ANNENBERG
BROWN UNIVERSITY

EdWorkingPaper No. 22-583

Design and Analytic Features for Reducing Biases in Skill-Building Intervention Impact Forecasts

Daniela Alvarez-Vargas
University of California, Irvine

Sirui Wan
University of California, Irvine

Lynn S. Fuchs
Vanderbilt University

Alice Klein
WestEd

Drew H. Bailey
University of California, Irvine

Despite policy relevance, longer-term evaluations of educational interventions are relatively rare. A common approach to this problem has been to rely on longitudinal research to determine targets for intervention by looking at the correlation between children's early skills (e.g., preschool numeracy) and medium-term outcomes (e.g., first-grade math achievement). However, this approach has sometimes over—or under—predicted the long-term effects (e.g., 5th-grade math achievement) of successfully improving early math skills. Using a within-study comparison design, we assess various approaches to forecasting medium-term impacts of early math skill-building interventions. The most accurate forecasts were obtained when including comprehensive baseline controls and using a combination of conceptually proximal and distal short-term outcomes (in the nonexperimental longitudinal data). Researchers can use our approach to establish a set of designs and analyses to predict the impacts of their interventions up to two years post-treatment. The approach can also be applied to power analyses, model checking, and theory revisions to understand mechanisms contributing to medium-term outcomes.

VERSION: June 2022

Suggested citation: Alvarez-Vargas, Daniela, Sirui Wan, Lynn S. Fuchs, Alice Klein, and Drew H. Bailey. (2022). Design and Analytic Features for Reducing Biases in Skill-Building Intervention Impact Forecasts. (EdWorkingPaper: 22-584). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/v8j6-2797

Design and Analytic Features for Reducing Biases in Skill-Building Intervention Impact Forecasts

Daniela Alvarez-Vargas[1], Sirui Wan[1], Lynn S. Fuchs[2], Alice Klein[3], & Drew H. Bailey[1]

[1]University of California, Irvine
[2]Vanderbilt University
[3]WestEd

Accepted for Publication at Journal of Research Educational Effectiveness on 04/21/2022

Author Note

The Number Knowledge Tutoring research was supported by 2 R01 HD053714 and Core Grant U54HD083211 from the Eunice Kennedy Shriver National Institute of Child Health & Human Development to Lynn S. Fuchs at Vanderbilt University. The content is solely the responsibility of the authors and does not necessarily represent the official views of the Eunice Kennedy Shriver National Institute of Child Health & Human Development or the National Institutes of Health. The Pre-K Mathematics research was supported by the Institute of Education Sciences, U.S. Department of Education through Grant R305K050004 to Alice Klein and Prentice Starkey at WestEd. The opinions expressed are those of the authors and do not represent views of the U.S. Department of Education. Drew Bailey is funded by a Jacobs Foundation Fellowship. Daniela Alvarez-Vargas is supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE-1839285. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation

Correspondence concerning this article should be addressed to Daniela Alvarez-Vargas, University of California, Irvine School of Education 401 E. Peltason Drive, Suite 3200, Irvine, CA 92617. Electronic mail may be sent to Dalvare5@uci.edu.

Abstract

Despite policy relevance, longer-term evaluations of educational interventions are relatively rare. A common approach to this problem has been to rely on longitudinal research to determine targets for intervention by looking at the correlation between children's early skills (e.g., preschool numeracy) and medium-term outcomes (e.g., first-grade math achievement). However, this approach has sometimes over—or under—predicted the long-term effects (e.g., 5th-grade math achievement) of successfully improving early math skills. Using a within-study comparison design, we assess various approaches to forecasting medium-term impacts of early math skill-building interventions. The most accurate forecasts were obtained when including comprehensive baseline controls and using a combination of conceptually proximal and distal short-term outcomes (in the nonexperimental longitudinal data). Researchers can use our approach to establish a set of designs and analyses to predict the impacts of their interventions up to two years post-treatment. The approach can also be applied to power analyses, model checking, and theory revisions to understand mechanisms contributing to medium-term outcomes.

Keywords: prediction, forecasting, non-experimental, intervention, evaluation

Design and Analytic Features for Reducing Biases in Skill-Building Intervention Impact Forecasts

Effective educational policy depends on evidence from the medium—to—long-term impacts of a proposed educational program or intervention (Martin et al., 2018). However, research on the medium and long term impacts of educational evaluations is scarce, difficult, and costly to conduct (Philips et al., 2017; Watts, Bailey, & Li, 2019). One solution to this problem has been to identify promising short-term outcomes (e.g., preschool numeracy) that are the most strongly correlated to later skills (e.g., first grade math achievement) using longitudinal observational data. However, forecasts about the medium-term impacts of interventions based on correlational analyses sometimes over-estimate (Bailey et al., 2018) or under-estimate (Li et al., 2020) the observed experimental impacts measured at medium-term follow up.

Experimental evaluations of skill-building interventions that successfully increased children's early math skills have not yielded the expected medium-term impacts that correlational work predicted. Instead, many successful early math interventions evaluated with randomized controlled trials (RCTs) demonstrate effect sizes – on multiple cognitive and achievement outcomes– that decrease by half or more in effect size units by just a year after program completion (for review, see Bailey et al., 2018; Li et al., 2020). Contradicting what we would expect from persistent and strong associations between early math skills and later math skills (e.g., Duncan, 2007; Jordan et al., 2012). These findings raise concerns about the usefulness of nonexperimental estimates for designing interventions for early academic skills to support children's later skill-development.

Our aim is to identify preferable intervention study designs – specifying how the measurement of covariates, the types of assessments used to measure skills, and the regression

models used– can best forecast the medium-term impacts of math skill-building interventions when short-term experimental data and medium-term nonexperimental data is available. We examine the performance of different forecasting approaches by comparing forecasted estimates to observed experimental benchmarks from two RCTs. We use the term forecasting, to conceptualize this exercise as an attempt to predict a future unknown event using data from previous observed events. Specifically, we attempt to predict the intervention impact two years after program completion (as if it was an unobserved event) using pretests scores and posttest scores (the previous observed events) to determine how well our forecasts correspond with the and the observed medium-term treatment impact.

To forecast medium-term effects, we run two series of regressions (shown in Figure 1 Panel B): we first regress a short-term outcome on a randomly assigned treatment to estimate the short-term treatment impact (Figure 1 $a_{experimental}$). Second, we regress a medium-term outcome (defined here as two years post-intervention[1]) on the short-term outcome using only data from the experimental control group to model a nonexperimental relation using the regression coefficient on the short-term outcome (Figure 1 $b_{nonexperimental}$). Third, we multiply the short-term treatment impact from the first regression by the relation between the short and medium-term outcome from the second regression to calculate our forecast ($a_{experimental} * b_{non\text{-}experimental}$). Lastly, we compare our forecasted estimate to the observed experimental benchmark from the original RCT (Figure 1 $c_{experimental}$). We try to identify best practices for using different measures of short and medium term outcomes and different covariates to reduce bias in the nonexperimental estimates of the effects of earlier skills on later math skills.

[1] Intervention designers may view impacts measured after two years of end-of-treatment as long-term impacts since the interventions were optimized to improve students' outcomes for up to one-year after end-of-treatment. On the other hand, many proposed benefits of early math instruction relate to children's longer-term outcomes. We find merit in both of these arguments and do not attempt a thorough critique of either of them here but see Bailey et al. (2020) and commentary by Schneider and Bradford (2020) for discussion of both views.

Forecast accuracy requires that a) the experimental treatment effect on the short-term outcome is unbiased (Figure 1 $a_{experimental}$), b) the causal relationship between the short-term outcome and the medium-term outcome is unbiased (Figure 1 $b_{non\text{-}experimental}$), and c) the short-term outcome measures capture the full causal pathway from the treatment to the medium-term outcomes. While random assignment can address bias in the $a_{experimental}$ path, the $b_{nonexperimental}$ path may be biased if common cause variables are omitted. For example, students' underlying ability may cause observed test scores in short and medium term tests, but controlling for a pretest score that provides a measure of students' ability can address some of the confounding that may occur between test scores.[2] However, the relations of early and later skills are often derived from nonexperimental data (i.e., public longitudinal datasets) to estimate a range of plausible impacts for power-analysis, to determine which skills to target with educational interventions, to predict the longer-run impacts of a policy change, or to estimate plausible impacts for theory testing and revision.

Since researchers evaluate the magnitude of experimental impacts on early skills relative to the relation of those early skills on indicators of later achievement. We provide a set of approaches that can leverage the strengths of experimental and nonexperimental data to enhance these types of forecasts and return to a specific example on these uses in the discussion section. Our work contributes new knowledge to current applied work in program evaluation (e.g., in calculating power to detect medium-term effects), intervention design (e.g., for identifying promising short-term outcomes to train and for power analysis for detecting longer-run impacts), to funding organizations interested in forecasting the effects of proposed interventions on student

[2] We would like to thank an anonymous reviewer for their suggestions for how to word this section, which substantially improved its clarity.

achievement years after the end of treatment, and for researchers and policy analysts attempting to model future program benefits.

How Can We Use Previous Research to Forecast Intervention Impacts?

A set of early skills that statistically predicts children's later academic achievement may represent a set of targets for potential intervention, with the intent that improving early skills would produce greater educational and economic returns to children. However, while longitudinal nonexperimental studies provide large nationally representative samples to develop and test theories about human development, they are limited by a lack of causally informative research designs, leaving any estimated relations between programs or skills on children's life outcomes susceptible to omitted variable bias (Bloom, Michaeloupoulos, and Hill, 2002). Thus, interventions targeting the skills that statistically predict later achievement may not necessarily produce the benefits predicted by these statistical models.

Experimental designs address the omitted variable bias problem by randomly assigning children to an intervention, so that on average, measured and unmeasured variables are equally distributed between treatment and control groups, allowing for an unbiased estimate of a causal impact. However, conducting randomized experiments to assess the causal impacts of an intervention on early skills and the effect of early skills on later skills is expensive. This is an important reason why evaluations of interventions that target specific skills and then follow participants for many years after treatment are scarce. One approach to forecasting the medium-term outcome of an intervention is using the short-term experimental outcomes and the correlations between short and medium outcomes from longitudinal data.

Forecasts are often implicitly made when predictive relations between a preceding variable and a later variable – calculated by regressing the later variable on the preceding

variable and covariates – are used to justify the potential usefulness of intervening on the preceding variable to improve the later variable (Grosz et al., 2020; Reinhart et al., 2013; Robinson et al., 2013). An impactful example of such a forecast comes from Duncan et al (2007) where nonexperimental estimates are used to argue that improvements to early math skills should yield higher levels of later academic achievement across domains.

We can extend this forecasting approach by combining short-run experimental impacts with internal or external estimates of the association between short-term and medium-term outcomes. For example, Deming (2009) calculated the estimated impact of Head Start participation on an index of outcomes in young adulthood and then multiplied this by the relation of the index of young adult outcomes on wages in adulthood from a separate cohort to project the impact on adulthood wages for Head Start participants. This is one way to mitigate the costs of collecting administrative data to observe the impact of programs on adult earnings. However, achievement-to-earnings correlations are substantially reduced when adjusting for a range of plausible confounders, highlighting the importance of how these effects are estimated (Watts, 2020).

