UNIVERSITY OF CALIFORNIA,

IRVINE

Three Extensions to Evaluating Educational Interventions

DISSERTATION

submitted in partial satisfaction of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in Education

by

Daniela Alvarez-Vargas

Dissertation Committee:
Professor Drew Bailey, Chair
Professor Greg Duncan
Associate Professor June Ahn

2023

DEDICATION

For my parents who worked tirelessly to give me the best opportunities, for my siblings who sparked deep curiosity and joy in human development that brought me here. Thank you all for bringing purpose to my life and to my career, thank you for showing me the beauty of human development. To my husband who helped ground me and supported me along the way. For all my family in the United States of America, Colombia, and across the world. Thank you for always believing in me and allowing me to redefine what it means to be a doctoral scholar and a scientist.

For the true lifelong friendships, I have developed across my graduate career. I will never forget you. For the future generations of scholars of color and of marginalized backgrounds that are deserving of the best educational and professional opportunities. For communities that have been marginalized, I hope to serve you and devote my time to rectifying injustice and promoting equity.

For my mentors, Dr. Drew H. Bailey, Dr. Greg Duncan, Dr. Andres Bustamante, Dr. June Ahn, Dr. Jade Jenkins, and Dr. Katherine Rhodes. All who pushed me beyond the bounds of comfort and into higher levels of reasoning. Thank you for bringing your true selves, fun, and joy to the work that you do.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

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. 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.

This research uses data from the Baby's First Years study. Research reported in this publication was supported by the Eunice Kennedy Shriver National Institute of Child Health and Human Development of the National Institutes of Health under Award Number R01HD087384. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. This research was additionally supported by the US Department of Health and Human Services, Administration for Children and Families, Office of Planning, Research and Evaluation; Andrew and Julie Klingenstein Family Fund; Annie E. Casey Foundation; Arnold Ventures;

Arrow Impact; BCBS of Louisiana Foundation; Bezos Family Foundation, Bill and Melinda Gates Foundation; Bill Hammack and Janice Parmelee, Brady Education Fund; Chan Zuckerberg Initiative (Silicon Valley Community Foundation); Charles and Lynn Schusterman Family Philanthropies; Child Welfare Fund; Esther A. and Joseph Klingenstein Fund; Ford Foundation; Greater New Orleans Foundation; Heising-Simons Foundation; Holland Foundation; Jacobs Foundation; JPB Foundation; J-PAL North America; Lozier Foundation; New York City Mayor's Office for Economic Opportunity; Perigee Fund; Robin Hood Foundation; Robert Wood Johnson Foundation; Russell Sage Foundation; Sherwood Foundation; Valhalla Foundation; Weitz Family Foundation; W.K. Kellogg Foundation; and three anonymous donors.

I thank Routledge Taylor and Francis Group for permission to include the copyrighted materials included in Chapter 1 as part of my dissertation. This chapter was previously published in the *Journal of Research on Educational Effectiveness.*

VITA

Daniela Alvarez-Vargas

EDUCATION

Expected: Ph.D., Education

5/2023 University of California Irvine

Advisor: Dr. Drew Bailey

Specialization: Human Development in Context

Dissertation: Three Methodological Extensions to Educational Evaluations

2014- B.A., Psychology, minor in statistics

2018 Florida International University, Miami, FL

Advisor: Dr. Shannon Pruden

HONORS AND AWARDS

2023 Latino Excellence and Achievement Dinner (LEAD)

Celebrating graduate student excellence and achievement

2022 Division of Teaching Innovation and Excellence Summer Fellowship ($5,000)

To dedicate 200 hours over the summer to improve applied regression course by incorporating hybrid virtual and in-person activities and lesson plans.

2022 Pedagogical Fellowship ($2,500)

Competitively selected campus-wide award for demonstrating excellence and potential in pedagogical practice in higher education.

2019 National Science Foundation Graduate Research Fellowship ($101,988)

National recognition as a promising scholar and three years of graduate funding

2018 Competitive Edge Summer Research Program ($5,000)

Funding for advanced entry into graduate training and enrollment in an eight-week summer research and professional development program

2018 UC Irvine Diversity Recruitment Fellowship ($1,000)

Financial award for academic record, accomplishments, and promise

2018 Nevin Graduate Endowment Fellowship ($1,000)

Financial Award "based on proven grit, leadership, academic record, and promising potential"

PUBLICATIONS

J9 **Alvarez-Vargas, D**., Braithwaite, D. W., Lortie-Forgues, H., Moore, M. M., Castro, M., Wan, S., Martin, E.A., Bailey, D. H. (2020, May 12). Hedges, mottes, and baileys: Causally ambiguous statistical language can increase perceived study quality and policy relevance. *Plos One*. July 2023. *In press*.

J8 Hye Rin, L., Xin, T., **Alvarez-Vargas, D**., Yang, J.S., Bailey, D., Simpkins, S., Safavian, N., Gaspard, H., Salmela-Aro, K, Moeller, J., Eccles, J.S., & Wigfield, A. (2023) Networks and Directed Acyclic Graphs: Initial Steps to Efficiently Examine Causal Relations between Expectancies, Values, and Prior Achievement. Current Psychology. *In press*.

J7 **Alvarez-Vargas, D.,** Lopez Perez, J.P*., Bermudez, V., Beltrán-Grimm, S., Santana, E., Begolli, K., Bustamante, A. S. (2023) Evidence-Based Designs for Physically Active and Playful Math Learning. *Theory Into Practice*. https://doi.org/10.1080/00405841.2023.2202131

J6 **Alvarez-Vargas, D**., Wan, S., Fuchs, L.S., Klein, A., & Bailey, D.H. (2022). Design and Analytic Features for Reducing Biases in Skill-Building Intervention Forecasts*. Journal of Research on Educational Effectiveness* https://doi.org/10.1080/19345747.2022.2093298.

J5 Wan, S., **Alvarez-Vargas, D.,** & Bailey, D. H. (2022) Toward a Causally Informative Fit Index of Longitudinal Models: A Within-Study Design Approach. *Developmental Psychology*.

J4 Goeke, M., Muller, A., **Alvarez-Vargas, D.,** & Van Steenis, E.J. (2022) Using Mediating Artifacts to Push for Greater Equity in Research Practice Partnerships. *The Assembly*.

J3 Bustamante, A. S., Begolli, K., **Alvarez-Vargas, D**., Bailey, D. H., & Richland, L. (2021). Fraction Ball: Playful and Physically Active Fraction and Decimal Learning. *Journal of Educational Psychology*. https://doi.org/10.1037/edu0000714

J2 Bailey, D. H., Jenkins, J. M., & **Alvarez-Vargas, D**. (2020). Complementarities between early educational intervention and later educational quality? A systematic review of the sustaining environments hypothesis. *Developmental Review, 56*, 100910. https://doi.org/10.1016/j.dr.2020.100910

J1 **Alvarez-Vargas**, D., Abad, C. & Pruden, S.M. Spatial anxiety mediates the sex difference in adult mental rotation test performance. *Cognitive Research: Principles and Implications* 5, 31 (2020). https://doi.org/10.1186/s41235-020-00231-8

SELECTED MANUSCRIPTS UNDER REVIEW

R2 **Alvarez-Vargas, D**., Begolli, K.N., Choc, M., Farag, L.M., Bailey, D.H., Richland, L., & Bustamante, A.S. Fraction ball impacts on student and teacher math talk and behavior. September 2022. Under review.

CONFERENCE PRESENTATIONS

*=Student mentee co-author

C14 Alvarez-Vargas, D., Bailey, D., Duncan, G., Gennetian, L., Halpern-Meekin, S., Magnuson, K., & Noble, K (2023, March) *A Within-Study Comparison of Experimental and Nonexperimental Estimates of Income Effects on Child and Maternal Outcomes* [Symposium] Society for Research on Child Development, Salt Lake City, UT.

C13 Lopez Perez, J.P*., Bermudez, V., **Alvarez-Vargas, D**., & Bustamante, A.S (2022, May) *Student's Engagement and Motivation in Number Ball* [Proposal Pitch] Undergraduate Research Opportunities Program, University of California, Irvine, CA.

C12 **Alvarez-Vargas, D**., Choc, M*., Bermudez, V., Begolli , K., & Bustamante , A.S (2022, April) *Fraction Ball as a Case Study of Centering Teacher's Voices for*

Intervention Design [Round table presentation]. American Education Research Association, San Diego, CA.

C11 **Alvarez-Vargas, D**., Bustamante , Begolli , K., Bailey, D.H. & Richland, L.E. (2022, April) *Fraction Ball: Using Rational Number Language in a Playful Context* [Paper Symposium]. Society for Research in Child Development Special Topics Meeting: Learning through Play and Imagination, St. Louis, MO.

C10 Hall, L., Rengel, M., Bowley, H., **Alvarez-Vargas, D**., Abad, C., Overton, D., & Pruden, S.M. (April, 2022) *Diversity and quantity of parent-child spatial talk: The roles of prosocial talk and negative talk.* [Poster Presentation] Cognitive Development Society, Madison, WI.

C9 Overton, D., Hall, L., Rengel, M., Bowley, H., **Alvarez-Vargas, D**., Abad, C., & Pruden, S.M. (April, 2022*) Parent and Child Math Language and Relations to Child Spatial Ability*. [Poster Presentation] Cognitive Development Society, Madison, WI.

C8 **Alvarez-Vargas, D**., Choc, M., Bermudez, V., Begolli , K., & Bustamante , A.S (2021, July) *Fraction Ball as a Case Study of Centering Teacher's Voices for Intervention Design* [Half-Baked Idea Presentation]. Rising Education Scholars Helping to Advance Partnerships and Equity (RESHAPE), Online due to COVID-19.

C7 **Alvarez-Vargas, D**., Wan, S., Fuchs, L., Klein, A., & Bailey, D.H (2021, March) *Design and Analytic Features for Reducing Biases in Skill-Building Intervention*

Impact Forecasts. Society for Research in Child Development Special Topics Meeting: Learning through Play and Imagination, Online due to COVID-19.

C6 Lee, H., **Alvarez-Vargas, D.**, Tang, X., Bailey, D.H., Yang, J., Safavian, N., Gaspard, H., Simpkins, S., Salmela-Aro, K., Eccles, J.S., & Wigfield, A. (2021, August) *Examining Students' Expectancies and Values with Networks and Directed Acyclic Graphs*. Symposium accepted at the 2021 annual meeting of the European Association for Research on Learning and Instruction, Online due to COVID-19.

C5 Hall, L., Montgomery, A., **Alvarez-Vargas, D**., Abad, C., & Pruden, S.M. (April, 2020*) Parent-child spatial language and spatial activities: Does drawing elicit the same amount of spatial language as block play?* [Virtual Poster Presentation] International Conference on Infant Studies, Glasgow, Scotland.

C4 **Alvarez-Vargas , D**., Wan, S., & Bailey, D.H (2020, March) *Everything in Moderation: Using Proximal and Distal Measures to Forecast the Long-term Impacts of Math Interventions* [Poster session]. Society for Research on Educational Effectiveness, Arlington, VA, United States. https://www.sree.org/conference-program (Conference canceled)

C3 **Alvarez-Vargas , D**., Wan, S., & Bailey, D.H (2019, November) *Everything in Moderation: Using Proximal and Distal Measures to Forecast the Long-term Impacts of Math Interventions.* RO1 Consortium at Laguna Hills Inn, Irvine, CA.

C2 **Alvarez-Vargas , D**., Wan, S., & Bailey, D.H (2019, September) *Everything in Moderation: Using Proximal and Distal Measures to*

Forecast the Long-term Impacts of Math Interventions. First Year Graduate Student Poster Presentations at the University of California, Irvine, CA.

C1 **Alvarez-Vargas , D**., & Bailey, D.H (2018, August) *The Role of Sustaining Environments in the Persistence of Educational Intervention Impacts*. Competitive Edge Summer Research Symposium at University of California, Irvine, CA.

GRADUATE STUDENT RESEARCH ASSISTANTSHIPS

Baby's First Years

Funders: See https://www.babysfirstyears.com/funding for complete list
PI: Greg Duncan, University of California, Irvine
Role on the project: Graduate Student Researcher; Publication in preparation entitled "*A Within-Study Comparison of Experimental and Observational Estimates of Income Effects on Child Development and Maternal Well-Being"*. Assisted team with data cleaning and figure production.

Fraction Ball

Funder: EF+Math Program of the Advanced Education Research and Development Program
PI: Andres S. Bustamante, University of California, Irvine
Role on the project: Graduate Student Researcher; Co-designed and implemented a math intervention for 4th- 6th grade students that is a set of basketball related games on a court that was redrawn to reflect fractions and decimals instead of the traditional three-point line. Contributed to grant application.

Methods for Producing Causally Informative Estimates of the Long-Run Impacts of Early Math Interventions

Funder: Jacobs Foundation Research Fellowship

Award # 2018-128802

PI: Drew H. Bailey, University of California, Irvine

Role on the project: Graduate Student Researcher; Developed and analyzed three separate statistical methods to forecast the long-term outcomes of intervention treatment impacts. Methods included separate mathematical equations where the estimates from ordinary least square regression were used to make predictions.

TEACHING EXPERIENCE

Workshops

7/2022 Teaching Assistant Professional Development (TAPDP) workshop. Designed and implemented a two-day (8 hours) discipline-specific pedagogical training for incoming graduate students to learn how to be effective and equitable teaching assistants.

1/2020 Alvarez-Vargas, D. *Introduction to Tidyverse*. Workshop led at the Introduction to R for Educational Data Science Workshop Series, Irvine, CA.

Statistical Consultant

1/-8/2022 Funded by the UCI School of Education as a Statistical Consultant for other Ph.D. students. [Virtual Remote]

Teaching Assistant

9/ – 12/2021	EDUC 124A&B: *Multicultural Education,* School of Education, University of California, Irvine [Virtual Remote & In-Person]
8/ – 9/2020	EDUC 202: *Outcomes of Assessment and Schooling,* School of Education, University of California, Irvine [In-Person]
1/ – 3/2020	EDUC 322B: *Math for Elementary School,* School of Education, University of California, Irvine [Virtual Remote]
9/ – 12/2019	EDUC 30: *21st Century Literacy Skills,* School of Education, University of California, Irvine [In-Person]
9/– 12/2018	EDUC 124: *Multicultural Education,* School of Education, University of California, Irvine [In-Person]

Reader

1/ - 3/2019	EDUC 106: *Introduction to Early Childhood Education,* School of Education, University of California, Irvine [In-Person]

SERVICE

Reviewer

2021 – 2022	Society for Research on Educational Effectiveness Poster and Symposium Submissions
2019	Ad-hoc Journal Reviewer American psychologist (co-reviewed with Dr. Drew H. Bailey) Journal of Learning Disabilities (co-reviewed with Dr. Drew H. Bailey)

Undergraduate Mentorship

2022-Present Ethan Shenting, Undergraduate Research Opportunity Program

2021-Present Jessica Paola Lopez, Undergraduate Research Opportunity Program

2019 Vanessa Llamas, Volunteer

2019 Jaylene Rios, Volunteer

2019-2020 Cesar Lopez De Lara Salgad, Independent Study Undergraduate

2019-2020 John Gomez, Volunteer

2019-2020 Priscilla Monique Molina, Independent Study Undergraduate

2019-2020 Consuelo Rojas, Independent Study Undergraduate

2019-2021 Yazmin Torres Ramirez, Independent Study Undergraduate

2019-2020 Jamie Moreno, Independent Study Undergraduate

2019-2020 Deisy Modesta Flores, Independent Study Undergraduate

2019-2020 Josue Adrian Hernandez-Aguilar, Volunteer

2019-2020 Marsha Choc, Independent Study Undergraduate

2019-2020 Jazmin Garcia Yescas, Independent Study Undergraduate

2019-2020 Crista Siboney Urena Hernandez, Independent Study Undergraduate

2019-2020 Liliana Hernandez, Independent Study Undergraduate

Ph.D. Admissions Committee, School of Education, UCI

2021-2022 Served as DECADE representative for the doctoral students admissions committee. Reviewed, organized, and advocated for under-represented minority students to faculty for recommendation for admission.

DECADE Graduate Student Representative for Climate Council, School of Education, UCI

2018 Attended meetings and collaborated on a draft of a strategic plan to address and resolve issues of climate in the School of Education.

Diverse Educational Community and Doctoral Experience (DECADE), School of Education, UCI

2018 Write and design a quarterly newsletter showcasing faculty, post-docs, graduate students, and events in the School of Education at UCI to encourage community building. Maintained team drive to provide graduate students with academic and non-academic resources. Peer mentor for incoming first-year graduate students.

PROFESSIONAL DEVELOPMENT ACTIVITIES

2022-2023 Pedagogical Fellowship from the Division of Teaching Excellence and Innovation, University of California, Irvine

Received 100 hours' worth of training across the 390A & B course series to build excellence in pedagogical practice in higher education

01/2022 Certificate in Preparing for a Faculty Career from the Graduate Division, University of California, Irvine

Professional development for entering the academic job market

12/2019 Course: University Teaching: Concepts and Practices (EDUC 226)

I learned about course design and instruction at the university level and created a syllabus for inclusive teaching in an intro to statistics in education course

7/2018 Completion of the Division of Teaching Excellence & Innovation Course Design Certificate Program

ABSTRACT OF THE DISSERTATION

Three Extensions to Evaluating Educational Interventions

by

Daniela Alvarez-Vargas

Doctor of Philosophy in Education

University of California, Irvine, 2023

Professor Drew Bailey, Chair

Educational intervention research addresses multiple barriers that under-sourced communities face in education. I test three methods to improve education program design and evaluation to develop programs that have substantial and lasting impacts on student outcomes. In my dissertation, I review current challenges to educational program evaluation research and present three different extensions to the methodologies researchers use to improve their practice. To address these challenges, I draw upon perspectives and methodological innovations from adjacent fields including causal inference methods from policy research, measures from implementation science, and practical insights from research-practice partnerships.

I contribute three separate studies as potential extension to different parts of our process to discovering what works, for whom, and when. Study 1 describes methods to over-come the challenge of conducting long-term follow up evaluations by empirically testing different intervention design features and analytical decisions for forecasting medium-term impacts of early skill-building interventions in mathematics. I empirically

test different study designs and analytical approaches to determine which combinations improve the accuracy of forecasting the medium-term impacts of math interventions using nonexperimental data.

Study 2 embraces the challenge of identifying what mechanism or lever to intervene on and how much nonexperimental data can inform theory of change to design interventions. In study 2, I use the data from the Baby's First Years Study a longitudinal randomized controlled trial (RCT) of an unconditional cash-gift given to the mothers of newborn children living in poverty. I compare nonexperimental estimates of the impact of income on child development and maternal well-being–using data from the control group–to the experimental estimates from the RCT. This study helps us understand and document the importance and difficulty of formulating interventions from theory and nonexperimental data. I discuss the implications of this work for attempting to craft evidence-based policy based on corresponding experimental and non-experimental estimates.

Study 3 embraces the challenge of program effect variation across different contexts. I conduct semi-structured interviews and a comparative case analysis to capture evidence about what adaptations and modifications are made to math programs when researchers iterate through program design and implementation. I describe when the adaptations are made, who decides on making the adaptation, what the adaptation is, where it is adapted, and how it is adapted using the Framework for Reporting Adaptations and Modifications-Enhanced (FRAME-IS; Miller et al., 2021) expanded framework for reporting adaptations. I develop a set of guiding principles that can inform researchers about what adaptations to anticipated when designing and implementing math programs

INTRODUCTION

Educational research seeking to influence policy making and educational practice faces multiple methodological challenges. In this dissertation I focus on three methodological extensions that researchers, grant funders, and policy makers can leverage to address these challenges.

The first challenge is the lack of research that has evaluated the medium- to long-term impacts of educational programs (Philips et al., 2017; Watts, Bailey, & Li, 2019). Doing this work is further complicated by the cost of funding long-term evaluations in educational settings where turnover and attrition is high. In addition, conducting research replicating programs and scaling them up can be risky, if the program does not have the desirable effects resulting in little reward for the researchers. Lastly, the large scale randomized controlled trials of efficacious programs that have been scaled-up often result in smaller effects with little information on the variation of the effects (Lortie-Forgues & Inglis, 2019). One way to circumvent the costs and time of long-term follow-up is to develop valid accurate predictions of the long-term effects. However, in the cases where this has been done the correlational analyses sometimes over-predict the observed experimental impacts measured at medium-term follow up (Bailey et al., 2018). To address this first challenge I conducted study 1 to test different methods of calculating predictions/ forecasts using multiple different forms of measurements and analytical decisions.

The second challenge to conducting research that is relevant for policy is centers around the mixed evidence generated from nonexperimental and experimental studies on how different factors influence child development. Nonexperimental data from cross-

sectional and longitudinal studies is often used to develop theories of child development, these theories are then used as the rationale for what kinds of studies to fund to test potential policy relevant interventions. A problem with this cyclical approach to identifying levers of influence is that it is very difficult to tease apart the omitted variables bias from nonexperimental data. To bypass this challenge policy researchers have developed within-study comparisons (Cook, Shadish, & Wong, 2008; LaLonde, 1986; Michalopoulos, Bloom, & Hill, 2004; Steiner & Wong, 2018) to determine what methodological approaches and analytical decisions approximate the estimates derived from experimental work. Following this tradition, I sought out to evaluate how well we can approximate the experimental estimates of an observed impact of income on multiple maternal and child outcomes , when only using nonexperimental data.

The third challenge to producing useful and relevant educational program evaluations is the disconnect between the regulatory research process of developing an efficacious program and scaling it up and the practice based process of preferring research that is developed and tested under local conditions within contexts and students like their own. To engage in more relevant research multiple groups have recommended engaging in collaborative research with schools and practitioners (Donovan, Snow, & Daro, 2013; Fishman, Penuel, Allen, Cheng, & Sabelli, 2013; Tseng, 2012). However, there is a dearth of research demonstrating how working in collaboration with schools and practitioners yield better outcomes. A challenge to getting this evidence is the time that it takes to build and conduct collaborative work. Another challenge is that researchers conducting collaborative work do not often capture the same outcome measures as researchers conducting efficacy and effectiveness trials due to differences in research goals. To gain some insight into the

ways that researchers modify and adapt educational programs to better fit local contexts I conduct a series of semi-structured interviews with researchers across different methodological approaches and research goals. Although I cannot quantitatively evaluate how these adaptations resulted in greater math test scores, the qualitative reports from researchers reveal a set of regularities in the opportunities and challenges of making these adaptations that can serve as guiding principles for future work.

Study 1

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.,5^{th}-grade math achievement) of successfully improving early math skills. I hypothesized that:

(1) Forecasts using demographic and pretest covariates should reduce the bias from estimating the causal impact of an increase in an early math skill on later math skills.

(2) Estimates from forecasts that assume that early math skills influence later math skills through the partially overlapping pathways (overlapping mediators) would yield smaller, more accurate forecasts than estimates from forecasts that assume that early math skills influence later math skills through fully independent pathways.

(3a) Using the end-of-treatment outcomes that are conceptually proximal (closely aligned) with the intervention to calculate forecasts will yield over-estimated treatment impacts.

(3b) Using the end-of-treatment outcomes that are conceptually distal (less closely aligned) with the intervention to calculate forecasts will yield under-estimated treatment impacts.

(3c) The most accurate forecasts will be calculated by using a combination of end-of-treatment outcomes that are conceptually proximal and distal to the treatment. Combining both kinds of end-of-treatment outcomes will reduce over-alignment bias – because conceptually distal outcomes contain variance unlikely to be affected by the intervention, which may be shared with the underlying skills targeted by the intervention.

Results showed that the most accurate forecasts are based on non-experimental approaches with comprehensive baseline controls and a combination of posttests conceptually proximal and conceptually distal to the intervention. When comprehensive baseline controls were not employed, over-prediction was also a problem. We discuss the theoretical and practical implications of this work.

Study 2

Evidence from observational and quasi-experimental methodologies show that educational achievement scores are higher for children in higher than lower income families and there is good reason to think that these associations are partly causal. However, correlations between income and early childhood behaviors and outcomes may also reflect the impact of confounds influencing both income and development (e.g., maternal educational opportunities). Further, it is possible that cash transfers are a qualitatively different treatment from naturally occurring increases in income. In both cases, the income gradient may not approximate the effects of cash transfer payments on child developmental outcomes. We hypothesized that nonexperimental would be upwardly

biased in comparison to the experimental estimates due to potential influence of omitted variables bias.

Using the experimental data from a randomized experiment, I find that the income gradients estimated from observational data correspond weakly with the observed causal impacts of an unconditional cash gift. I also find that experimental estimates are, on average, not significantly larger than the nonexperimental estimates, in contrast to what we would expect from omitted variables bias (Meyer, 1997). Weak associations may result from qualitative differences between a cash-gift and stable income, or from imprecisely estimated experimental impacts. I review potential sources of bias and discuss the implications of this work for attempting to craft evidence-based policy based on experimental and non-experimental estimates.

Study 3

Evidence-based educational innovations sometime fail to generalize to different contexts and situations or to maintain the costs and benefits when implemented at a bigger scale (List, Suskind, & Supplee, 2021). There is a collective effort to better understand the mechanisms that contribute to successful educational innovations (Sabol et al., 2022; Bryan, Tipton, and Yeager, 2021) and to sustain them in authentic educational settings. Of importance is the empirical identification of how the contextual variation can inform innovation design and inform theory on what it is about a context that contributes to children's development. Weiss, Bloom, & Brock, (2014) developed the *Conceptual Framework for Studying Variation in Program Effects, Treatment Contrasts, and Implementation* to guide research on variation of program impacts and to determine the sources of this variation for analysis and program improvement. This framework highlights

the complementary goals of studying program implementation with the goals of estimating program effectiveness, to capture sources of variation and more systematically collect data that can yield insights about program outcomes. However, the extent to which individual and contextual factors influence program effects is not operationalized in the framework and often it is difficult to assess this systematically.

To address this challenge, I incorporate the Framework for Reporting Adaptations and Modifications to Evidence-Based Implementation Strategies (FRAME-IS; Miller et al., 2021) from implementation science work in public health to complement the *Conceptual Framework for Studying Variation in Program Effects, Treatment Contrasts, and Implementation.* Specifically, this conceptual framework could be improved by further operationalizing and capturing how context specific adaptations to educational innovation implementation influences the variation of impacts of educational innovations. Currently, the conceptualization and measurement of how the variation in program effects are moderated by a changing local context can be vague and underspecified. Typically, the client and context characteristics that are measured as moderators depend on data that is readily available, moderators based on theory, or moderators based on the data collection norms of the field. Under specification of adaptations complicates the conceptual replication of math programs and can obscure important sources of heterogeneity across program impacts. The interviews and case comparisons I conduct leverage the FRAME-IS to operationalize the individual and contextual factors that have influenced math program implementation in previous studies. In doing so, I describe how researchers modified their program design theories and implementation models to adjust to different contexts.

The combination of both theories contributes to a more concrete approach to obtaining about the contextual influences on math program effects. Currently, the conceptualization and measurement of how the variation in program effects are moderated by a changing local context can be vague and underspecified. The under specification of adaptations complicates the conceptual replication of math programs and can obscure important sources of heterogeneity across program impacts. Moreover, the description of how adaptations arise and are handled in program implementation can alleviate the potential challenges that researchers seeking to replicate that program may face in other contexts. Overall, this work provides a set of guiding principles and examples of a systematic process of documenting and reporting the education innovation adaptations and modifications that are made during the research process from design to implementation, to evaluation. Which is currently not a norm in educational research reporting.

STUDY 1: 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, relative to the number of interventions under consideration (Philips et al., 2017; Watts, Bailey, & Li, 2019), because it is difficult and costly to conduct. One solution to this problem has been to rely on longitudinal correlational research to determine optimal targets for short-term 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, predictions about the medium-term impacts of interventions based on correlational analyses sometimes over-predict the observed experimental impacts measured at medium-term follow up (Bailey et al., 2018).

Experimental evaluations of skill-building interventions have successfully increased children's early math skills, but they have not yielded the expected medium-term impacts that correlational work predicted. Instead, impact estimates from randomized control trials (RCTs) of early math interventions decrease by half or more within just a year after the end of implementation. These findings raise concerns about the usefulness of non-experimental estimates for designing interventions for early academic skills to enhance children's skill-development. The current study seeks to determine how study design and analytical features can reduce bias in non-experimental estimates of the effects of earlier skills on later math skills.

We use experimental data from an early math intervention with end-of-treatment and medium-term (defined as two years post-intervention) impacts to determine how different analytical approaches applied to a combination of the end-of-treatment impacts

can yield accurate forecasts of the medium-term impacts. We use the control group data to model the application of these approaches to data from a non-experiment. First, we forecast the effects of the intervention on children's math achievement two years later, conditional on end-of-treatment outcomes, using various design and analytic features. Then, we explore how these specifications relate to the accuracy of our forecasts.

Overall, this work will help identify preferable design and analytic specifications to forecast the medium-term impacts of math skill-building interventions using longitudinal correlational data more accurately. We hope that researchers will further test the usefulness of these approaches when they combine experimental data with pre-existing non-experimental data (i.e., public longitudinal datasets) for calculating an estimated range of plausible impacts for power-analysis, determining which of kind of skills to target with educational interventions, making predictions about the longer-run impacts of a program change that is yet to be observed, and comparing estimated plausible impacts for theory testing and revision. We return to a specific explanation on these uses in the discussion section.

How Can We Use Previous Research to Forecast Intervention Impacts?

Non-experimental methods are widely used to identify variables that might be manipulated (e.g., early academic skills) to produce a desirable change in a later outcome (e.g., later academic achievement or educational attainment). A set of early skills that statistically predicts children's later academic achievement may represent a set of targets for potential intervention. However, while longitudinal non-experimental 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

estimated effects of 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 distributing measured and unmeasured variables equally, on average, between children who receive an intervention and children who do not, allowing for an unbiased estimate of a causal impact. However, conducting randomized experiments to assess potential causal impacts 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 leveraging the widely available nonexperimental longitudinal data and experimental designs is forecasting the medium-term outcomes of treatment effects using a combination of end-of-treatment experimental outcomes and correlations between end-of-treatment outcomes and medium-term outcomes from longitudinal data. Although the current study is concerned with quantitative forecasts, 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.

Quantitative forecasts are sometimes made by combining short-run experimental impacts with external estimates of the association between short-run and longer-run outcomes. A recent paper by Watts (2020) reviews some of this work and suggests that researchers should be careful when using the student achievement-to-earnings correlation to project the long-term effects of educational interventions. Some examples of such calculations include the work by Chetty et al., (2011) using students' random assignment to kindergarten classrooms to estimate how class quality raised average tests scores and adulthood earnings, as implied by the association between test scores and earnings. In another 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 estimated effect 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.

Recently, more methods for forecasting intervention impacts, which use a combination of information about short- and longer-run experimental impacts are being developed. 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 using a surrogate index of the difference in employment rates and earnings between the experimental and control group, in the first 1.5 years after the end of the program, they could forecast the mean impact on employment rates after nine years quite well. Thus, there is a need to explore which analytical decisions improve accuracy of forecasts using both experimental and longitudinal data can be improved to predict child skill development and medium-to-long-term program impacts.

Ideally, such an approach 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. In the current study, we forecast impacts by multiplying the treatment effect of an intervention on an end-of-treatment outcome by the regression coefficient of the medium -term outcome regressed on the same end-of-treatment outcome. However, several sources of bias may influence the accuracy of these forecasts, below we describe potential sources of bias and methods that can be used to reduce bias.

