorman, B. R., Francis, D. J., Davidson, K., Harm, M., & Griffin, J. (2004). Variability in text features in six grade 1 basal reading programs. *Scientific Studies in Reading*, *8*, 167–197. doi:10.1207/s1532799xssr0802_4

Foorman, B. R., Francis, D. J., Fletcher, J. M., Schatschneider, C., & Mehta, P. (1998). The role of instruction in learning to read: Preventing reading failure in at-risk children. *Journal of Educational Psychology*, *90*, 37–55. doi:10.1037/0022-0663.90.1.37

Foorman, B. R., Petscher, Y., & Herrera, S. (2018). Unique and common effects of oral language in predicting reading comprehension in grades 1-10. *Learning and Individual Differences*, *63*, 12–23. doi:10.1016/j.lindif.2018.02.011

Foorman, B. R., Petscher, Y., Stanley, C., & Herrera, S. (2017). Latent profiles of reading and language and their association with standardized reading outcomes in kindergarten through tenth grade. *Journal of Research on Educational Effectiveness*, *10*, 619–645. doi:10.1080/19345747.2016.1237597

Freud, S. (1961). The aetiology of hysteria. In J. Strachey (Ed. and Trans.), *The standard edition of the complete psychological works of Sigmund Freud* (Vol. 3, pp. 189–224). London, England: Hogarth Press. (Original work published 1924)

Frey, E., & Rogers, T. (2014). Persistence: How treatment effects persist after interventions stop. *Policy Insights from the Behavioral and Brain Sciences*, *1*, 172–179. doi:10.1177/2372732214550405

Funder, D. C., & Ozer, D. J. (2019). Evaluating effect size in psychological research: Sense and nonsense. *Advances in Methods and Practices in Psychological Science*, *2*, 156–168.

García, J. L., Heckman, J. J., Leaf, D. E., & Prados, M. J. (2016). *The life-cycle benefits of an influential early childhood program* (NBER Working Paper No. 22993). Cambridge, MA: National Bureau of Economic Research. doi:10.3386/w22993

Gelber, A., & Isen, A. (2013). Children's schooling and parents' behavior: Evidence from the Head Start Impact Study. *Journal of Public Economics*, *101*, 25–38.

Gersten, R., Newman-Gonchar, R., Haymond, K., & Dimino, J. (2017). *What is the evidence base for Response to Intervention in reading in grades 1–3?* (REL 2016-129). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Southeast. Retrieved from https://files.eric.ed.gov/fulltext/ED573686.pdf

Gibbs, C., Ludwig, J., & Miller, D. L. (2011). *Does Head Start do any lasting good?* (NBER Working Paper No. 17452). Cambridge, MA: National Bureau of Economic Research. doi:10.3386/w17452.

Gonzalez, J. E., Pollard-Duradola, S., Simmons, D. C., Taylor, A. B., Davis, M. J., Kim, M., & Simmons, L. (2011). Developing low-income preschoolers' social studies and science vocabulary knowledge through content-focused shared book reading. *Journal of Research on Educational Effectiveness*, *4*, 25–52. doi:10.1080/19345747.2010.487927

Gormley, W. T., Jr., Phillips, D., & Anderson, S. (2018). The effects of Tulsa's Pre-K Program on middle school student performance. *Journal of Policy Analysis and Management*, *37*, 63–87.

Goyer, J. P., Garcia, J., Purdie-Vaughns, V., Binning, K. R., Cook, J. E., Reeves, S. L., . . . Cohen, G. L. (2017). Self-affirmation facilitates minority middle schoolers' progress along college trajectories. *Proceedings of the National Academy of Sciences, USA*, *114*, 7594–7599.

Gramlich, E. M. (1990). *A guide to benefit-cost analysis*. Englewood Cliffs, NJ: Prentice-Hall.

Greenberg, M. T., & Abenavoli, R. (2017). Universal interventions: Fully exploring their impacts and potential to produce population-level impacts. *Journal of Research on Educational Effectiveness*, *10*, 40–67.