Given these concerns, new forecasting methods have been developed to overcome the limitation of waiting to observe long-term impacts. Athey and colleagues (2019) test the accuracy of forecasting the impacts of a randomly assigned job assistance program – California's Greater Avenues to Independence conducted in the 1980s – on participant employment rates and earnings 9-years after the program's end. The authors found that by gathering measures of the impact of an employment program on employment rates and earnings across in the first 1.5 years into a surrogate index, they could forecast the mean impact on employment rates after nine years. Yet, there is a need to explore which analytical decisions improve forecast accuracy using both

experimental and nonexperimental data to predict child skill development and medium to long term program impacts. Ideally, identifying different approaches to forecast accurately would allow educational researchers to bypass the substantial time delay and resources required to observe the medium-term impacts of an educational intervention when evaluating interventions or making policy decisions. Below we describe potential sources of bias and methods to improve forecast accuracy.

Threats to Accurate Forecasts

We describe potential biases in intervention forecasts in the context of forecasting the impact of a first-grade math intervention with follow-up at third grade, using the short-term first-grade outcomes to predict the treatment impacts on children's math achievement at the end of third grade, as the medium-term outcomes. We also detail how threats to accuracy relate to the real-world limitations of evaluating interventions at scale. In Figure 2, we demonstrate our causal assumptions using directed acyclical graphical (DAG) notation (Pearl, 2009) where directed lines represent the causal impact of one variable on another. In contrast to Figure 1, where we show simplified analytical model forecasting, in Figure 2 we explain our conceptual assumptions about the ways in which measured and unmeasured variables may bias the forecast calculation.

In Figure 2, solid lines reflect measured relations between variables and dashed lines reflect assumed unmeasured relations. We conceptualize the causal impact of a treatment on a *short-term outcome test score* as the measured impact on *Skill 1* at time *T1* (the skill intended to be targeted by the intervention). Then, we assume Skill *1 T1* (end of first grade) to influence the same *Skill 1* at time *T2* (end of third grade), which we measure with the *medium-term outcome test score*. For this assumption to be true, *Skill 1 T1* must be a truly causal mediator of *Skill 1 T2*, whereby the causal effect of the treatment on *Skill 1 T2* is occurring through *Skill 1 T1*. In our

example, Skill 1 represents the same skill measured at an earlier and later time point making this a safe assumption to make. However, there is the possibility that there exist other mediators through which the treatment may impacts *Skill 1 T2*, which we discuss further below. In the simplest case, we would forecast the impact of the intervention on *Skill 1 T2* by multiplying the observed treatment impact on the short-term outcome test score by the estimated relation of a 1-unit change in the short-term outcome test score on the medium-term outcome test score (as shown in Figure 1 Panel B). In some cases, the estimated relation between short-term outcome test scores on the medium-term outcome test scores could be obtained from an external nonexperimental longitudinal dataset that includes similar tests and age ranges as the intervention study of interest. However, in the current study, we estimate the effect of short-term outcome test scores on the medium-term outcome test scores internally using the data from the control group of the randomized controlled trial (RCT). We continue to walkthrough Figure 2 in the following sections.

Sources of Over-Prediction

Multiplying short-term outcome treatment impacts by the short-to-medium term outcome relation, using only data from the experimental control group, may bias our forecast of the medium-term outcome treatment impacts upward for two reasons: omitted variable bias and over-alignment. Omitted variable bias can occur when forecasting does not account for an unmeasured variable that influences the correlation between both the short-term outcome and the medium-term outcomes. For example, unmeasured stable individual and environmental variables such as student working memory or family income plausibly cause student math test scores in first and third grade to some extent. Therefore, a forecast of third-grade math skills will over-state the impact of first grade skills if factors like student working memory are not

controlled for (Bailey et al., 2018). Omitted variable bias is shown in Figure 2 with dashed arrows pointing towards both *Skill 1 T1* and *Skill 1 T2*; we expected these to upwardly bias the relation of *Skill 1 T1* on *Skill 1 T2* in external nonexperimental datasets as well as in the control group of the experimental dataset that we use.

To reduce omitted variable bias, researchers may include an extensive set of individual and environmental covariates in their specifications, however the desirable covariates may not always be available in nonexperimental datasets.

The second potential cause for over-predicting medium-term outcomes is the over-alignment (or the extent of content overlap) of outcome measures with the content that was taught in the intervention. A test is over-aligned if it measures content taught in the intervention (e.g., fact memorization) that reflects a shallower understanding of the material than observed in similarly scoring children who did not receive the intervention (What Works Clearinghouse, 4.0). For example, a test would be over-aligned with an intervention to the extent that it measures student's memorization of this exact math problem 3 x 2 = 6; if this specific item was taught repeatedly in the intervention. In Figure 2, over-alignment is shown as the dashed line from treatment to the short-term outcome test score. Over-alignment would occur when the treatment increases a student's short-term test score (e.g., answering 3 X 2 = 6 correctly) but does not have the same impact on *Skill 1 T1* (e.g., being able to multiply). Over-alignment upwardly biases the estimated impact of the treatment on the short-term outcome test score which inflates impacts relative to the actual unobserved treatment impacts on *Skill 1 T1*.

Overstated improvements on high stakes testing may reflect score inflation and inappropriate test preparation (Koretz, 2001). Thus, educational interventionists have the difficult task of identifying conceptually proximal assessments that accurately measure the

specific knowledge targeted by and gained from the intervention without relying too much on material that is repeatedly presented during the intervention.

Sources of Under-Prediction

In some cases, addressing over-alignment bias may lead to underestimating the impact of a treatment on the medium-term outcome due to under-alignment bias. To address over-alignment bias, the What Works Clearinghouse recommends using outcome measures that are "broadly educationally relevant" (p.79) to capture a broad and comprehensive measure of skill change. However, interventionists often raise a valid concern with this approach; an under-aligned measure with content that is conceptually distal to the intervention may fail to capture growth. For example, if a multiplication intervention focuses on children's conceptual understanding of the multiplication procedure but assess the impact of the intervention with measures of multiplication, division, and geometry, the impacts on the broader measure may be less sensitive to children's growth of conceptual understanding of the multiplication procedure. In Figure 2, the dash-and-dot arrow from treatment to an unmeasured *Skill 2 T1* (conceptual understanding of multiplication in this example) reflects what occurs when the treatment influences a skill that is not measured with a short-term test score. Under-alignment downwardly biases the estimated impact of the treatment on the medium-term outcome by not failing to capture other contributing skills accurately with short-term outcome measures.

An important distinction between measures that are conceptually proximal to an intervention is that they may be well-aligned or over-aligned measures. We would not expect impacts on over-aligned measures to transfer to broader conceptually distal math assessments. For example, a student's memorization of a few math facts is not indicative of conceptually understanding multiplication thus a proximal measure of math fact memorization may not be

measuring the same thing in the treatment (e.g., memorizing math facts) and control group (e.g., arithmetic fluency). However, impacts on a proximal well-aligned measure of efficient multiplication strategies might forecast gains on a distal assessment of math knowledge, because the development of efficient multiplication may contribute to later math learning, and thus the proximal measure may not over-predict longer-run impacts in the presence of strong baseline covariates.

Current Study

The goals of the current study are to estimate the net direction of bias, its approximate magnitude, and how different approaches best reduce bias in our forecasts to better inform the design and study of effective interventions. Although we focus on math interventions, we believe this general approach can inform efforts to forecast the impacts of interventions in other areas of educational research. We examine the following research questions: (1) how do design features, specifically the inclusion of demographic and cognitive pretests, influence the accuracy of forecasts? (2) How do different analytical approaches to forecast the impact of early math skills and later math skills influence the accuracy of forecasts? (3) How do analytical decisions about the types of measures used to assess outcomes influence the accuracy of forecasts?

Hypotheses

Prior to addressing our research questions, we developed the following hypotheses of the specific design features and analytic decisions that we would expect to bias our forecasts of medium-term outcomes conditional on short-term outcomes.

(1) Using demographic and pretest covariates should reduce forecast bias. Estimates of the causal impact of an early math skill on a later math skill should approach the experimental benchmark when a full set of covariates is included in the model.

(2) Estimates from forecasts that assume that early math skills influence later math skills through the partially overlapping pathways (overlapping mediators) will yield smaller, more accurate forecasts than estimates from forecasts that assume that early math skills influence later math skills through fully independent pathways. Since increasingly complex mathematical concepts continue to build on basic number competencies modeling them separately might "double count" short-term impacts that manifest in more than one short-term outcome measure such as measuring the ability to add by using word problems or number sentences. We explain the two alternative modeling approaches in the analytical strategy.

(3a) Using the short-term outcomes that are conceptually proximal (closely aligned) with the intervention to calculate forecasts will yield over-estimated treatment impacts. Since conceptually proximal measures consist of items closely related to the narrower skills taught during the intervention, these skills will show more optimistic improvements than if we were to consider the complex impacts of all the untrained math skills that impact medium-term math achievement.

(3b) Using the short-term outcomes that are conceptually distal (less closely aligned) with the intervention to calculate forecasts will yield under-estimated treatment impacts. If we fail to measure the true extent of skill growth post-intervention by measuring a skill too broadly, we may expect a smaller impact in the medium-term math achievement than that which is observed.

(3c) We hypothesize that the most accurate forecasts would be calculated by using two short-term outcomes, one that is conceptually proximal and one that is conceptually distal to the treatment. Using both kinds of outcomes could serve to bracket the forecast since an optimal measure would consist of both a (1) proximal measure to capture variance in the skills targeted

by the intervention and a (2) distal measure to reduce over-alignment bias by accounting for relevant underlying skills not targeted by the intervention.

Method

Data Design

We conducted a secondary analysis of the Number Knowledge Tutoring (NKT) data. The NKT data were collected as part of a randomized controlled trial assessing a tutoring program's effects on first graders' emerging simple arithmetic competence (Bailey, 2019; Fuchs, Geary, et al., 2013). Students were randomly assigned within classrooms to either one of two treatment arms, where students received one-on-one tutoring on the conceptual basis for arithmetic paired with either speeded (treatment 1) or non-speeded practice (treatment 2), or to the control group who received business-as-usual instruction.