Threats to Accurate Forecasts

We describe potential biases in intervention forecasts in the context of forecasting the impact of a first-grade math intervention for at-risk students that followed students until third grade, using the end of first-grade treatment outcomes to predict the treatment impacts on children's math achievement at the end of third grade. We also detail how threats to accuracy relate to the real-world limitations of evaluating interventions at scale. In Figure 1, we demonstrate our causal assumptions using directed acyclical graphical (DAG) notation (Pearl, 2009; Cunningham, 2021) where directed lines represent the causal impact of one variable on another in a solid line when it is a measured relation in a dashed line when it is an unmeasured relation. DAG notation allows us to represent our assumptions about causal relations to determine where and to what extent we should expect bias to interfere with the accuracy of our forecasts. Figure 1 demonstrates how we conceptualize the causal impact of a *treatment* (during first grade) on an end-of-treatment outcome called *Skill 1* at time *T1* (the end-of-treatment impact estimate). In this figure, *Skill 1 T1* is the skill targeted by the intervention. *Skill 1 T1* is expected to influence the same

Skill 1 at time *T2* (end of third grade), which we refer to as the medium-term outcome. *Skill 1 T1* and *Skill 1 T2* are both shown in ovals because they represent latent skills that are not directly observed. Rather, we observe test scores, depicted by rectangles, which are caused by a combination of these latent skills and other factors. Thus, we measure *Skill 1 T1* with the *end-of-treatment outcome test score* and *Skill 1 T2* with the *medium-term outcome test score*. In the simplest case, we would forecast the impact of the intervention on *Skill 1 T2* by multiplying the observed treatment impact on *Skill 1 T1* by the estimated effect of a 1-unit change in *Skill 1 T1* on *Skill 1 T2.* This estimate could be obtained from an existing non-experimental longitudinal dataset that covers similar constructs and age ranges as the intervention study of interest. In the current study, we estimate the effect of *Skill 1 T1* on *Skill 1 T2* using the data from the control group of a randomized controlled trial (RCT).

Sources of Over-Prediction

Multiplying first-grade impacts with first-to-third grade correlations may bias third grade impacts of interventions 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 both the end-of-treatment outcome and the medium- or long-term outcome. For example, unmeasured stable individual and environmental variables plausibly impact child math achievement in first and third grade. Therefore, a forecast of third-grade math skills will over-state the impact of first grade skills if unmeasured stable individual or environmental factors exert a positive influence on first and third grade skills (Bailey et al., 2018). Omitted variable bias is shown in Figure 1 with dotted arrows pointing from an oval towards *Skill 1 T1* and *Skill 1 T2*; we expected these to upwardly bias the estimated effect of *Skill 1 T1* on *Skill 1 T2.* 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 non-experimental datasets, or it may be too costly or difficult to collect.

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 is over-aligned with an intervention to the extent that it measures student's memorization that 3 x 2 = 6 because this specific item was taught repeatedly in the intervention. As shown in Figure 1, over-alignment occurs when the *treatment* increases a student's ability on the *end-of-treatment outcome test score* (e.g., answering 3 X 2 = correctly) that does not have the same impact on the latent ability (e.g., understanding multiplication), shown as *Skill1 T1*, that the *end-of-treatment outcome test score* measures.

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. Over-alignment bias is shown in Figure 1 as the dotted arrow from treatment to the *end-of-treatment outcome test.* We expect over-alignment to upwardly bias the estimated impact of the *treatment* on *Skill 1 T1* by inflating the *end-of-treatment outcome test score* impacts relative to the actual

treatment impacts on *Skill 1 T1* reflected in the test scores from similar non-experimental samples.

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 in multiplication knowledge if a multiplication intervention focuses on children's conceptual understanding of the multiplication procedure, which may help them remember the procedure for longer, but assess the impact of the intervention only with multiplication problems and do not include an assessment of conceptual understanding of the multiplication procedure. In Figure 1, the dash-and-dot arrow from *treatment* to *Skill 2 T1* (conceptual understanding of multiplication in this case) reflects under-alignment bias, which we expected to under-estimate the impact of treatment on Skill 1 T2 (later multiplication performance) by omitting the impact of *Skill 2 T1* on *Skill 1 T2*. If we use the end-of-treatment impacts of an intervention to forecast the medium-run impact on *Skill 1 T2*, we would under-estimate the impacts of the intervention because the impact on Skill *2 T1* was unmeasured.

An important distinction between measures that are conceptually proximal to an intervention is that they may be well-aligned or over-aligned measures. The difference is that we would not expect impacts on over-aligned measures to transfer to conceptually

distal broader math assessments because a student's memorization of a few math facts are not indicative of conceptually understanding multiplication thus these tests would not be measuring the same thing and should not be compared. However, impacts on a proximal well-aligned measure that narrows in on students conceptual understanding of multiplication might forecast gains on a more distal broader based assessment of math knowledge, because a conceptual understanding of 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 end-of-treatment impacts.

(1) Using demographic and pretest covariates should reduce the bias from estimating the causal impact of an increase in an early math skill on later math skills. Estimates of the causal impact of an early math skill on a later math skill should approach the observed causal impact without surpassing it when confounding variables like prior knowledge are controlled.

(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. Math achievement is contingent on numerous math skills that interact with one another; since they reflect some of the same factors, modeling them separately might "double count" end-of-treatment impacts that manifest in more than one end-of-treatment outcome measure. We explain the two alternative modeling approaches in the analytical strategy.

(3a) Using the end-of-treatment 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 end-of-treatment 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 will be calculated by using a combination of end-of-treatment outcomes that are conceptually proximal and distal to the treatment. Combining both kinds of end-of-treatment outcomes will reduce over-alignment bias – because conceptually distal outcomes contain variance unlikely to be affected by the intervention, which may be shared with the underlying skills targeted by the intervention. It will also reduce under-alignment bias, because conceptually proximal measures will capture variance in the skills targeted by the intervention.

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 end-of-treatment targets to train and for power analysis for detecting longer-run impacts), to funding organizations interested in forecasting the effects of proposed interventions on student achievement years after the end of treatment, and for researchers and policy analysts attempting to model future program benefits.

METHODS

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 end-of-treatment outcomes (7%) or all the medium-term outcomes (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, Geary, et al., 2013). Our sample is thus similar but not identical to the Fuchs, Geary, et al. (2013) sample as we included students that had at least end-of-treatment 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 in order 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 end-of-treatment outcomes, collected at the end of first grade, are measures of latent student *skill 1* at time *T1* as show in Figure 1. The end-of-treatment measures include measures that are both conceptually proximal and 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[1]. 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 (shown in Figure 2 Panel A path $C_{Experimental}$) by regressing each medium-term outcome on each of the two treatment conditions while controlling for child demographic and math pretest covariates using

[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 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.

classroom level clustered standard errors. Second, we calculate the experimental impact on each end-of-treatment outcome (Figure 2 Panel B path $a_{Experimental}$) by regressing each end-of-treatment 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 estimated effect of each end-of-treatment outcome on each medium-term outcome (Figure 2 Panel B path $b_{Non\text{-}experimental}$) by regressing each medium-term outcome on each end-of-treatment outcome, using only the control group data. The $b_{Non\text{-}experimental}$ 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 end-of-treatment impact (Figure 2 Panel B path $a_{Experimental}$) by the estimated effect of end-of-treatment test scores on third-grade test scores (Figure 2 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 that attempts 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 long-run 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). However, because much of the previous literature does not use explicit quantitative forecasts, it was not obvious which way to combine the experimental and non-experimental estimates. 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.

The first approach we chose to model assumes that only a single measure was collected at the end-of-treatment, we show this in Figure 3 Model A and hereafter refer to this approach as forecasting using a single end-of-treatment outcome. In model A we estimate paths *a* and *b* using multivariate regressions where: Path a_1 in Figure 3 Model A is the regression coefficient of treatment TRT_{iG1} on each end-of-treatment outcome EOT_1

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

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, and μ_{G1} is the classroom level residual since students are clustered in classrooms. Path b_1 is the regression coefficient of the end-of-treatment outcome EOT_{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 EOT_{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) lastly the model will also include distal pretests. The forecasted impact of each end-of-treatment outcome is calculated by multiplying path *a* and *b* for each combination of the 8 end-of-treatment outcomes predicting each of the 5 medium term outcomes for each of the two treatment arms, resulting in 80 forecasts that are shown in Figure 4.

The second approach assumes that a medium-term outcome is independently influenced by different end-of-treatment 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 $a_{1...n}$ and $b_{1...n}$ in the same way as model A (exact model estimates are shown in Supplementary Table 2). However, the forecast for each medium-term outcome is calculated by multiplying paths a and b are for each of the 8 end-of-treatment 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 shown in Figure 5 Plot A.

The third approach assumes that an intervention can impact a medium-term outcome through multiple dependent mediators with overlapping paths of influence from the end-of-treatment 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 $a_{1...n}$ and $b_{1...n}$ using multivariate regressions where:

Path a_1 in Figure 3 Model C is the regression coefficient of treatment TRT_{iG1} on each end-of-treatment outcome EOT_1

$$EOT_{iG1} = \beta_0 + \beta_1 TRT_{iG1} + \beta_2 X_{iG1} + \beta_3 OEOT_{iG1} + \epsilon_{iG1} + \mu_{G1}$$

Where, $OEOT_{iG1}$ is a vector containing all the end-of-treatment outcome measures. The b paths in Figure 3 are the regression coefficients on all of the end-of-treatment outcomes $OEOT_{iG1}$ as predictors of each medium-term outcome MTO_{iG3}

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

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 3). The forecasts for Model C are shown in 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 on Figure 3 with the inclusion of pretest covariates that can account for omitted variables that confound the association between the short-term and medium-term outcomes performed best. However, because Model A showed more variability, we find Model C to be a more promising approach– the assumptions implicit in this approach are theoretically appropriate for estimating the impact of an intervention on numerous math skills that are expected to be reflected to various degrees in different outcome measures.

We tested if over-estimation bias from conceptually proximal measures and under-estimation bias from conceptually distal measures could be reduced using three heuristics. The three heuristics we tested were (1) forecasting using the conceptually proximal end-of treatment outcome with the smallest treatment impact, (2) forecasting using the

conceptually distal end-of treatment outcome with the largest treatment impact, and (3) forecasting using the average of both the conceptually proximal end-of treatment outcome with the smallest treatment impact and the conceptually distal end-of treatment 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 forecasts with the most accurate results we calculate a measure of bias from the causal estimate 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, Geary, 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) in categorizing the end-of-treatment 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 end-of-treatment 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 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 α = .86. The *Wide Range Achievement Test–3 Arithmetic* (WRAT-Arithmetic; Wilkinson, 1993) subtest measured students 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

We excluded 45 students that did not complete at least one end-of-treatment outcome (7.0%) and an additional 90 students that did not complete least one medium-term outcome (14.1%). 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 Table 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 estimated effect of an end-of-treatment outcome on a medium-term outcome calculated without covariates. By

not accounting for common confounding variables that exert a positive influence on both end-of-treatment 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 end-of-treatment 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. The majority of the over-estimated forecasts were calculated using conceptually proximal end-of-treatment measures (shown in green). 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 end-of-treatment measures (shown in blue). The average of all these forecasts (0.123 SD) is shown in Table 2 Column (2), this is the value of the red dot which is more than double the experimental benchmark of 0.052 SD. The exact values of each forecast plotted and the bias from the experimental benchmark are shown in 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 are controlled.

Forecasting Approaches

The simplest methodological approach to forecasting is making predictions conditional on a single end-of-treatment outcome, as conceptually shown in Figure 3 panel A. Each marker on Figure 4 plot B (circles and triangles) reflect a single combination of one of the 8 end-of-treatment outcomes predicting one of the 5 long-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 end-of-treatment 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 end-of-treatment 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 end-of-treatment 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 effects of the end-of-treatment 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 end-of-treatment outcome to predict the medium-term outcome yielded the most accurate forecasts. In comparison to the other approaches, using a single end-of-treatment 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, end-of-treatment 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 end-of-treatment outcomes that are more conceptually proximal with the intervention will yield over-estimated forecasts whereas the conceptually distal end-of-treatment outcomes would yield under-estimated forecasts. Figure 4 plot B demonstrates that the proximal measures (green 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 end-of-treatment 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 end-of-treatment measures over-estimate treatment impacts more than conceptually distal end-of-treatment 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 by of the 37 forecasts that over-estimated the medium-term treatment impacts in the NKT study 29 were from the speeded condition while 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 we would find that if we calculate forecasts using the exact same end-of-treatment 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.

In contrast, in the PKM study, the forecasts were less sensitive to the conceptual proximity of the end-of-treatment measures as both the proximal and distal measures over-predicted the experimental benchmark (Supplementary Table 11). We believe over-prediction could be due, in part, to residual bias from omitted variables in the PKM study because of fewer baseline measures and plausibly noisier baseline pretests in the younger PKM sample in contrast to the NKT sample.

Conceptual 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 conditional on pre-K end-of-treatment outcomes. 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 1, Plot B). Furthermore, we found similar patterns of accuracy using the three different approaches to forecast medium-term outcomes. As shown in Supplementary Figure 2, 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 end-of-treatment 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.

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 end-of-treatment impacts in both datasets were of similar average magnitudes (0.34 in NKT and 0.40 in PKM, Table 1 and Supplementary Table 7), the forecasts for each of the end-of-treatment outcomes in the PKM dataset under full controls (0.35 and 0.48; Supplementary Table 8) 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 8, 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 end-of-treatment outcome. However, this approach was not validated in the conceptual replication, where omitted variables bias was apparently 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 end-of-treatment outcomes. These results demonstrate that in this particular case, early math skills influenced later math skills via largely overlapping pathways. Interestingly, using two end-of-treatment outcomes based on their theoretical alignment with the intervention yielded more accurate forecasts than using all end-of-treatment outcomes assuming multiple dependent

pathways. We hypothesize this may be due to the additional omitted variables that confound the relation between end-of-treatment outcomes and medium-term outcomes.

By assessing multiple measures in the NKT as end-of-treatment 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 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 which was 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. 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 end-of-treatment 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 non-experimental 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 end-of-treatment 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 end-of-treatment 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 end-of-treatment 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 possible 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.

Limitations

We demonstrate study design and analytic approaches that yield the most accurate forecasts for early math skill interventions. However, this work would benefit from extensions to other areas of intervention targeting early skills to improve later skills, such as literacy interventions. In the current analysis, the time delay between the end of the intervention and the medium-term outcomes that were forecasted was two years, so we have yet to understand how accurately we can forecast even longer-run outcomes. It would be useful to understand how accurate forecasts are by implementing this approach with datasets showing larger long-term impacts. It is important to note that the medium-term outcomes measured in third grade, two years after the end-of-treatment, may also be affected by unmeasured peer effects (Xu, Zhang, & Zhou., 2020). Future work might

consider whether composition of treated peers affects later learning (e.g., Jenkins et al., 2018) and whether such peer effects might be forecastable using non-experimental data. Another practical limitation to the generalizability of our work is attempting to reduce omitted variables bias when a large and diverse set of demographic and pretest covariates, such as those used in our sample, are not available. One approach is to model the stability of stable individual and environment traits using multiple waves of data (Bailey et al., 2018). In cases where neither a rich set of covariates or multiple waves of measurement are available, different approaches must be developed to determine optimal forecast accuracy, or to estimate a confidence interval for the forecast.

Potential Uses

Three research applications involve 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 end-of treatment 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.

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 use measurement invariance testing to determine whether theorized constructs are changed at the levels hypothesized by the interventionist, and if not, whether building partial invariance between the treatment and control groups might further improve forecasts. 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 non-experimental 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-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 (Waller & Meehl, 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 non-experimental work.

Tables

Table 1.1

Number Knowledge Tutoring Treatment Impacts on End-of-Treatment and Medium-Term Outcomes

Outcome	Speeded v. Control Estimate (SE)	Non-Speeded v. Control Estimate (SE)
End-of-Treatment 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 or not 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 1.2

Average Forecasts Using Three Approaches and Resulting Bias

	Experimental Benchmark	Forecast Independent Single EOT Outcome				Multiple Independent EOT Outcome		Multiple Dependent EOT 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

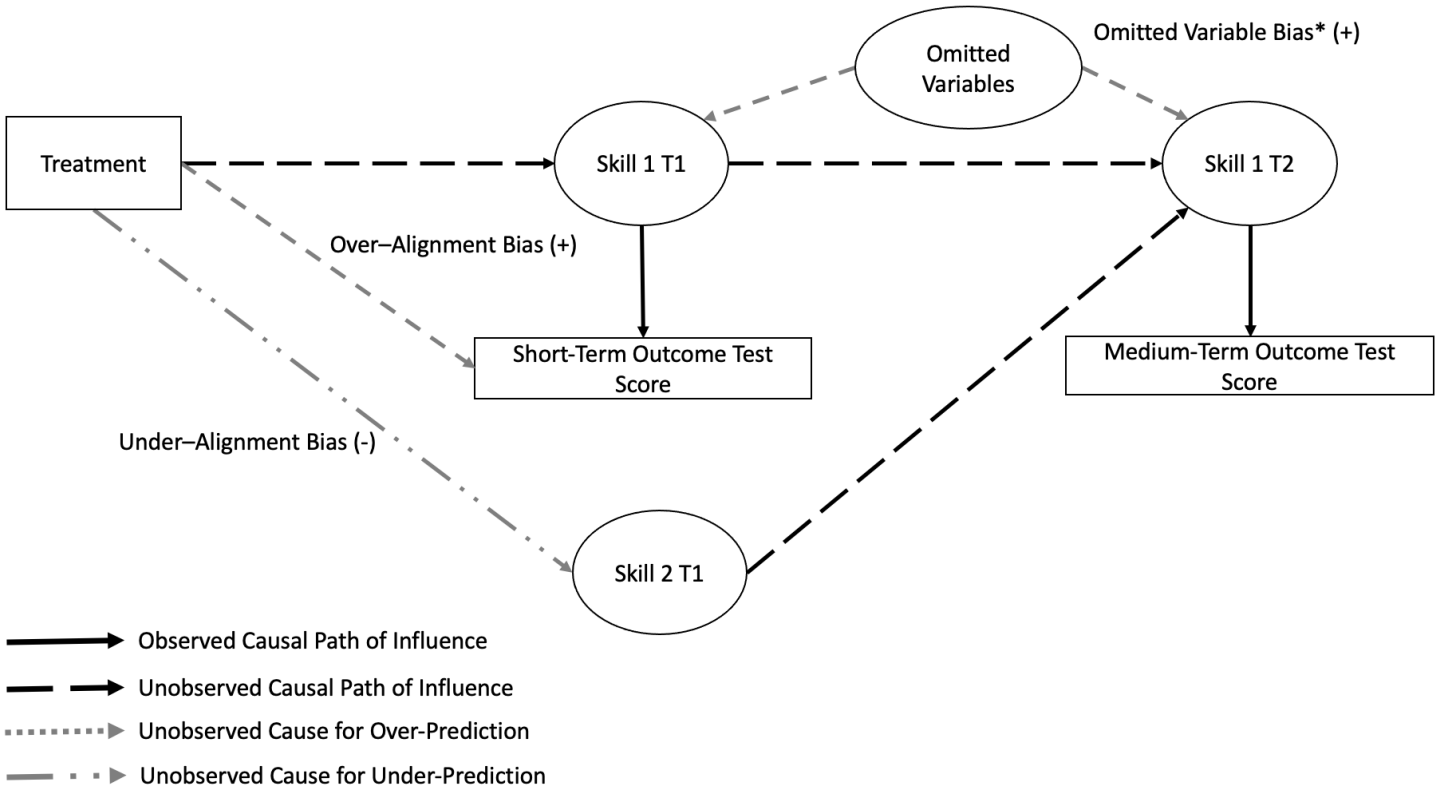
KeyMath-Numeration	0.07	0.123	0.054	0.056	-0.01	0.444	0.41	0.138	0.074
Full Covariates	X			X		X		X	

Note. EOT = End-of-treatment. 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 end-of-treatment 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 end-of-treatment and medium-term outcomes.

Figures

Figure 1.1

Sources of Bias in Forecasting Medium-Term Intervention Impacts

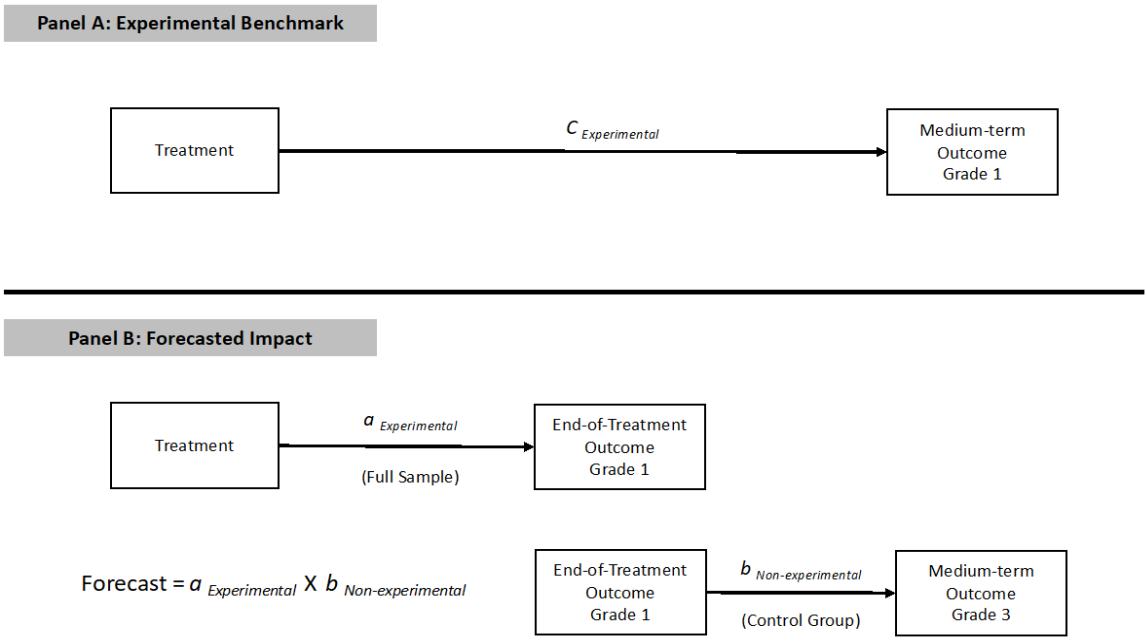


Note. Directed Acyclical Graph and Structural equation modeling notation is used in this figure to represent the causal paths that we expected to be influencing the key variables in our forecast. * For simplicity the potential influence of omitted variables bias is only shown to impact Skill 1 T1 and Skill

1 T2, however this bias would also be expected to impact Skill2 T1 and Skill2 T2. Similarly, under-alignment bias is represented by a single alternative unmeasured skill (skill 2 T1) however, we would expect under-alignment to influence measurement error at time T2. We also expect that Skill 2 T1 would influence Skill 1 T1, yet we do not include this relation in our model to simplify the explanation of this bias on later skills.

Figure 1.2

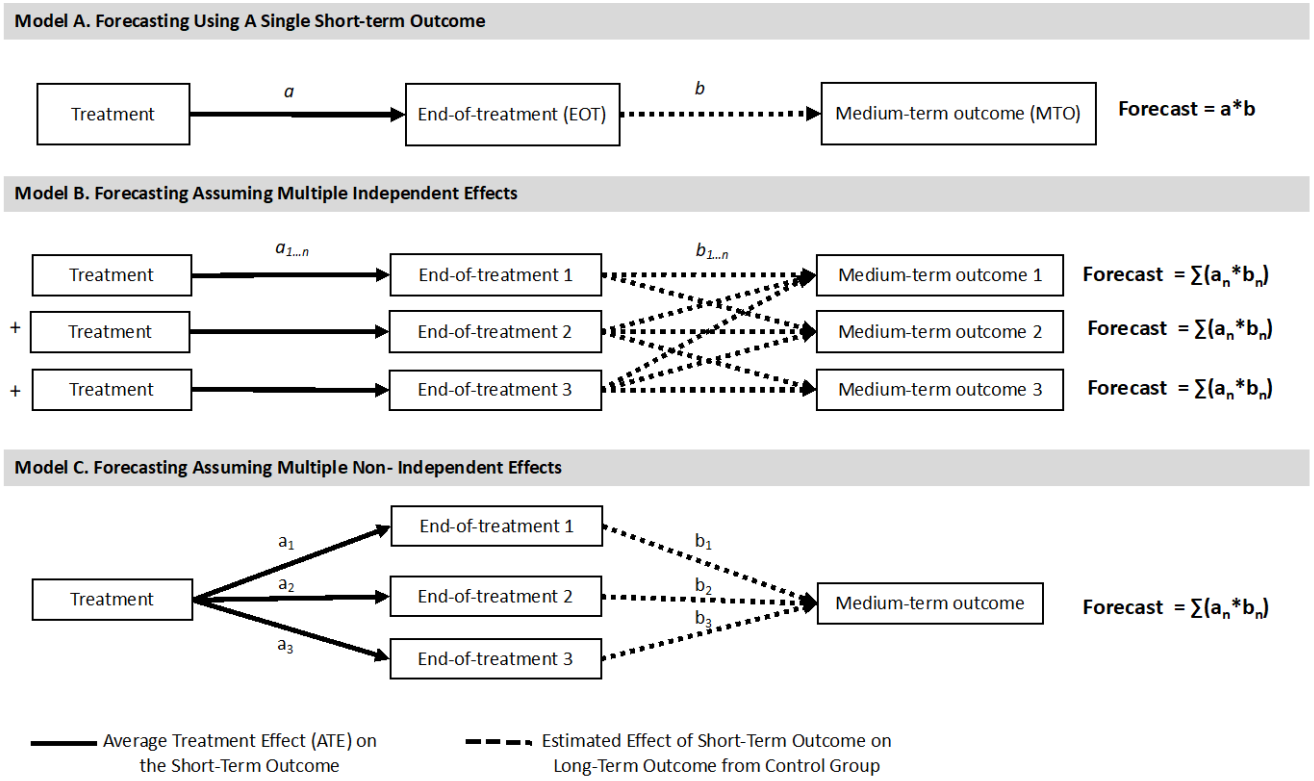
Conceptual Framework of Within Study Comparison of Number Knowledge Tutoring



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 end-of-treatment 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 estimated effect (regression coefficient) of a end-of-treatment outcome on a medium-term outcome using the control group data. Third, we calculate forecasts by multiplying the treatment effect on the end-of-treatment outcome by the estimated effect of the end-of-treatment 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 1.3

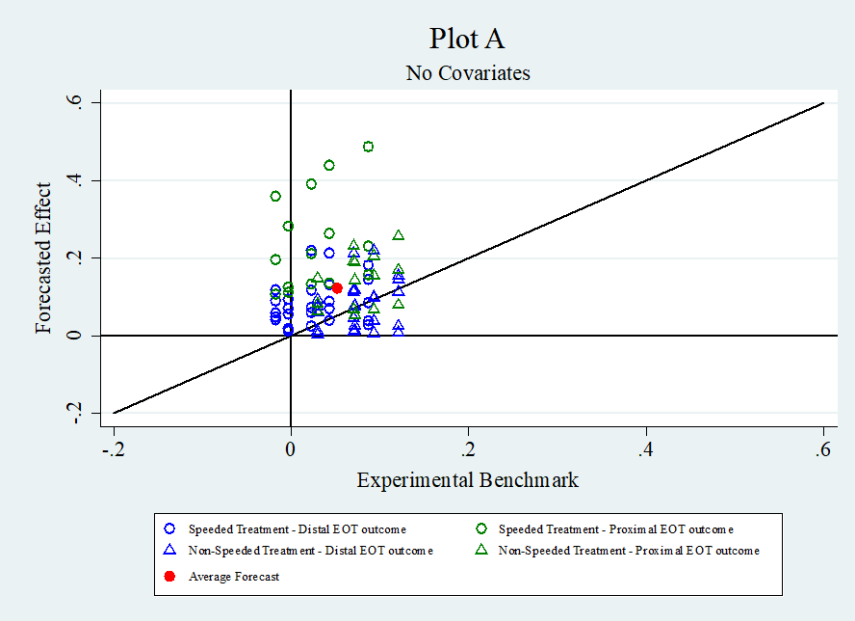
Conceptual Models of Forecasting Methods

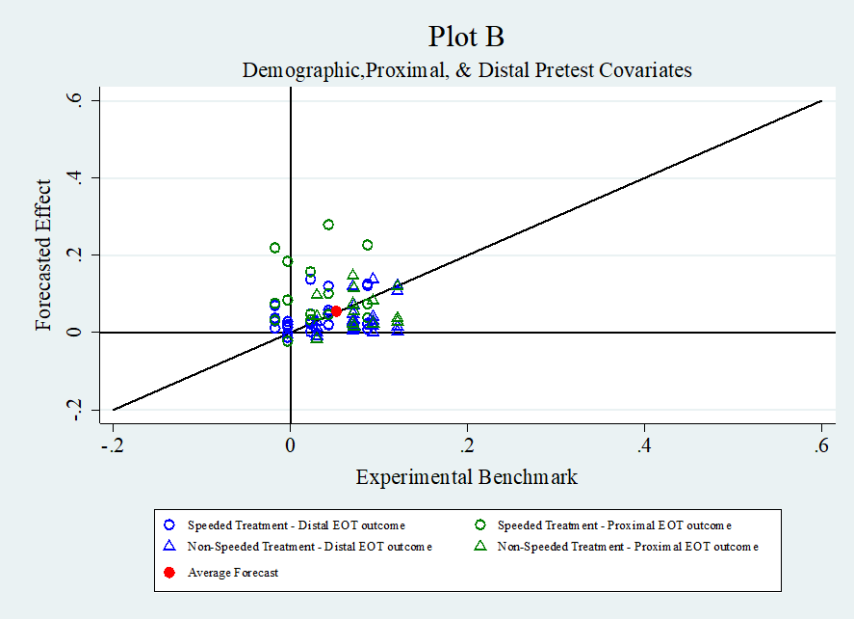


Note. Three different approaches to calculating forecasts are shown. Panel A shows how we forecast a single medium-term outcome using a single end-of-treatment outcome; the treatment impact on the end-of-treatment outcome is multiplied by the regression coefficient of regressing the medium-term outcome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

Figure 1.4

Replicating and addressing omitted variables bias



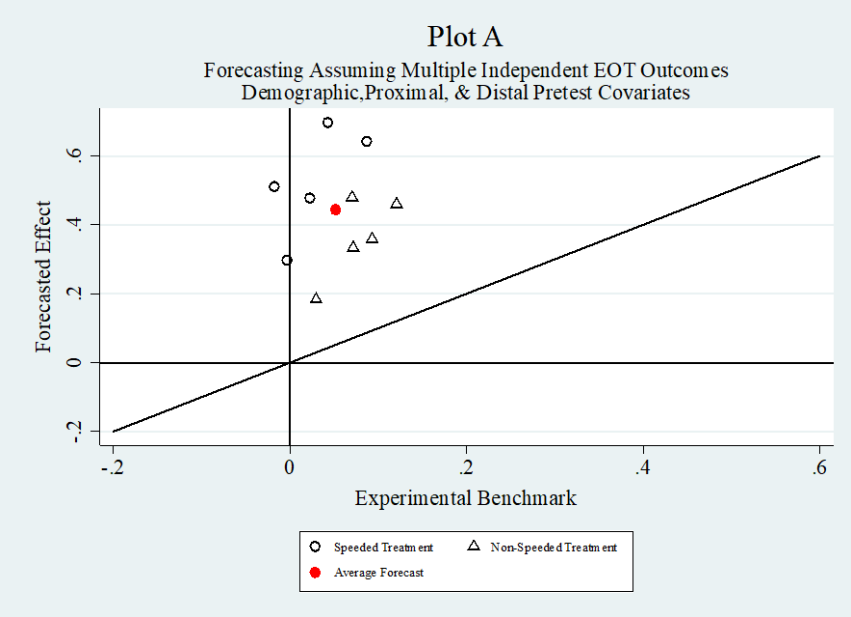


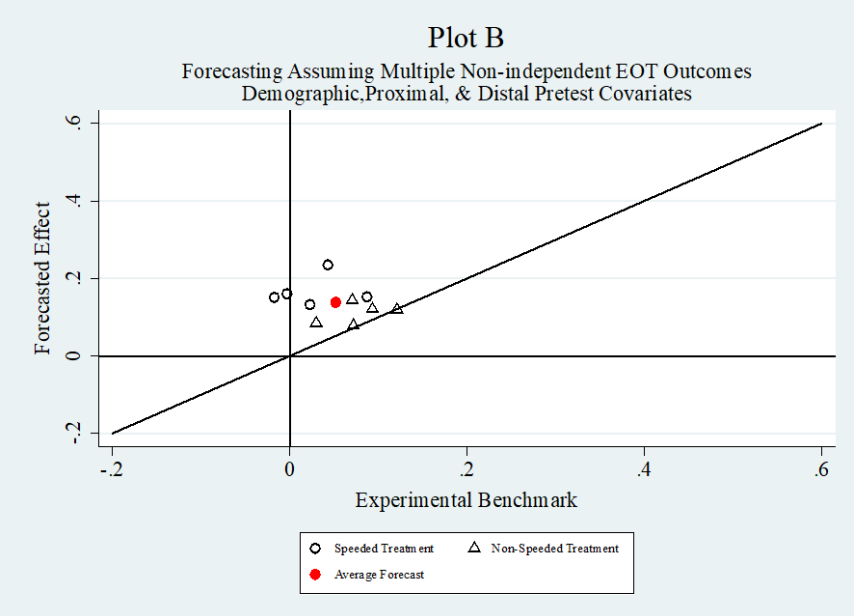
Note. EOT= End-of-Treatment. Each marker on the plots represents a forecast calculated using a single end-of-treatment outcome to predict each single medium-term outcome within each treatment. Forecasts

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 red circle, this is calculated as the average of all the forecasts in the same plot. Blue markers indicate forecasts calculated with distal end-of-treatment outcomes and green markers indicate forecasts calculated with proximal end-of-treatment outcomes.