Gunnsteinsson, S., Adhvaryu, A., Christian, P., Labrique, A., Sugimoto, J., Shamim, A. A., & West, K. P., Jr. (2019). Protecting infants from natural disasters: The case of vitamin A supplementation and a tornado in Bangladesh (NBER Working Paper No. 25969). Cambridge, MA: National Bureau of Economic Research.

Harackiewicz, J. M., & Priniski, S. J. (2018). Improving student outcomes in higher education: The science of targeted intervention. *Annual Review of Psychology*, *69*, 409–435. doi:10.1146/annurev-psych-122216-011725

Hattie, J., Biggs, J., & Purdie, N. (1996). Effects of learning skills interventions on student learning: A meta-analysis. *Review of Educational Research*, *66*, 99–136.

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

Heckman, J. J. (2017, February 23). With supports, the impact of high-quality preschool does not fade out [Letter to the editor]. *The Washington Post*. Retrieved from https://www.washingtonpost.com/opinions/with-supports-the-impact-of-high-quality-preschool-does-not-fade-out/2017/02/23/fd39f95a-f91a-11e6-aa1e-5f735ee31334_story.html

Heckman, J. J., & Karapakula, G. (2019). *Intergenerational and intragenerational externalities of the Perry Preschool Project* (NBER Working Paper No. 25889). Cambridge, MA: National Bureau of Economic Research. doi:10.3386/w25889

Heckman, J. J., Moon, S. H., Pinto, R., Savelyev, P. A., & Yavitz, A. (2010). The rate of return to the HighScope Perry Preschool Program. *Journal of Public Economics*, *94*, 114–128.

Heckman, J. J., Stixrud, J., & Urzua, S. (2006). The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior. *Journal of Labor Economics*, *24*, 411–482.

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.

Holmes Erickson, H., Mills, J., & Wolf, P. (2019). *The effect of the Louisiana Scholarship Program on college entrance*. SSRN. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3376236

Hoover, W. A., & Gough, P. B. (1990). The simple view of reading. *Reading and Writing*, *2*, 127–160.

Hoynes, H., Schanzenbach, D. W., & Almond, D. (2016). Long-run impacts of childhood access to the safety net. *American Economic Review*, *106*, 903–934.

Hurwitz, M., Mbekeani, P. P., Nipson, M. M., & Page, L. C. (2017). Surprising ripple effects: How changing the SAT score-sending policy for low-income students impacts college access and success. *Educational Evaluation and Policy Analysis*, *39*, 77–103. doi:10.3102/0162373716665198

Hyman, J. (2017). Does money matter in the long run? Effects of school spending on educational attainment. *American Economic Journal: Economic Policy*, *9*, 256–280.

Jackson, C. K. (2018). What do test scores miss? The importance of teacher effects on non–test score outcomes. *Journal of Political Economy*, *126*, 2072–2107.

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

Jenkins, J. M., Watts, T. W., Magnuson, K., Gershoff, E. T., Clements, D. H., Sarama, J., & Duncan, G. J. (2018). Do high-quality kindergarten and first-grade classrooms mitigate preschool fadeout? *Journal of Research on Educational Effectiveness*, *11*, 339–374.

Jensen, A. R. (1998). *The g factor: The science of mental ability*. Westport, CT: Praeger.

Johnson, R. C., & Jackson, C. K. (2019). Reducing inequality through dynamic complementarity: Evidence from head start and public school spending. *American Economic Journal: Economic Policy*, *11*, 310–349. doi:10.1257/pol.20180510.

Kang, C., Duncan, G. J., Clements, D., Sarama, J., & Bailey, D. H. (2019). The roles of transfer of learning and forgetting in the persistence and fadeout of early childhood mathematics interventions. *Journal of Educational Psychology*, *111*, 590–603.

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, USA*, *112*, 4612–4617. doi:10.1073/pnas.1417106112

Kim, J., Hemphill, L., Troyer, M., Thomson, J., Jones, S., LaRusso, M., & Donovan, S. (2016). Engaging struggling adolescent readers to improve reading skills. *Reading Research Quarterly*, *52*, 357–382. doi:10.1002/rrq.171

Kitchens, K. E., Gormley, W., & Anderson, S. (2020). Do better schools help to prolong early childhood education effects? *Journal of Applied Developmental Psychology*, *66*, 101092.