Participants

The sample includes 639 first-grade students from 40 schools and 233 classrooms in a southeastern metropolitan district who were evaluated as at-risk for having persistent math difficulties. Further description of the study participant recruitment and screening is available in Appendix A. We excluded 138 students who completed the pretests but did not complete all the short-term outcomes (n= 45; 7%) or all the medium-term outcomes (n=90; 14%). The remaining analytical sample consisted of 501 students that were mostly African American (70%), followed by white/Caucasian (19%), Hispanic (7%), and students of another race or who did not indicate a race (3%) who were grouped together as we cannot determine why the race indicator was missing. Half of the participants were male, most received free or reduced priced lunch (80%), and a few learned English as a second language (2%). Our analytical sample has a higher (2%) proportion of African American children and a smaller proportion of white children (1%),

mixed/other race children (2%), children receiving free or reduced-price lunch (4%) and English language learners (1%) than the original study sample (Fuchs et al., 2013). Our sample is thus similar but not identical to the Fuchs et al. (2013) sample as we included students that had at least one short-term outcome completed and at least one medium-term outcome completed. There were 17 cases of missing data for free-or-reduced price lunch and race, 421 cases had missing data on years that they received special education, and 20 cases had missing data on whether the student learned English as a second language. We created a separate variable as an indicator for missing cases to include the cases in all analyses. Each student with missing data for classroom was coded to have a unique identifier for classroom, such that we could cluster their standard errors at the classroom level.

Procedure

Students in both treatment groups were tutored one-to-one on the same content for 30-minute sessions three times a week for 16 weeks totaling 48 tutoring sessions from late October to March. The key difference between the treatment groups was the activity conducted during the last five minutes of the tutoring session. In the speeded practice condition students were encouraged to use the more efficient counting strategies to quickly answer math problems shown on flashcards within 90 seconds. In the non-speeded practice condition students were encouraged to use multiple different counting strategies (e.g., number lists, arithmetic principles, efficient strategies, manipulatives) to arithmetic problems presented in the form of a game and the tutor corrected any mistakes. A more detailed description of the study has been provided in Fuchs, Geary, et al. (2013).

The short-term outcomes, collected at the end of first grade, are measures of latent student skill 1 at time T1 as show in Figure 2. The short-term measures include different

assessments that are either conceptually proximal or distal to the content taught in the intervention. The medium-term outcomes were collected at the beginning of third grade, they represent an observable measure of latent student skill 1 at time T2 as shown in Figure 1. These medium-term outcomes include one measure that is conceptually proximal (e.g., Facts correctly Retrieved) to the intervention content and four measures that are conceptually distal (e.g. Number Sets, Wide Range Achievement Test–3 Arithmetic, Number Line, and Key-Math Numeration) to the intervention content. We further refer to conceptually proximal measures as outcome measures that assess skills that were closely related to the content that was taught to the treatment group. We further refer to conceptually distal measures as outcome measures that assess broad domain skills that consists of some, but not all, of the skills taught in the intervention.

Analytic Strategy

We used a within study comparison design (shown in Figure 2) to determine how well our forecasts of medium-term intervention impacts approximated the experimental benchmarks observed from the NKT program. We define medium-term impacts as the longest-run intervention impacts that were measured, which in this case were two years after end-of-treatment. All measures were standardized in control group standard deviations allowing comparisons of changes across time to the counterfactual condition. First, we estimated the experimental benchmarks ($c_{experimental}$ in Figure 2 Panel A path) by regressing each medium-term outcome on each of the two treatment conditions while controlling for child demographic and math pretest covariates using classroom level clustered standard errors. Second, we calculate the experimental impact on each short-term outcome (Figure 1 Panel B path $a_{Experimental}$) by regressing each short-term outcome on each of the two treatment conditions while controlling for

child demographic and math pretest covariates using classroom level clustered standard errors. The $a_{experimental}$ and $c_{experimental}$ regression coefficients and standard errors are shown in Table 1 in separate columns for each treatment condition. Third, we calculate the relation between each short-term outcome on each medium-term outcome (Figure 1 Panel B path $b_{non\text{-}experimental}$) by regressing each medium-term outcome on each short-ter, outcome, using only the control group data. The $b_{nonexperimental}$ paths were estimated differently based on the analytical approach (which we describe below in the model specification section), and the regression coefficients and standard errors are shown in Supplementary Table 3 and 4 in separate columns testing the sensitivity of these estimates to the addition of demographic and pretest covariates.

Fourth, we multiplied the short-term impact (Figure 1 Panel B path $a_{Experimental}$) by the relation between short-term test scores and medium-term test scores (Figure 1 Panel B path $b_{Non\text{-}experimental}$) to calculate the forecast. Calculating the forecasts entails numerous regression model specifications made at the researcher's discretion. We model alternative decisions about the covariates and measures used to explore how different analytical decisions relate to forecast accuracy.

Model Specifications

We attempt to identify an approach to forecasting to address the problem that developmental psychologists and educational program evaluators often encounter: "What is our best estimate for the longer-run impacts of an intervention, based on a pattern of observed or hypothetical short-term impacts of the intervention and the pattern of (partial) correlations between our short—and longer—term outcome measures?" In this case, long-run impacts of this hypothetical intervention have not already been observed in previous implementations (as required by Athey and colleagues, 2019). After discussion, we identified three conceptually

different variations of this approach, shown in Figure 3, that could be tested for their usefulness for forecasting the impact of an intervention on medium-term outcomes.

Model A: Forecasting Using a Single Short-term Outcome

The first approach we chose to model assumes that only a single short-term outcome was measured at the end-of-treatment, we show this in Figure 3 Model A and hereafter refer to this approach as forecasting using a single short-term outcome. In model A we estimate paths *a* and *b* using multiple linear regressions where: Path a_1 in Figure 3 Model A is the regression coefficient of treatment TRT_{iG1} on each short-term outcome STO_{iG1}

$$STO_{iG1} = \beta_0 + \beta_1 TRT_{iG1} + \beta_2 X_{iG1} + \epsilon_{iG1} + \mu_{G1}$$

where *i* represents individual students in *G1* first-grade classrooms, X_{iG1} is a vector of student demographic covariates and pretests, ϵ_{iG1} is a child level residual, μ_{G1} and is the classroom level residual since students are clustered in classrooms. Path b_1 is the regression coefficient of the short-term outcome STO_{iG1} on each medium-term outcome MTO_{iG3}

$$MTO_{iG3} = \beta_0 + \beta_1 EOT_{iG1} + \beta_2 X_{iG1} + \epsilon_{iG1} + \mu_{G1}$$

A key difference in the *b* path estimation is that these are estimated with only the control group data, and we estimate the impact of STO_{iG1} on MTO_{iG3} with the stepwise inclusion of covariates where X_{iG1} will include (1) no covariates, (2) covariates only, (3) covariates and proximal pretests, and (4) covariates, proximal, and distal pretests. The forecasted impact of each short-term outcome is calculated by multiplying path *a* and *b* for each combination of the 8 short-term outcomes predicting each of the 5 medium term outcomes for each of the two treatment arms, resulting in 80 forecasts that are plotted in Figure 4. Figure 4 presents the magnitude of the experimental benchmark on the x-axis and the forecasted impact on the y-axis; the diagonal line represents the trend we would expect if there were perfect correspondence

between the two. If forecasts are over-estimating the experimental benchmark, they will fall above the diagonal line; however, if they under-estimate the experimental benchmark they will fall below the diagonal line.

Model B: Forecasting Assuming Multiple Independent Effects

The second approach assumes that a medium-term outcome is independently influenced by different short-term outcomes, we show this in Figure 3 Model B and hereafter refer to this approach as forecasting assuming multiple independent effects. In Model B we estimate paths $a_{1\ldots n}$ and $b_{1\ldots n}$ in the same way as model A (exact model estimates are shown in Supplementary Table 3). However, the forecast for each medium-term outcome is calculated by multiplying paths a and b are for each of the 8 short-term outcomes, and then summed. This procedure is repeated for each of the 5 medium term outcomes. Because there are two treatment arms and 5 medium-term outcomes, this calculation yields 10 forecasts, which are plotted in Figure 5 Plot A. Both plots in Figure 5 follow the same plot formatting conventions as those in Figure 4.

Model C: Forecasting Assuming Multiple Non-Independent Effects

The third approach assumes that an intervention can impact a medium-term outcome through multiple dependent mediators with overlapping paths of influence from the short-term outcomes to the medium-term outcomes, we show this in Figure 3 Model C and hereafter refer to this approach as forecasting assuming multiple non-independent effects. In model C we estimate paths $a_{1\ldots n}$ and $b_{1\ldots n}$ using multiple linear regressions where Path a_1 in Figure 3 Model C is the regression coefficient of treatment TRT_{iG1} on each short-term outcome STO_{iG1}

$$STO_{iG1} = \beta_0 + \beta_1\, TRT_{iG1} + \beta_2\, X_{iG1} + \beta_3\, OSTO_{iG1} + \epsilon_{iG1} + \mu_{G1},$$ where $OSTO_{iG1}$

is a vector containing all the other short-term outcome measures? The *b* paths in Figure 3 are the regression coefficients on all the short-term outcomes $OSTO_{iG1}$ as predictors of each medium-term outcome MTO_{iG3}.

$$MTO_{iG3} = \beta_0 + \beta_1\ OSTO_{iG1} + \ \beta_2\ X_{iG1} + \epsilon_{iG1} + \mu_{G1}$$

Thus, Model C differs from Model B in that all the medium-term outcomes are simultaneously included in each regression equation to account for their covariance (exact model estimates are shown in Supplementary Table 4). The forecasts for Model C are shown in Figure 5 Plot B. In summary, forecasts from Figure 3 Model A are shown in Figure 4 and Supplementary Table 5 columns 2 and 3, forecasts from Figure 3 Model B are shown in Figure 5 Plot A and Supplementary Table 5 column 4, and forecasts from Figure 3 Model C are shown in Figure 5 Plot B and Supplementary Table 5 column 5. Of all the forecasts shown, Models A and C in Figure 3 performed best when pretest covariates – which account for omitted variables that confound the association between the short-term and medium-term outcomes– were included.

We tested if over-estimation bias from conceptually proximal measures and under-estimation bias from conceptually distal measures could be reduced using three heuristics. First, we consulted with the intervention's designer and classified conceptual proximity based on the extent to which the content in the measures overlapped with the content of the intervention. Then, we tested three heuristics (1) forecasting using the conceptually proximal short-term outcome with the smallest treatment impact, (2) forecasting using the conceptually distal short-term outcome with the largest treatment impact, and (3) forecasting using the average of both the conceptually proximal short-term outcomes with the smallest treatment impact and the conceptually distal short-term outcome with the largest treatment impact. It is important to note that the over- or under-prediction problem that we explore in this work may have potentially

different implications for educational interventions than fadeout. For example, fadeout implies that appropriate post-treatment supports may be necessary for sustaining impacts. On the other hand, the clearest implications for over- or under-prediction are methodological (e.g., including more baseline covariates and a range of outcome measures), rather than applied from a practitioner perspective. Still, understanding the sources of over- and under-prediction may be useful for improving practitioners' understanding of the mechanisms through which the long-run impacts of educational interventions emerge.