Figure 1.5

Forecasting with Multiple End-of-treatment Outcomes





Note. EOT= End-of-Treatment. Each marker on the plots represents a forecast calculated using all the end-of-treatment 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 red circle, this is calculated as the average of all the forecasts in the same plot.

SUPPLEMENTARY MATERIAL

Supplementary Table 1.1

Number Knowledge Tutoring Descriptive Statistics and Baseline equivalence Split by Treatment Group

	Control			Speeded			Non-Speeded			Speeded v. Control	Non-Speeded v. Control	Speede v. Non- Speede
	N	***M***	***SD***	***N***	***M***	***SD***	***N***	***M***	***SD***	***p***	***p***	***p***
Student Demographics												
Age in Years	173	6.48	0.36	161	6.50	0.37	167	6.46	0.38	0.59	0.70	0.37
Male	173	48%		161	53%		167	48%		0.38	0.99	0.38
African American	173	73%		161	69%		167	67%		0.44	0.25	0.72
White	173	17%		161	21%		167	21%		0.31	0.32	0.97
Hispanic	173	6%		161	7%		167	8%		0.69	0.48	0.76
Other or missing	173	4%		161	2%		167	4%		0.43	0.83	0.56
Free or Reduced-Price Lunch	173	88%		161	84%		167	77%		0.29	0.01**	0.10
English as a Second Language	173	3%		161	2%		167	2%		0.82	0.51	0.67
Missing sex	173	1%		161	0%		167	0%		0.34	0.33	
Missing Free or Reduced-Price Lunch	173	1%		161	0%		167	0%		0.34	0.33	
Missing race	173	1%		161	0%		167	0%		0.34	0.33	
Missing English as a Second Language	173	1%		161	1%		167	0%		0.96	0.33	0.31
Pretests (Fall 1st Grade)												
Arithmetic Combinations	173	12.39	7.13	161	12.64	7.71	167	12.65	6.71	0.76	0.73	0.99
Double-Digit Calculation	173	0.42	0.98	161	0.49	1.11	167	0.43	0.91	0.55	0.98	0.56
Facts Correctly Retrieved	173	1.38	2.07	161	1.51	1.91	167	1.40	2.16	0.54	0.93	0.61
Number Sets	173	-0.53	0.72	161	-0.52	0.83	167	-0.48	0.67	0.88	0.49	0.63

Story Problems	173	1.69	1.87	161	1.74	1.59	167	1.83	1.88	0.81	0.51	0.65
WRAT-Arithmetic	173	88.75	11.9	161	89.32	12	167	90.05	12.72	0.66	0.33	0.59
Number Line	173	-26.36	6.62	161	-25.76	7.01	167	-25.55	6.03	0.42	0.24	0.77
KeyMath-Numeration	173	97.86	10.4	161	97.05	9.87	167	97.81	10.56	0.47	0.97	0.50
End-of-treatment Outcomes (Spring of 1st Grade)												
Arithmetic Combinations	173	22.25	11.4	161	33.45	14.2	167	28.29	11.52	0.00***	0.00***	0.00***
Double-Digit Calculation	173	1.85	2.79	161	4.20	3.95	167	3.62	3.23	0.00***	0.00***	0.14
Facts Correctly Retrieved	173	3.80	3.35	161	5.20	3.86	167	4.51	3.72	0.00***	0.06	0.1
Number Sets	173	-0.80	1.27	161	-0.37	1.35	167	-0.39	1.16	0.00**	0.00**	0.92
Story Problems	173	2.84	2.28	161	3.35	2.48	167	3.59	2.50	0.05	0.00**	0.37
WRAT-Arithmetic	173	92.05	15.1	161	96.88	13.4	167	98.00	14.95	0.00**	0.00***	0.48
Number Line	173	-21.37	7.57	161	-20.32	7.26	167	-20.71	7.03	0.2	0.41	0.62
KeyMath-Numeration	173	100.23	10.2	161	101.30	9.09	167	101.59	10.44	0.31	0.23	0.79
Medium-term Outcome (Spring 3rd grade)												
Facts Correctly Retrieved	173	8.47	3.69	161	8.52	3.52	167	8.67	3.58	0.90	0.61	0.69
Number Sets	173	-0.68	1.82	161	-0.48	1.78	167	-0.37	1.31	0.31	0.07	0.52
WRAT-Arithmetic	173	90.83	14.92	161	91.25	13.36	167	93.22	15.09	0.78	0.14	0.21
Number Line	173	-12.02	5.90	161	-11.85	5.60	167	-11.3	5.90	0.78	0.23	0.35
KeyMath-Numeration	173	97.08	10.65	161	97.67	9.72	167	98.53	10.09	0.60	0.20	0.43

Note. * p<0.05 ** p<0.01 *** p<0.001. P-values are drawn from two-tailed t-test group comparisons. Raw scores are reported for all outcome measures.

Supplementary Table 1.2

Number Knowledge Tutoring Correlation Table of End-of-Treatment and Medium-term Outcome Measures

	Variables	1	2	3	4	5	6	7	8	9	10	11	12
End-of-treatment Outcomes													
1	Arithmetic Combinations	–											
2	Double-Digit Calculation	0.65***	–										
3	Facts Correctly Retrieved	0.58***	0.42***	–									
4	Number Sets	0.59***	0.47***	0.53***	–								
5	Story Problems	0.41***	0.36***	0.34***	0.45***	–							
6	WRAT-Arithmetic	0.45***	0.36***	0.43***	0.44***	0.36***	–						
7	Number Line	0.33***	0.26***	0.25***	0.33***	0.30***	0.33***	–					
8	KeyMath-Numeration	0.41***	0.28***	0.37***	0.46***	0.39***	0.65***	0.40***	–				

Medium-term Outcome

9	Facts Correctly Retrieved	0.26***	0.18***	0.26***	0.20***	0.19***	0.20***	0.12**	0.20***	–			
10	Number Sets	0.45***	0.32***	0.37***	0.54***	0.36***	0.40***	0.28***	0.40***	0.34***	–		
11	WRAT-Arithmetic	0.40***	0.29***	0.31***	0.39***	0.31***	0.56***	0.26***	0.49***	0.39***	0.48***	–	
12	Number Line	0.33***	0.24***	0.24***	0.32***	0.28***	0.33***	0.45***	0.42***	0.18***	0.33***	0.35***	–
13	KeyMath-Numeration	0.42***	0.26***	0.31***	0.43***	0.41***	0.56***	0.38***	0.64***	0.26***	0.48***	0.58***	0.45***

Note. * p<0.05 ** p<0.01 *** p<0.001. N=501.

Supplementary Table 1.3

Estimated Effects Assuming Independent Pathways ($b_{Non\text{-}experimental}$)

Medium-Term Outcome (Spring of 3rd Grade)	End-of-Treatment Outcome (Spring of 1st Grade)	None	Demographics	Demographics & 1st grade pretests	Demographics, 1s - & cross- grade pretests
Facts Correctly Retrieved	Arithmetic Combinations	0.30*** (0.08)	0.29*** (0.08)	0.19 (0.10)	0.19 (0.10)
	Double-Digit Calculation	0.14 (0.08)	0.11 (0.09)	-0.03 (0.09)	-0.03 (0.09)
	Facts Correctly Retrieved	0.32*** (0.07)	0.31*** (0.08)	0.22* (0.09)	0.22* (0.09)
	Number Sets	0.22** (0.08)	0.19* (0.08)	0.04 (0.09)	0.04 (0.09)
	Story Problems	0.26*** (0.07)	0.23** (0.08)	0.10 (0.09)	0.10 (0.09)
	WRAT-Arithmetic	0.28*** (0.07)	0.31*** (0.09)	0.09 (0.13)	0.09 (0.13)
	Number Line	0.12 (0.08)	0.10 (0.08)	-0.01 (0.09)	-0.01 (0.09)
	KeyMath-Numeration	0.18* (0.08)	0.17 (0.10)	-0.12 (0.12)	-0.12 (0.12)
Number Sets	Arithmetic Combinations	0.51*** (0.08)	0.49*** (0.08)	0.24* (0.11)	0.24* (0.11)
	Double-Digit Calculation	0.29*** (0.07)	0.25** (0.08)	0.05 (0.07)	0.05 (0.07)
	Facts Correctly Retrieved	0.40*** (0.08)	0.41*** (0.09)	0.19 (0.10)	0.19 (0.10)
	Number Sets	0.56*** (0.08)	0.55*** (0.08)	0.39*** (0.10)	0.39*** (0.10)
	Story Problems	0.40*** (0.06)	0.39*** (0.07)	0.10 (0.08)	0.10 (0.08)
	WRAT-Arithmetic	0.43*** (0.06)	0.61*** (0.07)	0.36*** (0.10)	0.36*** (0.10)
	Number Line	0.27** (0.08)	0.26** (0.09)	0.09 (0.10)	0.09 (0.10)
	KeyMath-Numeration	0.38*** (0.08)	0.50*** (0.10)	0.23 (0.13)	0.23 (0.13)
WRAT-Arithmetic	Arithmetic Combinations	0.41*** (0.06)	0.38*** (0.06)	0.17* (0.08)	0.17* (0.08)
	Double-Digit Calculation	0.26*** (0.06)	0.21*** (0.06)	0.04 (0.06)	0.04 (0.06)
	Facts Correctly Retrieved	0.34*** (0.07)	0.29*** (0.07)	0.12 (0.07)	0.12 (0.07)
	Number Sets	0.36*** (0.08)	0.30*** (0.07)	0.15* (0.07)	0.15* (0.07)
	Story Problems	0.34*** (0.07)	0.30*** (0.08)	0.11 (0.10)	0.11 (0.10)
	WRAT-Arithmetic	0.65*** (0.06)	0.62*** (0.08)	0.41*** (0.10)	0.41*** (0.10)
	Number Line	0.23* (0.09)	0.23** (0.08)	0.03 (0.07)	0.03 (0.07)
	KeyMath-Numeration	0.58*** (0.07)	0.50*** (0.08)	0.26** (0.10)	0.26** (0.10)

Number Line	Arithmetic Combinations	0.38*** (0.06)	0.37*** (0.05)	0.23* (0.09)	0.23* (0.09)
	Double-Digit Calculation	0.24*** (0.06)	0.20** (0.06)	0.09 (0.09)	0.09 (0.09)
	Facts Correctly Retrieved	0.27** (0.09)	0.21* (0.10)	0.08 (0.10)	0.08 (0.10)
	Number Sets	0.28*** (0.07)	0.26*** (0.07)	0.12 (0.09)	0.12 (0.09)
	Story Problems	0.27*** (0.08)	0.24** (0.08)	0.05 (0.08)	0.05 (0.08)
	WRAT-Arithmetic	0.35*** (0.07)	0.45*** (0.08)	0.21 (0.12)	0.21 (0.12)
	Number Line	0.45*** (0.10)	0.42*** (0.10)	0.31* (0.12)	0.31* (0.12)
	KeyMath-Numeration	0.40*** (0.06)	0.49*** (0.07)	0.29*** (0.08)	0.29*** (0.08)
KeyMath-Numeration	Arithmetic Combinations	0.46*** (0.07)	0.44*** (0.07)	0.29*** (0.08)	0.29*** (0.08)
	Double-Digit Calculation	0.33*** (0.07)	0.27*** (0.06)	0.12 (0.07)	0.12 (0.07)
	Facts Correctly Retrieved	0.35*** (0.07)	0.29*** (0.07)	0.12 (0.07)	0.12 (0.07)
	Number Sets	0.41*** (0.06)	0.34*** (0.06)	0.18* (0.08)	0.18* (0.08)
	Story Problems	0.41*** (0.08)	0.34*** (0.08)	0.10 (0.10)	0.10 (0.10)
	WRAT-Arithmetic	0.63*** (0.06)	0.60*** (0.07)	0.36*** (0.09)	0.36*** (0.09)
	Number Line	0.37*** (0.07)	0.34*** (0.07)	0.19** (0.06)	0.19** (0.06)
	KeyMath-Numeration	0.68*** (0.06)	0.65*** (0.07)	0.49*** (0.09)	0.49*** (0.09)

Note. N = 173 because estimates are calculated using only the control group to model estimates as non-experimental regression coefficients. Standard errors in parentheses. Standard errors were adjusted for clustering at the classroom level * p<0.05 ** p<0.01 *** p<0.001

Supplementary Table 1.4

Estimated Effects Assuming Dependent Pathways ($b_{Non\text{-}experimental}$)

Medium-Term Outcome (Spring of 3rd Grade)	End-of-Treatment Outcome (Spring of 1st Grade)	None	Demographics	Demographics & 1st grade pretests	Demographics, 1s & cross-grade pretests
Facts Correctly Retrieved	Arithmetic Combinations	0.16 (0.09)	0.20 (0.11)	0.20 (0.12)	0.20 (0.12)
	Double-Digit Calculation	-0.11 (0.09)	-0.15 (0.09)	-0.14 (0.10)	-0.14 (0.10)
	Facts Correctly Retrieved	0.19* (0.09)	0.20 (0.10)	0.20 (0.10)	0.20 (0.10)
	Number Sets	-0.05 (0.10)	-0.06 (0.10)	-0.06 (0.11)	-0.06 (0.11)
	Story Problems	0.15 (0.08)	0.13 (0.09)	0.11 (0.10)	0.11 (0.10)
	WRAT-Arithmetic	0.21 (0.12)	0.20 (0.15)	0.10 (0.17)	0.10 (0.17)
	Number Line	-0.01 (0.09)	-0.02 (0.09)	-0.03 (0.09)	-0.03 (0.09)
	KeyMath-Numeration	-0.11 (0.10)	-0.12 (0.11)	-0.20 (0.12)	-0.20 (0.12)
Number Sets	Arithmetic Combinations	0.23* (0.11)	0.17 (0.11)	0.06 (0.11)	0.06 (0.11)
	Double-Digit Calculation	-0.09 (0.07)	-0.13 (0.08)	-0.14 (0.08)	-0.14 (0.08)
	Facts Correctly Retrieved	0.02 (0.09)	0.03 (0.10)	0.04 (0.10)	0.04 (0.10)
	Number Sets	0.36** (0.12)	0.34** (0.12)	0.34** (0.11)	0.34** (0.11)
	Story Problems	0.06 (0.06)	0.03 (0.07)	-0.03 (0.08)	-0.03 (0.08)
	WRAT-Arithmetic	0.07 (0.10)	0.19 (0.13)	0.21 (0.15)	0.21 (0.15)
	Number Line	0.06 (0.08)	0.04 (0.09)	0.07 (0.09)	0.07 (0.09)
	KeyMath-Numeration	0.04 (0.11)	0.11 (0.11)	0.10 (0.13)	0.10 (0.13)
WRAT-Arithmetic	Arithmetic Combinations	0.11 (0.09)	0.11 (0.10)	0.06 (0.10)	0.06 (0.10)

	Double-Digit Calculation	-0.10 (0.07)	-0.12 (0.07)	-0.12 (0.07)	-0.12 (0.07)
	Facts Correctly Retrieved	0.03 (0.07)	0.02 (0.07)	0.04 (0.08)	0.04 (0.08)
	Number Sets	-0.00 (0.08)	0.01 (0.08)	0.04 (0.08)	0.04 (0.08)
	Story Problems	0.03 (0.07)	0.03 (0.08)	0.04 (0.09)	0.04 (0.09)
	WRAT-Arithmetic	0.44*** (0.09)	0.44*** (0.11)	0.37** (0.12)	0.37** (0.12)
	Number Line	-0.03 (0.08)	0.00 (0.08)	-0.02 (0.08)	-0.02 (0.08)
	KeyMath-Numeration	0.25** (0.09)	0.20* (0.09)	0.15 (0.10)	0.15 (0.10)
Number Line	Arithmetic Combinations	0.13 (0.10)	0.15 (0.10)	0.12 (0.12)	0.12 (0.12)
	Double-Digit Calculation	-0.00 (0.08)	-0.04 (0.08)	-0.05 (0.10)	-0.05 (0.10)
	Facts Correctly Retrieved	0.02 (0.10)	-0.07 (0.11)	-0.03 (0.10)	-0.03 (0.10)
	Number Sets	0.04 (0.09)	0.05 (0.09)	0.07 (0.10)	0.07 (0.10)
	Story Problems	0.01 (0.08)	-0.02 (0.07)	-0.02 (0.08)	-0.02 (0.08)
	WRAT-Arithmetic	0.02 (0.11)	0.09 (0.12)	0.05 (0.13)	0.05 (0.13)
	Number Line	0.31* (0.12)	0.27* (0.12)	0.27 (0.14)	0.27 (0.14)
	KeyMath-Numeration	0.18* (0.09)	0.25** (0.09)	0.20* (0.09)	0.20* (0.09)
KeyMath-Numeration	Arithmetic Combinations	0.11 (0.09)	0.16 (0.09)	0.18* (0.08)	0.18* (0.08)
	Double-Digit Calculation	-0.02 (0.07)	-0.03 (0.06)	-0.05 (0.07)	-0.05 (0.07)
	Facts Correctly Retrieved	-0.03 (0.06)	-0.06 (0.06)	-0.04 (0.07)	-0.04 (0.07)
	Number Sets	0.04 (0.06)	0.05 (0.06)	0.08 (0.07)	0.08 (0.07)
	Story Problems	0.06 (0.08)	0.02 (0.09)	-0.01 (0.09)	-0.01 (0.09)
	WRAT-Arithmetic	0.20* (0.08)	0.17 (0.09)	0.13 (0.09)	0.13 (0.09)
	Number Line	0.08 (0.05)	0.08 (0.06)	0.10 (0.06)	0.10 (0.06)
	KeyMath-Numeration	0.43*** (0.08)	0.42*** (0.09)	0.40*** (0.09)	0.40*** (0.09)

Note. N = 173 because estimates are calculated using only the control group to model estimates as non-experimental regression coefficients. Standard errors in parentheses. Standard errors were adjusted for clustering at the classroom level * p<0.05 ** p<0.01 *** p<0.001

Supplementary Table 1.5

Raw Forecasts Using Three Approaches and Resulting Bias

Outcomes		Experimental Benchmark	Independent Single EOT Outcome				Multiple Independent EOT Outcome		Multiple Dependent EOT Outcomes	
		1	2		3		4		5	
		Full Covariates			Full Covariates		Full Covariates		Full Covariates	
Medium-term	End-of-Treatment	Estimate	Forecast	Raw Bias	Forecast	Raw Bias	Forecast	Raw Bias	Forecast	Raw Bias
Non-Speeded Practice										
Facts Correctly Retrieved	Number Line	0.030	0.004	-0.026	0.000	-0.030	0.184	0.154	0.085	0.055
Facts Correctly Retrieved	WRAT-Arithmetic	0.030	0.094	0.064	0.029	-0.001	0.184	0.154	0.085	0.055
Facts Correctly Retrieved	KeyMath-Numeration	0.030	0.012	-0.018	-0.008	-0.038	0.184	0.154	0.085	0.055
Facts Correctly Retrieved	Story Problems	0.030	0.074	0.045	0.027	-0.003	0.184	0.154	0.085	0.055
Facts Correctly Retrieved	Arithmetic Combinations	0.030	0.149	0.119	0.097	0.067	0.184	0.154	0.085	0.055
Facts Correctly Retrieved	Double-Digit Calculations	0.030	0.082	0.052	-0.016	-0.046	0.184	0.154	0.085	0.055
Facts Correctly Retrieved	Facts Correctly Retrieved	0.030	0.063	0.033	0.042	0.013	0.184	0.154	0.085	0.055
Facts Correctly Retrieved	Number Sets	0.030	0.060	0.030	0.012	-0.017	0.184	0.154	0.085	0.055
Number Line	Number Line	0.072	0.014	-0.058	0.010	-0.062	0.334	0.263	0.079	0.007
Number Line	KeyMath-Numeration	0.072	0.027	-0.045	0.019	-0.052	0.334	0.263	0.079	0.007
Number Line	WRAT-Arithmetic	0.072	0.118	0.046	0.070	-0.002	0.334	0.263	0.079	0.007

Number Line	Arithmetic Combinations	0.072	0.189	0.118	0.116	0.044	0.334	0.263	0.079	0.007
Number Line	Story Problems	0.072	0.078	0.007	0.015	-0.056	0.334	0.263	0.079	0.007
Number Line	Facts Correctly Retrieved	0.072	0.054	-0.018	0.016	-0.055	0.334	0.263	0.079	0.007
Number Line	Number Sets	0.072	0.077	0.005	0.032	-0.039	0.334	0.263	0.079	0.007
Number Line	Double-Digit Calculations	0.072	0.144	0.072	0.056	-0.016	0.334	0.263	0.079	0.007
KeyMath-Numeration	WRAT-Arithmetic	0.070	0.213	0.142	0.120	0.050	0.480	0.410	0.144	0.074
KeyMath-Numeration	Number Line	0.070	0.011	-0.059	0.006	-0.065	0.480	0.410	0.144	0.074
KeyMath-Numeration	KeyMath-Numeration	0.070	0.045	-0.025	0.032	-0.038	0.480	0.410	0.144	0.074
KeyMath-Numeration	Double-Digit Calculations	0.070	0.194	0.123	0.074	0.004	0.480	0.410	0.144	0.074
KeyMath-Numeration	Facts Correctly Retrieved	0.070	0.068	-0.002	0.024	-0.047	0.480	0.410	0.144	0.074
KeyMath-Numeration	Number Sets	0.070	0.112	0.042	0.049	-0.021	0.480	0.410	0.144	0.074
KeyMath-Numeration	Story Problems	0.070	0.118	0.048	0.028	-0.043	0.480	0.410	0.144	0.074
KeyMath-Numeration	Arithmetic Combinations	0.070	0.231	0.161	0.147	0.077	0.480	0.410	0.144	0.074
WRAT-Arithmetic	WRAT-Arithmetic	0.093	0.219	0.126	0.138	0.045	0.359	0.266	0.121	0.028
WRAT-Arithmetic	KeyMath-Numeration	0.093	0.039	-0.055	0.017	-0.076	0.359	0.266	0.121	0.028
WRAT-Arithmetic	Number Line	0.093	0.007	-0.086	0.001	-0.092	0.359	0.266	0.121	0.028
WRAT-Arithmetic	Facts Correctly Retrieved	0.093	0.067	-0.026	0.024	-0.069	0.359	0.266	0.121	0.028
WRAT-Arithmetic	Double-Digit Calculations	0.093	0.155	0.062	0.024	-0.069	0.359	0.266	0.121	0.028
WRAT-Arithmetic	Story Problems	0.093	0.097	0.004	0.030	-0.063	0.359	0.266	0.121	0.028
WRAT-Arithmetic	Arithmetic Combinations	0.093	0.206	0.113	0.083	-0.010	0.359	0.266	0.121	0.028

Supplementary Table 5 *(continued)*

Outcomes		Experimental Benchmark	Independent Single EOT Outcome				Multiple Independent EOT Outcome		Multiple Dependent EOT Outcomes	
		1	2		3		4		5	
		Full Covariates			Full Covariates		Full Covariates		Full Covariates	
Medium-term	**End-of-Treatment**	**Estimate**	**Forecast**	**Raw Bias**	**Forecast**	**Raw Bias**	**Forecast**	**Raw Bias**	**Forecast**	**Raw Bias**
WRAT-Arithmetic	Number Sets	0.093	0.100	0.007	0.041	-0.052	0.359	0.266	0.121	0.028
Number Sets	Number Line	0.121	0.008	-0.113	0.003	-0.118	0.460	0.340	0.120	-0.001
Number Sets	KeyMath-Numeration	0.121	0.025	-0.096	0.015	-0.105	0.460	0.340	0.120	-0.001
Number Sets	WRAT-Arithmetic	0.121	0.145	0.024	0.122	0.001	0.460	0.340	0.120	-0.001
Number Sets	Arithmetic Combinations	0.121	0.257	0.136	0.119	-0.002	0.460	0.340	0.120	-0.001
Number Sets	Facts Correctly Retrieved	0.121	0.079	-0.042	0.037	-0.084	0.460	0.340	0.120	-0.001
Number Sets	Double-Digit Calculations	0.121	0.170	0.049	0.028	-0.092	0.460	0.340	0.120	-0.001
Number Sets	Story Problems	0.121	0.114	-0.007	0.028	-0.092	0.460	0.340	0.120	-0.001
Number Sets	Number Sets	0.121	0.155	0.034	0.107	-0.014	0.460	0.340	0.120	-0.001
Speeded Practice										
WRAT-Arithmetic	Number Line	0.023	0.025	0.003	0.003	-0.020	0.478	0.455	0.133	0.110
WRAT-Arithmetic	KeyMath-Numeration	0.023	0.060	0.037	0.027	0.004	0.478	0.455	0.133	0.110
WRAT-Arithmetic	WRAT-Arithmetic	0.023	0.220	0.197	0.138	0.116	0.478	0.455	0.133	0.110

WRAT-Arithmetic	Facts Correctly Retrieved	0.023	0.133	0.111	0.048	0.025	0.478	0.455	0.133	0.110
WRAT-Arithmetic	Double-Digit Calculations	0.023	0.212	0.189	0.033	0.011	0.478	0.455	0.133	0.110
WRAT-Arithmetic	Arithmetic Combinations	0.023	0.391	0.369	0.158	0.135	0.478	0.455	0.133	0.110
WRAT-Arithmetic	Story Problems	0.023	0.073	0.050	0.023	0.000	0.478	0.455	0.133	0.110
WRAT-Arithmetic	Number Sets	0.023	0.118	0.095	0.049	0.026	0.478	0.455	0.133	0.110
Number Sets	KeyMath-Numeration	0.087	0.039	-0.048	0.024	-0.063	0.642	0.555	0.152	0.066
Number Sets	Number Line	0.087	0.029	-0.058	0.009	-0.078	0.642	0.555	0.152	0.066
Number Sets	WRAT-Arithmetic	0.087	0.146	0.059	0.122	0.035	0.642	0.555	0.152	0.066
Number Sets	Double-Digit Calculations	0.087	0.231	0.144	0.039	-0.048	0.642	0.555	0.152	0.066
Number Sets	Story Problems	0.087	0.086	-0.001	0.021	-0.066	0.642	0.555	0.152	0.066
Number Sets	Number Sets	0.087	0.182	0.095	0.126	0.039	0.642	0.555	0.152	0.066
Number Sets	Facts Correctly Retrieved	0.087	0.157	0.071	0.074	-0.013	0.642	0.555	0.152	0.066
Number Sets	Arithmetic Combinations	0.087	0.488	0.401	0.227	0.140	0.642	0.555	0.152	0.066
Facts Correctly Retrieved	Number Line	-0.003	0.012	0.016	-0.001	0.002	0.297	0.301	0.160	0.164
Facts Correctly Retrieved	WRAT-Arithmetic	-0.003	0.094	0.098	0.029	0.032	0.297	0.301	0.160	0.164
Facts Correctly Retrieved	KeyMath-Numeration	-0.003	0.018	0.022	-0.012	-0.009	0.297	0.301	0.160	0.164
Facts Correctly Retrieved	Number Sets	-0.003	0.071	0.074	0.014	0.018	0.297	0.301	0.160	0.164
Facts Correctly Retrieved	Story Problems	-0.003	0.056	0.059	0.021	0.024	0.297	0.301	0.160	0.164
Facts Correctly Retrieved	Arithmetic Combinations	-0.003	0.283	0.286	0.185	0.188	0.297	0.301	0.160	0.164
Facts Correctly Retrieved	Double-Digit Calculations	-0.003	0.112	0.115	-0.022	-0.019	0.297	0.301	0.160	0.164

Facts Correctly Retrieved	Facts Correctly Retrieved	-0.003	0.125	0.129	0.085	0.088	0.297	0.301	0.160	0.164

Supplementary Table 1.5 *(continued)*

Outcomes		**Experimental Benchmark**	**Independent Single EOT Outcome**				**Multiple Independent EOT Outcome**		**Multiple Dependent EOT Outcomes**	
		1	**2**		**3**		**4**		**5**	
		Full Covariates			**Full Covariates**		**Full Covariates**		**Full Covariates**	
Medium-term	**End-of-Treatment**	**Estimate**	**Forecast**	**Raw Bias**	**Forecast**	**Raw Bias**	**Forecast**	**Raw Bias**	**Forecast**	**Raw Bias**
KeyMath-Numeration	WRAT-Arithmetic	0.043	0.213	0.170	0.120	0.078	0.697	0.655	0.235	0.192
KeyMath-Numeration	KeyMath-Numeration	0.043	0.070	0.027	0.050	0.007	0.697	0.655	0.235	0.192
KeyMath-Numeration	Number Line	0.043	0.040	-0.003	0.020	-0.023	0.697	0.655	0.235	0.192
KeyMath-Numeration	Number Sets	0.043	0.132	0.089	0.058	0.015	0.697	0.655	0.235	0.192
KeyMath-Numeration	Double-Digit Calculations	0.043	0.264	0.221	0.101	0.058	0.697	0.655	0.235	0.192
KeyMath-Numeration	Facts Correctly Retrieved	0.043	0.136	0.093	0.047	0.004	0.697	0.655	0.235	0.192
KeyMath-Numeration	Arithmetic Combinations	0.043	0.440	0.397	0.279	0.237	0.697	0.655	0.235	0.192
KeyMath-Numeration	Story Problems	0.043	0.089	0.046	0.021	-0.022	0.697	0.655	0.235	0.192
Number Line	WRAT-Arithmetic	-0.018	0.118	0.136	0.070	0.088	0.511	0.529	0.151	0.169
Number Line	KeyMath-Numeration	-0.018	0.041	0.059	0.030	0.048	0.511	0.529	0.151	0.169
Number Line	Number Line	-0.018	0.048	0.066	0.033	0.051	0.511	0.529	0.151	0.169
Number Line	Number Sets	-0.018	0.090	0.108	0.038	0.056	0.511	0.529	0.151	0.169

Number Line	Double-Digit Calculations	-0.018	0.196	0.214	0.076	0.093	0.511	0.529	0.151	0.169
Number Line	Arithmetic Combinations	-0.018	0.360	0.378	0.220	0.237	0.511	0.529	0.151	0.169
Number Line	Facts Correctly Retrieved	-0.018	0.107	0.125	0.032	0.050	0.511	0.529	0.151	0.169
Number Line	Story Problems	-0.018	0.059	0.077	0.012	0.029	0.511	0.529	0.151	0.169

Note. Full covariates includes Students' age, sex, race, eligibility for free-or-reduced priced lunch, status of learning English as a second language, all the missing indicator variables and the pretest scores for all measures. Raw Bias = (Forecast – Experimental Benchmark). The average experimental benchmark is 0.052 SD.