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

Knudsen, E. I. (2004). Sensitive periods in the development of the brain and behavior. *Journal of Cognitive Neuroscience*, *16*, 1412–1425.

Kraft, M. A. (2019). Teacher effects on complex cognitive skills and social-emotional competencies. *Journal of Human Resources*, *54*, 1–36.

Kraft, M. A. (2020). Interpreting effect sizes of education interventions. *Educational Researcher*, *49*, 241–253. doi:10.3102/0013189X20912798

LaBerge, D., & Samuels, S. J. (1974). Toward a theory of automatic information processing in reading. *Cognitive Psychology*, *6*, 293–323.

Lave, C. A., & March, J. G. (1993). *An introduction to models in the social sciences*. Lanham, MD: University Press of America.

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

Levin, H. M. (1988). Cost-effectiveness and educational policy. *Educational Evaluation and Policy Analysis*, *10*, 51–69.

Levin, H. M., Belfield, C., Muennig, P., & Rouse, C. (2007). *The costs and benefits of an excellent education for all of America's children*. New York, NY: Teachers College.

Lewin, K. (1952). Group decision and social change. In T. M. Newcomb & E. L. Hartley (Eds.), *Readings in social psychology* (2nd ed., pp. 330–344). New York, NY: Holt.

Li, W., Duncan, G. J., Magnuson, K., Schindler, H. S., Yoshikawa, H., & Leak, J. (2017). Timing in early childhood education: How cognitive and achievement program impacts vary by starting age, program duration, and time since the end of the program (EdWorkingPaper 20-201). Providence, RI: Annenberg Institute at Brown University. doi:10.26300/5tvg-nt21

Linn, R. L. (2000). Assessments and accountability. *Educational Researcher*, *29*(2), 4–16.

Linn, R. L., Graue, M. E., & Sanders, N. M. (1990). Comparing state and district test results to national norms: The validity of the claims that "Everyone is above average." *Educational Measurement: Issues and Practice*, *9*(3), 5–14.

Lipsey, M. W., Farran, D. C., & Durkin, K. (2018). Effects of the Tennessee Prekindergarten Program on children's achievement and behavior through third grade. *Early Childhood Research Quarterly*, *45*, 155–176.

List, J. A., Momeni, F., & Zenou, Y. (2019). Are estimates of early education programs too pessimistic? Evidence from a large-scale field experiment that causally measures neighbor effects (IFN Working Paper No. 1293). Retrieved from https://www.ifn.se/wfiles/wp/wp1293.pdf

Liu, J., & Loeb, S. (2019). Engaging teachers: Measuring the impact of teachers on student attendance in secondary school. *Journal of Human Resources*. Advance online publication. doi:10.3368/jhr.56.2.1216-8430R3

Lucas, R. E. (2007). Adaptation and the set-point model of subjective well-being: Does happiness change after major life events? *Current Directions in Psychological Science*, *16*, 75–79.

Ludwig, J., & Miller, D. L. (2007). Does Head Start improve children's life chances? Evidence from a regression discontinuity design. *The Quarterly Journal of Economics*, *122*, 159–208.

Martorell, P., & Mariano, L. T. (2018). The causal effects of grade retention on behavioral outcomes. *Journal of Research on Educational Effectiveness*, *11*, 192–216.

McCoy, D. C., Yoshikawa, H., Ziol-Guest, K. M., Duncan, G. J., Schindler, H. S., Magnuson, K., . . . Shonkoff, J. P. (2017). Impacts of early childhood education on medium-and long-term educational outcomes. *Educational Researcher*, *46*, 474–487.

McShane, B. B., Böckenholt, U., & Hansen, K. T. (2016). Adjusting for publication bias in meta-analysis: An evaluation of selection methods and some cautionary notes. *Perspectives on Psychological Science*, *11*, 730–749.