Bias Calculation

To identify the most accurate forecasts we estimate bias by subtracting the forecasts for each medium-term outcome from the experimental benchmark. In this calculation, the most accurate forecasts should yield a degree of total bias closer to 0. We follow Shadish, Clark, and Steiner (2008) in measuring absolute bias as the absolute difference between each forecast (Figure 2 Panel B) and the experimental benchmark (Figure 2 Panel A). Additionally, we calculate the average bias of the forecasts used to predict each medium-term outcome for each treatment. Lastly, we measure the accuracy of the forecasts of each medium-term outcome for each treatment as the average bias squared.

Measures

Students' age, sex, race, eligibility for free-or-reduced priced lunch, English learner status, and pretest scores for all measures were included as baseline covariates. A more detailed description of measures appears in Fuchs et al. (2013). Supplementary Table 1 lists the descriptive statistics for all the covariates included in our models split by condition. We follow Bailey and colleagues (2020) and the intervention designer's guidance in categorizing the short-term and medium-term mathematics outcome measures as measuring skills that are either

conceptually proximal or distal to the intervention. Conceptually proximal measures assess skills that were closely related to the content that was taught to the treatment group. Conceptually distal measures reflect assessments of a broad domain that consists of some, but not all, of the skills taught in the intervention. All these measures were used both as separate indicators and grouped as proximal or distal indicators to determine which combination of short-term outcomes would best forecast the treatment impact on the medium-term outcomes.

Conceptually Proximal Measures

The First-Grade Mathematics Assessment Battery (Fuchs, Hamlett, & Powell, 2003) was used to measure students' ability to add and subtract with digits from 5-12 with the Arithmetic Combinations subtests (Cronbach's α =.96) and with double digits like 28 + 48 with and without regrouping with the Double-Digit subtests (Cronbach's α = .94). It should be noted that, although we classify this measure as proximal, it was less proximal than the other measures in this category, because many students did not reach the lessons that addressed double-digit calculations, and instruction regarding double-digit calculation was minimal. The main difference between the treatment arms was that during the last 5 minutes of the speeded practice condition students played a game to meet or beat their score where they had 90 seconds addition (answers less than equal to 18) and subtraction problems (minuends less than equal to 18), whereas the non-speeded condition played non-speeded games on the same pool or arithmetic problems as in the speeded condition. Thus, children in the speeded condition answered more problems in the same timeframe. The Facts Correctly Retrieved assessment (from Geary et al., 2007) tests children's ability to answer simple addition problems verbally without the use of a pencil or paper and the use of efficient counting strategies. This measure is proximal to both the speeded and non-speeded treatment conditions because the efficient counting strategy was taught

and used in both treatment conditions. The total score is the amount of addition problems the students solved without using the counting fingers strategy. Overall, these three proximal measures broadly sampled first-grade mathematical content closely aligned with the intervention treatment arms which included units on addition and subtraction with problem sets of numbers from 5-12, adding double digit numbers from 10 to 19, and generating and solving story problems using addition and subtraction.

Conceptually Distal Measures

Measure of mathematics content not directly taught during the intervention and broader achievement tests were included as distal outcome measures. We categorize these measures as distal to the intervention because although they include some simple arithmetic, they also include broader mathematic problems to gauge performance relative to other students in older and younger grades. Thus, these tests measure skill in domains that were not explicitly taught in the intervention. The Number Sets Test (Geary, Bailey, & Hoard, 2009) measured students' speed and accuracy in operating with small numerosities of objects and linking them to the corresponding Arabic numeral. The test-retest reliability for the number sets test is .89 (Bailey et al., 2018) and this measure has been found to predict individual differences in math achievement more strongly than reading achievement (Geary, 2011), however it assesses a much broader numerosity construct than what was taught during the intervention. The Story Problems measure consists of 14-word problems that are read out-loud to students and requires them to combine, compare, or change two quantities to solve a simple arithmetic problem. Students have 30 seconds to answer the story problems and they can ask for the story to be re-read until they answer (Jordan & Hanich, 2000). This measure has a Cronbach's a = .86. The Wide Range Achievement Test–3 Arithmetic (WRAT-Arithmetic; Wilkinson, 1993) subtest measured

students' ability to answer calculation problems that increase in difficulty. Although, WRAT-Arithmetic contains a few items that are proximal to the content taught in 1st grade they also cover content that spans across multiple grades making them less sensitive to treatment effects and more distal to the intervention. KeyMath–Numeration (Connolly, 1998) was used to measure students' ability to orally respond to questions about identifying, sequencing, and relating numerals; problems were presented with increasing difficulty. Lastly, the Number Line Estimation 0-100 (Siegler & Booth, 2004) measured students understanding of relative numeric magnitudes. The percent absolute error from the position on the number line that the response is supposed to be is calculated for each student where lower score indicates better performance. To simplify the comparison between all the measures the scores were reverse coded so that higher numbers indicated better performance.

Results

Baseline Equivalence

Little's MCAR test did not provide strong evidence ($\chi 2 = 249.11$, $df = 249$, $p = 0.87$) to reject the null hypothesis that data are not missing completely at random (Little, 1988). Further, we conduct additional sensitivity analyses using case wise deletion and find results were robust to this estimation strategy (Supplementary Figure 4). Demographic information and test scores are shown in Supplementary Table 1 split by experimental condition. Students across the three experimental conditions did not significantly differ in baseline measured with the exception that more students were eligible for free-or-reduced price lunch in the control group than in the non-speeded practice group.

Replicating and Addressing Omitted Variable Bias

We hypothesized that using demographic and pretest covariates should reduce forecast inaccuracy caused by omitted variables bias by accounting for measures of confounding variables. In this study we are not concerned about omitted variables confounding treatment and outcomes, because treatments are randomly assigned, and pretest scores are available. However, associations between end of treatment skills and later skills are plausibly confounded by skills and environments that affect learning during this period but are not affected by the interventions. We modeled omitted variables bias by forecasting the medium-term impact of an intervention using the relation between a short-term outcome on a medium-term outcome calculated without covariates. By not accounting for common confounding variables that exert a positive influence on both short and medium term outcomes, such as previous knowledge, we illustrate the importance of addressing omitted variables bias. To demonstrate this, we plot our forecasted impacts on the y-axis and compare these to the experimental benchmarks on the x-axis in Figure 4. If the forecasts land on a value that is above the diagonal line this would indicate an over-estimation of the experimental benchmark, if the forecast falls below the diagonal line, this reflects an under-estimation of the experimental benchmark.

Figure 4, plot A shows forecasts calculated with a single short-term outcome and without any controls. The triangles and circles positioned above the diagonal line reflect over-estimated forecasts that predicted a treatment impact of 0.20 SD or more when the observed experimental benchmark reflected a treatment impact close to zero. Most of the over-estimated forecasts were calculated using conceptually proximal short-term measures (shown in black). Some forecasts landed along the diagonal line and others below the diagonal line demonstrating under-estimation. Most of the forecasts that landed below the diagonal were calculated by conceptually distal short-term measures (shown in gray). The average of all these forecasts (0.123 SD) is

shown in Table 2 Column (2), this is the value of the black square which is more than double the experimental benchmark of 0.052 SD. We only show the average of all the forecasts on the table for simplicity to allow for comparison across the three regression specification models (see Supplementary table 5 column 2). As we hypothesized excluding demographic and pre-tests yields largely over-estimated treatment impacts for most, but not all medium-term outcomes.

In contrast, once we include all the demographic and pretest covariates forecasts were reduced by 55% and approximated the experimental benchmark demonstrating a decrease in omitted variable bias (see Figure 4 plot B). All the 80 forecasts on this plot decreased when we introduced the covariates. If this were due to a reduction in noise, we would expect the forecast differences to go in different directions, however, we found that once we account for demographic and pretest covariates, the estimated forecasts were all reduced. As shown in Table 2 column (3) the average forecast is 0.056 SD which better approximates the experimental benchmark of 0.052 SD. Furthermore, the average forecast bias for each medium-term outcome is smaller than the average forecast bias in Table 2 column (2), except for forecast bias for the three conceptually distal measures: Number Sets, WRAT-Arithmetic, and the Number Line. The changes in average forecast bias hold across both treatment groups. However, three forecasts over-estimate the experimental benchmark by more than 0.20 SD, demonstrating that large errors are still present. Overall, we confirm our hypothesis that forecasts of the impact of an early math skill on a later math skill approach the experimental benchmark with a comprehensive set of baseline pretests and demographic variables are controlled.

Forecasting Approaches

The simplest methodological approach to forecasting is making predictions conditional on a single short-term outcome, as conceptually shown in Figure 3 panel A. Each marker on

Figure 4 plot B (black circles and triangles) reflect a single combination of one of the 8 short-term outcomes predicting one of the 5 medium-term outcomes including all covariates, when the average of all the forecasts 0.056 SD (Table 2 column 3) best approximates the average of all experimental benchmark (Table 2 column 1) of 0.052 SD. The values of each forecast are shown in Supplementary Table 5 column 3.

The second approach, shown conceptually in Figure 3 Panel B, assumes that the short-term outcomes are independent of each other and separately influence the medium-term outcome. In Figure 5 plot A, the 10 markers on the plot reflect the overall forecast for each medium-term outcome calculated as the sum of all the forecasts calculated from each short-term outcome. Even with a full set of covariates the forecasts under- and over-estimate the experimental benchmark by 0.20 SD to more than 0.60 SD. For simplicity, we consider the average forecast, shown on Table 2 column (4), is 0.444 SD which is 8.5 times larger than the experimental benchmark of 0.052 SD. This approach over-estimates all the medium-term outcomes, the raw forecast values and bias are shown in Supplementary Table 5 column (4).

The third approach, shown conceptually in Figure 3 panel C, assumes that the short-term outcomes are dependent on each other and together influence the medium-term outcomes. In Figure 5 plot B, the 10 markers on the plot reflect the forecast for each medium-term outcome calculated as the sum of all the estimated relations of the short-term outcomes. The average forecast, shown on Table 2 column (5), is 0.138 SD which is 2.7 times larger than the experimental benchmark of 0.052 SD. The raw forecast values and bias are shown in Supplementary Table 5 column (5).

By comparing the average forecasts for each of the three approaches to the average experimental benchmark we find that using a single short-term outcome to predict the medium-

term outcome yielded the most accurate forecasts. In comparison to the other approaches, using a single short-term outcome yielded 62 out of 80 forecasts within 0.20 SD of the observed experimental benchmark. The exact forecast values, mean bias, absolute bias, and accuracy calculated by using this method are shown in Supplementary Table 5 separately for each treatment, short-term outcome, and medium-term outcome. In contrast, forecasting assuming multiple non-independent effects yielded 9 out of 10 forecasts within 0.20 SD. As hypothesized, calculating forecasts assuming multiple non-independent pathways explaining the causal link between early math skills and later math skills yielded more accurate forecasts than forecasts assuming multiple independent causal pathways. This suggests it is important to model math development as contingent on numerous math skills that are mutually dependent.