Supplementary Table 1.6

Forecasting Using a Single Short-Term Outcome

Treatment	Medium-term Outcome	End-of-Treatment Outcome	ATE	Estimated Effect	Experimental Benchmark	Forecast	Mean Bias	Absolute Bias	Accuracy
Speeded	Facts Correctly Retrieved	Double-Digit Calculations	0.810	-0.027	-0.003	-0.022	0.040	-12.233	0.006
Non-Speeded	Facts Correctly Retrieved	Double-Digit Calculations	0.594	-0.027	0.030	-0.016	-0.007	-0.229	0.001
Speeded	Facts Correctly Retrieved	KeyMath-Numeration	0.103	-0.120	-0.003	-0.012	0.040	-12.233	0.006
Non-Speeded	Facts Correctly Retrieved	KeyMath-Numeration	0.066	-0.120	0.030	-0.008	-0.007	-0.229	0.001
Speeded	Facts Correctly Retrieved	Number Line	0.108	-0.011	-0.003	-0.001	0.040	-12.233	0.006
Non-Speeded	Facts Correctly Retrieved	Number Line	0.031	-0.011	0.030	0.000	-0.007	-0.229	0.001
Non-Speeded	WRAT-Arithmetic	Number Line	0.031	0.025	0.093	0.001	-0.048	-0.518	0.004
Non-Speeded	Number Sets	Number Line	0.031	0.086	0.121	0.003	-0.063	-0.524	0.006
Speeded	WRAT-Arithmetic	Number Line	0.108	0.025	0.023	0.003	0.037	1.644	0.004
Non-Speeded	KeyMath-Numeration	Number Line	0.031	0.188	0.070	0.006	-0.010	-0.147	0.002
Speeded	Number Sets	Number Line	0.108	0.086	0.087	0.009	-0.007	-0.076	0.005
Non-Speeded	Number Line	Number Line	0.031	0.307	0.072	0.010	-0.030	-0.416	0.002
Speeded	Number Line	Story Problems	0.216	0.054	-0.018	0.012	0.082	-4.615	0.011
Non-Speeded	Facts Correctly Retrieved	Number Sets	0.277	0.044	0.030	0.012	-0.007	-0.229	0.001

Speeded	Facts Correctly Retrieved	Number Sets	0.326	0.044	-0.003	0.014	0.040	-12.233	0.006
Non-Speeded	Number Sets	KeyMath-Numeration	0.066	0.233	0.121	0.015	-0.063	-0.524	0.006
Non-Speeded	Number Line	Story Problems	0.287	0.054	0.072	0.015	-0.030	-0.416	0.002
Non-Speeded	Number Line	Facts Correctly Retrieved	0.197	0.083	0.072	0.016	-0.030	-0.416	0.002
Non-Speeded	WRAT-Arithmetic	KeyMath-Numeration	0.066	0.260	0.093	0.017	-0.048	-0.518	0.004
Non-Speeded	Number Line	KeyMath-Numeration	0.066	0.293	0.072	0.019	-0.030	-0.416	0.002
Speeded	KeyMath-Numeration	Number Line	0.108	0.188	0.043	0.020	0.044	1.035	0.008
Speeded	Facts Correctly Retrieved	Story Problems	0.216	0.095	-0.003	0.021	0.040	-12.233	0.006
Speeded	KeyMath-Numeration	Story Problems	0.216	0.097	0.043	0.021	0.044	1.035	0.008
Speeded	Number Sets	Story Problems	0.216	0.099	0.087	0.021	-0.007	-0.076	0.005

Supplementary Table 1.6 *(continued)*

Treatment	Medium-term Outcome	End-of-Treatment Outcome	ATE	Estimated Effect	Experimental Benchmark	Forecast	Mean Bias	Absolute Bias	Accuracy
Speeded	WRAT-Arithmetic	Story Problems	0.216	0.105	0.023	0.023	0.037	1.644	0.004
Non-Speeded	KeyMath-Numeration	Facts Correctly Retrieved	0.197	0.120	0.070	0.024	-0.010	-0.147	0.002
Speeded	Number Sets	KeyMath-Numeration	0.103	0.233	0.087	0.024	-0.007	-0.076	0.005
Non-Speeded	WRAT-Arithmetic	Facts Correctly Retrieved	0.197	0.122	0.093	0.024	-0.048	-0.518	0.004
Non-Speeded	WRAT-Arithmetic	Double-Digit Calculations	0.594	0.041	0.093	0.024	-0.048	-0.518	0.004
Speeded	WRAT-Arithmetic	KeyMath-Numeration	0.103	0.260	0.023	0.027	0.037	1.644	0.004
Non-Speeded	Facts Correctly Retrieved	Story Problems	0.287	0.095	0.030	0.027	-0.007	-0.229	0.001

Non-Speeded	KeyMath-Numeration	Story Problems	0.287	0.097	0.070	0.028	-0.010	-0.147	0.002
Non-Speeded	Number Sets	Double-Digit Calculations	0.594	0.048	0.121	0.028	-0.063	-0.524	0.006
Non-Speeded	Number Sets	Story Problems	0.287	0.099	0.121	0.028	-0.063	-0.524	0.006
Non-Speeded	Facts Correctly Retrieved	WRAT-Arithmetic	0.337	0.086	0.030	0.029	-0.007	-0.229	0.001
Speeded	Facts Correctly Retrieved	WRAT-Arithmetic	0.338	0.086	-0.003	0.029	0.040	-12.233	0.006
Speeded	Number Line	KeyMath-Numeration	0.103	0.293	-0.018	0.030	0.082	-4.615	0.011
Non-Speeded	WRAT-Arithmetic	Story Problems	0.287	0.105	0.093	0.030	-0.048	-0.518	0.004
Speeded	Number Line	Facts Correctly Retrieved	0.391	0.083	-0.018	0.032	0.082	-4.615	0.011
Non-Speeded	KeyMath-Numeration	KeyMath-Numeration	0.066	0.489	0.070	0.032	-0.010	-0.147	0.002
Non-Speeded	Number Line	Number Sets	0.277	0.117	0.072	0.032	-0.030	-0.416	0.002
Speeded	Number Line	Number Line	0.108	0.307	-0.018	0.033	0.082	-4.615	0.011
Speeded	WRAT-Arithmetic	Double-Digit Calculations	0.810	0.041	0.023	0.033	0.037	1.644	0.004
Non-Speeded	Number Sets	Facts Correctly Retrieved	0.197	0.190	0.121	0.037	-0.063	-0.524	0.006
Speeded	Number Line	Number Sets	0.326	0.117	-0.018	0.038	0.082	-4.615	0.011
Speeded	Number Sets	Double-Digit Calculations	0.810	0.048	0.087	0.039	-0.007	-0.076	0.005
Non-Speeded	WRAT-Arithmetic	Number Sets	0.277	0.149	0.093	0.041	-0.048	-0.518	0.004
Non-Speeded	Facts Correctly Retrieved	Facts Correctly Retrieved	0.197	0.216	0.030	0.042	-0.007	-0.229	0.001
Speeded	KeyMath-Numeration	Facts Correctly Retrieved	0.391	0.120	0.043	0.047	0.044	1.035	0.008
Speeded	WRAT-Arithmetic	Facts Correctly Retrieved	0.391	0.122	0.023	0.048	0.037	1.644	0.004

Supplementary Table1.6 *(continued)*

Treatment	Medium-term Outcome	End-of-Treatment Outcome	ATE	Estimated Effect	Experimental Benchmark	Forecast	Mean Bias	Absolute Bias	Accuracy
Speeded	WRAT-Arithmetic	Number Sets	0.326	0.149	0.023	0.049	0.037	1.644	0.004
Non-Speeded	KeyMath-Numeration	Number Sets	0.277	0.178	0.070	0.049	-0.010	-0.147	0.002
Speeded	KeyMath-Numeration	KeyMath-Numeration	0.103	0.489	0.043	0.050	0.044	1.035	0.008
Non-Speeded	Number Line	Double-Digit Calculations	0.594	0.093	0.072	0.056	-0.030	-0.416	0.002
Speeded	KeyMath-Numeration	Number Sets	0.326	0.178	0.043	0.058	0.044	1.035	0.008
Non-Speeded	Number Line	WRAT-Arithmetic	0.337	0.208	0.072	0.070	-0.030	-0.416	0.002
Speeded	Number Line	WRAT-Arithmetic	0.338	0.208	-0.018	0.070	0.082	-4.615	0.011
Speeded	Number Sets	Facts Correctly Retrieved	0.391	0.190	0.087	0.074	-0.007	-0.076	0.005
Non-Speeded	KeyMath-Numeration	Double-Digit Calculations	0.594	0.125	0.070	0.074	-0.010	-0.147	0.002
Speeded	Number Line	Double-Digit Calculations	0.810	0.093	-0.018	0.076	0.082	-4.615	0.011
Non-Speeded	WRAT-Arithmetic	Arithmetic Combinations	0.500	0.166	0.093	0.083	-0.048	-0.518	0.004
Speeded	Facts Correctly Retrieved	Facts Correctly Retrieved	0.391	0.216	-0.003	0.084	0.040	-12.233	0.006
Non-Speeded	Facts Correctly Retrieved	Arithmetic Combinations	0.500	0.194	0.030	0.097	-0.007	-0.229	0.001
Speeded	KeyMath-Numeration	Double-Digit Calculations	0.810	0.125	0.043	0.101	0.044	1.035	0.008
Non-Speeded	Number Sets	Number Sets	0.277	0.386	0.121	0.107	-0.063	-0.524	0.006
Non-Speeded	Number Line	Arithmetic Combinations	0.500	0.231	0.072	0.116	-0.030	-0.416	0.002
Non-Speeded	Number Sets	Arithmetic Combinations	0.500	0.239	0.121	0.119	-0.063	-0.524	0.006
Non-Speeded	KeyMath-Numeration	WRAT-Arithmetic	0.337	0.356	0.070	0.120	-0.010	-0.147	0.002

Speeded	KeyMath-Numeration	WRAT-Arithmetic	0.338	0.356	0.043	0.120	0.044	1.035	0.008
Non-Speeded	Number Sets	WRAT-Arithmetic	0.337	0.361	0.121	0.122	-0.063	-0.524	0.006
Speeded	Number Sets	WRAT-Arithmetic	0.338	0.361	0.087	0.122	-0.007	-0.076	0.005
Speeded	Number Sets	Number Sets	0.326	0.386	0.087	0.126	-0.007	-0.076	0.005
Non-Speeded	WRAT-Arithmetic	WRAT-Arithmetic	0.337	0.409	0.093	0.138	-0.048	-0.518	0.004
Speeded	WRAT-Arithmetic	WRAT-Arithmetic	0.338	0.409	0.023	0.138	0.037	1.644	0.004
Non-Speeded	KeyMath-Numeration	Arithmetic Combinations	0.500	0.294	0.070	0.147	-0.010	-0.147	0.002
Speeded	WRAT-Arithmetic	Arithmetic Combinations	0.950	0.166	0.023	0.158	0.037	1.644	0.004

Supplementary Table 1.6 *(continued)*

Treatment	Medium-term Outcome	End-of-Treatment Outcome	ATE	Estimated Effect	Experimental Benchmark	Forecast	Mean Bias	Absolute Bias	Accuracy
Speeded	Facts Correctly Retrieved	Arithmetic Combinations	0.950	0.194	-0.003	0.185	0.040	-12.233	0.006
Speeded	Number Line	Arithmetic Combinations	0.950	0.231	-0.018	0.220	0.082	-4.615	0.011
Speeded	Number Sets	Arithmetic Combinations	0.950	0.239	0.087	0.227	-0.007	-0.076	0.005
Speeded	KeyMath-Numeration	Arithmetic Combinations	0.950	0.294	0.043	0.279	0.044	1.035	0.008

*Note.*Forecasts are sorted from smallest to largest. All estimates were calculated using regressions with full covariates which includes students' age, sex, race, eligibility for free-or-reduced priced lunch, status of learning English as a second language, all the missing indicator variables and the pretest scores for all measures. ATE= Average treatment effect on EOT, Estimated Effect = Regression coefficient of regressing TEMA scores in Grade 1 on EOT, Experimental Benchmark = Observed treatment effect of Pre-K Mathematics intervention on TEMA scores in Grade 1, Raw Bias = (Forecast – Experimental Benchmark). The following are calculated as the mean of each observation for each treatment and Medium-Term Outcome observation: Absolute Bias = (Raw Bias /Experimental Benchmark, Accuracy = (Raw Bias* Raw Bias).

Supplementary Table 1.7

Average Forecasts Using Three Heuristics

	Experimental Benchmark	Forecast Independent Single EOT Outcome				Small Proximal Measure		Large Distal Measure		Average Small Proximal & Large Distal	
	(1)	(2)		(3)		(6)		(7)		(8)	
Medium-term Outcome	Estimate	Average Forecast	Average Bias	Average Forecast	Average Bias	Forecast	Bias	Forecast	Bias	Forecast	Bias
Speeded Treatment											
Facts Correctly Retrieved Number	-0.003	0.123	0.100	0.056	0.04	0.084	0.088	0.029	0.032	0.057	0.060
Sets	0.087	0.123	0.083	0.056	-0.007	0.074	-0.013	0.122	0.035	0.098	0.011
WRAT-Arithmetic	0.023	0.123	0.131	0.056	0.037	0.048	0.025	0.138	0.116	0.093	0.070
Number Line	-0.018	0.123	0.145	0.056	0.082	0.032	0.050	0.070	0.088	0.051	0.069

KeyMath-Numeration	0.043	0.123	0.130	0.056	0.044	0.047	0.004	0.120	0.078	0.084	0.041
Non-speeded Treatment											
Facts Correctly Retrieved	0.03	0.123	0.037	0.056	-0.007	0.042	0.013	0.029	-0.001	0.036	0.006
Number Sets	0.121	0.123	-0.002	0.056	-0.063	0.037	-0.084	0.122	0.001	0.080	-0.041
WRAT-Arithmetic	0.093	0.123	0.018	0.056	-0.048	0.024	-0.069	0.138	0.045	0.081	-0.012
Number Line	0.072	0.123	0.016	0.056	-0.03	0.016	-0.055	0.070	-0.002	0.043	-0.028
KeyMath-Numeration	0.07	0.123	0.054	0.056	-0.01	0.024	-0.047	0.120	0.050	0.072	0.001
Full Covariates	X			X		X		X		X	

Note. EOT = End-of-treatment. 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 end-of-treatment 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 end-of-treatment and medium-term outcomes.

Supplementary Table 1.8

Pre-K Mathematics Descriptive Statistics and Baseline equivalence Split by Treatment Group

	All			Control			Treatment			
	N	***M***	***SD***	***N***	***M***	***SD***	***N***	***M***	***SD***	***p***
Demographics										
	55									
California	8	0.51	0.50	261	0.54	0.50	297	0.49	0.50	0.26
	55									
Kentucky	8	0.49	0.50	261	0.46	0.50	297	0.51	0.50	0.26
	55									
Head Start	8	0.55	0.50	261	0.57	0.50	297	0.53	0.50	0.36
	55									
State Funded Public Pre-K	8	0.45	0.50	261	0.43	0.50	297	0.47	0.50	0.36
	55									
English Assessment	8	0.90	0.31	261	0.89	0.32	297	0.91	0.29	0.43
	55									
Bilingual Assessment	8	0.10	0.31	261	0.11	0.32	297	0.09	0.29	0.43

	55									
Male	8	0.46	0.50	261	0.42	0.49	297	0.51	0.50	0.04
	55									
African American	8	0.15	0.36	261	0.16	0.37	297	0.14	0.35	0.38
	55									
Hispanic	8	0.19	0.39	261	0.19	0.39	297	0.19	0.39	0.99
	55									
Asian American	8	0.03	0.16	261	0.03	0.16	297	0.02	0.15	0.81
	55									
White	8	0.51	0.50	261	0.53	0.50	297	0.50	0.50	0.52
	55									
Mixed Ethnicity or Other	8	0.12	0.32	261	0.09	0.28	297	0.14	0.35	0.04
Pre-K Pretests										
	55									
CMA Age	8	4.45	0.28	261	4.45	0.28	297	4.46	0.28	0.80
	55									
CMA Total Correct	8	1.78	0.97	261	1.84	0.96	297	1.73	0.97	0.19
	55									
CMA Mean Proportion Correct	8	1.77	0.97	261	1.84	0.96	297	1.72	0.98	0.16

	55								
TEMA Age	8	4.45	0.28	261	4.45	0.28	297	4.46	0.28	0.92
TEMA Raw Score	55								
	8	1.42	0.99	261	1.41	0.99	297	1.44	0.99	0.73
WJIII Age	55								
	8	4.52	0.27	261	4.52	0.27	297	4.52	0.28	0.72
WJIII Letter-Word Identification	55								
	8	14.34	0.97	261	14.32	0.98	297	14.36	0.97	0.59
WJIII Understanding Directions	55								
Pictures	8	35.64	1.00	261	35.56	1.00	297	35.71	1.00	0.09
WJIII Spelling	55								
	8	15.39	1.01	261	15.43	1.02	297	15.36	1.01	0.47
Pre-K Posttests										
CMA Age	55								
	8	5.03	0.29	261	4.95	0.28	297	5.10	0.28	0.00
CMA Total Correct	55								
	8	2.95	1.04	261	2.59	0.99	297	3.27	0.98	0.00
CMA Mean Proportion Correct	55								
	8	2.96	1.04	261	2.60	0.99	297	3.27	0.98	0.00

TEMA Age	55									
	8	5.03	0.29	261	4.95	0.28	297	5.10	0.28	0.00
TEMA Raw Score	55									
	8	2.07	1.03	261	1.87	0.99	297	2.24	1.04	0.00
WJIII Age	55									
	8	5.04	0.29	261	4.97	0.27	297	5.11	0.28	0.00
WJIII Letter-Word	55									
	8	15.25	1.00	261	15.20	0.99	297	15.30	1.00	0.24

Supplementary Table 8 *(continued)*

		All			Control			Treatment		
	N	***M***	***SD***	***N***	***M***	***SD***	***N***	***M***	***SD***	***p***
WJIII Understanding Directions	55									
Pictures	8	39.91	0.96	261	39.88	0.95	297	39.93	0.97	0.53
	55									
WJIII Spelling	8	15.80	0.96	261	15.77	1.01	297	15.83	0.92	0.48
First-Grade Follow-up										
	55									
TEMA Age	8	7.00	0.29	261	6.96	0.29	297	7.03	0.29	0.00

	55									
TEMA Raw Score	8	4.56	0.98	261	4.53	1.00	297	4.60	0.95	0.40

Note. Only students with completed pretests, post-tests, and first-grade follow-up tests were included in the analytic sample.

CMA= TEMA = Test of Early Mathematics Ability, WJIII= Woodcock Johnson III

Supplementary Table 1.9

Pre-K Mathematics Correlation Table of End-of-Treatment and Medium-term Outcome Measures

	Variable	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
1	Pre-K Pretests WJIII Letter-Word Identification	–									
2	Pre-K Pretests WJIII Understanding Directions Pictures	0.32***	–								
3	Pre-K Pretests WJIII Spelling	0.49***	0.31***	–							
4	Pre-K Pretests CMA Total Correct	0.49***	0.48***	0.43***	–						
5	Pre-K Pretests TEMA Raw Score	0.60***	0.43***	0.53***	0.69***	–					
6	Pre-K Posttest WJIII Letter-Word Identification	0.68***	0.34***	0.45***	0.49***	0.58***	–				

	Pre-K Posttest WJIII Understanding Directions Pictures	0.26***	0.43***	0.25***	0.41***	0.42***	0.29***	–			
8	Pre-K Posttest WJIII Spelling	0.53***	0.34***	0.60***	0.50***	0.57***	0.66***	0.35***	–		
	Pre-K Posttest CMA Total Correct	0.36***	0.43***	0.38***	0.62***	0.59***	0.51***	0.40***	0.54***	–	
10	Pre-K Posttest TEMA Raw Score	0.57***	0.42***	0.49***	0.65***	0.76***	0.67***	0.39***	0.63***	0.71***	–
11	First-Grade TEMA Raw Score	0.41***	0.34***	0.35***	0.52***	0.56***	0.51***	0.35***	0.49***	0.56***	0.67***

Note. N= 558*.* p < .05, ** p < .01,*

**** p < .001*

Supplementary Table 1.10

Pre-K Mathematics Treatment Impacts on Pre-K Posttests and First-Grade Follow-Up

Outcome	Estimate (SE)
Pre-K Posttests CMA Total Correct	0.67 (.10) ***
Pre-K Posttests TEMA Raw Score	0.40 (.09) ***
First-Grade TEMA Raw Score	0.04 (.07)

Note. $p < .05$, ** $p < .01$, *** $p < .001$.* Estimates were obtained from multivariate regression models controlling for state (CA vs. KY), the type of preschool program children (Head Start vs. State Preschools), child gender, child ethnicity, language of child assessment, pre-K school site and classroom (for pre-K posttest estimates), and first-grade school site and classroom (for first-grade follow-up estimates). The following pretests were use d as controls for each estimate CMA total number correct, CMA mean proportion correct, TEMA total number correct, Woodcock-Johnson III (WJIII) Letter-Word Identification W-score, Woodcock-Johnson III Understanding Directions Pictures W-score, Woodcock-Johnson III Spelling W-score. Standard errors were clustered at the grade 1 classroom level for each wave estimated.

Supplementary Table 1.11

Pre-K Mathematics First-Grade Follow-Up Measures Regressed on Pre-K Posttests for Control Group

		No Covariates	Demographic Covariates	Domain General Pretest & Demographic Covariates	Demographic, & Domain General Pretest Posttest Covariates
Outcome	*Predictor*				
Assuming Independent Pathways					
First-Grade TEMA Raw Score	CMA Total Correct	0.58 *** (0.05)	0.61*** (.05)	0.52*** (.06)	0.35*** (.07)
First-Grade TEMA Raw Score	TEMA Raw Score	0.66 *** (0.05)	0.65*** (.05)	0.64*** (.06)	0.48*** (.09)
Model C: Assuming Non-Independent Pathways					
First-Grade TEMA Raw Score	CMA Total Correct	0.24 *** (0.07)	0.32*** (.07)	0.31*** (.07)	0.26*** (.07)
	TEMA Raw Score	0.49 *** (0.07)	0.44 *** (.07)	0.44*** (.08)	0.40*** (.09)

Note. p* < .05*, ** p* < .01*, *** p* < .001. Standard error is shown in parentheses under regression coefficient. Standard errors were adjusted for clustering at the 1st grade classroom level. Regression estimates are reflect the b-paths shown in the conceptual models in Figure 3.

Supplementary Table 1.12

Pre-K Mathematics Raw Forecasts Using Three Approaches and Resulting Bias

			Model A: Independent Single EOT Outcome				Model B: Multiple Independent EOT Outcome		Model C: Multiple Dependent EOT Outcomes	
			1		2		3		4	
					Full Covariates		Full Covariates		Full Covariates	
EOT		Experimental Benchmark	Forecast	Raw Bias	Forecast	Raw Bias	Forecast	Raw Bias	Forecast	Raw Bias
CMA	Proximal	0.040	0.389	0.35	0.228	0.19	0.420	0.38	0.334	0.29
TEMA	Distal	0.040	0.264	0.22	0.192	0.15	0.420	0.38	0.334	0.29

Note. EOT = End-of-Treatment Outcome, Experimental Benchmark = Observed treatment effect of Pre-K Mathematics intervention on TEMA scores in Grade 1, Raw Bias = (Forecast – Experimental Benchmark). We provide forecasts based on the conceptual models in Figure 3.

Supplementary Table1. 13

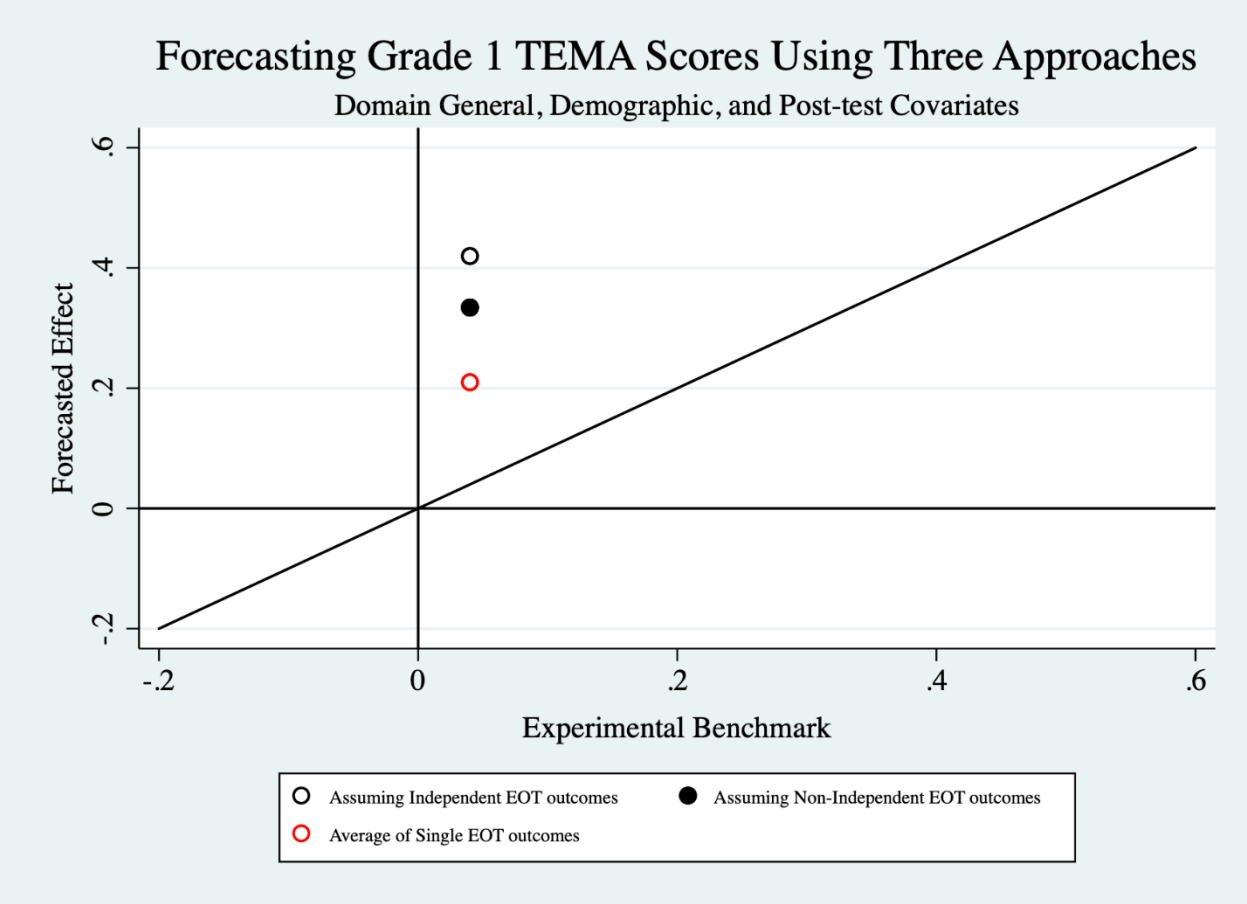
Pre-K Mathematics Bias in Using Single End-of-Treatment Outcome Approach

EOT Outcome	ATE	Estimated Effect	Experimental Benchmark	Forecast	Raw Bias	Absolute Bias	Accuracy
CMA	0.67	0.58	0.04	0.23	0.19	5.69	0.04
TEMA	0.40	0.66	0.04	0.19	0.15	4.80	0.02

Note. EOT = End-of-treatment outcomes, ATE= Average treatment effect on EOT, Estimated Effect = Regression coefficient of regressing TEMA scores in Grade 1 on EOT, Experimental Benchmark = Observed treatment effect of Pre-K Mathematics intervention on TEMA scores in Grade 1, Raw Bias = (Forecast – Experimental Benchmark), Absolute Bias = (Raw Bias /Experimental Benchmark, Accuracy = (Raw Bias* Raw Bias)

Supplementary Figure 1.1

Forecasting Using Three Theoretically Based Heuristics

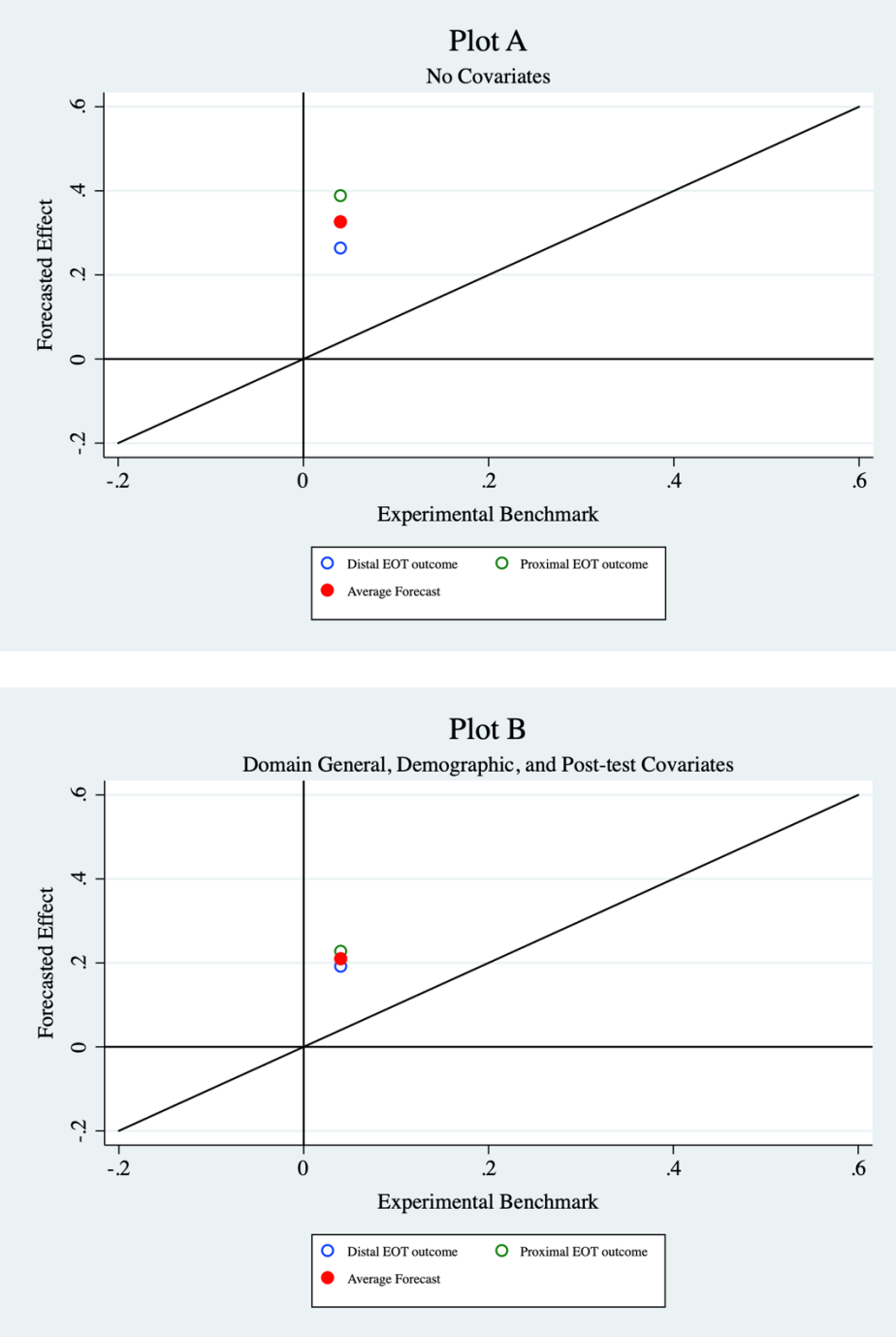


Note. EOT= End-of-Treatment. Each marker on the plot reflects a forecast for each medium-term calculated from one of three heuristics: (1) forecasting all medium-term outcomes using the proximal measure with the smallest treatment impact, (2) forecasting all medium-term outcomes using the distal measure with the largest treatment impact, (3) forecasting using both forecasting all medium-term outcomes both the proximal measure with the smallest treatment impact and the distal measure with the largest treatment impact. 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 using the small proximal and big distal end-of-treatment measures (0.041), shown by the red dot, accurately predicts the average treatment impact on all the medium-term outcomes.

