Abstract Title Page

Title: Everything in Moderation: Using Proximal and Distal Measures to Forecast the Long-term Impacts of Math Interventions

Authors and Affiliations:
Daniela Alvarez-Vargas, University of California, Irvine
Sirui Wan, University of California, Irvine
Drew Bailey, University of California, Irvine

Abstract (1157/1000 words excluding appendix)

Background:

Interventionists identify short-term intervention targets on the basis of their potential for long-term effects, which are often inferred from partial correlations between early and later academic skills. This is exemplified in a well-known study by Duncan et al (2007) which identified school entry math, reading, and attention skills as being the strongest predictors of achievement in the third grade. The long-term treatment impacts of interventions targeting these skills, however, are likely often under- or over-predicted, based on their short-run effects and these partial correlations. This over-prediction problem occurs when strong associations in non-experimental data between early psychological characteristics and later achievement outcomes are confounded by stable unmeasured variables (Bailey et al., 2018). Over-prediction may also occur when outcome measures are closely aligned to the skills that were targeted by the intervention and not taught to the control group (Slavin, 2008). If there is over-alignment the skills measured after treatment may be more reflective of how well the student learned the material from the intervention rather than of the underlying latent ability the intervention was targeting (Koretz, 2005). In contrast, outcome measures that are mis-aligned with the skills taught by the intervention may fail to adequately measure the targets the intervention meant to treat, thus underestimating the long-term impact of the intervention.

Purpose:

We use a within-study design to evaluate varying measurement features and analytical designs to calculate forecasts of long-term treatment effects and assess how closely they approximate the observed treatment impacts. We show how omitted variable bias, over-alignment bias from the use of proximal outcome measures, and under-alignment bias from the use of distal outcome measures, contribute to inaccurate forecasts of long-term treatment impacts.

Data:

We use data from the Number Knowledge Tutoring program that followed 639 students from 40 schools and 227 classrooms from a southeastern metropolitan district from first to third grade (Fuchs et al., 2013). Students at-risk of low academic performance were randomly assigned to either a control group, tutoring with speeded practice, or tutoring with non-speeded practice. 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. The key difference between the treatment groups was the activity conducted during the last five minutes of the tutoring session where students answered math problems through non-speeded games allowing the use of manipulatives or through the speeded game where students had 90 seconds to answer math problems on flash cards. The control group received business as usual instruction.

Measures:

An important characteristic of this dataset is that it included more than one proximal and distal measure at each of the three waves of data collection. The following proximal measures assess skills that were closely related to the content that was taught to the treatment group. The First-Grade Mathematics Assessment Battery (Fuchs, Hamlett, & Powell, 2003) a measure of student's ability to add and subtract with the Arithmetic Combinations and the Double-Digit subtests. The Number Sets Test (Geary, Bailey, & Hoard, 2009) measured speed and accuracy in operating with small numerosities of objects and linking them to the corresponding Arabic numeral. The Story Problems (Jordan & Hanich, 2000) assessed students' ability to understand and respond to arithmetic word problems. The Facts Correctly Retrieved (Geary et al., 2007) measured of children's addition strategy students use and accuracy.

In contrast, distal measures reflect assessments of a broad domain that consists of some, but not all, of the skills taught in the intervention. Additionally, these measures include items that become increasingly harder during the test administration allowing comparisons in math gains across grades. The Wide Range Achievement Test–3 Arithmetic (WRAT-Arithmetic; Wilkinson, 1993) subtest measured the ability to orally answer calculation problems, the Number Line Estimation 0-100 (Siegler & Booth, 2004) measured students understanding of relative numeric magnitudes, and *KeyMath–Numeration* (Connolly, 1998) measures students ability to orally respond to questions about identifying, sequencing, and relating numerals at increasingly difficulty.