Melby-Lervåg, M., Redick, T. S., & Hulme, C. (2016). Working memory training does not improve performance on measures of intelligence or other measures of "far transfer" evidence from a meta-analytic review. *Perspectives on Psychological Science*, *11*, 512–534.

Murphy, M. C., & Walton, G. M. (2013). From prejudiced people to prejudiced places: A social-contextual approach to prejudice. In C. Stangor & C. Crandall (Eds.), *Frontiers in social psychology series: Stereotyping and prejudice* (pp. 181–203). New York, NY: Psychology Press.

Murre, J. M., & Dros, J. (2015). Replication and analysis of Ebbinghaus' forgetting curve. *PLOS ONE*, *10*(7), Article e0120644. doi:10.1371/journal.pone.0120644

Obradović, J., Burt, K. B., & Masten, A. S. (2010). Testing a dual cascade model linking competence and symptoms over 20 years from childhood to adulthood. *Journal of Clinical Child and Adolescent Psychology*, *39*, 90–102. doi:10.1080/15374410903401120

Ou, S. R., Arteaga, I., & Reynolds, A. J. (2019). Dosage effects in the child-parent center PreK-to-3rd grade program: A re-analysis in the Chicago longitudinal study. *Children and Youth Services Review*, *101*, 285–298.

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

Pashler, H., Rohrer, D., Cepeda, N. J., & Carpenter, S. K. (2007). Enhancing learning and retarding forgetting: Choices and consequences. *Psychonomic Bulletin & Review*, *14*, 187–193.

Penner, A. M., Domina, T., Penner, E. K., & Conley, A. M. (2015). Curricular policy as a collective effects problem: A distributional approach. *Social Science Research*, *52*, 627–641. doi:10.1016/j.ssresearch.2015.03.008

Perfetti, C. (2007). Reading ability: Lexical quality to comprehension. *Scientific Studies of Reading*, *11*, 357–383. doi:10.1080/10888430701530730

Perfetti, C., & Stafura, J. (2014). Word knowledge in a theory of reading comprehension. *Scientific Studies of Reading*, *18*, 22–37. doi:10.1080/10888438.2013.827687

Phillips, D. A., Lipsey, M. W., Dodge, K. A., Haskins, R., Bassok, D., Burchinal, M. R., . . . Weiland, C. (2017). Puzzling it out: The current state of scientific knowledge on pre-kindergarten effects. A consensus statement, in *The current state of scientific knowledge on pre-kindergarten effects* (pp. 19–30). Washington, DC: Brookings Institution. Retrieved from https://www.brookings.edu/wp-content/uploads/2017/04/duke_prekstudy_final_4-4-17_hires.pdf

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

Protzko, J. (2016). Does the raising IQ-raising g distinction explain the fadeout effect? *Intelligence*, *56*, 65–71.

Puma, M., Bell, S., Cook, R., Heid, C., Broene, P., Jenkins, F., . . . Downer, J. (2012). *Third grade follow-up to the Head Start Impact Study: Final report* (OPRE Report No. 2012-45). Retrieved from http://eric.ed.gov/?id=ED539264

Ritchie, S. J., Bates, T. C., & Deary, I. J. (2015). Is education associated with improvements in general cognitive ability, or in specific skills? *Developmental Psychology*, *51*, 573–582.

Ritchie, S. J., & Tucker-Drob, E. M. (2018). How much does education improve intelligence? A meta-analysis. *Psychological Science*, *29*, 1358–1369.

Roberts, B. W., Luo, J., Briley, D. A., Chow, P. I., Su, R., & Hill, P. L. (2017). A systematic review of personality trait change through intervention. *Psychological Bulletin*, *143*, 117–141.

Roberts, G., Quach, J., Spencer-Smith, M., Anderson, P. J., Gathercole, S., Gold, L., . . . Wake, M. (2016). Academic outcomes 2 years after working memory training for children with low working memory: A randomized clinical trial. *JAMA Pediatrics*, *170*(5), e154568.

Rohrer, D., & Pashler, H. (2010). Recent research on human learning challenges conventional instructional strategies. *Educational Researcher*, *39*, 406–412.