Supplementary Figure 1.2

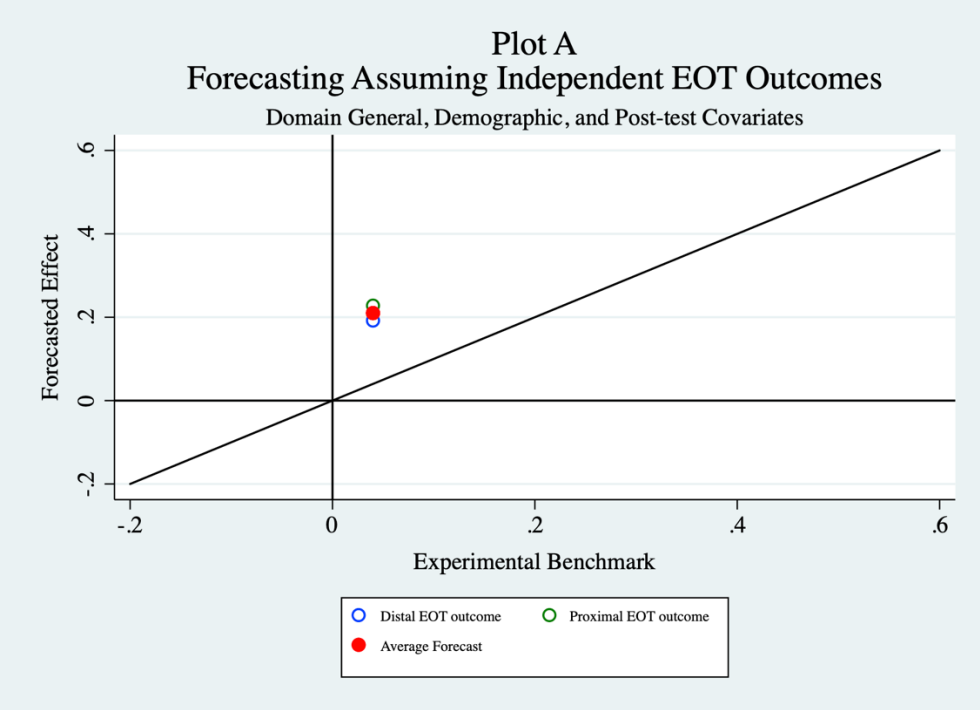
REPLICATING AND ADDRESSING OMITTED VARIABLES BIAS USING PRE-K MATHEMATICS DATA

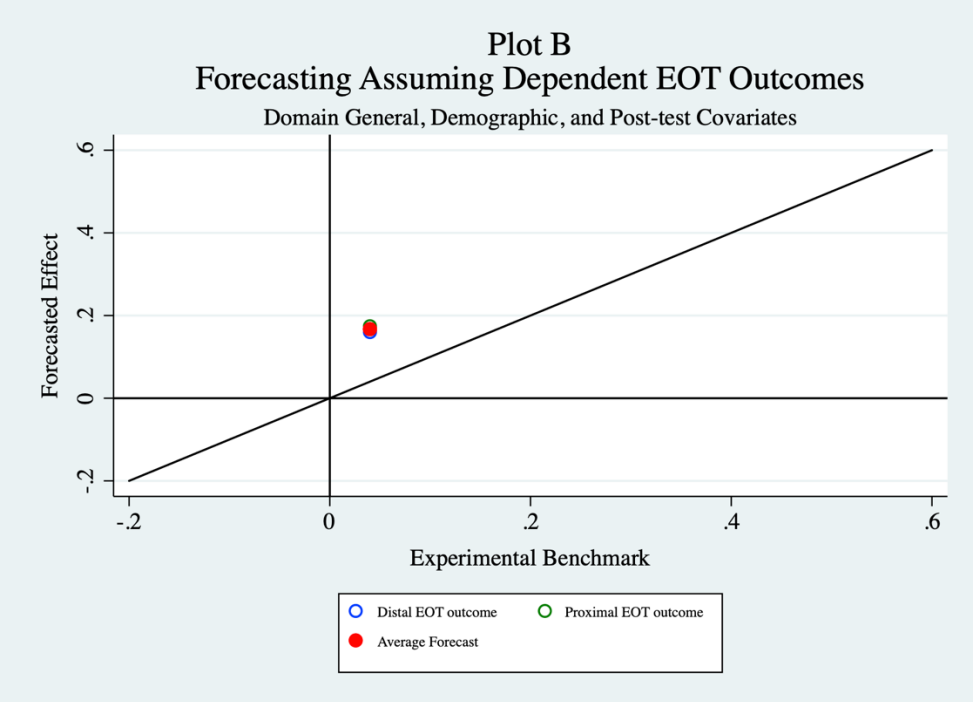


Note. Note. EOT= End-of-Treatment. Each marker on the plots represents a forecast calculated using a single short-term outcome to predict grade 1 TEMA scores. The average forecast is shown in a red circle, this is calculated as the average of all the forecasts in the same plot. Blue markers indicate forecasts calculated with conceptually distal EOT outcomes (TEMA scores at end of Pre-k) and green markers indicate forecasts calculated with conceptually proximal EOT outcomes (CMA scores at end of Pre-k).

Supplementary Figure 1.3

Forecasting with Multiple Short-term Outcomes Using Pre-K Mathematics Data

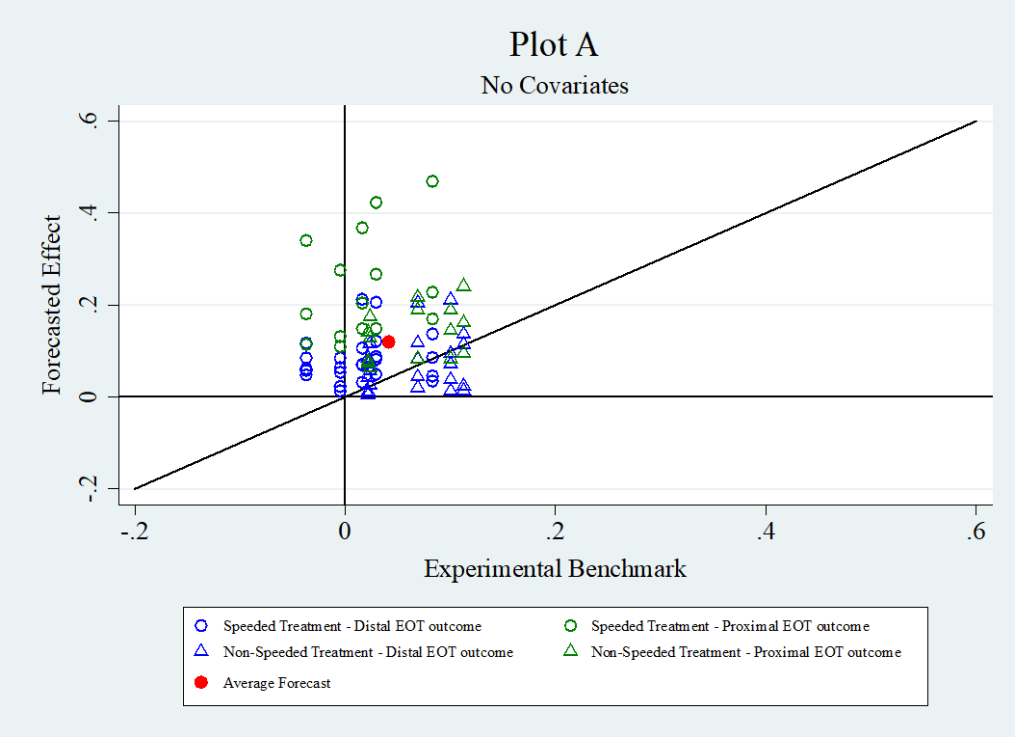


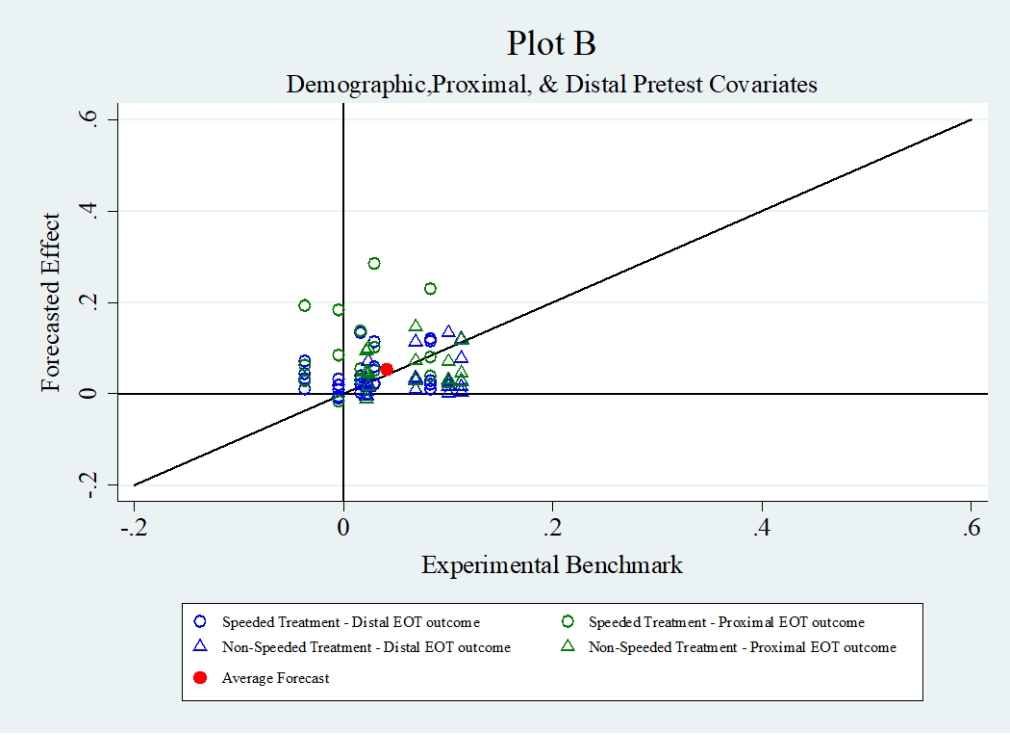


Note. EOT= End-of-Treatment. Each marker on the plot represents a forecast calculated using a combination of both the EOT outcomes to predict grade 1 TEMA scores. The forecasts that were calculated as the average of using both single EOT outcomes to calculate grade 1 TEMA scores I the most accurate and is shown in a hollow red circle.

Supplementary Figure 1.4

Number Knowledge Tutoring Forecasts Robustness Check of Including Full Sample in Forecast Analysis





Appendix A

Number Knowledge Tutoring Participants

Recruitment

The intervention includes four cohorts from four consecutive years. The Vanderbilt Institutional Review Board and the school district permitted the researchers to contact principals from the school district and ask for permission to share the opportunity to participate in the intervention to first-grade teachers. Teachers and parents were consented, and students provided their assent to participate in the study. Out of a total of 4,141 children, 3,051 received a consent form and 2,806 of those that consented were present on the day that they were screened for the intervention.

Screening

Students were screened in September, right at the beginning of the school year, to determine if they were at-risk of having persistent math difficulty using three measures the *First-Grade Test of Computational Fluency* (Fuchs, Hamlett, & Fuchs, 1990), the *First-Grade Test of Mathematics Concepts and Applications* (Fuchs et al., 1990), and the *Word Identification Fluency test* (Fuchs, Fuchs, & Compton, 2004). A latent factor approach was taken to classify student performance on all three tests into three performance strata: high, average, and at-risk (Fuchs et al., 2013). This approach took into account student performance in both reading and mathematics as students who have the most persistent difficulties with mathematics also exhibit difficulties in reading (Geary et al., 2011; Jordan & Montani, 1997; Murphy, Mazzocco, Hanich, & Early, 2007). The 639 students in the at-risk strata were randomly assigned to receive

tutoring with speeded practice, tutoring with non-speeded practice, or continue their regular math instruction without tutoring. The cut-off score used to determine the at-risk students in the first cohort was used in all subsequent cohorts. Students in the high and average strata were considered to be not at-risk for persistent math difficulty and they continued math instruction as usual. For the current study we did not include students not at-risk of math difficulties because our goal is to test methods for best approximating the experimental impact on students randomly assigned to receive (vs. not receive) the treatment, and students not at-risk were not eligible for the intervention.

Pre-K Mathematics

Design

We conducted a secondary analysis of the *Pre-K Mathematics* data. The data were collected as part of a randomized controlled trial assessing the impact of the Pre-K Mathematics intervention on children's mathematics knowledge (Starkey & Klein, 2012). The design included one treatment arm , where children received the *Pre-K Mathematics* curriculum, and one control group arm, where children received their business-as-usual math instruction without the support of an intentional math curriculum.

Participants

The sample includes 744 preschool children from 94 classrooms in 63 schools in a metropolitan area of California and a rural area of Kentucky. We excluded 79 children that did not complete least one end-of-treatment outcome (10.6%) and an additional 107 children that did not complete least one medium-term outcome (14.4%). Little's MCAR test (1988) found these missing data to be completely at random (Little, 1988). The remaining analytical sample

consisted of 558 children that were mostly White / Caucasian (51%), followed by Hispanic/Latino (19%), African American (15%), and children of another race or those without a race indicator in the dataset (12%)[2]. Slightly less than half (46%) of the participants were male, 10% were bilingual (English-Spanish), 51% were from preschools in California, and 55% of the children attended a Head Start preschool program. Descriptive statistics on our analytical sample, shown in Supplementary Table 6, deviate slightly from the original sample as we constrained our sample to exclude students without Pre-K post-test, or G1 follow-up data.

Procedures

The intervention assessed the impact of the published *Pre-K Mathematics* curriculum (Klein, Starkey, & Ramirez, 2002) in treatment classrooms for the purpose of mathematically enriching children's preschool and home learning environments. Since *Pre-K Mathematics* is a supplemental curriculum, treatment teachers added implementation of the math curriculum to their existing classroom curricula (e.g., *Creative Curriculum*).

The *Pre-K Mathematics* curriculum consists of classroom math activities with manipulatives that teachers conduct in small groups (typically 4 children) and home math activities with materials for parents to conduct with their children (in English or Spanish) in parent-child dyads. A teacher's manual provides a curriculum plan, which links the small-group classroom activities to the home activities, and an implementation schedule tailored to each preschool program. The curriculum units, consisting of sets of mathematically related activities,

[2]Children did not have a race indicator if the parents left that question blank on the consent form.

targets the development of number sense, arithmetic operations , spatial sense and geometry, pattern unit construction and duplication, informal measurement and data. These mathematical concepts and skills were selected to prepare children for the clusters in the Common Core State Standards in mathematics in kindergarten.

The implementation schedule called for treatment teachers to present 26 small-group math activities over the course of the intervention year at the rate of approximately one new math activity per week. Since each activity was conducted with small groups of children twice per week for 15-20 minutes per group, the optimal curriculum dosage per child was 52 small-group sessions. In addition, parents of treatment children received 16 *Pre-K Mathematics* home activities over the course of the school year at the rate of a new activity every 1-2 weeks. In contrast, teachers in the control group provided children with business-as-usual math instruction that did not include an intentional early math curriculum.

A trainer-of-trainers model of professional development was used in this study to enable programs to implement *Pre-K Mathematics* under conditions of routine educational practice. Program trainers (curriculum coaches) attended a multi-day trainers' institute where they learned the *Pre-K Mathematics* curriculum as well as how to train and provide on-going support to their teachers as they implemented the curriculum in their classrooms. Then preschool teachers in the treatment group were trained through a series of multi-day *Pre-K Mathematics* workshops conducted by the coaches and project trainers. In addition to the workshops, treatment teachers received on-site support from their coaches to implement the curriculum with fidelity. Coaches made fidelity observations and support visits to treatment teachers' classrooms approximately twice a month over the course of the intervention year.

These visits were not made to control teachers' classrooms, because they were not implementing *Pre-K Mathematics*.

Measures

Children's age, sex, race, bilingual status, and pretest scores for all measures were included as baseline covariates. To examine the impact of the *Pre-K Mathematics* intervention on children's mathematical knowledge at the end of the pre-K year (end-of-treatment), the mathematics measures need to be sufficiently broad to assess the diverse knowledge structures comprising early mathematical cognition. Therefore, two measures of early mathematical knowledge were used in this study, the *Child Math Assessment* (*CMA*; Milburn et al., 2019; Starkey, P. & Klein, A., 2012) and the Test of Early Mathematics Ability (*TEMA-3*; Ginsburg & Baroody, 2003). In addition, several *Woodcock-Johnson Tests of Achievement III* (WJIII ; Woodcock, McGrew,& Mather, 2001) were administered at pretest to measure children's mental abilities and serve as baseline controls for overall cognitive functioning. These subtests included the WJIII Letter-Word Identification, Understanding Directions (Pictures), and Spelling.

The CMA measures preschool children's informal mathematical knowledge across a broad range of concepts and skills that are developing during this period, including number, arithmetic operations, space and geometry, informal measurement and patterns. The *TEMA-3* is a standardized instrument that measures both informal and formal (symbolic) mathematical knowledge, and it can be used with children ages 3 through 8. The TEMA-3 is narrower in scope than the CMA and focuses exclusively on number and operations. Both the *CMA* and the *TEMA-3* were administered in pre-K prior to the treatment (pretest) and at the end-of-treatment (posttest). However, only the *TEMA-3* was administered at the two-year follow-up, because it

assesses math content that is relevant to first grade. Although end-of-treatment impacts have been reported in Starkey, Klein, DeFlorio, and Beliakoff (2020), we report our analysis of the end-of-treatment impacts on first grade (G1) follow-up in Supplementary Table 7.

The mathematics measures in this study were categorized as conceptually proximal or distal based on their alignment with the mathematical content taught in the intervention. Accordingly, the *CMA* was categorized as a conceptually proximal measure, because it assesses a broad range of math skills which align conceptually with the content taught in the intervention. On the other hand, the *TEMA-3* was categorized as a conceptually distal measure, because it does not align as closely and assesses only a subset of the early math skills taught in the intervention.

Appendix B

Design Assumptions Under the Causal Replication Framework

We conceptualize our within-study comparison as a prospective replication, as defined by the Causal Replication Framework (Steiner, Wong, & Anglin, 2018), meaning that we meet all the identification and estimation assumptions required with the exception of one which is under investigation in our current study – the assumption that we have an unbiased estimation of a causal estimand. We draw inspiration from Shadish et al. (2008) in our empirical design to determine which design features and analytical approaches would best approximate the causal estimand if we did not have a strong identification.

Our study design meets the treatment and outcome stability assumptions that there is no hidden variation in treatment conditions, because treatment and control conditions were well-defined and implemented simultaneously (A1.1), there is no variation in outcome measurement and administration time (A1.2), there is no selection across treatment arms that may affect potential outcomes (A1.3), and because students received one-to-one tutoring, this limited the opportunity for peer, spillover, or carryover effects (A1.4). In both studies the researchers are estimating the same causal estimand – ATE(A2.1), data for both studies were collected from participants at the same place and time thus, effect-generating mechanisms were unlikely to vary across sites or time. Initial evaluations found little evidence of treatment impact moderators (Bailey et al., 2020; Fuchs et al, 2013) (A2.2), randomization into study 1 and 2 ensured that baseline equivalence across conditions was met with the exception of a small difference in students eligible for free-or-reduced price lunch across treatment arms (A2.3), and there is identical distribution of setting variables since students in the treatment and control

groups were randomized within classrooms (A2.4). In study 2, we used observational data from the control group to estimate the potential impacts of early math skills on later math skills assuming that our covariates are able to account for confounding bias, this is an assumption violation that we systematically induce to test its effect on the treatment effects heterogeneity between the two studies(A3-not met). We also violate the assumption that we have unbiased estimations of the causal estimand in study 2 as we systematically vary the parametric models to determine which parametric models yield the most accurate forecasts (A4-not met). Lastly, our re-analysis of the original RCT data found no reporting errors of the results, thus we are able to meet assumption (A5).

In this prospective replication design, we are able to test the design features and analytical decisions that do not replicate the observed medium-term outcome impacts and are able to describe three sources of bias that should be addressed in future forecasting studies: omitted variables bias, over-alignment bias, and under-alignment bias.

STUDY 2: WITHIN-STUDY COMPARISONS OF EXPERIMENTAL AND OBSERVATIONAL ESTIMATES OF INCOME IMPACTS ON CHILD HEALTH AND MATERNAL WELL-BEING OUTCOMES

INTRODUCTION

Income gradients are associations in observational data between income and mothers' and children's outcomes (Pamuk, 1985; Wilkinson & Pickett, 2008; Williams, 1999; Deaton, 2002; Adda et al., 2009). Much of the research assessing impacts of income on child and adult health relies on income gradients, but these can be biased by the omission of other sources of disadvantage that have effects on income and mothers' and children's outcomes. Evidence from experimental and quasi-experimental research indicates that income has a causal influence on some aspects of child health and maternal well-being, with larger effects in children and mothers living in poverty (Adler et al., 1994; Case et al., 2002; National Academies of Sciences, Engineering, and Medicine (NASEM), 2019). However, results are not definitive across outcomes. To probe the sources of these mixed results, we examine the correspondence between income gradients on child health and maternal well-being outcomes with experimental impacts of a cash transfer experiment income from the same sample. We test the correspondence of income gradients and experimental cash transfer impacts to determine the potential usefulness of income gradients for making predictions about what we should expect from cash transfer policies. We also consider the implications for using data from cash transfer experiments to test theories of child health and development.

The experimental evaluation of the Baby's First Years (BFY) intervention is the first study in the United States to measure the impact of an unconditional cash gift on child and maternal health and well-being. By using these data, we can test whether income gradients estimated from nonexperimental data correspond with the observed causal impacts of an unconditional cash gift of $3,760 a year. If income gradient estimates approximate the

experimental impacts, we can test the usefulness of potential strategies to reduce bias in nonexperimental estimates of household income on child and maternal outcomes and test the validity of forecasts based in these estimates. If income gradient estimates over-estimate the experimental impacts, one possible explanation is that confounding variables could upwardly bias the income gradient estimates making them consistently larger than the experimental impacts. A second possibility is that disagreement between experimental impacts and income gradients could result from differences between the income constructs (e.g., a cash gift delivered via debit card v. permanent income boost) and the measurement of the construct. For example, measurement error could arise from mother's inaccurate recall of the household annual income which would contribute to measurement error in the experimental and nonexperimental estimates reducing their correspondence. A third possibility is that experimental impact is imprecisely estimated relative to their magnitude which could create a weak correlation between experimental and nonexperimental estimates. This is because the range of income in the BFY RCT ($4,000) is far more constrained than the range of income gradients drawn from a wider income distribution ($100,000 or more).

The multiple potential causes for disagreement between experimental and nonexperimental estimates could serve as challenges to using income experiments and income gradients as mutually informative sources of evidence. In the current study, we find some evidence that construct differences might contribute to the non-alignment between experimental impacts and income gradients, but that experimental estimates are too variable to make strong conclusions about whether experimental impacts and income gradients reflect mostly the same or different processes. We encourage future evaluations

of experimental studies to incorporate income gradient estimates into their power analyses and reporting as useful benchmarks to highlight cases where the observed experimental estimates and the nonexperimental gradient align or differ dramatically."

Why Income Gradients?

Income gradients are estimated associations between a family's annual income and multiple child- and parent-level health, economic, behavioral, and educational outcomes derived from nonexperimental studies. When a family's household income is at or below poverty there is a consistent positive association with a variety of risks from childhood to adulthood such as severely decreased health (National Academies of Sciences, Engineering, and Medicine, 2019). The United States Census Bureau (2017) defines poverty using poverty thresholds for a given family size and family income. A family is determined to be living in poverty if their pre-tax income is below the poverty threshold for a family of that size, based on assumptions about the minimum resources required to meet a family's basic needs, including food, clothing, utilities, and housing costs. For example, the poverty threshold for a family of four with two children in 2019 is $25,750[3]. Therefore, poverty constrains access to nutritional, medicinal, and educational resources (Huston, McLoyd, & Garcia Coll,1994) leaving families vulnerable to a marginalization that maintains a cycle of living in scarcity and adversity (Jäntti, 2009; Rueda & Conejero, 2020) that results in lower educational and economic attainment and lower mental and physical health.

[3] *Source*: https://aspe.hhs.gov/topics/poverty-economic-mobility/poverty-guidelines/prior-hhs-poverty-guidelines-federal-register-references/2019-poverty-guidelines

There are two kinds of processes that may explain how poverty impedes mothers' and children's power to determine their directions in life. One hypothesis is that living in poverty constrains the investments that parents can make to support children's health and well-being, often referred to as the investment pathway (Halpern-Meekin et al., 2015; Becker, 2009). A second hypothesis is that poverty affects the well-being and mental health of parents, relationship strain, and parenting harshness, which adversely affects children, otherwise known as the stress pathway (Evans, 2004; Mistry, Vandewater, Huston, McLoyd, 2002). Considering both pathways, income supplementation policy can serve as a direct solution to poverty by disrupting the cycle of scarcity and adversity by empowering families to supply their needs. Below we review how income gradients have been consistently demonstrated for childhood and adulthood health and well-being across time and various populations.

For children, the income is consistently positively related to health (Currie & Stabile, 2002) and access to health services (Chen, Martin, & Matthews, 2006; Larson & Halfon, 2010). In contrast the income is negatively associated to lower birth weights (Finch, 2003), illness, disability, and mortality (Evans, Wolfe, & Adler, 2012; Milligan and Stabile, 2011). Specifically, children living in poverty consistently demonstrate greater levels of behavioral problems and lower reading and math achievement test scores in comparison to children from middle-income families (Mayer, 1997; Duncan, Magnuson, Kalil, Ziol-Guest, 2011). Moreover, neuroscientific research has reported some differences in the neural networks between children who grow up in poverty, such as the amount of grey matter in their brains (Hanson et al., 2015), posing concerns about poverty creating a vulnerability in children's learning and behavior that might persist into adulthood (Hackman & Farrah,

2009; Noble et al., 2015). Although there is disagreement on whether the income gradient for child health steepens as children get older (West, 1997; Currie et al., 2004; Burgess et al., 2004), the severity of income impacts on child health is the worst for children living in conditions of persistent poverty (Curtis et al., 2001; Milligan and Stabile, 2011). Considering the trends in the literature, we would predict that an $1,000 increase to a family's annual income to result in a .02 SD increase in child overall health, a .03 SD increase in not being diagnosed with a disability, a .11 to .16 SD decrease in conduct disorders or indirect aggression scores, a .014 increase in never experiencing hunger due to lack of food, and a -.001 decrease in being injured in the last 12 months (Milligan & Stabile, 2011).Though most research on income gradients focuses on the general adult population, these impacts also reflect the conditions in which mothers living in poverty may be exposed to such as increased exposure to violence, crime, trauma, and lower physical (Cutler & Mooney, 2011) and mental health (Ridley et al., 2020). There are consistent negative associations between living in poverty and issues with mental health such as increased rates of depression, anxiety, and suicide (Milligan & Stabille, 2011; Ridley, Rao, Schilbach, & Patel, 2020; Sareen et al., 2011). Cash transfer experiments have shown that increases in income can improve mental health through the reduction of stress or worry, especially for pregnant women who gained more control over household expenditures (Bharanov et al., 2017). Bastagli et al (2021) supports this evidence with a review of 201 cash transfer experiments conducted across the world which also demonstrate significant impacts on nutrition and health with increased access to health services. Based on the literature, we would predict that an $1,000 increase to a family's annual income would result in a -.10 SD to -.20 SD decrease in maternal depression and a -

.001 SD decrease in maternal health, based on previous estimates from changes to a child benefit program (Milligan & Stabile, 2011). These estimates are quite different from each other, perhaps indicating that effects of child benefits, cash transfers, and the cash gifts vary dramatically across outcomes or reflecting large variability in experimental estimates.

Potential Confounders

The most prevalent approach to modeling the effect of income on child and maternal outcomes is to statistically adjust for numerous potential confounding factors. Confounding factors are variables that can plausibly cause both income and the outcomes of interest, these might explain the consistent associations between household income and child and adulthood. For example, a variable like maternal mental health can impact parent income, child health, and maternal health. However, modeling a factor like maternal mental health with nonexperimental data presents two statistical challenges (1) the difficulty in accurate measurement and (2) difficulty in modeling the association between maternal mental health, the mother's ability to work, and care for their child, especially as these relations can be bidirectional and variable across time. If child and maternal outcomes are not directly influenced by changes to mothers' incomes in adulthood (e.g., because they are influenced by mothers' circumstances before they reach adulthood), they likely confound the relation between current income and child and maternal outcomes.