Addressing Over- and Under-Alignment Bias

We hypothesized that the short-term outcomes that are more conceptually proximal with the intervention will yield over-estimated forecasts whereas the conceptually distal short-term outcomes would yield under-estimated forecasts. Figure 4 plot B demonstrates that the proximal measures (black markers) have the highest forecasts. However, these both over-estimate and under-estimate the experimental benchmark. The highest forecast value shown in Supplementary Table 5 column 3 is 0.279 SD the lowest is -0.022 SD, when the experimental benchmark is 0.052 SD. Similarly, the conceptually distal short-term measures over-estimate and under-estimate the experimental benchmark, but to a lesser extent, with the highest forecast being -0.012 SD and the lowest being 0.138 SD. Therefore, in line with our hypothesis, conceptually proximal short-term measures over-estimate treatment impacts more than conceptually distal short-term measures. Additionally, most conceptually distal measures under-estimated the treatment impacts. However, some conceptually proximal measures and some distal measures

both over-estimate and under-estimate the experimental benchmark. Of the three different heuristics we modeled in Supplementary Figure 1, we find that by calculating forecasts with a combination of one conceptually proximal measure and one conceptually distal measure, shown by the orange markers, we estimate the treatment impact within .10 SD from the observed experimental benchmark.

Regarding treatment, we find that among the 37 forecasts that over-estimated the medium-term treatment impacts in the NKT study, 29 were from the speeded condition and 8 were from the non-speeded condition. The opposite trend was true in the 42 forecasts that were under-estimated, where 10 were from the speeded condition and 32 were from the non-speeded condition. This finding suggests that the outcome measures were more closely aligned with the speeded treatment condition than with the non-speeded treatment condition, thus we tended to over-estimate forecasts for the speeded condition and under-estimate forecasts in the non-speeded condition.

We hypothesized that if we calculate forecasts using the exact same short-term and medium-term outcomes, we would over-estimate the impact, if the tests were proximal, and under-estimate the impact if the tests were distal. However, we found that in the speeded condition, forecasts using the same tests longitudinally over-estimated the forecast regardless of conceptual proximity. Overall, the average forecasts using the same tests are 0.076 SD and the average forecasts using different tests are 0.056 SD when the experimental benchmark is 0.052 SD. Therefore, the different interventions showed different evidence of over-alignment bias, with forecasts of the speeded practice impact showing more evidence of over-alignment bias than forecasts from the non-speeded practice condition. The finding that different activities in the last five minutes of treatment sufficiently yielded different patterns of impact forecasts calculated

from the exact same measures implies that over-alignment is an important factor to consider when forecasting.

Replication

We replicated the analysis using data from a study of the Pre-K Mathematics (PKM) intervention (Starkey et al., 2020) to determine if our hypotheses were supported. The PKM data were collected as part of a randomized controlled trial examining the effects of an early math curriculum on pre-K children's mathematical knowledge. Children were assessed with pre- and post-tests in pre-K and again at the end of first grade allowing us to conduct a within study comparison to compare forecasts of first grade impacts (medium-term outcome) conditional on pre-K end-of-treatment outcomes (short-term outcome). Details about the study sample and measures are available in the online supplementary material.

Like the NKT dataset, the PKM data demonstrated that by accounting for confounding variables such as demographics, general ability pretests, and math pretests, forecast bias was reduced by 41% and approximated the experimental benchmark demonstrating a decrease in omitted variable bias (see Supplementary Figure 2, Plot B). Furthermore, we found similar patterns of accuracy using the three different approaches to forecast medium-term outcomes. As shown in Supplementary Figure 3, we found that calculating forecasts assuming multiple non-independent causal pathways yielded a more accurate forecasts than assuming multiple independent causal pathways (see Supplementary Table 9 for estimate comparison). Though the PKM was limited to two short-term outcomes – one conceptually proximal to the intervention and one conceptually distal – we still found that using the heuristic of forecasting using the average of both measures yielded the most accurate forecast of 0.21 SD when the experimental

benchmark in this dataset was 0.04 SD, meaning there was still an upward bias by 0.17 SD (see Supplementary Figure 5).

The PKM study forecasts were less sensitive to the conceptual proximity of the short-term measures as both the proximal and distal measures over-predicted the experimental benchmark (Supplementary Table 12). We believe over-prediction could be due, in part, to estimation error caused by omitted variables in the PKM study. The PKM study had fewer baseline measures and plausibly noisier baseline pretests due to the younger sample in contrast to the NKT sample.

Although we found support for hypotheses 1 and 2, both the more proximal and more distal measures led us to over-estimate the experimental benchmark in the PKM data. Several sources of evidence suggest that omitted variable bias remained a major concern in the PKM reanalysis. First, although the short-term impacts in both datasets were of similar average magnitudes (0.34 in NKT and 0.40 in PKM, from Table 1 and Supplementary Table 10, respectively), the forecasts for each of the short-term outcomes in the PKM dataset under full controls (0.35 and 0.48; Supplementary Table 11) would have been the second and sixth largest forecasts in the NKT dataset (Supplementary Table 2). Second, whereas the magnitudes of the forecasts leveled off after adding the first set of pretests within the NKT dataset (Supplementary Table 2, last 2 columns) suggesting that key confounds had been successfully accounted for by pretests, they continued to drop in the PKM dataset (Supplementary Table 11, last 2 columns) suggesting the potential for additional drops if more pretests had been available. We return to the implications of these discrepancies in the discussion section. Overall, these findings support the importance of including pretest measures that are conceptually proximal to the skills that the intervention is designed to improve to reduce bias in forecasting medium-term outcomes.

Discussion

In the present study we demonstrated prevailing threats to forecasts accuracy due to omitted variables bias, measurement over-alignment, and measurement under-alignment. We modeled the direction and magnitude of bias finding that demographic variables that are correlated to the pretests and post-tests of the skills measured are necessary covariates but not sufficient to improve the predictive accuracy of our forecasts. Furthermore, we found that over-alignment and under-alignment influenced both forecast over-estimation and under-estimation with patterns favoring over-estimation for proximal measures and under-estimation for distal measures, however these were not as consistent as we hypothesized and, in some cases, proximal measures under-predicted while distal measures over-predicted outcomes. In an exploratory analysis, the most accurate forecasts were calculated using both a single conceptually proximal and distal short-term outcome. However, this approach was not validated in the replication, where omitted variables bias was not fully reduced.

Forecast models based on assumptions of early math skills influencing later math skills through independent direct causal pathways yielded severe over-estimations. Forecast models that assumed mutually dependent direct causal pathways were more accurate, yet not as accurate as models using one or two short-term outcomes. These results demonstrate that in this case, early math skills influenced later math skills via largely overlapping pathways. Interestingly, using two short-term outcomes based on their theoretical alignment with the intervention yielded more accurate forecasts than using all short-term outcomes assuming multiple dependent pathways. We hypothesize this may be due to the additional omitted variables that confound the relation between short and medium term outcomes.

By assessing multiple measures in the NKT as short-term outcomes, we found that the measures that were most conceptually proximal to the intervention over-estimated, while measures more distal to the intervention most often underestimated the experimental benchmark. However, this pattern differed in the PKM dataset (Starkey et al., 2020). Although the CMA was more conceptually proximal to the intervention than the TEMA-3 in that analysis, we found that both measures over-estimated the medium-term treatment impact. The variation in accuracy across the two studies partially reflects the real-world constraints of gathering sufficient measures from interventions to forecast medium-term impacts. Still, results suggest that researchers should be wary of forecasting (or making claims about the importance of an intervention for future outcomes) based on a single proximal assessment, particularly in the absence of comprehensive baseline statistical controls. We attempt to reconcile these findings below.

Explaining Different Findings in the Two Datasets

One major difference in findings across the two datasets was that when we forecast using the combination of one conceptually proximal measure and one conceptually distal measure the NKT forecasts were reasonably accurate, on average, within 0.10 SD of the experimental impact. However, in the PKM dataset (Starkey et al., 2020), this approach yielded a less accurate forecast of 0.21 SD, being 0.17 SD bigger than the experimental impact of 0.04 SD. This discrepancy appears to be at least partially explained by greater omitted variable bias in the PKM dataset, although we cannot rule out estimation error as a contributor as well. There are significant differences in the two datasets that may help explain the differences in forecast accuracy. First, the PKM intervention evaluated the impact of a curriculum intervention for all pre-K children, in contrast, the NKT intervention evaluated the impact of a tutoring program

targeting a narrower population of at-risk children. These differences in intervention designs reflect real-world constraints that precluded the PKM study from being able to collect as many pretests and short-term outcome measures as the NKT study. In the PKM evaluation, entire preschool classrooms had to be tested before the intervention began in such a way that limited class-time interruptions. Further, PKM children were two years younger than NKT children. Thus, the PKM evaluation was limited to five measures of children's cognitive skills at baseline. This contrasted with the NKT, which tested only a subset of students from each classroom individually and collected fourteen measures at baseline. The lower number of baseline pretests, coupled with the likely assumption that baseline pretests in the younger PKM sample are noisier than in the older NKT sample, raises the possibility that we could not account for residual bias from omitted variables in the PKM data as well as we could in the NKT data.

Taken together, findings point to the importance of considering multiple competing biases in forecasting. The differences between the two datasets correspond to real-world constraints. Results suggest that nonexperimental longitudinal studies designed for theory development and testing should (1) be concerned with strong baseline measures of children's domain general cognitive skills (Geary, 2011), and 2) consider a mix of specific cognitively informed assessments (which might stand in as "proximal" measures for an interventionist hoping to forecast medium-term effects based on a hypothesized developmental model and plausible short-term impact effect size) and broad achievement measures (which will likely serve as "distal" measures of achievement for any educational intervention). If a comprehensive set of baseline measures is available, averaging across forecasts from proximal and distal short-term outcome measures may balance biases from over- and under-alignment, as suggested by our reanalysis of the NKT data. If a comprehensive set of baseline measures is not available, the

results of our reanalysis of PKM data (Starkey et al., 2020) suggest that distal measures with smaller forecasted short-term impacts will yield more accurate forecasts of medium-term impacts.

Tentatively, we hypothesize that omitted variable bias is a harder problem to solve in preschool aged children because of the difficulty of giving a comprehensive battery of pretest assessments and more measurement error. Whereas under-alignment might be more concerning in later grades, when skills may be more differentiated from each other. However, we do not offer a strong confirmatory test of this hypothesis in this paper.