Roisman, G. I., & Fraley, R. C. (2013). Developmental mechanisms underlying the legacy of childhood experiences. *Child Development Perspectives*, *7*, 149–154.

Ross, L., & Nisbett, R. E. (1991). *The person and the situation: Perspectives of social psychology*. New York, NY: McGraw-Hill.

Rossin-Slater, M., & Wüst, M. (2016). *What is the added value of preschool? Long-term impacts and interactions with a health intervention* (NBER Working Paper No. 22700). Cambridge, MA: National Bureau of Economic Research. doi:10.3386/w22700

Rothstein, J. (2017). *Revisiting the impacts of teachers* (IRLE Working Paper No. 101-17). Retrieved from http://irle.berkeley.edu/files/2017/Revisiting-the-Impacts-of-Teachers.pdf

Rubin, D. B. (2004). Direct and indirect causal effects via potential outcomes. *Scandinavian Journal of Statistics*, *31*, 161–170.

Sacerdote, B. (2007). How large are the effects from changes in family environment? A study of Korean American adoptees. *The Quarterly Journal of Economics*, *122*, 119–157. doi:10.1162/qjec.122.1.119

Sala, G., & Gobet, F. (2018). Cognitive training does not enhance general cognition. *Trends in Cognitive Sciences*, *23*, 9–20.

Sawyer, A. M., Borduin, C. M., & Dopp, A. R. (2015). Long-term effects of prevention and treatment on youth antisocial behavior: A meta-analysis. *Clinical Psychology Review*, *42*, 130–144.

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

Steele, C. M. (1997). A threat in the air: How stereotypes shape intellectual identity and performance. *American Psychologist*, *52*, 613–629. doi:10.1037/0003-066X.52.6.613

Stipek, D., Franke, M., Clements, D., Farran, D., & Coburn, C. (2017). PK-3: What does it mean for instruction? *Social Policy Report*, *30*(2). Retrieved from https://files.eric.ed.gov/fulltext/ED581657.pdf

Stuart, E. A., Bell, S. H., Ebnesajjad, C., Olsen, R. B., & Orr, L. L. (2017). Characteristics of school districts that participate in rigorous national educational evaluations. *Journal of Research on Educational Effectiveness*, *10*, 168–206. doi:10.1080/19345747.2016.1205160

Takacs, Z. K., & Kassai, R. (2019). The efficacy of different interventions to foster children's executive function skills: A series of meta-analyses. *Psychological Bulletin*, *145*, 653–697.

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

Taylor, R. D., Oberle, E., Durlak, J. A., & Weissberg, R. P. (2017). Promoting positive youth development through school-based social and emotional learning interventions: A meta-analysis of follow-up effects. *Child Development*, *88*, 1156–1171.

te Nijenhuis, J., Jongeneel-Grimen, B., & Armstrong, E. L. (2015). Are adoption gains on the g factor? A meta-analysis. *Personality and Individual Differences*, *73*, 56–60.

Thaler, R. H., & Sunstein, C. R. (2008). *Nudge: Improving decisions about health, wealth, and happiness*. New Haven, CT: Yale University Press.

Thapar, A., Collishaw, S., Pine, D. S., & Thapar, A. K. (2012). Depression in adolescence. *The Lancet*, *379*, 1056–1067. doi:10.1016/S0140-6736(11)60871-4

Thistlethwaite, D. L., & Campbell, D. T. (1960). Regression-discontinuity analysis: An alternative to the ex post facto experiment. *Journal of Educational Psychology*, *51*, 309–317.

Thomas, M. S., Fedor, A., Davis, R., Yang, J., Alireza, H., Charman, T., . . . Best, W. (2019). Computational modelling of interventions for developmental disorders. *Psychological Review*, *5*, 693–726.

Thompson, O. (2018). Head Start's long-run impact evidence from the program's introduction. *Journal of Human Resources*, *53*, 1100–1139.