Experimental and Quasi-Experimental Literatures on Income Impacts

The extent of bias from confounding variables in nonexperimental estimates can be identified by contrasting them with experimental estimates to triangulate plausible statistical bounds of the estimated impacts of income on various outcomes, while accounting for different sources of bias (Hernán, 2018). However, the experimental and

nonexperimental literatures on income gradients demonstrate mixed income effects that may result from multiple sources of bias (Mayer, 1997; Ross & Mirowsky, 1999). Nonexperimental income gradient literatures consistently report that higher income raises children's academic test scores, increases completed schooling, and reduces the prevalence health and behavioral problems (Blau,1995; Mayer, 1997). These gradients are supported by quasi-experiments like the casino disbursements to Native American youth (Akee et al. 2010; Wolf et al., 2012) which found reductions in adolescent depression disorders, anxiety disorders, conduct or oppositional disorders, criminal behavior, and drug use. Similarly, income increases in Earned Income Tax Credits have also demonstrated positive impacts on maternal mental and physical health, showing reductions in the amount of risky levels of biomarkers like blood pressure and cholesterol levels (Evans & Garthwaite, 2014). In Canada, a $500 income increases reduced children's levels of aggression and mother's reports of depression (Milligan & Stabile, 2011). In contrast to health, the association between household income and academic test scores is only found for individuals whose households received the income boost during early childhood (Duncan et al., 2010; Ziol-Guest, et al., 2012). Therefore, impacts are further complicated by the outcomes measured and the point in children's development at which income increases occur.

Experimental evidence from the Negative Income Tax (NIT) experiments partially supports some of the trends in income impacts. The NIT experiments found positive impacts on the family income, child health, and academic achievement in elementary school for families across four sites North Carolina, Iowa, and Gary, Indiana. However, these impacts did not hold for high school students in North Carolina or for students in Iowa (Maynard and Murnane, 1979). A commonality across all these types of studies is the

positive impacts on children's early health and academic development and the mixed impacts on their development in adolescence (Duncan, Magnuson, Votruba-Drzal, 2017) and impacts on measures of other behavioral outcomes that are not educational tests. Based on these trends, Duncan, Morris, and Rodrigues (2011) used data from a set of anti-poverty experiments to estimate that a $3,500 boost to a low-income family's annual income would be projected to yield a .20 SD gain in their children's early educational achievement test scores.

Although there is some consistency between income gradient estimates and experimental impacts on early childhood health and maternal health, there are also notable Limitations in the experimentally derived estimates are due to the external validity of experiments. For example, some experimental cash-transfers have focused on specific minoritized racial / ethnic groups (Maynard & Murnane,1979) and living within a service-rich economy (Akee et al., 2010).This complicates the relevance of experimental comparisons to nonexperimental estimates from nationally representative samples, especially to those who may live within more rural or less service-rich economies. The validity tradeoff between nonexperimental income gradients literature (more generalizable, but lower internal validity) and experimental impacts (less generalizable, but high internal validity) provides the opportunity to understand the extent to which impact magnitudes are similar or different in predictable ways and to reconcile mixed evidence across these literatures.

Within-Study Comparisons

Within-study comparisons is the method through which we can estimate the correspondence between experimental and nonexperimental estimates to shed light on the

usefulness of nonexperimental methods for approximating a causal benchmark and identify optimal to use these different methods (Cook, Shadish, & Wong, 2008; LaLonde, 1986; Michalopoulos, Bloom, & Hill, 2004; Steiner & Wong, 2018). Cook, Shadish, and Wong (2008) explained that conduct a within-study comparison with high-technical merit one must meet certain criteria, such as compare two counterfactual groups and estimate the same causal estimand. In addition, group assignment should not be a confounding variable that is related to the outcomes, to ensure this we use data from an RCT that randomly assigned participants within the same geographical locations to comparison groups and ensured that all participant outcomes were measured in the same way. Moreover, the BFY RCT meets all the criteria for a well-implemented experiment with no treatment cross-over or differential attrition. We present estimate comparisons as method differences in statistical significance patterns, causal estimates in original or standardized metrics, and as percent differences when estimates are not close to zero. Although the authors were not blind to the experimental impacts the ITT estimates were analyzed independently and reported by our colleagues (Gennetian et al., 2022; Magnuson et al., under review; Sperber et al., in preparation) prior to when we conducted our nonexperimental analysis. Lastly, for estimates to align, we would have to assume that an unconditional cash gift reliably delivered through a debit card (experimental estimate) is comparable to a permanent cash boost (nonexperimental estimate).

We draw on examples from LaLonde (1986) who conducted a within-study comparison to determine whether regression adjusted estimates approximated experimental impacts in the national supported work program on adult earnings. LaLonde found that econometric models did not replicate the causal benchmark. Similarly,

Weidmann and Miratrix (2021) conducted a within study comparison of 14 studies to understand the distribution of selection bias of schools into various types of educational interventions (i.e., professional development, curricular changes). They find that regression adjusted models approximated the experimental benchmarks reasonably well, within .11 standard deviations. This finding implies that nonexperimental studies of educational interventions like those studied by Weidmann and Miratrix may be useful for choosing among promising interventions and for informing theories of the development of education-related outcomes. Lastly, Harding et al (2021) used a within-study design to determine which estimation approaches and sample restrictions replicated the experimental impacts of reducing neighborhood poverty on adult economic outcomes using a randomized experiment of housing vouchers. Non-experimental estimates were consistently larger than experimental impacts, and none of the factors tested (i.e., measures, treatment size, heterogeneity, treatment duration, mobility, etc.) could account for these discrepancies. Indicating that omitted variables bias was driving the observed associations between neighborhood effects on adult outcomes, whereby some unmeasured factors about the individuals that select into and out of certain neighborhoods also carry these advantages or disadvantages into adulthood. These findings caution that neighborhood gradients should be carefully used to inform research on the causal effects of neighborhoods. We anticipate similar cautionary tales will be true for income gradients.

Knowing the extent to which income gradients reflect causal effects would make the non-experimental literature more useful and policy relevant. If we find that nonexperimental approaches approximate experimental impacts, we can use the specification of the nonexperimental approach to further test effects of income on different

outcomes. This exercise can inform forecasts of longer-run experimental impacts or inform of an integrated theory of how poverty affects child and maternal health. On the other hand, if we find that non-experimental estimates reliably overestimate experimental impacts, we can inform the field that using income gradients for making forecasts about the effects of hypothetical policy changes on important child and maternal well-being outcomes should be interpreted with appropriate caution. Alternatively, if we find that non-experimental estimates under-estimate the experimental impacts this may reveal that non-experimental approaches may be insufficient to inform theory or can inform lower bound estimates of the causal links between income and child and maternal outcomes.

Potential Sources of Incongruence Between Experimental and Correlational Estimates

Nonexperimental estimation of income gradients may yield biased results if the assumptions for causal identification are not met in the field (Wong, Steiner, and Anglin, 2018). To best approximate experimental estimated using nonexperimental data Heckman and colleagues (Heckman & Hotz, 1989; Heckman, Ichimura, Smith, & Todd, 1998; Heckman, Ichimura, & Todd, 1997) recommend the use of rich covariates, drawing comparisons from the same local markets, and measuring dependent variables in the same way. Although these methods can help to reduce bias, they have not been able to fully eliminate it. We use the approximation strategies recommended by Wong et al., (2018) and Heckman and colleagues and hypothesize that additional potential sources of incongruence might include: (1) confounding bias in the identification of the causal estimand from nonexperimental design, (2) construct impurity biasing the comparison across estimation methods, and (3) the imprecision of experimental estimates biasing the correspondence

with the income gradient estimates, making the comparison less informative. We contrast the expected estimate comparisons across these three potential sources of incongruence in Table 1 and explain our expectations in greater detail below.

Confounding

First, estimates may be biased by confounding variables that are not adjusted for or measured at baseline. In general, we would expect the income gradient to over-estimate causal effects of an income increase on developmental outcomes, because the most plausible omitted variables (e.g., parent education) are thought to influence outcomes in the same direction as income effects. To explore how omitted variables may upwardly bias the nonexperimental income gradient estimates, we compare the differences between experimental and nonexperimental estimates of income effects when we calculate the nonexperimental estimates using a multivariate regression with different covariate specifications. We hypothesize that if income gradients are substantially reduced with the inclusion of additional covariates, then there is reason to believe that part of the incongruence between experimental and non-experimental estimates is due to omitted variables that we have not adjusted for. Additionally, if we find that the income gradient estimates are consistently larger than the experimental impacts, this would be indicative of omitted variables bias.

Construct Impurity

Second, estimates may be biased by qualitative differences between measured income and cash transfers, such that an income gradient reflects a stable income whereas an experimental cash-gift is a short-term income increase. If these two types of income have different social meanings (Bandelj, Wherry, Zelizer, 2017), they may be utilized

differently by mothers, which may be reflected in predictable discrepancies between income gradients and experimental impacts even in the absence of confounders (Sykes, Križ, Edin, Halpern-Meekin, 2015). For example, a family with a new baby living in poverty may spend a cash gift of $1,000 to pay hospital fees or debt, to purchase an infant car seat, or to place a down payment on a car; this may differ from their expenses in the future assuming a reliable additional $1,000 per month income. If this is true, then income gradients may be less useful for forecasting the impacts of policies incorporating steady income increases on longer-run effects. In these cases, both the income gradients and experimental impacts would be policy-relevant, but for different kinds of policies. For example, an unconditional cash gift that is branded as "4MyBaby", as in BFY, is meant to encourage expenditures that benefit the baby, as opposed to an earned income tax credit, which is meant to encourage adults to work. Indeed, parents engage in different behaviors that may not be economically optimal to ensure benefits for their children, such as saving for college but not paying off credit card debt and incurring more interest, even if such acts may cost them more in the long run (Soman & Ahn, 2011).

If this problem causes experimental and observational estimates to diverge, they may diverge in predictable ways. A reasonable prediction is that there will be greater incongruence across estimates of income impacts on one-time purchases (such as a crib, car seat, and infant medical care), compared to frequently occurring purchases (such as diapers, clothing, and baby formula). This hypothesis assumes that children's goods that require a one-time purchase will be more responsive to a monthly cash gift than an annual income boost of $1000 based on parent budgeting planning. This hypothesis predicts that the experimental estimate is greater than the non-experimental estimate for one-time

purchases than for recurring purchases. If this difference is large, this may suggest that this experimental evaluation is not as externally valid as a stable income increase across a longer-period of time may induce different forms of investments.

Imprecise Estimation of Experimental Impacts

Third, income experiments may yield relatively imprecise estimates if they increase income to a small fraction of the degree in which it varies in the population, making income experiments less useful for studying the potential impacts of income transfers. In this case, income gradients may still be useful for generating estimates of plausible impacts to conduct power analyses for other income experiments. However, such a finding would indicate that income experiments are less helpful for distinguishing among the likely relative income impacts across various outcomes (e.g., will impacts be larger on maternal depression or on child health?). Which would be useful for building a stronger theory of the effects of income on child and maternal health. In the current study, we find that the range of standardized estimated standard errors for the experimental estimates are several times greater than the range of nonexperimental estimates (see Table A4). The median standard error for experimental impacts standardized to the control group standard deviation is 0.0187 per $1,000 increase, while the standard error for the nonexperimental estimates is 0.0043 per $1,000 increase, less than one-fourth the size of the experimental standardized standard error amount.

Current Study

In the current study we ask: (1) Do income gradients approximate experimental impacts in magnitude and direction? (2) Are income gradients calibrated across outcomes, such that larger gradients predict larger impacts? The original experimental impacts from

the Baby's First Years RCT are reported in Gennetian et al., 2022 (family investment outcomes), Magnuson et al., 2022 (family stress outcomes and processes), Yoo et al., 2021 (maternal substance use), Sperber et al., forthcoming (child health and sleep), and Escueta, Gennetian, et al., forthcoming (physical abuse). The experimental estimates reported in this study are not the original sources for the BFY study wave 1 impacts and should be interpreted secondarily to the estimates in the studies.

METHOD

Data

The data were drawn from the Baby's First Years (BFY) study, a multi-site randomized controlled trial in which 1,000 mothers with newborn infants were randomly assigned to receive a monthly cash gift of either $20 (control group, n=600) or $333 (treatment group, n=400) each month for the first 40 months of the child's life. Thus, the compensation difference between the low-cash and high-cash gift groups is a $3,760 higher annual income. The purpose of the RCT was to estimate how such an income increase could impact the lives of mothers and their newborns living in poverty. The BFY study has been preregistered Duncan et al (2022), details about the study design can be found in Noble et al. (2021), and public use data and documentation files were developed by Magnuson et al. (2020). In this study we use the publicly available data from the first wave of data collected during the first year of the child's life during which mothers received the cash gift each month for 12 months. The intent-to-treat impacts have been reported across several different studies, baseline balance across multiple maternal demographic and income measures between the high-cash and low-cash gift groups was originally reported in Noble, Magnuson, Gennetian, et al. (2021), however we and show these estimates in Appendix

Table A1 for variables in the present study. The experimental impacts are shown in Table 3 column 1, these were separately estimated in the current study, but the original impacts are reported in Gennetian et al. (2021), Magnuson et al. (under review), and Sperber et al., (in preparation).

Participants

From May 2018 through June 2019, BFY recruiters visited 12 hospitals in 4 metropolitan areas (New York City, the greater New Orleans metropolitan area, the greater Omaha metropolitan area, the Twin Cities) to recruit mothers who had just given birth and reported being under the poverty threshold. The mothers had to be at least 18 years or older, speak English or Spanish, and have a newborn not requiring intensive care to participate. If mothers consented to a baseline questionnaire, they were offered the opportunity to receive an unconditional cash gift that would be disbursed via a debit card each month on the day of the child's birth. These debit cards were branded with a "4MyBaby" logo and activated in the hospital. Of the 1,051 mothers who qualified to participate in a baseline survey, 1,003 agreed to receive the cash gift. After three mothers subsequently withdrew, 1,000 mothers comprised the final sample. Further descriptions of participant enrollment and study design are available in Noble et al. (2021).

In the following year, from July 2019 to July 2020, interviewers successfully contacted 931 of the 1,000 mothers around the time of their child's first birthday. On March 14, 2020, the COVID-19 pandemic halted ongoing in-person interviews (n=605), which were quickly adapted to be administered over the phone (n=326). At this time, two-thirds of infants had reached their first birthday. Overall, 93% of the participants completed the measures we analyze in the present study. Further information on the interview and field

contact procedures are available in the public data deposit (https://www.icpsr.umich.edu/web/DSDR/studies/37871) and on the study's website (www.babysfirstyears.com).

Statistical Analysis

To determine the extent to which nonexperimental income gradients approximate the causal impacts of income on child development and maternal well-being outcomes, we analyze the baseline and age 1 data from the BFY study. This large-scale RCT provides the optimal experimental research design to compare the experimental impact of income on child and maternal outcomes to the correlational evidence that would have otherwise only been available through a cross-sectional or longitudinal study design. In the current study we use a within-study comparison (Shadish, Cook, & Wong, 2008; Lalonde, 1986) to determine how well we can approximate the observed treatment impact of income on child development and maternal well-being outcomes.

As shown in Figure 1, we first use the data from participants in the treatment and control groups to estimate the intent-to-treat (ITT) estimand (β_{ex} in equation 1) using ordinary least squares (OLS) multiple regressions with baseline covariates for each outcome.:

(1) $Outcome_i = \beta_0 + \beta_{ex}Trt_i + \beta_2 Cov_i + e_i$ Then, we isolate data from the participants in the control group to estimate nonexperimental income gradients (shown as equation 2 β_{ne}) by regressing each outcome on income while controlling for the same baseline covariates included in the ITT model.

$$(2)\ Outcome_i = \beta_0 + \beta_{ne}Income_i + \beta_2 Cov_i + e_i$$

We follow the same estimation strategy employed by the original BFY randomized controlled trial to estimate the experimental impact of a $3,760 increase in annual income. Following the preregistration plan, the experimental estimate is obtained by regressing each outcome measure on the treatment indicator and the four site indicators, since randomization was within site. All experimental estimates are adjusted for the following baseline demographic child and family characteristics: mother's age, mother's years of completed schooling, household income, net worth, general health, depression, race and ethnicity, marital status, number of adults in the household, number of other children born to the mother, number of cigarettes smoked per week during pregnancy, number of alcoholic drinks consumed during pregnancy, father living with the mother; child's sex, birth weight, gestational age at birth, age of child in months during data collection, and whether the interview was conducted over the phone or in person at Age 1 (because of the COVID-19 Pandemic). The use of these baseline controls increases the precision of the experimental estimates, we also adjust the standard errors using robust variance estimation techniques.

For the non-experimental estimates we use only the data from the participants in the control group to conduct multiple linear regressions on each outcome while controlling for the same baseline covariates, excluding the income quintile covariates. A survey of demographic characteristics was collected at baseline prior to participant randomization to determine group equivalence. The low-cash and high-cash groups meet baseline equivalence on all the preregistered control variables as shown by the data documentation released on the Inter-university Consortium for Political and Social Research by Magnuson and colleagues (2018-2021) we provide this table in Supplementary Table 1.

By using the data that is exclusively from the control group in the BFY RCT, we model the estimated effect of income on each outcome that would have been estimated if we only had longitudinal or cross-sectional data available as in an nonexperimental study. We scale income in two separate ways (1) as thousands of dollars and (2) the log of income in thousands of dollars to estimate the income gradients in a bivariate and multiple linear regression models as shown in Table 3. The outcome measures we used are described in further detail below.

The raw experimental estimates reflect the impact of an additional $3,760 per year, and the raw linear income gradient estimates reflect changes in outcomes for each $1,000 increase in annual income. To make the experimental and nonexperimental gradient estimates comparable, we rescaled the experimental coefficients (shown Table 2 column 1) to reflect a $1,000 income increase per year by multiplying them by 0.27 (because $1,000/ $3,760 = 0.27). Log estimates were also transformed (columns 6 & 8) to reflect the average impact of an additional $1,000 per year. We rescaled the log estimate to reflect a model-implied increase in $Outcome_i$ associated with a $1000 in average annual income. To rescale we subtracted the estimated impact of a log income increase on the average control group income plus $1000 and subtracted that by the log of the average control group income: β_1*log(23.41) - β_1*log(22.41), which equals β_1*(log(23.41)-log(22.41)), which equals a more simplified equation β_1*.0437. Therefore, we transformed all the log nonexperimental estimates and standard errors by multiplying them by .0437. All the columns that are shaded in grey (columns 2, 3, 4, 6, and 8) reflect comparable estimates from various model specifications.

Measures

The BFY measures and analysis plan were preregistered at ClinicalTrials.gov under Identifier: NCT03593356, here we review the measures collected when the newborn child reached age 1. All the original intent-to-treat (ITT) impacts have been published, or will soon be published, by the BFY PIs and colleagues. In the current study we focus on the correspondence between observed experimental impacts and estimated nonexperimental impacts, therefore we reserve interpretations of the original ITT estimand to the publications cited alongside each of these measures.

We selected 12 of the pre-registered child and maternal health and well-being outcomes that yielded either statistically significant experimental impacts or non-experimental income gradients to maximize variation in experimental and non-experimental estimates. Among the outcome variables selected were measures of family investments including the: Purchases for Child Since Birth Index an additive index of purchased items for the child from birth to age 1,
child expenditures as a continuous dollar amount for expenditures made for the child in in the past 30 days, the frequency of parent-child activities, mothers' perceptions of neighborhood safety, an index of maternal economic stress, an index of the number of social services a mother receives, an additive index of mothers' reported anxiety, and whether the mother and child have ever experienced homelessness. All the ITT impacts of an unconditional cash transfer on these family investment measures have been reported by Gennetian and colleagues (2022). In addition, we include outcome measures of family stress including an additive index of maternal hope and agency and the occurrence of domestic violence. The original impacts of the unconditional cash transfer on the family

stress model are reported in Magnuson et al (*under review*). We also include a measure of the child's overall health and child's sleep disturbances, for which the original impacts are reported in Sperber et al., (in preparation). Negative outcomes were reverse coded such that higher estimates reflected desirable outcomes in the directions hypothesized by the BFY primary investigators. For example, measures such as child health and sleep disturbances were measured on scales that indicated worse outcomes for higher numbers. Therefore, the following measures were reverse coded: economic stress, social services use, maternal anxiety, child's overall poor health, sleep disturbances, whether mother and child have experienced homelessness, and whether mother has experienced physical abuse.

Analytic Approach

The correspondence of experimental income impacts and income gradients may differ based on how family income is measured. For example, child items purchased may be more strongly correlated with a measure of parental income that is not including additional income from other members of the household who don't contribute to child expenses. To test the robustness of our results to different operationalizations of income, we compare the regression coefficients from regressions including different measures of income. As shown in Table 3, we compare the mother's income, the father's income, the combined parental income, and the average overall household income including all household members across treatment groups. The average income for the household, the mother, and their spouse did not vary significantly across treatment groups. We show the estimated impacts of an income increase on Table 3, where the estimates shown in grey columns are comparable. The first column contains the ITT estimate of the BFY study, the second column contains the same estimate that is re-scaled to reflect the impact of a $1,000

increase in annual income rather than the full $3,760 to facilitate experimental comparisons with the nonexperimental estimates. Non-experimental estimates of the effect of a $1,000 annual income increase are shown in columns 3 through 6. In column 3, estimates were modeled in a bivariate regression with no covariates, and in column 4 estimates were modeled in multiple linear regressions with baseline covariates. Then, in columns 5 to 6 we show the estimated impacts of the log function of income. Because the log function demonstrates the expected impact of about $60,000, we rescaled this estimate to reflect the expected impact of a $1,000 increase per year. The rescaled log income estimates are modeled without covariates in column 5 and with covariates in column 6.

RESULTS

We used an OLS bivariate regression to estimate significant differences in income at baseline and age 1 between treatment groups in Table 2. All these estimates were originally reported in the BFY Baseline User Guide (Magnuson et al., 2020), the BFY experiment design paper (Noble et al., 2021), and ITT impacts publications (Gennetian et al., 2022; Magnuson et al., 2022; Sperber et al., *in preparation*). To examine how the estimated impacts of income on various child and maternal health outcomes correspond, we present all these estimates in Table 2 and make comparisons about the direction, magnitude, and statistical significance patterns in the estimates. Table 2 summarizes comparable rescaled experimental estimates (column 2), estimates from linear gradients (columns 3 & 4), and rescaled log gradient estimates (columns 5 & 6). Overall, there was weak correspondence between the experimental impacts and the linear and log income gradients on all 12 preregistered outcomes.

We illustrate the differences in these model estimates in Figure 2 where we focus on the household income gradient[4] predicted estimates for the impact of a $1,000 increase in annual income on child-focused expenditures. The first plot shows the raw trend of child-focused expenditures calculated using the *lowess()* plotting function from the *stats* library available in R (Cleveland, 1981) which computes and connects smooth lowess lines to approximate a non-linear income gradient on child-focused expenditures using a locally weighted regression. The Lowess lines shown on the left-hand plot of Figure 2 in orange indicate that there is no trending income gradient on child focused expenditures in the raw data[5]; the right-hand plot closes into the area between the two vertical dotted lines which reflect the average income in the control group and the expected increase in child expenditures if we were to increase this income by $1,000 per year. The right panel shows

---

[4] We replicate Table 3 in the Appendix tables A2 and A3 with only income reported as mother and spouse earnings to determine the robustness of the income gradient to different forms of income measures. Estimates are similar when income gradients are estimated from total household income in comparison to income gradients from mom and spouse earnings. Mother and spouse earnings do not systematically approach more accurate correspondence with the experimental impacts. The gradient estimates using only maternal income corresponded more closely with measures of purchases for the child since birth, maternal agency, and maternal anxiety than gradients estimated with only the spousal income or household income. Gradients estimated with both maternal and spouse income corresponded more closely with child-focused expenditures and neighborhood safety than gradients estimated with the household income. Gradients based on household income were the most accurate in estimating social services receipt.

[5] Lowess line depicted is calculated using the RStudio version 1.4.1103 base command lowess(). The values shown in the right hand plot are the same as those shown in the left hand plot, however the constrained range of the x and y axis slightly distort the smoothness of the line.

the model-implied impacts of the intervention and illustrates that in comparison to the observed experimental impact (shown by the red line). All the nonexperimental impact estimates are much lower than the experimental estimate.

Are the Magnitudes of Experimental and Nonexperimental Estimates Similar?

Although we expected nonexperimental gradients to overpredict experimental estimates, we found the opposite. Half of the nonexperimental estimates were smaller than the magnitude of the experimental impact. Social services were an exception, wherein linear estimate magnitudes and direction corresponded exactly (compare columns 2 & 4 in Table 2; β_{ex} = 0.021, β_{ne} = 0.021**) and log estimates closely approximated (compare columns 2 & 6 in Table 2; β_{ex} = 0.021, β_{ne} = 0.017**).

To compare estimates across different outcomes and approaches we standardized the estimates by dividing each one by the corresponding control group standard deviation. Standardized estimates are reported in Appendix Table A4. Additionally, we use the percent difference formula ($\%difference = \frac{\beta_{nonexperimental} - \beta_{experimental}}{|\beta_{experimental}|} X\ 100\%$) presented in Steiner & Wong (2018) and originated by Wilde & Hollister (2007) to express the difference in estimates in terms of the experimental estimates magnitude. There was widespread over- and under- estimation by the estimated linear gradients (shown in Appendix Table A5) where some outcomes were underestimated by 115% and others over-estimated by 144%, with a median underestimation of 69%. Because the median linear nonexperimental estimate was approximately 0 (specifically 0.0038) we find that the average experimental estimate was underestimated by approximately 100%.

In the same manner, log income gradient estimates over- and under- estimated the experimental impact (shown in Appendix Table A6) where some outcomes were

underestimated by -179% and others over-estimated by 101%, with a median underestimation of -83%. Since the median log nonexperimental estimate was approximately 0.03%, we find that the average experimental estimate was underestimated by approximately 100%.

Are the estimates in the same direction?

Only 33% of the linear and 42% of the log gradients were in the same direction as the experimental impacts. Across multiple income modeling specifications and statistical estimation approaches, less than half of the experimental and nonexperimental estimates corresponded in their directions.

Is There Correspondence Between the Estimates Standardized to Scale of the Outcome?

To compare the correspondence of estimates for all the measured outcomes we plotted the control group standardized estimates in Figures 3 and 4. In Figure 3, we plot the linear regression estimates without covariates (left) and with covariates (right). The nonexperimental linear regression estimate magnitude is reflected on the y-axis and the experimental impacts is reflected in the x-axis. If there were perfect correspondence between the experimental and non-experimental estimates, we would expect these to align with the 45-degree dotted line. The same plotting conventions are used in Figure 4 to compare the nonexperimental log income gradients rescaled to reflect a $1,000 annual income increase. In summary, the plots in Figures 3 and 4 show that the experimental impacts are much more variable than the nonexperimental estimates. We test the robustness of our results and explore possible explanations for this finding below.

Robustness Checks

The correspondence of experimental and nonexperimental income gradients may differ when estimating gradients in a selective sample of mothers of newborns living in poverty in comparison to a nationally representative sample of parents with newborns. To test whether our nonexperimental income gradients are representative (given that the BFY sample was selected based on mothers' incomes), we also compare our nonexperimental estimates to estimated effects calculated using the Panel Study of Income Dynamics (PSID) – the longest-running panel study of household income dynamics in the US. This dataset provides a larger sample with a wider range of household employment, income, time use, and the levels of poverty to test the robustness of income gradients on outcomes that are comparable to the child and maternal outcomes collected in BFY. The PSID is a suitable dataset to represent the vast diversity of families in the US to understand how their income may relate to their lived experiences, constraints, and opportunities. We replicate our analysis of income gradients with the 2013 cohort of family units (FU) reporting at least 1 newborn within 2011-2013 (n = 892). We were able to match four of the twelve outcomes in BFY to outcomes measured in the PSID. Estimated income gradients from the BFY control group and the estimated income gradients from the PSID sample can be compared in Appendix Table A7. Three of the four income gradients estimated in the BFY control group were larger in magnitude than the linear estimates in the PSID sample even when including covariates and constraining the sample to only include parents earning between 0 to 25 thousand dollars annually. In addition, the gradients estimated in the PSID sample had consistently smaller standard errors than the BFY gradients, likely due to the larger sample and range of incomes in PSID. Overall, we find that the PSID linear estimates were

negligibly different than the BFY control group linear gradient estimates, with differences in the regression coefficients ranging from -.013 SD to .013 SD. Therefore, the correspondence of experimental and nonexperimental income gradients was robust to different samples.

Possible Explanations for Discrepancies

Confounding

We estimate the potential influence of omitted variable bias on the nonexperimental income gradient estimates by comparing the estimates across models with different covariate specifications, such as comparing models that exclude the mother's demographic variables to models that include them. These are shown in Appendix Table A8. We find mixed results: Estimates of the items purchased for the child since birth, child focused expenditures in the past 30 days, and maternal agency grew larger with the inclusion of all covariates. For neighborhood safety, social services, parent-child activities, and maternal anxiety the estimates decreased; however, most differences were small, falling within a standard error of the unadjusted gradient. Estimates of the index of economic stress, child overall poor health, sleep disturbance, and dichotomous estimates of houselessness and physical abuse were consistent across different covariate specifications. Within the estimated gradients, only a few outcomes demonstrate a pattern of diminishing towards 0 with the inclusion of difference covariates, including the Beck Index of Maternal Anxiety, and Parent-child activities index. In summary, there was not a clear trend of decreasing observational estimates when covariates were adjusted.

Construct Impurity

We also test if the lack of correspondence may be due to qualitative differences in income in the experimental and nonexperimental arms of the within-study comparison. The stable income that income gradients reflect may be fundamentally different from receiving an unconditional cash-gift delivered each month on the child's birthdate on a debit card branded as "4MyBaby". If increasing income through cash-gift specifically branded as for the newborn child is not the same as ensuring a basic income, then we should expect to see greater impacts on child outcomes in the experimental estimates in contrast to the nonexperimental estimates as shown in Table 3.

We compare the differences between standardized experimental and nonexperimental estimates in raw units in Appendix Table A4 and in percentage differences in Appendix Tables A5 and A6. The median difference between experimental and nonexperimental estimates for child outcomes was -.0424 SD . The difference in estimates for maternal outcomes were on average -0.0048 SD, which is a very small difference. Across child and maternal outcomes, the difference in estimates range indicated that nonexperimental estimates were on average 99% smaller to 115% bigger than the observed experimental estimates. To exemplify the magnitude of these differences the discrepancy in purchases since birth indicates that for each additional $1,000 the gradient predicted 0.045 fewer items purchased for the child since birth, than what was experimentally observed. For the amount spent on child-focused expenditures in the past 30 days, the discrepancy in gradient estimates reflects that for each additional $1,000 of income the families would be expected to spend $15.33 less than what was experimentally observed on child expenditures. For the index of economic stress, the discrepancy in

gradient estimates reflects that for each additional $1,000 of income the families would be expected to have a -.0963 lower score, and instead we observed a 0.2466 score increase on the economic stress index, making the discrepancy a .1503 point difference on a 9 point scale. Altogether, these results do not indicate clear patterns of discrepancies being systematically smaller for child outcomes and larger for maternal outcomes.