Analyses:

We implement a within-study design (Cook, Shadish, and Wong, 2008) to test the effectiveness of study design and analytical approaches in forecasting long-term treatment impacts mathematical skills. Using data from a randomized control trial we make forecasts of the treatment impact on a long-term outcome with data gathered from the control group and compare it to the observed treatment impact estimated from the treatment group with a full set of demographic and pretest controls. The forecasts are estimated using the product from the estimated treatment impact on a short-term outcome and the estimated effect of a change in the short-term outcome on the long-term outcome. We assess forecast accuracy using three different analytical approaches: forecasting using a single short-term outcome (Figure 1a), assuming each short-term outcome has an independent causal impact on a long-term outcome (Figure 1b), and assuming that short-term outcomes share causal impacts on a long-term outcome (Figure 1c). We also assessed three different study designs to determine if using proximal measures, distal measures, or a combination of the two yielded more accurate forecasts.

Results:

Descriptive statistics for all the measures and demographic variables are shown in Table 1. Omitted variable bias is often addressed by statistically controlling for demographic variables and pretests. As shown in Figure 2, we find that by including all these as covariates the average forecast approximates the observed long-term treatment impact, but there are several remaining forecasts drastically overestimating the observed impact. We estimate forecasts using three different methods and find that estimating forecasts using a single short-term outcome yields, on average, the most accurate forecasts (shown in Figure 3). To improve forecast accuracy, we explore whether using specific proximal and distal short-term measures may yield better forecasts. In Figure 4, we plot forecasts calculated from proximal short-term measures with small treatment impacts (to reduce over-alignment bias) in green and distal short-term measures with large treatment impacts (to reduce under alignment bias) in blue. We find that the most accurate forecasts are the result of taking the average of proximal measures with small treatment impacts and distal measures with large treatment impacts.

Conclusions:

We find that omitted variable bias is substantially reduced when we include demographic variables and use the same measures in the pretests as in the posttests, yet over-alignment bias leads to over-estimated forecasts. Forecasting on the basis of a single short-term outcome is more accurate when we estimate the average forecasts using proximal short-term outcomes with the smallest treatment impacts and distal short-term outcomes with the biggest treatment impacts. We propose study design and analytical approaches to ensure that forecasts are within a .1 standard deviation unit from the observed treatment impact. Future work, however, is needed to replicate these findings and validate the accuracy of the suggested study design and analytic approaches. By forecasting long-term treatment impacts interventionists can estimate the power they have to determine and forecast the practical significance of their interventions.

Appendices

Appendix A. References

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

Connolly, A. J. (1998). KeyMath-Revised. Circle Pines, MN: American Guidance Service.

Cook, T. D., Shadish, W. R., & Wong, V. C. (2008). Three conditions under which experiments, and observational studies produce comparable causal estimates: New findings from within-study comparisons. Journal of Policy Analysis and Management, 27(4), 724–750. https://doi.org/10.1002/pam.20375

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

Fuchs, L. S., Geary, D. C., Compton, D. L., Fuchs, D., Schatschneider, C., Hamlett, C. L., … 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/a0030127

Fuchs, LS.; Hamlett, CL.; Powell, SR. First-Grade Mathematics Assessment Battery. L. S. Fuchs; 228 Peabody, Vanderbilt University, Nashville, TN 37203: 2003. doi: 10.1111/j.1540-5826.2008.01272.x

Geary DC, Bailey DH, Hoard MK. Predicting mathematical achievement and mathematical learning disability with a simple screening tool: The Number Sets Test. Journal of

Psychoeducational Assessment. 2009; 27:265–279. 10.1177/0734282908330592 [PubMed: 20161145]

Geary DC, Hoard MK, Byrd-Craven J, Nugent L, Numtee C. Cognitive mechanisms underlying achievement deficits in children with mathematical learning disability. Child Development. 2007; 78:1343–1359.10.1111/j.1467-8624.2007.01069.x [PubMed: 17650142]

Jordan NC, Hanich L. Mathematical thinking in second-grade children with different forms of LD. Journal of Learning Disabilities. 2000; 33:567–578.10.1177/002221940003300605 [PubMed: 15495398]

Koretz, D. (2005). Alignment, High Stakes, and the Inflation of Test Scores. Yearbook of the National Society for the Study of Education, 104(2), 99–118. https://doi.org/10.1111/j.1744-7984.2005.00027.x