Tipton, E., & Hedges, L. V. (2017). The role of the sample in estimating and explaining treatment effect heterogeneity. *Journal of Research on Educational Effectiveness*, *10*, 903–906. doi:10.1080/19345747.2017.1364563

Tipton, E., Yeager, D. S., Iachan, R., & Schneider, B. (2019). Designing probability samples to study treatment effect heterogeneity. In P. J. Lavrakas, M. W. Traugott, C. Kennedy, A. L. Holbrook, E. D. de Leeuw, & B. T. West (Eds.), *Experimental methods in survey research: Techniques that combine random sampling with random assignment* (pp. 435–456). Hoboken, NJ: Wiley.

Tough, P. (2019). *The years that matter most: How college makes or breaks us*. New York, NY: Houghton Mifflin Harcourt.

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.

U.S. Census Bureau. (2016). Table 8. In *Annual survey of school system finances*. Retrieved from https://www2.census.gov/programs-surveys/school-finances/tables/2016/secondary-education-finance/elsec16_sumtables.xls

Unterman, R., & Weiland, C. (2019). *Quantifying and predicting variation in the medium-term effects of oversubscribed prekindergarten programs*. New York, NY: MSRC. Retrieved from https://www.mdrc.org/publication/quantifying-and-predicting-variation-medium-term-effects-oversubscribed-prekindergarten

van Dis, E. A., van Veen, S. C., Hagenaars, M. A., Batelaan, N. M., Bockting, C. L., van den Heuvel, R. M., . . . Engelhard, I. M. (2019). Long-term outcomes of cognitive behavioral therapy for anxiety-related disorders: A systematic review and meta-analysis. *JAMA Psychiatry*, *77*, 265–273. doi:10.1001/jamapsychiatry.2019.3986

von Stumm, S., & Plomin, R. (2018). Monozygotic twin differences in school performance are stable and systematic. *Developmental Science*, *21*, e12694.

Walton, G. M., & Wilson, T. D. (2018). Wise interventions: Psychological remedies for social and personal problems. *Psychological Review*, *125*, 617–655.

Walton, G. M., & Yeager, D. S. (2020). Seed and soil: Psychological affordances in contexts help to explain where wise interventions succeed or fail. *Current Directions in Psychological Science*, *29*, 219–226. doi:10.1177/0963721420904453

Wanzek, J., Stevens, E., Williams, K., Scammacca, N., Vaughn, S., & Sargent, K. (2018). Current evidence on the effects of intensive early reading interventions. *Journal of Learning Disabilities*, *51*, 612–624. doi:10.1177/0022219418775110

Watts, T. W., Bailey, D. H., & Li, C. (2019). Aiming further: Addressing the need for high quality longitudinal research in education. *Journal of Research on Educational Effectiveness*, *12*, 648–658. doi:10.1080/19345747.2019.1644692

Watts, T. W., Ibrahim, D. A., Khader, A., Gandhi, J., & Raver, C. C. (2020). Exploring the effects of early childhood intervention on later school choice. *Educational Researcher*. Advance online publication. doi:10.3102/0013189X20935060

Wicherts, J. M. (2016). The importance of measurement invariance in neurocognitive ability testing. *The Clinical Neuropsychologist*, *30*, 1006–1016.

Wixted, J. T. (2004). The psychology and neuroscience of forgetting. *Annual. Review of Psychology*, *55*, 235–269.

Yeager, D. S., Hanselman, P., Muller, C., & Crosnoe, R. (2019). *Mindset × context theory: How agency and affordances interact to shape human development and group-based inequality*. Manuscript in preparation.

Yeager, D. S., Hanselman, P., Walton, G. M., Crosnoe, R., Muller, C. L., Tipton, E., . . . Dweck, C. S. (2019). A national experiment reveals where a growth mindset improves achievement. *Nature*, *573*, 364–369. doi:10.1038/s41586-019-1466-y

Yeager, D. S., Purdie-Vaughns, V., Hooper, S. Y., & Cohen, G. L. (2017). Loss of institutional trust among racial and ethnic minority adolescents: A consequence of procedural injustice and a cause of lifespan outcomes. *Child Development*, *88*, 658–676.

Yeager, D. S., & Walton, G. M. (2011). Social-Psychological Interventions in Education: They're not magic. *Review of Educational Research*, *81*, 267–301. doi:10.3102/003465431140599