Imprecise Measures of Experimental Impacts

We find that the experimental estimates were substantially more variable than the nonexperimental estimates, which may be partially responsible for the weak alignment between experimental and non-experimental estimates. This contrasts with the hypothesis that weakly aligned experimental and non-experimental estimates reflect the influences of different kinds of income on child development and maternal well-being outcomes. It is possible that estimates reflect largely the same underlying data generating processes, but the estimates are weakly related because the constructs that they are measured with introduce error into the measurement. To assess this possibility, we plotted the average of the linear and log estimates as a red triangle on each plot. The average of all the linear and log gradient estimates reveals that, on average, experimental impacts are very small but relatively unbiased.

To address whether the imprecision of experimental impacts influences their correspondence with the non-experimental linearly estimated gradients, we conducted a Bayesian analysis of their correlation (Matzke et al., 2017). An advantage of the Bayesian approach is that it is sensitive to the imprecision of measurement and produces a full posterior distribution for the correlation coefficient. The posterior distribution quantifies

how likely every possible correlation value is to be the true correlation.[6] This allows the impact of uncertainty in measurement to be expressed in terms of uncertainty about the true correlation coefficient. It also allows for what is known as the "attenuation" of correlation (Spearman, 1904; see also Matzke et a., 2017), which is the possibility that two measurements showing a modest correlation when imprecisely measured may in fact have a strong underlying true correlation that is being obscured by the imprecise measurement.

The main panel of Figure 5 shows the experimental and non-experimental estimates in a scatterplot, with the linear nonexperimental Bayesian estimates on the y-axis and the Bayesian experimental estimates on the x-axis. The error bars correspond to the uncertainty of measurement and the experimental estimates are substantially more imprecise. The inset panel shows the posterior distributions as a shaded area. We assumed a uniform prior distribution for the correlation coefficient, in which all correlations between -1 and +1 were a priori equally likely, as shown by the dotted line in the inset panel. To infer the posterior distribution, we applied the statistical model developed by Lee and Wagenmakers (2013, Ch 5.2), which is implemented in the JAGS graphical modeling language (Plummer, 2003). JAGS automates computational Bayesian inference based on Markov-chain Monte-Carlo methods (Gilks et al., 1996). The posterior distribution is shown by the shaded area in the inset panel. It shows that there is a large degree of uncertainty

---

[6] Correlations between average household income and various outcome measures for the full sample are shown in Appendix Table A9, correlations for income and all outcomes for the control group are shown in Appendix Table A10.

about the true correlation. Correlations in the range -0.50 to +0.80 are plausible true values. The most likely correlation (the mode of the posterior distribution) is around +0.50. However, the observed correlations between experimental and nonexperimental estimates from our linear and logit models ranged from r =.13 to .18. The reason the posterior distribution is higher than the correlation is that any estimation error in the experimental and nonexperimental estimates will bias the observed correlation downward, so our best estimate of the true correlation between population parameters will always be higher than the observed correlation. A value of .50 would indicate that experimental and non-experimental population parameters substantially reflect the same data-generating process, but also derive from unique processes. On the other hand, based on the small number of observed outcomes, the results shown in the posterior distribution are also consistent with a very small correlation. the uncertainty that remains about the true correlation, more evidence is needed to reach this conclusion.

We summarize the correlation between experimental impacts and nonexperimental estimates from linear and logit regressions in Table 3. Even if we were to assume that there is a strong correlation between the experimental and nonexperimental estimates of the impact of income on child development and maternal well-being, the BFY study yielded effect sizes and standard errors reveal that it is unlikely to produce estimates that closely correspond using these two estimation strategies. On the positive side, the experimental impacts may not be too far off, but the downside is that if we want to analyze the data to determine which effects of income are bigger than others then we might not have sufficient precision to do this with one year's worth of data. It may be unrealistic to expect that we can use these data to make fine grained distinctions between theories of income gradient

impacts on various child and maternal outcomes. This is difficult considering the imprecision of estimated experimental impacts when compared to their nonexperimental counterparts. Overall, this has important implications for the promise of a year-long income transfer experiment for informing theories of child development and maternal well-being.

DISCUSSION

We found little correspondence between experimental income impacts and nonexperimental estimates of income gradients. Both the linear and logarithmic income gradients over-estimated and under-estimated the causal impact of a $1,000 increase in annual income. We find that the income gradients estimated from observational data correspond weakly with the observed causal impacts of a year-long unconditional cash gift. Weak associations appear to substantially result from low precision in the experimental impact estimates, relative to the size of their magnitudes. Below we discuss the implications of this work for attempting to craft evidence-based policy based on experimental and non-experimental estimates.

Although we find little correspondence between the experimental and nonexperimental impacts, this is not sufficient evidence to conclude that either experimental or nonexperimental analyses are not mutually beneficial. Instead, we view the uncertainty reflected in these results as indicating that if we could obtain larger or more precisely estimated income impacts, we may see better calibration between experimental and nonexperimental estimates. The income gradients from the BFY and PSID datasets provide further evidence that the estimated impacts of a $1,000 increase to annual income are too small relative to the calculated standard error, making it difficult to

compare them to the experimental estimates. Although the literature indicates that income is correlated with the outcomes measured in this paper, it is plausible that the magnitudes of causal effects are either incorrect or reflect different kinds of economic inputs than cash transfers. Lack of correspondence may result if (1) unmeasured variables cause differences in both income and outcomes (confounding); (2) the measured outcomes actually cause income (i.e., people with better mental health are better able to secure employment; reverse causation), and/or (3) cash transfers associated with one's children are used in different ways, thus affecting mothers and children differently than other kinds of spending.

Lack of correspondence may also be in part due to measurement error in the income measures of experimental and non-experimental gradients, complicating the comparability of both estimates. The experimental income boost is measured precisely, but the control group income gradient may have more measurement error for multiple reasons. Though, this is likely true the measurement error in nonexperimental income would bias the estimated effect on the outcomes downward by a square root of the reliability of the income measure. If this is the case and nonexperimental gradients have a very low reliability of 0.5, then we could scale them up by 1/.7 = 1.4. However, as shown in Table A4 this scaling still does not approach the factor of differences in spread of the estimates which is 4x larger for experimental estimates vs. non-experimental estimates as shown in Figures 3 and 4. Therefore, income measurement error is unlikely to explain a large proportion of the discrepancies between experimental and non-experimental estimates.

Due to the precision constraints of measuring multiple maternal and child outcomes, this work indicates that the experimental impacts reported in BFY may not

provide sufficient information to make clear distinctions between impacts on separate outcomes. Our results are similar to those reported in a recent randomized control trial of the impacts of a one-time unconditional cash transfer delivered during the onset of the COVID-19 pandemic on individuals material hardships and mental health (Jaroszewicz, Jachimowicz, Hauser, & Jamison, 2022) . Jaroszewicz and colleagues (2022) found nonsignificant impacts on various outcomes although expert researchers and laypeople predicted impacts ranging from +0.16 to +0.65 SD, perhaps indicating that people expect experimental impacts to be larger than a careful analysis of gradients might lead them to expect. An implication of this pattern of findings is that experimental estimates that are particularly large in magnitude relative to their corresponding income gradients should be interpreted with caution, because they are estimated with far more error.

Future Directions

For nonexperimental and experimental estimates to be more mutually informative, larger samples or experimental impacts might be necessary. Following the sample receiving the unconditional cash-gift for a longer period might be useful if impacts might plausibly grow with a longer treatment period. Alternative approaches would include a more intensive intervention, where more than $4,000 are gifted annually, or collecting data from larger sample than 1,000 mothers. However, we acknowledge that both solutions would be very expensive, large scale policy experiments may be designed to answer some of these questions.

Conclusion

Identifying the causal effects of cash transfers for mothers with newborn children living in poverty is of great importance for welfare policy. This task is further complicated

by our findings that the experimental income impacts of a cash transfer experiment do not closely correspond closely to the income gradients often used to inform decision making in the past. By increasing the precision of the experimental estimate, we might improve our approximation experimental and nonexperimental gradients, and have sufficient precision to determine when these estimates do not correspond due to biases in the nonexperimental income gradients.

Assessing the correspondence would help to determine under what conditions nonexperimental can be modeled to predict the outcomes of child allowance policies and to test theories of how income influences child development

TABLES

Table 2.1

Hypotheses About Estimate Correspondence Split by Sources of Incongruence and by Outcome Measure Type

Outcome Variables	Income Construct Impurity	Confounding bias in nonexperimental estimand	Imprecision in Experimental Estimates
	Qualitative differences between measured stable income and cash transfers may result in predictable difference in how families spend their money.	Because confounding variables not adjusted for or measured at baseline would likely influence outcomes in the same direction as income.	No Correlation or Weak correlation across all outcome estimates. Could indicate nonexperimental estimate is unbiased
Child Focused Investment (source for impacts reported in Gennetian, et al., 2022)			
Items Purchased since birth Expenditures in last 30 days Parent-Child activities	Nonexperimental < Experimental	Nonexperimental > Experimental	No correlation or weak correlation
Maternal Focused Investment (source for impacts reported in Gennetian, et al., 2022)			
Social services receipt Economic stress Neighborhood safety Homeless or in group shelter	Nonexperimental > Experimental	Nonexperimental > Experimental	No correlation or weak correlation

Maternal agency			
Child Health (source for impacts reported in Sperber et al., forthcoming)			
Child Overall poor health Child sleep disturbance	Nonexperimental < Experimental	Nonexperimental > Experimental	No correlation or weak correlation
Maternal Health (source for impacts reported in Magnuson et al., 2022)			
Beck index of maternal anxiety Ever cut/bruised/seriously hurt by partner	Nonexperimental > Experimental	Nonexperimental > Experimental	No correlation or weak correlation

Table 2.2

Income Descriptive Statistics by Treatment Group

Originally reported in Noble et al., 2021; Gennetian et al., 2022, and Magnuson et al., 2022

Variable	All	Low-Cash	High-Cash	T-test *p-value*	
Household combined income - revised	21.61	21.89	21.21	0.52	
	(15.57)	(15.96)	(15.00)		
HH income (last calendar year)(2019 $)	23.00	23.69	22.03	0.20	
	(18.65)	(18.82)	(18.39)		
Average Yearly Income Age 0-1	22.00	22.41	21.41	0.30	
	(14.42)	(14.44)	(14.39)		
Mom Earned Average Yearly Income Age 0-1	9.74	9.72	9.78	0.91	
	(7.094)	(7.497)	(6.484)		
Spouse Earned Average Yearly Income Age 0-1	16.24	16.60	15.67	0.40	
	(11.92)	(12.38)	(11.15)		
Mom & Spouse Average Yearly Income Age 0-1	11.50	11.66	11.28	0.44	
	(7.10)	(7.46)	(6.55)		
Age 1 Outcome Variables					
Purchases for child since birth Index	4.85	4.80	4.91	0.40	*
	(2.00)	(2.05)	(1.92)		
Child-focused expenditure Index (amount in last 30 days)	332.59	311.73	362.38	0.02	*
	(319.5)	(283.2)	(363.5)		
Perceptions of neighborhood safety Index	4.30	4.38	4.19	0.03	*
	(1.35)	(1.32)	(1.37)		
Index of Economic Stress	-2.78	-2.68	-2.92	0.05	
	(1.87)	(1.81)	(1.94)		
Social Services Receipt Index	-2.88	-2.87	-2.89	0.9	
	(1.66)	(1.65)	(1.68)		
HOPE Maternal Agency scale	31.55	31.72	31.32	0.19	**
	(4.5)	(4.59)	(4.54)		
Beck Index of Maternal Anxiety	-5.14	-4.58	-5.94	0.01	**
	(7.38)	(6.57)	(8.34)		
Parent-Child Activities Index	10.49	10.29	10.78	0.01	
	(2.65)	(2.68)	(2.58)		
Child Overall Poor Health Index	-5.64	-5.56	-5.75	0.17	
	(2.11)	(2.03)	(2.22)		
PROMIS Sleep Disturbance (child)	-7.89	-8.02	-7.70	0.17	
	(3.41)	(3.44)	(3.36)		
Has been homelessness or in group shelter Indicator	-0.09	-0.08	-0.09	0.6	
	(0.28)	(0.28)	(0.29)		
Ever cut/bruised/seriously hurt by partner	-0.08	-0.08	-0.07	0.53	
	(0.27)	(0.28)	(0.25)		

Observations	931	548	383	485-931

Mean coefficients, Standard deviation in parentheses. Income is represented in in thousands in last calendar year 2019 dollars. Treatment assignment was 60/40 split to conserve 1 million dollars. The average household combined income age 0-1 in the control group is used as our point of reference in the remaining analyses (i.e., the estimated impact of increasing income by $1,000 for families averaging 22.41 thousand annually). Reverse coded variables have negative means.

Table 2.3

Comparing Experimental Estimates to Nonexperimental Income Gradients

Outcome Variables	Experimental Impact		OLS		OLS Income Log Transformed	
	Original	Scaled to $1,000	Income scaled in thousands of dollars		Scaled to $1,000	Scaled to $1,000
	(1)	(2)	(3)	(4)	(5)	(6)
Covariates Included	Yes	Yes	No	Yes	No	Yes
Child Items Purchased since birth Index	0.243+	0.066+	0.011*	0.012*	0.012*	0.012
	(.136)	(.0366)	(.006)	(.007)	(.006)	(.007)
Child-focused expenditure Index (last 30 days)	65.900**	17.793**	1.079	2.468**	1.416*	3.018**
	(23.088)	(6.234)	(.713)	(.857)	(.705)	(.855)
Perceptions of neighborhood safety Index	-0.170+	-0.046+	-0.001	0.003	-0.004	0.0005
	(.094)	(.0254)	(.004)	(.004)	(.004)	(.005)
Index of Economic Stress	-0.183	-0.050	0.017**	0.014**	0.017**	0.016
	(.129)	(.0347)	(.005)	(.006)	(.005)	(.006)

Social Services Receipt Index	0.080	0.022	0.021**	0.021**	0.018**	0.017**
	(.104)	(.028)	(.005)	(.005)	(.005)	(.005)
HOPE Maternal Agency scale	-0.349	-0.094	0.020	0.028*	0.028*	0.036**
	(.310)	(.084)	(.0144)	(.0151)	(.014)	(.015)
Beck Index of Maternal Anxiety	-1.657**	-0.447**	-0.021	0.007	-0.023	-0.002
	(.510)	(.138)	(.020)	(.021)	(.018)	(.020)
Parent-Child Activities Index	0.438*	0.118*	-0.006	-0.001	-0.004	0.001
	(.180)	(.049)	(.007)	(.009)	(.008)	(.009)
Child Overall Poor Health Index	-0.253+	-0.068+	0.005	0.011+	0.007	0.014**
	(.151)	(.041)	(.006)	(.007)	(.006)	(.007)
PROMIS Child Sleep Disturbance	0.350	0.094	-0.017+	-0.008	-0.017	-0.010
	(.230)	(.062)	(.010)	(.012)	(.009)	(.011)

Table 2 Continued						
Homeless or in group shelter Indicator	-0.025	-0.007	-0.002**	-0.002	-0.002**	-0.002
	(.0187)	(.005)	(.001)	(.001)	(.001)	(.001)
Ever cut/bruised/seriously hurt by partner	0.022	0.006	0.002*	0.002	0.002+	0.001
	(.024)	(.006)	(.0008)	(.0008)	(.0001)	(.0001)

Note. + p<0.10; * p<0.05; ** p<0.01. Estimated impact of additional $3,760 per year in shown in column 1, the rescaled impacts to compare to nonexperimental impacts are shown in column 2. Estimates in column 2 can be multiplied by 4 to simulate average impacts of EITC and CTC benefits. Regressions on last two variables reflect marginal effects – the average change in probability, evaluated at the mean value of the dependent variable. Covariates from baseline survey: Mother's age, Completed Schooling, Household Income, Net Worth, General Health, Mental Health, Race and Ethnicity, Marital Status, Number of adults in the household, Number of other children born to the mother, smoked during pregnancy, drank alcohol during pregnancy, Father living with the mother, Child's sex, Birth weight, Gestational age at birth, phone interview, child age at interview (in months). Log estimates were transformed (columns 6 & 8) to reflect the average impact of an additional $1,000 per year, we rescaled the log estimate by dividing it by the product of the average annual income in the control group (22.32K) multiplied by the natural log 2.718 (log / (22.32*2.718). SE of log transformed estimates (columns 6 & 8) are transformed by using the t-ratio (dividing the original log regression estimate by its standard error)

Table 2.4

Correlations Between Income Experimental Impacts and Linear and Log Estimates

	Estimate	1	2	3	4
1	Experimental Impact	–			
2	Linear Income Gradient (no covariates)	0.15	–		
3	Linear Income Gradient (with covariates)	0.13	0.92***	–	
4	Log Income Gradient (no covariates)	0.18	0.99***	0.91***	–
5	Log Income Gradient (with covariates)	0.14	0.90***	0.98***	0.92***

Note. Income estimates are from estimated income impacts on 12 child and maternal outcomes.* p<.05,** p<.10, *** p<.00

FIGURES

Figure 2.1

Within Study Comparison Design Using Baby's First Years Data

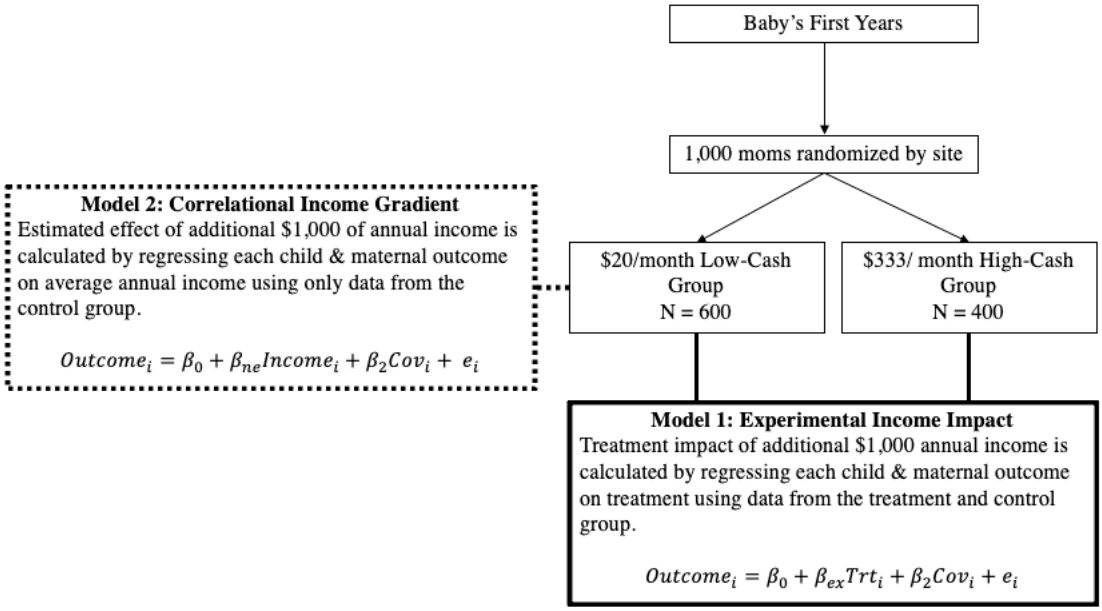
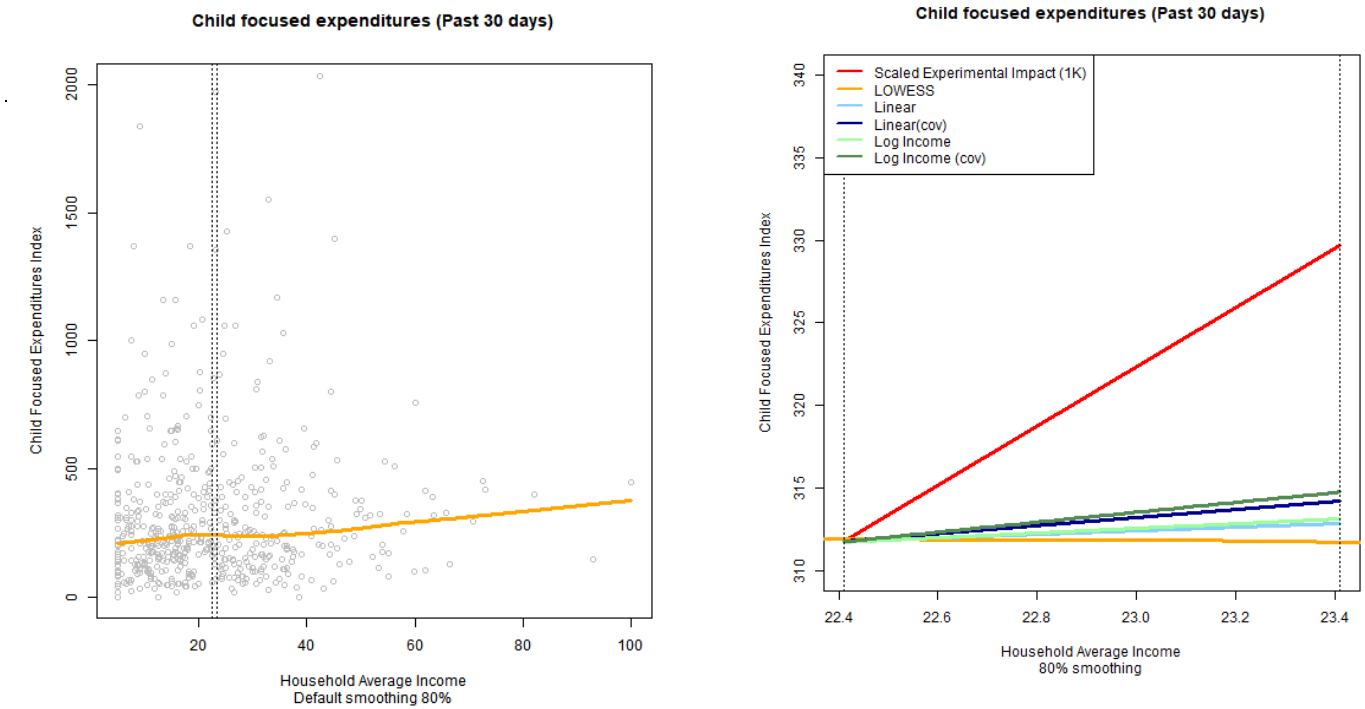


Figure 2.2

Example of Income Gradients Comparisons Using Child Focused Expenditures

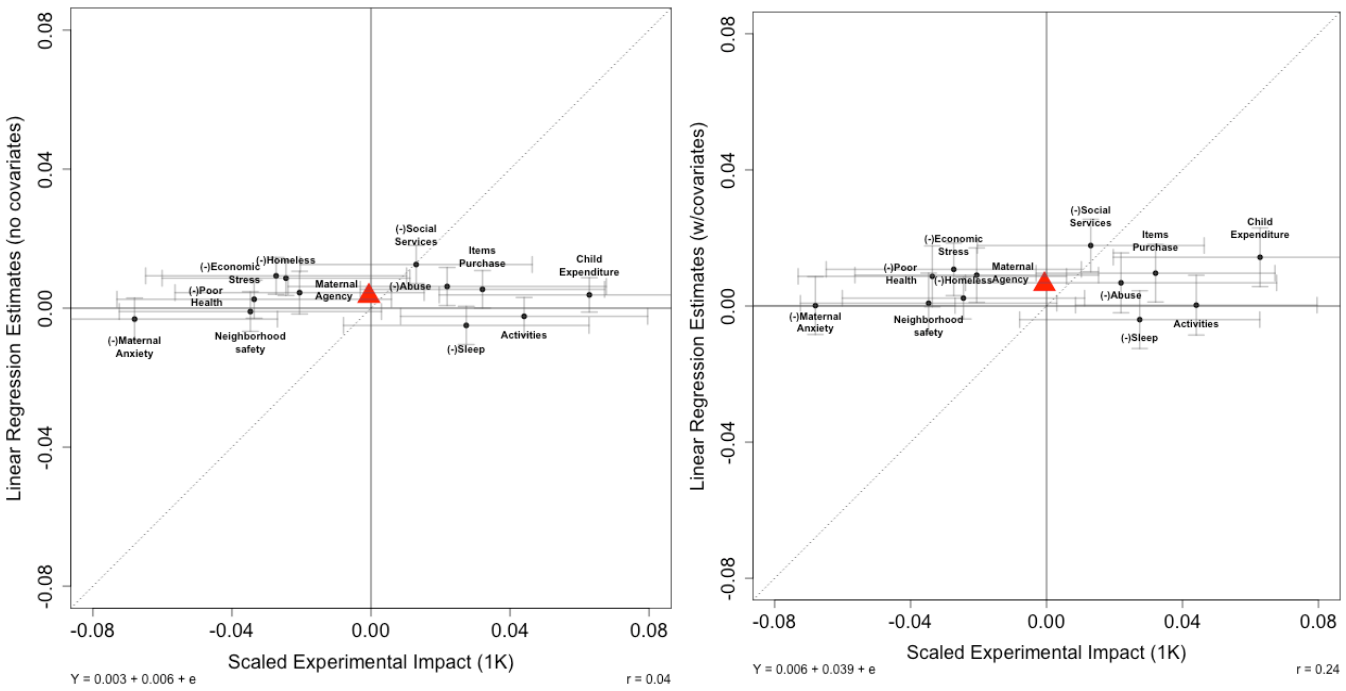
	Experimental Impact		OLS		OLS Income Log Transformed	
Outcome Variables	Original	Scaled to $1,000	Income scaled in thousands of dollars		Scaled to $1,000 increase at the mean	Scaled to $1,000 increase at the mean
Covariates Included	Yes	Yes	No	Yes	No	Yes
						1.126**



Note. Left plot shows the lowess line fit for the distribution of child-focused expenditures across all average annual income levels. *N* = 544 (4 missing income). The right plot shows average annual income in the control group of 22.41 K to illustrate the differences in income impacts estimated from the experiment (red line) in comparison to linear (blue lines) and log (green lines) regression-based income gradients. In the right-hand plot the intercept for each line, except the experimental slope, was fixed to begin at the control group income mean of $311.773

Figure 2.3

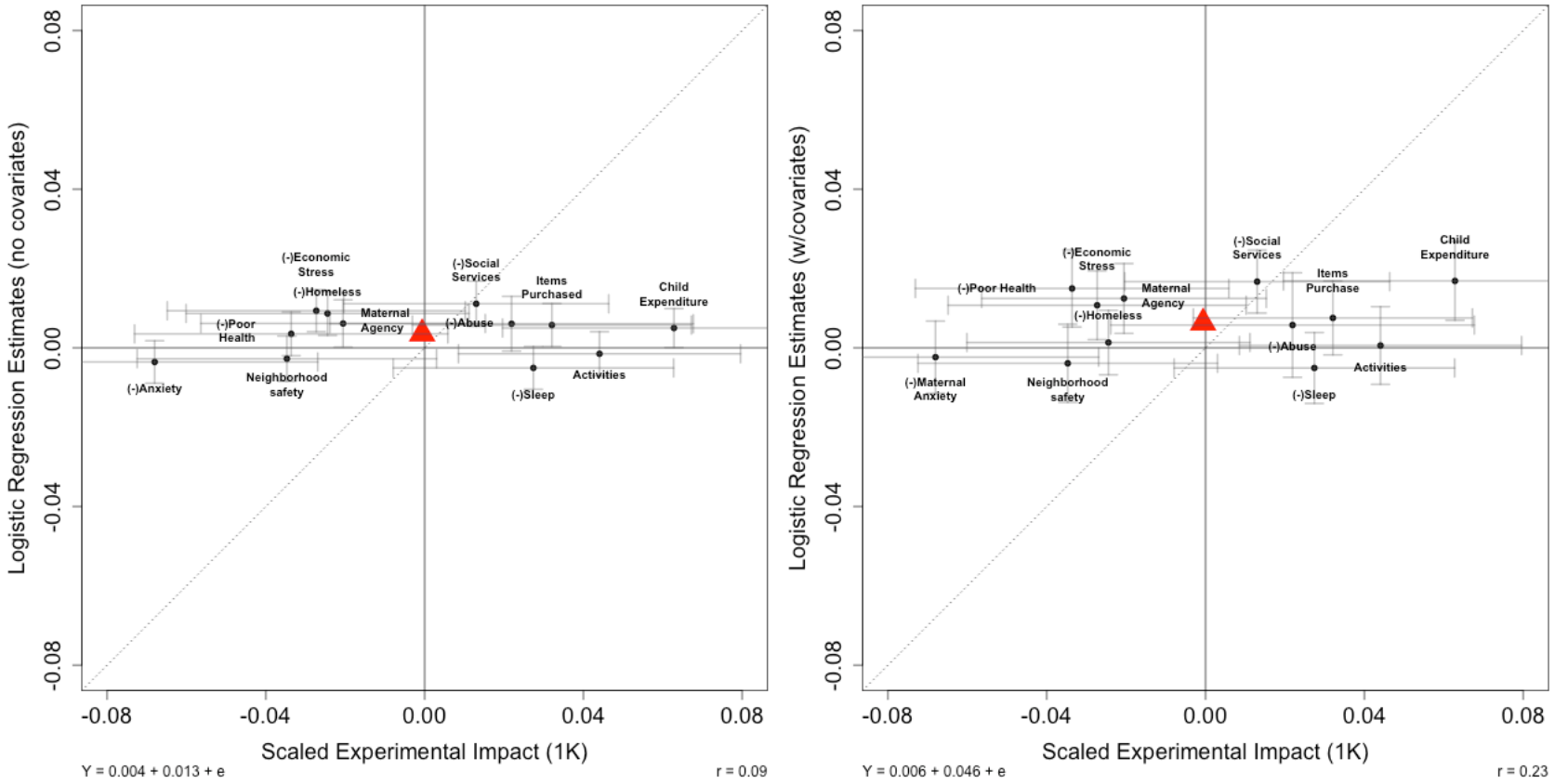
Comparing Linear Income Gradients and Experimental Impacts Using Control Group Standardized Outcomes



Note. Dotted line reflects where the markers should land for perfect correspondence. The blue line reflects the relation between non-experimental estimates when regressed on the experimental estimates. The regression equation is shown on the bottom-left of the plot and the correlation between estimates are shown in the bottom-right.