Potential Uses

There are at least three research applications of our presented approach: power-analysis, model checking, and theory revision. In our current research we estimated the treatment impact of the Number Knowledge Tutoring speeded practice on children's counting strategies measured by Facts Correctly Retrieved (0.39 SD, Table 1); then, we estimated the effect of a hypothetical 1 SD change to Facts Correctly Retrieved in first grade on Facts Correctly Retrieved in third grade using the control group data and full covariates (0.22 SD, Supplementary Table 3). Using the approach of forecasting using a single independent short-term outcome, we forecasted the treatment impacts 2-years after the end-of-treatment to be (0.39*0.22= 0.09 SD). For a researcher planning a similar intervention that projected an end of treatment impact of approximately 1 SD, this would justify a sample size adequate to detect a 0.09 SD effect size in third grade. The researcher might compare this forecast to another forecast based on a hypothetical intervention strategy that targets a different broader set of skills or children of different age groups. A researcher who estimates a model predicting later skills from earlier skills who finds an estimate substantially larger than .22 (perhaps closer to the zero-order

correlation between first and third grade Facts Correctly Retrieved scores) should consider whether omitted variables might be biasing this and other estimates in the model upward and might consider alternative estimation strategies for addressing them. Finally, when this method fails, it suggests the importance of theory revision. When, after observing longer-term impacts, forecasts were overly optimistic, this suggests the existence of omitted variables, some of which may be targeted by successful interventions. When forecasts are overly pessimistic, this may suggest that under-alignment is a concern and that a better understanding of the underlying mechanisms might improve theories of development within the skill domain(s) under study. For example, understanding the sources of over- and under-prediction may be useful for improving practitioners' understanding of the mechanisms through which the long-run impacts of educational interventions emerge.

Limitations

Our analyses suggest that, for the combination of interventions, outcomes, and estimation strategies under consideration, some forecasting approaches may be predictably more or less biased than others. However, it is also important to note that in this study, forecasts were not strongly calibrated with observed impacts within this range of observed impacts, as reflected by the weak association (r =.04) between forecasts and impacts in Figure 4 Plot B. We hypothesize this at least in part reflects the narrow range of longer-term impacts observed in the current study but note that our methods is likely less useful for making forecasts of impacts relative to each other than relative to other benchmarks (e.g., 0 or a forecast developed on the basis of proximal measures alone).

Future Directions

Additional methods may improve the accuracy of forecasting above and beyond the methods we have tested in the current study. One future direction of this work will be to further investigate the psychometric properties of the different measures used to calculate the forecasts to ensure that measures are comparable across groups and time and to assess whether models that allow for the possibility of group differences in measurement models yield more accurate forecasts. This approach would allow for identifying and modeling changes in multiple latent variables with different effects on proximal or distal measures to determine if this improves forecast accuracy.[3] In addition, although we think the current study adds value by demonstrating the importance of considering omitted variables and alignment for generating accurate forecasts, in using a within-study design approach, we did not establish the validity of this approach for use across datasets. For the approach to be most useful, it must be able to provide accurate forecasts when the units and settings in the nonexperimental dataset differ from those from the experimental dataset. Such findings would increase our confidence in our ability to transport forecasts generated from estimates in large longitudinal datasets to the population of interest. Although prior work suggests some regularity across datasets in the ratio of end-of-treatment impacts to later impacts of early math interventions (Bailey et al., 2018), the ability of these methods to capture systematic variation in patterns of impacts across units, treatments, and settings has not been investigated. This is an important direction for future work.

Implications

The practical significance of educational interventions is partially known only with additional work to determine how present findings compare to other interventions and their utility in promoting future outcomes. Improving the accuracy of our forecasts of the medium-

[3] We thank an anonymous reviewer for the idea to pursue this as a future direction.

term impacts using observed end-of-treatment impacts could lead to more efficient design and investment in educational interventions. Forecasting not only better informs policy decisions about what educational interventions to fund, it can also be adapted to inform statistical power calculations in intervention evaluation, to provide a risky test to corroborate theories of causal processes (Meehl & Waller, 2002), and to foster transparency in research communication to aide belief confirmation or revision (DellaVigna et al., 2019). We thus provide a simple approach to forecasting the treatment impact of early math skills on later math skills as a method in need of replication across different applications and contexts to improve the accuracy of forecasts utilizing experimental and nonexperimental work.

References

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 (No. w26463). National Bureau of Economic Research.

Bailey, D. H., Duncan, G. J., Cunha, F., Foorman, B. R., Yeager, D. S. (2020). Persistence and fade-out of educational intervention effects: Mechanisms and potential solutions. Psychological Science in the Public Interest, 21, 55–97. doi:10.1177/1529100620915848

Bailey, D. H., Duncan, G. J., Watts, T., Clements, D. H., & Sarama, J. (2018). Risky business: Correlation and causation in longitudinal studies of skill development. American Psychologist, 73(1), 81–94. https://doi.org/10.1037/amp0000146

Bailey, D. (2019). Explanations and Implications of Diminishing Intervention Impacts Across Time. In Cognitive Foundations for Improving Mathematical Learning (pp. 321–346). Elsevier. https://doi.org/10.1016/B978-0-12-815952-1.00013-X

Bloom, H. S., Michalopoulos, C., Hill, C. J., & Lei, Y. (2002). Can Nonexperimental Comparison Group Methods Match the Findings from a Random Assignment Evaluation of Mandatory Welfare-to-Work Programs? Washington, DC: Manpower Demonstration Research Corporation.

Connolly, A.J. (1988). KeyMath Revised. Circle Pines, MN: American Guidance Service.

DellaVigna, S., Pope, D., & Vivalt, E. (2019). Predict science to improve science. Science, 366(6464), 428-429. https://doi.org/10.1126/science.aaz1704

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

Duncan, G. J., Dowsett, C. J., Claessens, A., Magnuson, K., Huston, A. C., Klebanov, P., ... & Japel, C. (2007). School readiness and later achievement. Developmental psychology, 43(6), 1428.

Fuchs, L. S., Geary, D. C., Compton, D. L., Fuchs, D., Schatschneider, C., Hamlett, C. L., Deselms, J., Seethaler, P. M., Wilson, J., Craddock, C. F., Bryant, J. D., Luther, K., & Changas, P. (2013). Effects of First-Grade Number Knowledge Tutoring With Contrasting Forms of Practice. Journal of educational psychology, 105(1), 58–77. https://doi.org/10.1037/a0030127Geary D. C. (2011). Cognitive predictors of achievement growth in mathematics: a 5-year longitudinal study. Developmental Psychology, 47(6), 1539–1552. https://doi.org/10.1037/a0025510

Fuchs LS, Hamlett CL, Powell SR. First-Grade Mathematics Assessment Battery. L. S. Fuchs; 228 Peabody, Vanderbilt University, Nashville, TN 37203: 2003.

Geary DC. Cognitive predictors of individual differences in achievement growth in mathematics: A five-year longitudinal study. Developmental Psychology. 2011; 47:1539–1552. doi: 10.1037/a0025510.

Geary DC, Bailey DH, Hoard MK. (2009) Predicting mathematical achievement and mathematical learning disability with a simple screening tool: The Number Sets Test. Journal of Psychoeducational Assessment. 27:265–279.

Geary DC, Hoard MK, Byrd-Craven J, Nugent L, Numtee C. (2007) Cognitive mechanisms underlying achievement deficits in children with mathematical learning disability. Child Development. 78:1343–1359. doi: 10.1111/j.1467-8624.2007.01069. x.

Grosz, M. P., Rohrer, J. M., & Thoemmes, F. (2020). The taboo against explicit causal inference in nonexperimental psychology. Perspectives on Psychological Science, 15(5), 1243-1255.

Ginsburg, H. P., & Baroody, A. J. (2003). TEMA-3: Test of Early Mathematics Ability–Third Edition.

Jordan NC, Hanich L. Mathematical thinking in second-grade children with different forms of LD. Journal of Learning Disabilities. 2000; 33:567–578. doi: 10.1177/002221940003300605.

Klein, A., Starkey, P., & Ramirez, A. (2002). Pre-K Mathematics Curriculum. Glendale, IL: Scott Foresman.

Koretz, D. M., McCaffrey, D. F., & Hamilton, L. S. (2001). Toward a Framework for Validating Gains Under High-Stakes Conditions. American Psychological Association. https://doi.org/10.1037/e648282011-001

Li, Weilin, Greg J. Duncan, Katherine Magnuson, Holly S. Schindler, Hirokazu Yoshikawa, and Jimmy Leak. (2020). 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). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/5tvg-nt21

Little, R. J. (1988). A test of missing completely at random for multivariate data with missing values. Journal of the American statistical Association, 83(404), 1198-1202. https://doi.org/10.1080/01621459.1988.10478722

Martin, J., McBridge, T., Brims, L., Doubell, L., Pote, I., & Clarke, A. (2018). Evaluating early intervention programmes: Six common pitfalls, and how to avoid them. Retrieved from

Early Intervention Foundation website: http://www.eif.org.uk/publication/evaluating-early-intervention-programmes-six-common-pitfalls-and-how-to-avoid-them"

Meehl, P. E., & Waller, N. G. (2002). The Path Analysis Controversy: A new statistical approach to strong appraisal of verisimilitude. *Psychological methods*, *7*(3), 283.

Milburn, T. F., Lonigan, C. J., DeFlorio, L., & Klein, A. (2019). Dimensionality of preschoolers' informal mathematical abilities. Early Childhood Research Quarterly, 4 (2), 487-495. https://doi.org/10.1016/j.ecresq.2018.07.006

Pearl, J. (2009). Causality. Cambridge university press.

Phillips, D. A., Lipsey, M. W., Dodge, K. A., Haskins, R., Bassok, D., Burchinal, M. R., …Weiland, C. (2017). The current state of scientific knowledge on pre-kindergarten effects. Retrieved from Brookings website: https://www.brookings.edu/wp-content/uploads/2017/04/duke_prekstudy_final_4-4-17_hires.pdf

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

Robinson, D.H., Levin, J.R., Schraw, G. et al. On Going (Way) Beyond One's Data: A Proposal to Restrict Recommendations for Practice in Primary Educational Research Journals. Educ Psychol Rev 25, 291–302 (2013). https://doi.org/10.1007/s10648-013-9223-5

Schneider, B., & Bradford, L. (2020). What We Are Learning About Fade-Out of Intervention Effects: A Commentary. Psychological Science in the Public Interest, 21(2), 50-54.

Shadish, W. R., Clark, M. H., & Steiner, P. M. (2008). Can Nonrandomized Experiments Yield Accurate Answers? A Randomized Experiment Comparing Random and Nonrandom Assignments. Journal of the American Statistical Association, 103(484), 1334–1344.

Siegler, R. S., & Booth, J. L. (2004). Development of numerical estimation in young children. *Child development, 75*(2), 428-444.

Starkey, P., Klein, A., DeFlorio, L., & Beliakoff, A. (2020). Scaling Up the Pre-K Mathematics Intervention in Public Preschool Programs [Manuscript submitted for publication]. WestEd.

Starkey, P., & Klein, A. (2012). Scaling up the implementation of a pre-kindergarten mathematics intervention in public preschool programs. Final Report: IES Grant R305K050004. National Center for Educational Research, US Department of Education.

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(4), 648-658.

What Works Clearinghouse™ Standards Handbook (Version 4.0). (n.d.). 130.

Wilkinson GS. *Wide Range Achievement Test 3.* 3. Wilmington, DE: Wide Range; 1993.