Siegler, R. S., & Booth, J. L. (2004). Development of numerical estimation in young children. Child Develop- ment, 75, 428–444. https://doi.org/10.1111/j.1467-8624. 2004.00684.x

Slavin, R. E. (2008). Perspectives on Evidence-Based Research in Education—What Works? Issues in Synthesizing Educational Program Evaluations. Educational Researcher, 37(1), 5–14. https://doi.org/10.3102/0013189X08314117

Wilkinson, GS. *Wide Range Achievement Test 3 (WRAT3).* Wilmington, DE: Wide Range; 1993.

Appendix B. Tables and Figures

Table 1.
Descriptive Statistics

	Control			Speeded Practice			Non-Speeded Practice			ANOVA	Speeded vs. Control	Non-Speeded vs. Control
	N	Mean	SD	*N*	Mean	SD	*N*	Mean	SD	*p*	*p*	*p*
Demographics												
Age at pretest	224	6.47	.37	211	6.49	.37	204	6.47	.37	0.81	0.58	0.97
Sex (1=male)	223	0.50		207	0.53		203	0.48		0.55	0.59	0.68
Free or reduced lunch	222	0.88		207	0.85		201	0.80		0.09	0.33	0.03*
African American	224	0.72		211	0.67		204	0.66		0.36	0.30	0.17
White	224	0.17		211	0.22		204	0.21		0.50	0.25	0.41
Hispanic	224	0.07		211	0.06		204	0.08		0.67	0.82	0.52
Ethnicity: Other or Missing	224	0.04		211	0.05		204	0.05		0.80	0.71	0.51
ESL	222	0.03		206	0.02		201	0.02		0.73	0.43	0.68
Pretests												
First-Grade Content (Proximal Measures)												
Arithmetic Combinations	224	12.32	7.21	211	12.46	7.48	204	12.56	6.87	0.94	0.84	0.72
Double-Digit Calculation	224	0.42	0.91	211	0.46	1.08	204	0.38	0.88	0.68	0.71	0.59
Facts Correctly Retrieved	223	1.51	2.31	210	1.55	1.99	203	1.36	2.15	0.66	0.86	0.50
Number Sets	224	-0.51	0.70	211	-0.51	0.82	204	-0.48	0.67	0.92	0.96	0.68
Story Problems	224	1.68	1.76	211	1.76	1.56	204	1.80	1.79	0.75	0.60	0.48
Cross-Grade Content (Distal Measures)												
WRAT-Arithmetic	224	88.78	12.11	211	89.19	11.71	204	89.61	12.72	0.78	0.72	0.49
Number Line	224	26.38	6.21	211	25.71	7.27	204	26.06	6.19	0.57	0.31	0.59
KeyMath-Numeration	224	97.37	10.53	211	97.30	10.24	204	97.30	10.51	1.00	0.95	0.95
Attrition												
Attrition by posttest		0.04			0.06			0.06		0.63	0.42	0.38
Attrition by grade 2		0.16			0.11			0.14		0.43	0.19	0.58
Attrition by grade 3		0.17			0.14			0.13		0.41	0.29	0.23