Tables

Table 1

Number Knowledge Tutoring Treatment Impacts on Short and Medium Term Outcomes

Outcome	Speeded v. Control Estimate (SE)	Non-Speeded v. Control Estimate (SE)
Short-Term Outcome (Spring 1st Grade) *a Experimental*		
Proximal Content		
Arithmetic Combinations	0.95*** (0.10)	0.50*** (0.08)
Double-Digit Calculations	0.81*** (0.11)	0.59*** (0.09)
Facts Correctly Retrieved	0.39*** (0.10)	0.20 (0.10)
Distal Content		
Number Sets	0.33*** (0.10)	0.28** (0.09)
Story Problems	0.22* (0.10)	0.29** (0.10)
WRAT-Arithmetic	0.34*** (0.06)	0.34*** (0.07)
Number Line	0.11 (0.09)	0.03 (0.09)
KeyMath-Numeration	0.10 (0.07)	0.07 (0.08)
Medium-Term Outcome (Spring 3rd Grade) *c Experimental*		
Proximal Content		
Facts Correctly Retrieved	-0.00 (0.10)	0.03 (0.10)
Distal Content		
Number Sets	0.09 (0.10)	0.12 (0.08)
WRAT-Arithmetic	0.02 (0.09)	0.09 (0.09)
Number Line	-0.02 (0.09)	0.07 (0.10)
KeyMath-Numeration	0.04 (0.09)	0.07 (0.08)

Note. N= 501. * p < .05 ** p < .01 *** p < .001. Treatment groups were entered as dummy variables in which (Speeded = 1, Control = 0) and (non-Speeded = 1, Control = 0). Demographic controls are race/ethnicity, sex, free or reduced lunch status, and whether the student learned English as a Second Language. Missing demographic variables were coded as missing dummy variables and included as covariates. Participants were nested in grade 1 classrooms, so we used classroom level clustered standard errors. Standardized effects are in control group standard deviation units. Number line was reverse coded, so higher scores reflect stronger performance.

Table 2

Average Forecasts Using Three Approaches and Resulting Bias

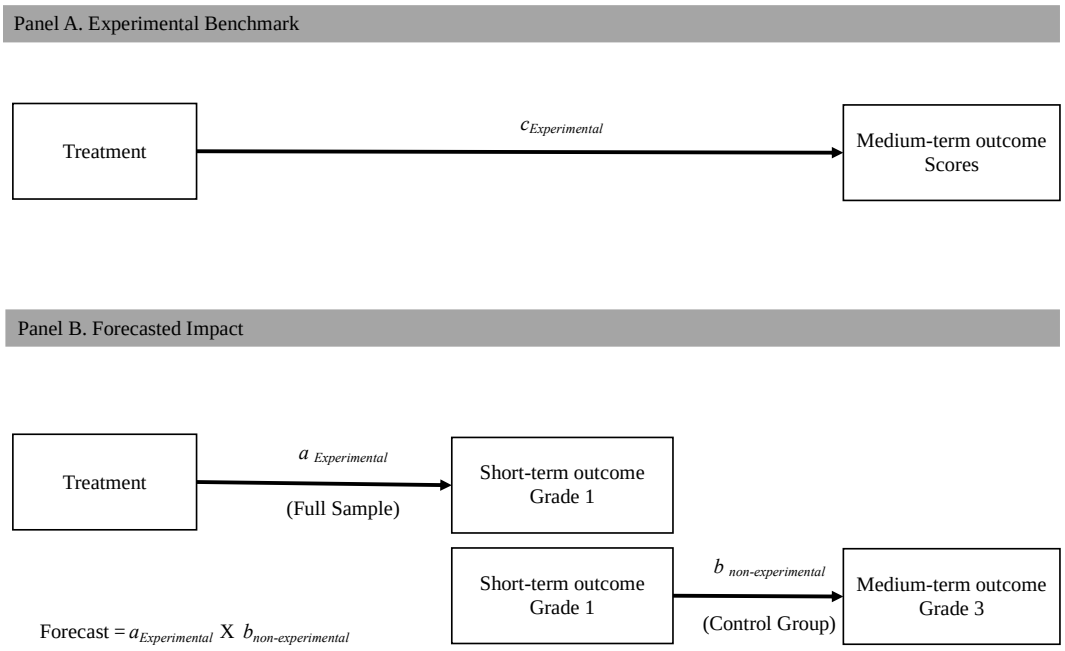
	Experimental Benchmark	Forecast Independent Single STO Outcome				Multiple Independent STO Outcome		Multiple Dependent STO Outcomes	
	(1)	(2)		(3)		(4)		(5)	
Medium-term Outcome	Estimate	Average Forecast	Average Bias	Average Forecast	Average Bias	Average Forecast	Average Bias	Average Forecast	Average Bias
Speeded Treatment									
Facts Correctly Retrieved	-0.003	0.123	0.100	0.056	0.04	0.444	0.301	0.138	0.164
Number Sets	0.087	0.123	0.083	0.056	-0.007	0.444	0.555	0.138	0.066
WRAT-Arithmetic	0.023	0.123	0.131	0.056	0.037	0.444	0.455	0.138	0.11
Number Line	-0.018	0.123	0.145	0.056	0.082	0.444	0.529	0.138	0.169
KeyMath-Numeration	0.043	0.123	0.130	0.056	0.044	0.444	0.655	0.138	0.192
Non-speeded Treatment									
Facts Correctly Retrieved	0.03	0.123	0.037	0.056	-0.007	0.444	0.154	0.138	0.055
Number Sets	0.121	0.123	-0.002	0.056	-0.063	0.444	0.34	0.138	-0.001
WRAT-Arithmetic	0.093	0.123	0.018	0.056	-0.048	0.444	0.266	0.138	0.028
Number Line	0.072	0.123	0.016	0.056	-0.03	0.444	0.263	0.138	0.007
KeyMath-Numeration	0.07	0.123	0.054	0.056	-0.01	0.444	0.41	0.138	0.074
Full Covariates	X			X		X		X	

Note. STO = Short-term outcome. Table compares observed treatment impacts on medium-term outcomes split by treatment, to forecasts calculated using four approaches (columns 2 to 5) and to heuristics applied to forecasting with a single short-term outcome (columns 6 to 8). In columns 2 to 5 the average forecast is shown as the total average of all the forecasts calculated using this approach for simplicity; Full table available in Supplementary Table 4. The average bias is also shown to demonstrate the average deviation of each forecast from the experimental benchmark, the bigger the bias the more inaccurate the forecast. In columns 6 to 8 the raw forecast is included instead because only one forecast was calculated using each heuristic for each medium-term outcome. Additionally, the raw bias is shown for each heuristic as forecast minus the experimental benchmark. The last row indicates the forecasts and heuristics estimated using all the covariates including demographic variables and pretests for all short and medium term outcomes.

Figures

Figure 1

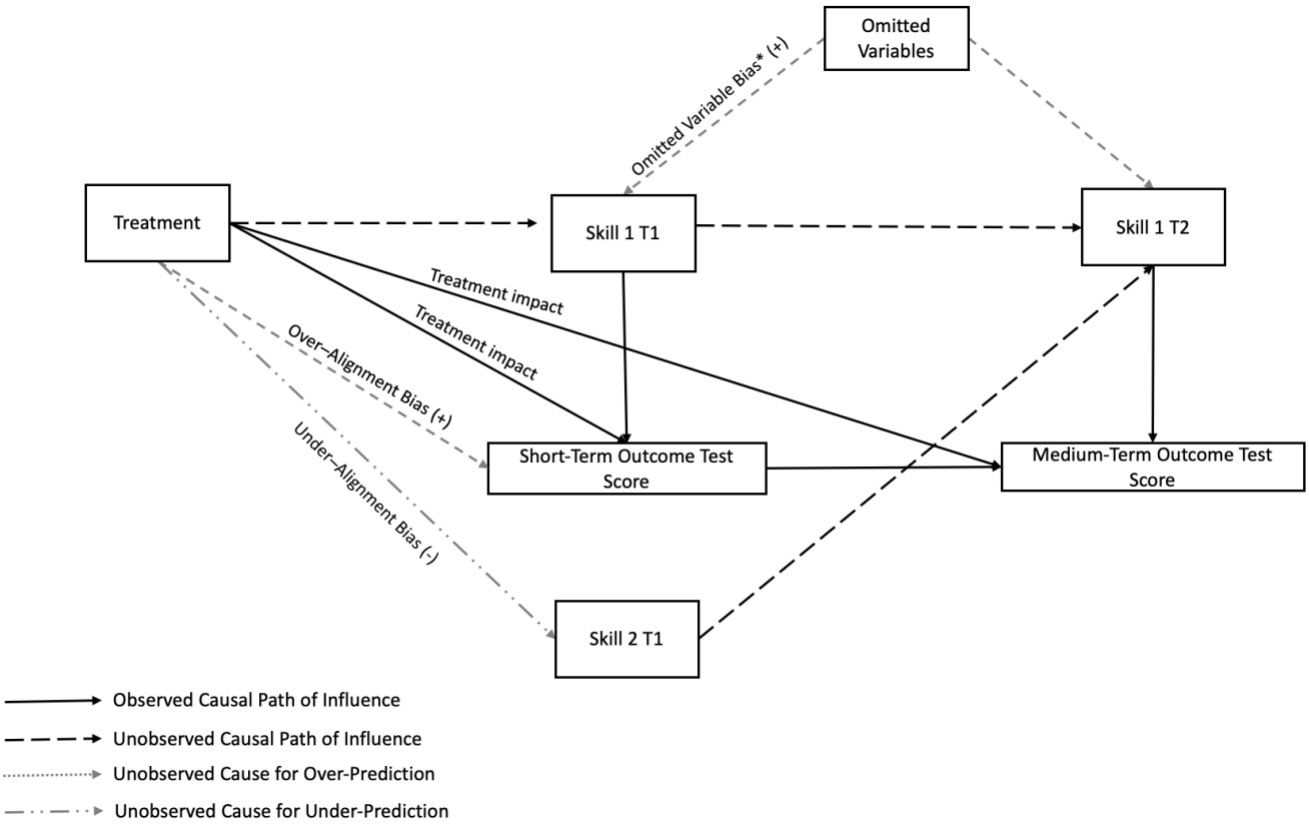
Conceptual Framework of Within Study Comparison of Number Knowledge Tutoring



Note. Directed Acyclical Graph notation is used to demonstrate the estimates we draw from separate groups within the same randomized control trial. First, we calculate the average treatment effect on the short-term outcome and on the medium-term outcome from the treatment and control groups. This is the expected impact of treatment on math skill growth at grade 1. Second, we calculate the relation (regression coefficient) of a short-term outcome on a medium-term outcome using the control group data. Third, we calculate forecasts by multiplying the treatment effect on the short-term outcome by the relation of the short-term outcome on the medium-term outcome. To complete this within study comparison, we compare the accuracy of our forecast to the observed experimental benchmark from the experimental evaluation.