Note. SD = standard deviation, N = number of students observed

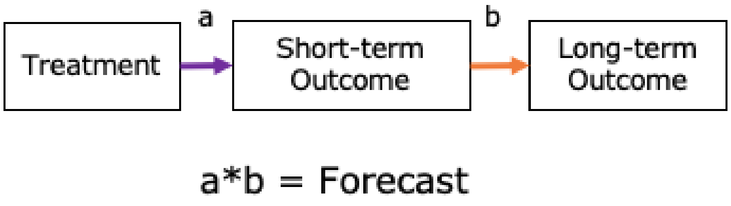

Figure 1a. Forecasting Using a Single Short-Term Outcome

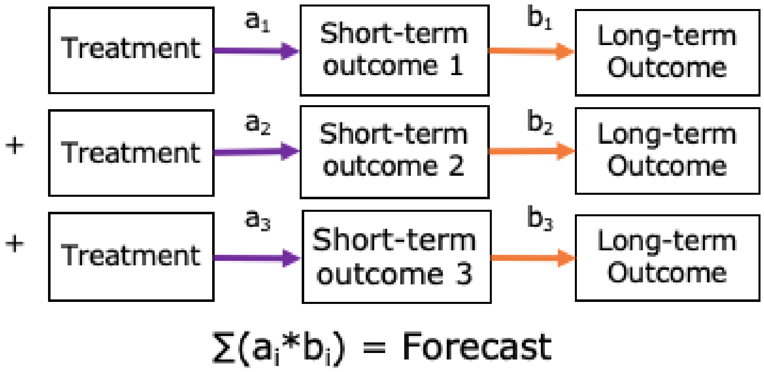

Figure 1b. Forecasting Assuming Multiple Independent Effects

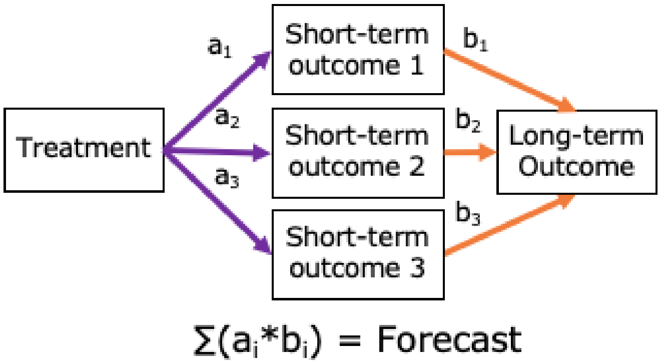

Figure 1c. Forecasting Assuming Multiple Non-Independent Effects

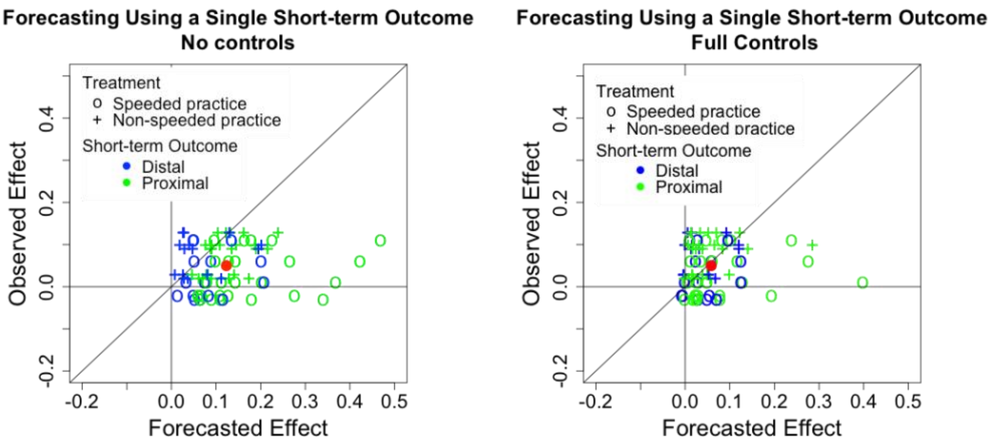

Figure 2. Omitted variable bias leads to over-estimated forecasts that are reduced with a full set of controls.

Note. The red dot on each plot represents the average forecast.