Figure 2

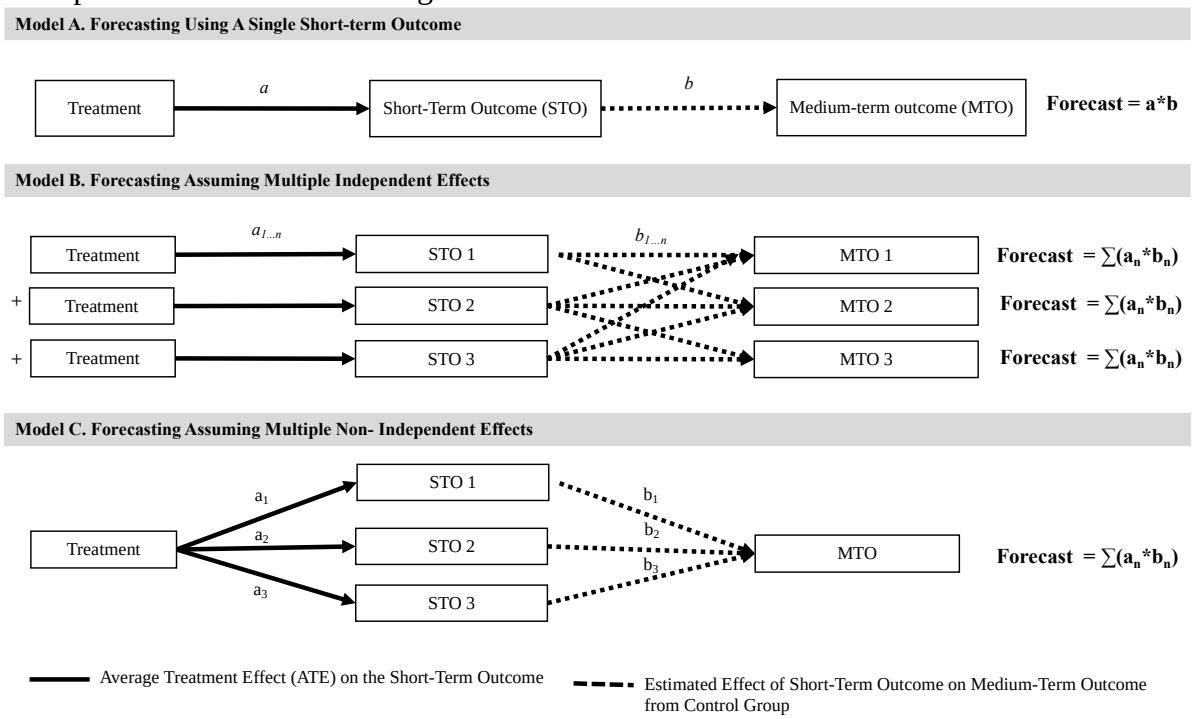
Sources of Bias in Forecasting Medium-Term Intervention Impacts



Note. Directed Acyclical Graph (DAG) notation is used in this figure to represent the causal paths that we expected to be influencing the key variables in our forecast. * For simplicity the potential influence of omitted variables bias is only shown to impact *Skill 1 T1* and *Skill 1 T2*, however this bias may also be expected to impact *Skill 2 T1* and *Skill 2 T2*. Similarly, under-alignment bias is represented by a single alternative unmeasured skill (skill 2 T1) however this bias may also be expected to impact *Skill 2 T1* and *Skill 2 T2*. Similarly, under-alignment bias is represented by a single alternative unmeasured skill (*Skill 2 T1*) however, *Skill 2* might also be conceptualized as measurement error in *Skill 1* at time 1.

Figure 3

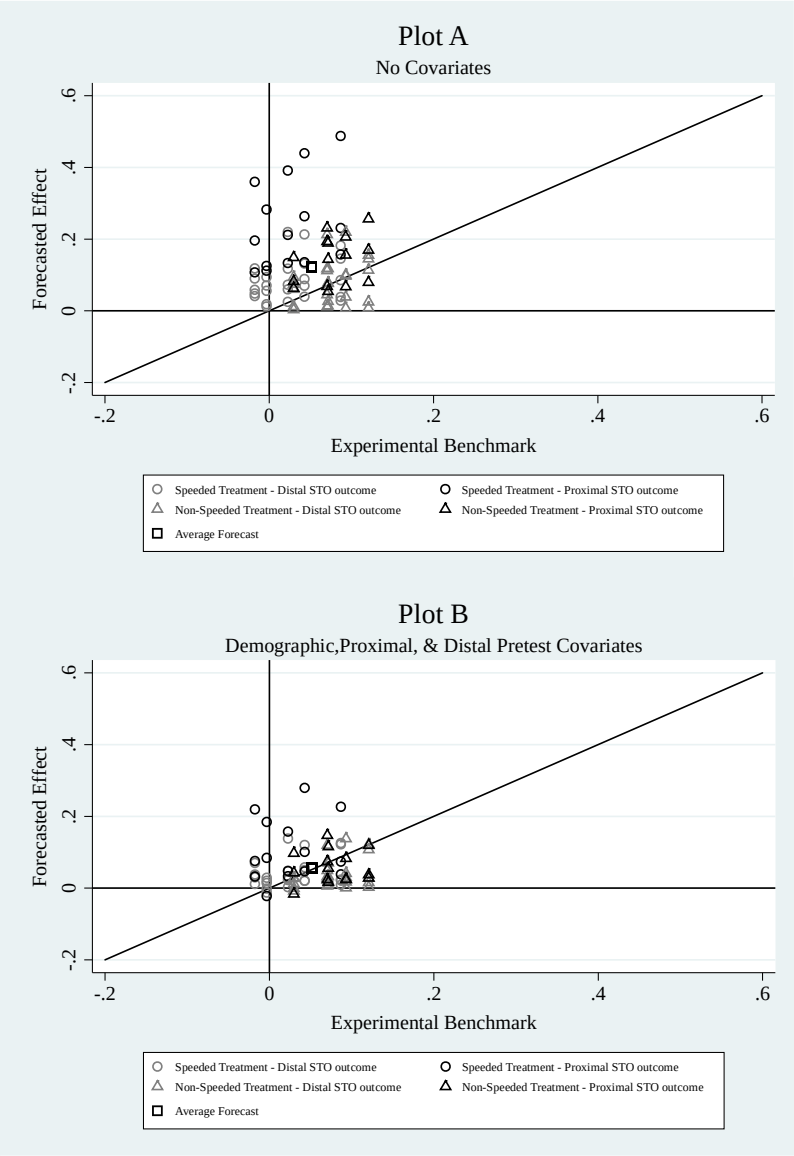
Conceptual Models of Forecasting Methods



Note. Three different approaches to calculating forecasts are shown. Panel A shows how we forecast a single medium-term outcome using a single short-term outcome; the treatment impact on the short-term outcome is multiplied by the regression coefficient of regressing the medium-term outcome on the short-term outcome from the control group data to reflect an estimated relation from nonexperimental data. Panel B shows how we forecast a single medium-term outcome using all the short-term outcomes assuming each short-term outcome independently impacts the medium-term outcome; the treatment impacts each short-term outcome is multiplied by the regression coefficient of regressing the medium-term outcome on each short-term outcome in a separate regression, with demographic and pretest covariates. Panel C shows how we forecast a single medium-term outcome using all the short-term outcomes assuming all the short-term outcomes share causal pathways to the medium-term outcome; the treatment impact each short-term outcome is multiplied by the regression coefficient of regressing the medium-term outcome on each short-term outcome when all the short-term outcomes are entered in the same regression model along with demographic and pretest covariates.

Figure 4

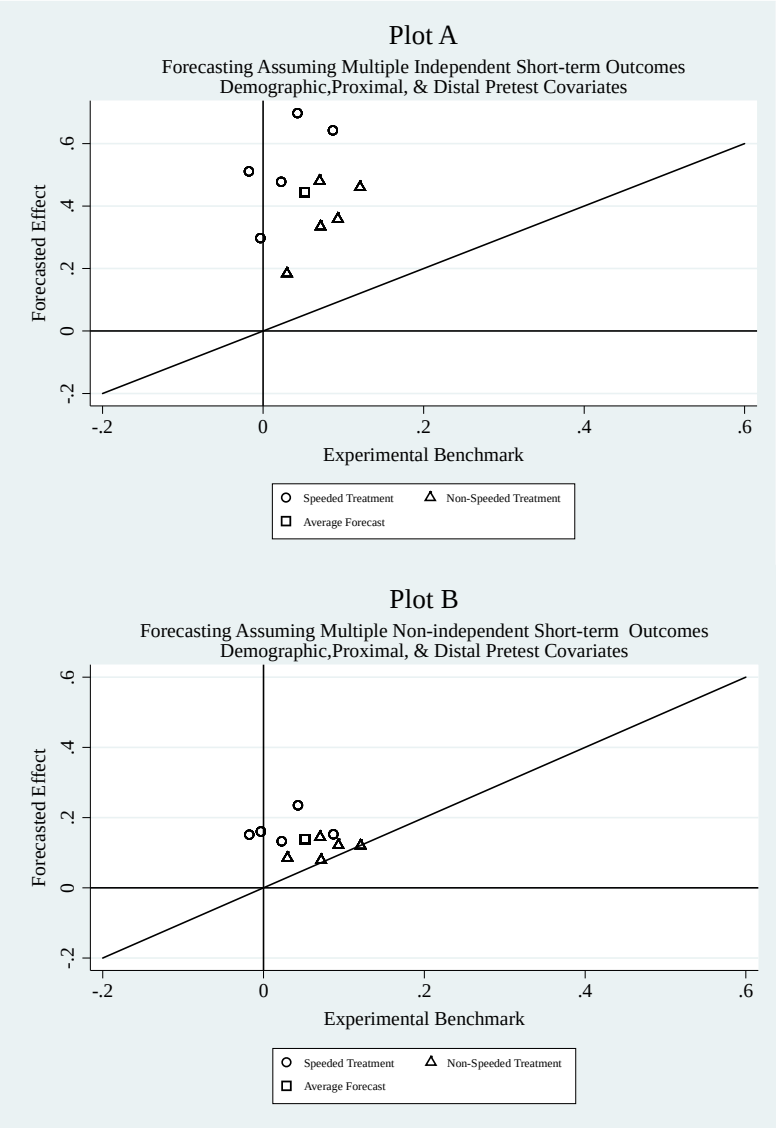
Replicating and Addressing Omitted Variables Bias



Note. STO= Short-term outcome. Each marker on the plots represents a forecast calculated using a single short-term treatment outcome to predict each single medium-term outcome within each treatment. Forecasts calculated from the speeded-treatment group are shown in circles, those from the non-speeded treatment group are shown in triangles. The average forecast is shown in a black square, this is calculated as the average of all the forecasts in the same plot. Gray markers indicate forecasts calculated with distal short-term outcomes and Black markers indicate forecasts calculated with proximal short-term treatment outcomes.

Figure 5

Forecasting with Multiple Short-term Outcomes



Note. Each marker on the plots represents a forecast calculated using all the short-term outcomes to predict each single medium-term outcome. Forecasts calculated from the speeded-treatment group are shown in circles, those from the non-speeded treatment group are shown in triangles. The average forecast is shown in a square circle, this is calculated as the average of all the forecasts in the same plot.