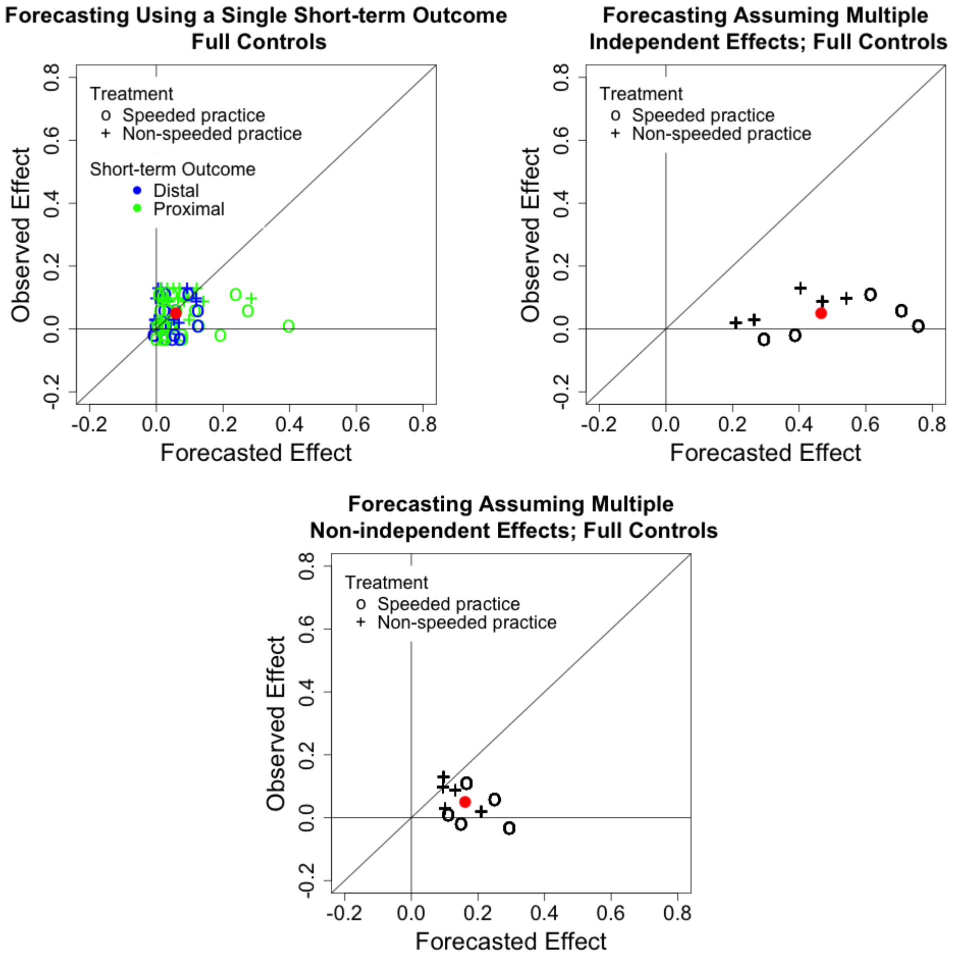

Figure 3. Forecasting Using Three Separate Methods

Note. The red dot on each plot represents the average forecast.

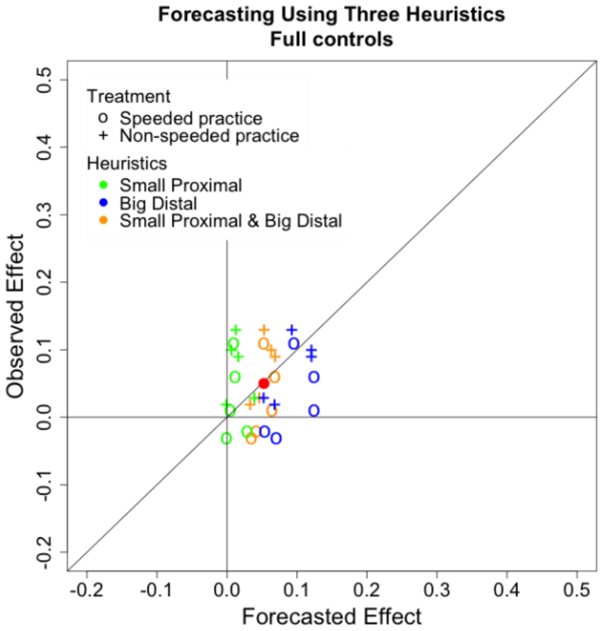

Figure 4. Forecasting Long-Term Treatment Impacts Using Proximal and Distal Short-Term Measures. *Note.* The red dot represents the average forecast.