Figure 2.4

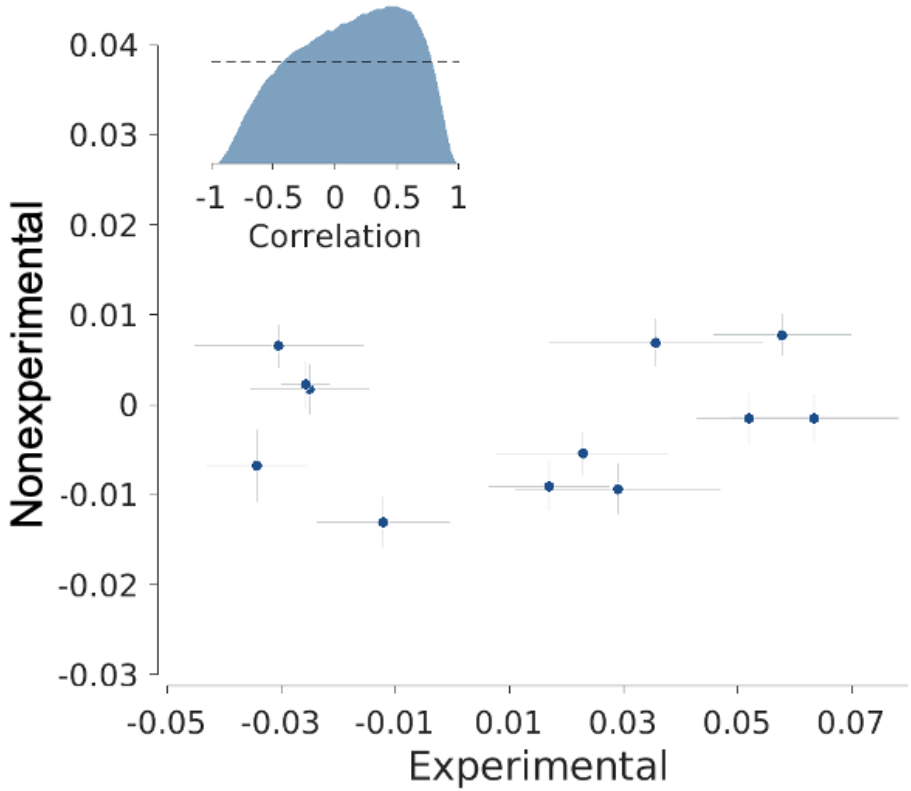
Comparing the Log Income Gradients and Experimental Impacts Using Control Group Standardized Outcomes



Note. Dotted line reflects where the markers should land for perfect correspondence. The blue line reflects the relation between non-experimental estimates when regressed on the experimental estimates. The regression equation is shown on the bottom-left of the plot and the correlation between estimates are shown in the bottom-right

Figure 2.5

Comparing Bayesian Correlation of Income Gradients and Experimental Impacts



Note. The relationship between the experimental and non-experimental estimates. The main panel shows a scatterplot of the measures, with error bars representing the imprecision of measurement. The inset panel shows the uniform prior distribution on the correlation coefficient (dotted line) and the inferred posterior distribution of the correlation coefficient (shaded area).

STUDY 3: LESSONS LEARNED FROM MATH PROGRAM ADAPTATIONS

INTRODUCTION

Various strands within the social sciences have been trying to understand how to bring promising programs to scale successfully (Baron, 2013; List, Suskind, & Supplee, 2021). As new programs are adopted and scaled their success depends on a delicate balance between maintaining fidelity to the key components of the program design (Dusenbury et al., 2003), while adapting the program to best fit the local context and meet the needs of teachers and students (Castro, Barrera, & Martinez, 2004). However, the question of how variation across adaptations influences the impacts of educational innovations remains. Weiss, Bloom, & Brock, (2014) developed the Conceptual Framework for Studying Variation in Program Effects, Treatment Contrasts, and Implementation to guide research on variation of program impacts and to determine the sources of this variation for analysis and program improvement. This framework highlights the complementary goals of studying program implementation with the goals of estimating program effectiveness, to capture sources of variation and more systematically collect data that can yield insights about program outcomes.

The framing that Weiss and colleagues (2014) provide is focused on increasing the accuracy and information surround an average treatment effect. This framework is iterative across instances of study design and implementation, such that the data collected during the implementation of the study are used thereafter to inform a subsequent study. Key to this framework is the experimental manipulation of a treatment, contrasted with some counterfactual. The key dimensions of treatment contrasts to measure include the content, quantity, quality, and conveyance of the treatment. However, the researcher must decide what aspects and components best characterize these dimensions. Theory and

previous research are important references for determining what to measure a priori. However, a key component that researchers may or may not anticipate is how the local context may respond to each of the dimensions of treatment contrasts.

In the current study, I incorporate the Framework for Reporting Adaptations and Modifications to Evidence-Based Implementation Strategies (FRAME-IS; Miller et al., 2021) from implementation science work in public health to complement the Conceptual Framework for Studying Variation in Program Effects, Treatment Contrasts, and Implementation. The FRAME-IS contributes an in-depth form of data collection wherein the processes taking place at the implementation site and amongst the researchers are captured. Using this form of data collection operationalizes and captures how adaptations to program implementation take shape and contributes to better designs for different populations. Currently, the conceptualization and measurement of how the variation in program effects are moderated by a changing local context can be vague and underspecified, requiring iterative testing to determine in what contexts the program works best on average. Typically, the client and context characteristics that are measured as moderators depend on data that is readily available, moderators based on theory, or moderators based on the data collection norms of the field. By capturing the various modifications that occur "on the ground" of a research study can clarify the various implications of changes, make explicit the assumptions of design conjectures, and enhance the take-up of the program being implemented (Bernal et al., 2012).

Capturing the adaptations that educational stakeholders make contributes to understanding the variation of experiences underlying program implementation and contribute to program effects. Moreover, understanding these experiences can reveal

important nuances about what works, for whom, and under what conditions. Currently, the extent to which different forms of adaptations influence the effectiveness of programs is understudied (Miller et al., 2021). This descriptive qualitative study is designed to operationalize the individual and contextual factors that have influenced math program implementation in previous rigorous studies. In doing so, I aim to gather and highlight insights on elements of program design and implementation that can be designed to better fit diverse educational contexts as they scale up. My commitment in doing this work is to highlight the ways that adaptations made in the local context can contribute to design theory and educational theory that is better suited to address the limitations and build on the strengths of complex and dynamic educational systems.

By combining these Conceptual Framework for Studying Variation in Intervention Effects and the FRAME-IS, we can better understand and refine educational theories and inform the future design of experiments to test new solutions to build on or mitigate contextual challenges. I contribute a summary of the adaptations that have occurred to researchers developing and evaluating math programs over the past two decades. I highlight the challenges and lessons learned that future researchers might anticipate as they design math curricula, instructional practices, and programs. More broadly, this study culminates in guiding principles that can informs theory and practice on how educational research can be responsive to contextual modification to better fit dynamic local contexts.

Adaptations in Educational Program Implementation

Quinn (2017) & Kim (2017) have experimentally evaluated the differential impacts of implementing a reading intervention that emphasized fidelity in contrast to a scaffolded sequence emphasizing fidelity and structured opportunities for adaptation. The authors

found that implementing a fidelity-focused intervention helped build teacher knowledge about the new intervention. Then, once the teachers had prior experience with the program, implementing the intervention with a structured adaptive approach that allowed teachers to make changes to improve intervention fit with the local context further contributed to improved teachers' self-reported literacy related learning and students' standardized reading comprehension scores.

The added value of teacher's developing adaptations to improve intervention fit with their local context underscores the importance of capturing these variations in program evaluation. Especially, considering the potential adaptations making interventions less effective (McLaughlin & Mitra, 2001). Furthermore, understanding various forms of adaptations at different levels can contribute to the development of frameworks for studying the variation of program effects (e.g., Weiss, Bloom, & Brock, 2014), which cannot account for the impacts of design conjectures or programmatic adaptations made within the local context if these adaptations are not systematically recorded and reported.

Adaptation is the process of modifying an intervention to align with the local situation or context without changing the core elements of the intervention (Zayas et al., 2012). A review of 293 educational reform projects found that throughout all of them, implementation reforms were shaped by the local context (Berman & McLaughlin, 1976). Adaptation is often necessary for developing educational programs that (1) account for ongoing processes and constraints in a complex system, (2) are designed to deal with implementation barriers, (3) understand the local variation and can identify factors influencing it, and (4) can be sustained by the local infrastructure (Datnow, Hubbard, & Mehan, 1998; Donovan, 2013; Bryck, 2014). However, adaptations are difficult to plan for

because they vary based on the goal of the research, this complexity obscures our understanding of how to design new programs and interventions for adaptation and how to empirically test the effectiveness of these adaptations.

FRAME-IS

The field of implementation science has developed and tested methods and strategies that promote the adoption and implementation of evidence-based programs (Eccles & Mittman, 2006). To better understand the impacts of adaptation processes on the effectiveness and sustainability of healthcare programs, Miller and colleagues (2021) developed the Framework for Reporting Adaptations and Modifications to Evidence-Based Implementation Strategies (FRAME-IS) to capture and evaluate the adaptations or modifications that are made to these implementation strategies in different populations, settings, and contexts. The FRAME-IS can be used an interview protocol or a spreadsheet format to build a database of the adaptations that take place as programs are implemented and scaled up. It serves as an organizing framework to capture and explicitly report the iterative development of evidence-based programs. This framework and approach to documenting the adaptations has grown in the public health fields since 2013, but it has not been applied to the development and evaluation of educational programs. Using this tool to document the changes made to new programs and interventions as they are implemented by researchers and practitioners can be crucial to determining the key components of interventions that make them successful and the components that can be adapted to better align with contextual needs.

The FRAME-IS can be used as a research tool to document how difference across populations and contexts arise across time in alignment with the propositions of the

Bioecological Model of Human Development (Bronfenbrenner & Morris, 2006) thus capturing the Process-Person-Context- Time (PPCT) model within each study of math program development and evaluation. At the study level, this can help researchers illustrate the multiple forces of influence on a program's effectiveness in their publications and reports to school stakeholders who often need to know under what conditions the math programs have achieved the desired outcomes. At a broader level, when multiple math program studies design research and capture data that gives specific meaning to the PCCT model this allows for meta-analytic analyses and tests of theories that can further reveal what math programs work for whom and under what contextual conditions the uncovering of these fine process details can contribute to more refined and specific theories that can be further tested. For example, suppose a math program is being implemented to the mathematical development of refugee children in a local school district across several years. It will be difficult to capture all the potential influential factors described in the a guiding developmental theory such as the Integrative Risk and Resilience Model (Suárez-Orozco et al., 2018). However, using the FRAME-IS and adapting it to capture sources of influence on program adaptations from the individual to the global ecological systems levels can help to operationally define what proximal and distal factors had the most notable effects on the program implementation. This is important to capture as programs more widely implemented so that researchers and implementers can attend to and anticipate the factors that have demonstrated the strongest influences in past iterations.

Science of Adaptation

Adaptation to new situations and environments is important for educational reform to be implemented and sustained (Datnow, 2002). In clinical psychology the cultural adaptation of evidence-based interventions (EBIs) has been shown to increase the relevance, acceptability, effectiveness, and sustainability of the intervention by its targeted population (Benish, Quintana, & Wampold, 2011; Griner & Smith, 2006). Theories informing the process of adaptation note the importance of working in collaboration with the target population and assessing the consequences of adaptation through empirical study (Ferrer-Wreder, Sundell, & Mansoory, 2012). Educational research can draw on the examples of successful cultural adaptations in clinical settings to inform our research process educational innovations (Cabassa & Baumann, 2013).

There are, however, important limitations to adaptation. One limitation is not being able to scale-up an innovation that is tailored to a very specific context such as if the innovation design is entirely contingent on the specificity and cannot define what is necessary about the situation for the innovation to be effective (Kelly, 2004). A second limitation is the possibility that researcher or practitioner intuitions about what may work and when can fail: judgments across many domains of expertise fail relative to very simple statistical prediction, although both are fallible (Dawes, Faust, & Meehl, 1989), and the relevant statistical information may not be available if the adaptation in question has not previously been studied. In addition, misdirection of collaborative research can result from imbalances in the division of power between multiple members (Wallerstein, 1999). Moreover, collaboration between researchers and practitioners is dynamic and difficult to document and empirically analyze, leaving much to be known about the process through

which these collaborative arrangements achieve their goals (Penuel & Hill, 2019). There is a need for empirical analysis assessing how collaborative processes such as the adaptation of new educational innovations co-designed in partnerships changes outcomes (Tseng, 2017).

Anticipating Change

This work is useful for researchers to be able to foresee or anticipate the kinds of adaptations that will be made to the programs they are designing and evaluating. Most educational programs are designed to eventually be effective enough to be implemented in real educational systems without direct researchers' intervention. Indeed, the goal of educational research is to identify the best practices for the school systems to implement and maintain whether it is focused on how teachers are trained, classroom instruction, or school service structures. However, due to the diversity of contextual factors and systemic changes that occur over time, no single program implementation will always look the same changes occur and components are modified, as researchers engage in iterative program development the lessons learned from adaptations are stored in their experiential knowledge.

When new researchers or implementers take up the program or similar math programs that full knowledge base may be difficult to transfer due to a phenomenon called the curse of knowledge (Newton, 1990). The curse of knowledge is a cognitive bias that makes it difficult for the expert to imagine what it is like for novice to not know what the expert knows. In this case, the researcher that has developed and implemented numerous iterations of a math program is biased against knowing what a novice researcher or implementer may need to know about the program to effectively use it in a new setting. If

key components and adaptations of the program are not clearly communicated, then the program may not be implemented with sufficient fidelity or integrity to the key components that make it helpful. For this reason, researchers write extensive reports of the research activities however, adaptations made during implementation do not always make it into these publications and reports remaining only in the expert knowledge of the person who made them. By using the FRAME-IS through interviews or reporting the adaptations in spreadsheets these can serve as one concrete form of information storage and sharing that can build transparency and trustworthiness of the research activities and how they may be relevant to any future implementation of the math programs. Without knowledge of this history, new researchers or implementers may be left to react to adaptations they did not anticipate in productive or unproductive ways, in a worst case scenario the lack of anticipation and planning for contextual adaptations can cost money, time, and threaten ongoing investment into the program.

Replication

The goal of multiple math programs is to identify a process, structure, or system that helps students learn and retain mathematical competencies above and beyond their current local contextual affordances. The math program is meant to be of service beyond the population it was developed in, and thus is expected to be trustworthy and replicable in different contexts and with different students. Specific definitions of replication vary, however, across the field of education the rate of replication is low (Makel & Plucker, 2014; Perry, Morris, & Lea, 2022). Often replication may not be the explicit goal of a study attempting to work in a partnership towards a very context specific problem. However, replication may still be a useful tool in this process to identify potential moderators, across

the different schools or classrooms embedded within a research-practice partnership. When replications fail to produce the same results, they are simultaneously helping develop new hypotheses and identifying systematic barriers, which help researchers understand for whom, and in what contexts the program or practice should work best (Kim, 2019). Moreover, these replications can further inform, theory, policy, and practice with using a causal replication framework (Steiner, Wong & Anglin, 2019).

Researcher Approaches and Tools to Math Program Development

Research in education can be very diverse with new educational practices and phenomenon being tested in laboratory spaces, homes, classrooms and throughout the school system. National funders such as the Institute of Educational Sciences (IES) categorize funding streams for different kinds of projects with their own explicit requirements and guidelines such as Developmental and Innovation studies, requiring cycles of development and implementation to ensure it is usable and feasible. In contrast Effectiveness studies are required to provide proof of prior efficacy and are expected to be implemented by a school without special support from the researchers or developers (IES, 2012). The goal of all this work is to improve education using evidence-based practices however, many of those practices or approaches have been evolving over time.

One current promising approach is intervention development and evaluation through the approach of research-practice partnerships to develop programs that are trusted and usable for the stakeholders of the research. By working in collaboration with teachers, students, and community's researchers can better understand of the factors influencing implementation at scale (Bassok, Markowitz, & Morris, 2021; p.6). Approaches from improvement science (Bryk, Gomez, Grunow, & LeMahieu, 2017) and design-based

research (Penuel, Fishman, et al., 2011; Anderson & Shattuck, 2012) are particularly well suited to guide research-practice partnerships with strategies to continuously collaborate on the design and improvement of the implementation of educational programs, practices, or interventions. An important aspect of the iterative improvement approaches of design-based research is the understanding of the local context informed by researchers and teachers (Anderson & Shattuck, 2012). Improvement science focuses on using network improvement communities of researchers and the users of the research to understand how to design and combine the multiple factors that make schools work through rapid testing of changes, collaborative examination of the results, and collaborative designing for further improvement (Bryk et al., 2017). Similarly, the Design Based Implementation Research (DBIR) approach focuses on problems that persist in education from the perspectives of the research users and is committed to iterate through multiple co-created designs to develop the capacity to sustain effective changes in the educational system (Fishman, Penuel, et al., 2013).

Although, multiple strategies for working effectively in research partnerships have been established (Farrell, Wentworth, & Nayfack, 2021) challenges to effectively maintaining a partnership remain. One dimension of effective partnerships is the contributions that they make to educational improvement, one way to capture this in collaborative work is to develop methods to capture and organize program adaptations that support the partners practice and help them achieve their goals.

I propose the use of an adaptation/modification framework and tool like the FRAME-IS to capture adaptations, such as the theories of change and models of implementation that change based on data collected and practitioner feedback. Though this

proposal is not equipped to answer the question of whether specific processes of adaptation yielded greater outcomes, we can still inform the field of the insights gained by researchers and the different forms of evidence that can be captured to evaluate how adaptations influence specific partner goals and educational outcomes. These are important contingencies that must be documented to communicate whether this program of research, theory of change, or implementation model can be used by other researchers and practitioners.

To gather understanding of how adaptations manifest in educational program research I focus on work with math interventions and programs because mathematics is a field in education where there are considerable obstacles in providing marginalized students with sufficient opportunities to learn and develop expertise. Work on designing and implementing math education programs are often designed to target historically marginalized students to improve their performance and promote the nation's competitiveness in the global market (U.S. Department of Education, 2008). In contrast, I frame my goal towards improving the mathematics opportunities afforded to historically marginalized students to emancipate and liberate them from historical stereotypes, economic hardships, and other forms of systemic oppression (Martin, 2009). I emphasize the role of adaptation to understand how adapting interventions to meet the needs of students and teachers may require the active involvement of the teachers and students themselves in the research.

Adaptations that increase the fit between the intervention and the target population can improve the outcomes of the program through better acceptability and engagement (Bernal & Domenech-Rodríguez, 2012). However, there are many advantages and

limitations to this kind of collaborative research which contribute to the difficulty of measuring their effectiveness in achieving their goals (Welsh, 2021). Key in the alignment of new evidence-based practices is identifying the active ingredients in the intervention that cannot be changed and identifying the needs of the community (Chambers, Glasgow, & Stange, 2013). The question of how adaptations and modifications to educational innovations have influenced the effects of educational innovations in field settings remains. By combining the iterative evaluation approach using a conceptual framework developed to study the variation of program effects (Weiss et al.,2014) with the FRAME-IS (Miller et al., 2022) methodology of systematically investigating how collaborative adaptation takes shape in dynamic contexts, educational researchers can better understand the processes that: (1) increase program acceptability and engagement, (2) influence the success of collaborative / partnership work (Henricks et al., 2017), (3) manifest during joint work between research and community partners, and (4) to document processes contributing towards improved scale-up and sustainability (Anderson & Shattuck, 2012; Farrell, 2021). The current study aims to document some regularities in the kinds of modifications commonly made during the development and implementation of math interventions, which I hope will inform future work on whether and how such adaptations might influence intervention efficacy and fit with diverse educational contexts and populations.

CURRENT STUDY

I draw on FRAME-IS, to characterize modifications to the implementation of evidence-based math programs. This framework will be applied to the field of education to standardize the collection and categorization of adaptations or modifications that are made

to educational innovations. I apply this method of extensive documentation to add specificity to the many ways that the characteristics of the implementing organization, the stakeholders, and the context can influence variation of implementation to better understand what works for whom, when, and under what conditions.

Research Questions

RQ1: What kinds of adaptations are made to educational innovations? RQ2: Are there regularities in these adaptations across different research goals such as between developmental studies v. efficacy studies? RQ3: How might regularities across research goals be planned and built into the original program design?

Interpretative Framework

In this work I use a post-positivist interpretative framework with elements of constructivism to understand conditions of math program development and implementation. In doing so, I co-construct meaning with each interviewee and value competing perspectives as both representing a truth. To this end, I highlight multiplicity of approaches, perspectives, and outcomes. My ontological inclination is to view the nature of reality as what emerges from participating in a community of researchers implementing math programs. Epistemologically, I view the findings that are co-created with participants as multiple ways of knowing, while also respectfully interrogating the values that researchers express to gain deeper understanding about how these values have been informed by their experiences.

METHOD

To exemplify how the FRAME-IS can contribute to the study of variation in math program effects research, I conducted a comparative case study of research teams that have developed or implemented math programs to understand if and how adaptations occurred. I conducted semi-structured interviews with researchers and publication content analysis to triangulate instances of the reported process of conducting adaptations and the goals and reasons for those adaptations. The semi-structure interview protocol was informed by the FRAME-IS codebook (Miller et al., 2021) and the interviews were conducted in a private Zoom room following participant agreement and virtual consent. The interview questions were shared ahead of time for transparency and to prime participants to reflect on their past research experiences.

Participants

University of California, Irvine self-exempt IRB approval was granted before recruiting participants. Participants were recruited in via email (shown in appendix A). Participants received a link to a Qualtrics Survey with a complete IRB-approved Study Info Sheet as the first page. The study sheet asked participants, prior to administering research procedures, to verify that they meet the eligibility criteria. If so, they indicated their willingness to participate in the research by clicking "Yes" at the bottom of the sheet. Following consent participants were they were redirected to the HubSpot meeting scheduling where they provided their name, email, and scheduled a meeting time at their earliest convenience. After scheduling this meeting an automatic confirmation email was sent including a simplified protocol of the interview questions that would be asked.

Participant names and audio recordings will be stored in separate protected files and linked through an anonymized identification number.

A total of 14 participants were interviewed once for 30 minutes to one hour using semi-structured interviews via zoom. Each interview is treated as a case study describing the research activities of each researcher, in one case there were two participants that completed a second interview. Therefore, there are 14 interviews total but 13 separate case studies. Table 1 demonstrates a breakdown of the participants research goals and approaches as described during the research conducted in the interviews as well as the number of changes they reported. These are in no way meant to be interpreted as statics characteristics of these researchers as many have developed new approaches and techniques across their research based on what is most fitting for the research goals and questions. Therefore, if one researcher is identified as an external evaluator in this study, this does not mean they continue to only do external evaluation in their current work, rather this is a simplified description of their role during the specific study discussed in the interview.

Data Preparation

Interview audio and transcripts were entered into MAXQDA software (VERBI Software, 2022). I reviewed and cleaned all the transcripts to immerse myself in the data and generate thematic memos which were applied to each transcript and used to develop summary case descriptions that were sent to each interviewed member to engage in the meaning making reflection. Interviewees sent back their case descriptions with tracked changes, all results incorporate these peer-revisions. Differing sources of evidence of adaptations (interview, journal articles, and case descriptions) were grouped into set to

allow for cross-case comparisons. Interview responses were used to qualify researchers' characteristics such as the goals of the studies that were adapted, the approaches used in the research, and the fields relevant to a researcher's identity. For interviewees with scientific publications, I coded for evidence that triangulated the codes I had applied to the interview regarding the researcher's personal characteristics as well as the nature of the adaptations they made throughout the process of their research. No other identifiable information was collected. To respect the wishes of some participants to remain anonymous, all participants were given pseudonyms.

Sampling Plan

Using grant funding databases from the Institutes of Educational Sciences (IES) and the National Science Foundation (NSF) I recruited principal investigators, evaluators, and program staff to participants to semi-structured interviews. I reviewed interventions that have met the Evidence-Based What Works Clearinghouse (WWC) standards demonstrating positive impacts on math performance. The WWC evidence standards require replication and scale-up ensuring that endorsed programs have been through various levels of design and implementation, making it more viable that various adaptations. Of the 225 intervention reports in the WWC database (https://ies.ed.gov/ncee/wwc/Publication#/FWWFilterId:1,ContentTypeId:1,SortBy:RevisedDate,SetNumber:1). I emailed 18 primary investigators and successfully scheduled 6 interviews. See Table 1 for participant description. I also reviewed the NSF-DRK12 funding database for continuing program design and testing grant (see website for more details https://www.nsf.gov/pubs/2020/nsf20572/nsf20572.htm). This included developmental math program that were approached using tools like improvement science (IS) or design-

based approach (DBIR). Of the 2,232 awards that resulted from this search (https://nsf.gov/awardsearch/simpleSearchResult?queryText=math&ActiveAwards=true) 16 focused on math program development. As programs are developed in partnership, I expect that adaptations or modifications may occur to meet the needs of school administrators, teachers, and students. However, it is possible that university or researcher developed math interventions are simply applied to new contexts without specific design changes to meet local needs. Only 2 interviews were successfully conducted prompting me to pursue a more representative sample using the following methods.

Following the direct emails to the principal investigators and interviews, I was able to recruit more participants through snowball sampling researchers that were within participants networks and evaluating or developing mathematics programs. Five more participants were recruited, two who identified as working in partnerships and three participants conducting IES funded evaluations.

Interviews

I reviewed the relevant publications for the studies that I planned to speak to interviewees about to prepare for each interview and to familiarize myself with any reports of adaptations or modifications for each project. Examples drawn from the literature were used during the interview to describe what I mean by modifications, adaptations, or key change points in the development and evaluation of an intervention to be expanded upon. During each interview, I began by asking that the interviewee tell me about their work so that I may capture their preferred language and use the same language throughout. The full interview protocol is shown in Appendix A. The goal of the retrospective interviews is to have investigators reflect on how adaptations came to take place in their programs as they

were implemented or scaled. I use the questions provided by the FRAME-IS coding manual as a reflection guide. Since the FRAME-IS, was developed for use in health care interventions, I make inductive modifications to the codebook as examples arise from the data that the codebook is not designed to capture. Following the coding manual of the FRAME-IS I use an deductive coding approach to summarize the types of adaptations that math intervention developers report. Moreover, I inquire about the outcome of these changes and gather lessons learned from each interviewee to inform how future program developers and evaluators can proactively plan to consider, potentially evaluate, and document the influence of adaptations. Overall, my aim is to describe the important role of adaptations and modifications in program development, implementation, and evaluation to provide specific case assertions that program developers can refer to as they engage in research.

When interviews reveal evidence of improvements made and share strategies that were productive or not productive, I summarize these as key components of math programs that can inform grant funders and researchers' practices. The variation in evidence of improvements contributes to the production of knowledge that can inform educational improvements more broadly. As I document the strategies or design conjectures that worked and didn't work for research groups and how these were reported I can gather how research groups share lessons learned to contribute to broader educational improvement. In addition, I ask participants about processes or outcomes they wish they would have measured to gather potential lessons learned that can contribute to other researchers planning process. However, as these interviews are retrospective there may be information that may have been forgotten, or not explicitly mentioned. To address

these issues, I also review publications and privately shared artifacts from the research teams such as autobiographical writings and technical reports. This extensive search may reveal potential regularities in adaptations made across research groups that can contribute to improvements in research planning by being built into the design of future math programs that are being iteratively implemented or evaluations that plan to scale-up.

Positionality

As a graduate student researcher who has worked for five years in the development and evaluation of math interventions, I bring in an insider perspective regarding the different assumptions researchers make and the challenges to doing this work. I have worked at the boundaries of conducting program evaluation to provide causal quantitative evidence of student improvement on math performance. I have also completed course work and worked in a developing research-practice partnership with a local school district, where the focus was on relationship building, working towards a problem of practice, and iteratively improving the math intervention leveraging practitioner expertise and prioritizing the contextual constraints and affordances. Through this work I have developed an appreciation for both research approaches, and I combine the language of both approaches to cross the boundaries of these research goals as I communicate and interpret the participants experiences. As I use my past experiences to build interpretation, I verbalize this openly to interviewees to ensure that their motives and perspectives are communicated accurately and not biased by my own experiences. I often resonate with the interviewees experiences and openly communicate this to build common understanding of adaptations, research challenges, and academic endeavors. This open communication has built trust and clarity to ensure that participants voices are clearly and accurately

represented in this work. Lastly, although I am coding from a framework, I have used open coding and memoing to gather the codes into grand themes that represent participants own voices and experiences in the results that follow.

Codebook

The codebook I will use will build from the FRAME-IS (Miller et al., 2021) components and additional characteristics about the study to contextualize the educational innovation, the members of the partnership, and the outcomes of their work on the innovation. The following descriptions operationalize and provide examples of the different characteristics I will code for in the studies reviewed. The full codebook is provided in Appendix B.

FRAME-IS Codebook Adaptation and Modifications

What is being adapted or modified in the educational innovation or its implementation process will be categorized as either being (1) content based if the content of the innovation or implementation strategy changes, (2) evaluation based a change is made to how the implementation is evaluated, (3) training based if there are changed made to how the innovation implementers are trained, and (4) context based if changes are made to the overall implementation strategy. The contextual based changes can include the format of delivery, the setting it is delivered in, who delivers the innovation, the target population, and any other potential contextual factors that arise in the analysis that were not originally accounted for. Deeper descriptions of the nature of the adaptations made will also be recorded from a list provided in the FRAME-IS. Some examples include whether the modification was to change the materials, to remove elements of the innovation, or to integrate a new strategy in the implementation. The description of these changes will be

saved as data to facilitate future analyses and rich descriptions of the adaptations made through partnerships. Lastly, I will record whether the adaptations made were in fidelity with the core elements or functions of the original educational innovation or if this is unknown.

The rationale of a modification is broken down into two separate codes. The first is to determine what is the goal of the adaptation. This will be noted from a pre-specified list, but if the goals stated by the researchers do not align with the pre-specified goals a rich description will be entered. Some examples of the pre-specified goals include: to increase the reach of the educational innovation, to increase the effectiveness, or to decrease disparities in delivery. The second code captures the level of the organization of partnership from which the rationale stems from such as to adapt to the sociopolitical level (i.e., existing policies), the organizational level (i.e., staffing capacity), implementer level, practitioner level, or the recipient level. Some of the reasons for adaptations at the implementer, practitioner, or recipient levels include but are not limited to cultural norms, racialized experiences, sexual orientation, accessibility, preferences, or time constraints.

When the adaptation occurred in the study will be coded as occurring pre-implementation, during implementation, during scale-up, or during an attempt to maintain or sustain the educational innovation. Furthermore, I will code for whether the adaptation was planned proactively, planned reactively in response to an unanticipated event, or if the adaptation just occurred without a formalized plan in a reaction to an unanticipated event which I will consider as more of a modification. If coding for when the adaptation occurred is not possible, I will email the lead author to try to gather this information from their reflections.

Who participates in the decision to make an adaptation or modification will be coded as including political leaders, principal investigators, administrators, managers, funders, researchers, implementation experts, practitioners, community members, and students or recipients of the innovation. However, if other roles to titles are reported, such as graduate students or parents, these will also be included to provide the most accurate description. Furthermore, I will code for who the adaptation is meant to be made for. For example, if administrators modify the language of the innovation to align with the language used by the students, I will code the decision makers as administrators and the students as the group that the adaptation is made for.

RESULTS

Each case was analyzed at the interviewee/project level, to compare the frequency of adaptations between research teams, in one case two researchers were involved and interviewed their codes were aggregated at the case level. This allows for a richer analysis of different study types within groups lead by the same primary investigators and how adaptations may vary based on the goals and stages of their different projects. Interviewees described the development or evaluation of either a single math program, multiple math programs, or changes occurring at different points of the developmental timeline of a math program (s) which ranged from three to seven years. To understand regularities across various forms of research we first aggregate the code frequencies of adaptations made to math programs at the interviewer level.

What kinds of adaptations / modifications are made to educational innovations?

The code summaries of the adaptations made to math programs focuses on two aspects (1) the process of undergoing that adaptation, and (2) the goals and reasons why the adaptations were made.

Process

Across all cases there were changes to the content of the math intervention or program whether that was a curriculum, a model for teacher professional development, or an online tutoring software (see Figure 3.1). Content changes (Figure 3.2) encompassed adding new elements to the program (54%), for example participant 11 noted

> *They [teachers implementing program] really struggled with not having an anchor chart with the Spanish and English word, for numerator and denominator …I said if you feel that strongly go ahead…I don't think an anchor chart is changing the fidelity.*

Elements were also frequently removed from the program (46%) such as grade-level content that was too advanced for the target group, especially when efficacious interventions were adapted to new schools. One element was removed based on the progression of research in education as participant 11 explained during a scale-up replication study:

> *Use a pen and strike it out. That was a content issue for us that, like it was like we're across the board. We are changing this, and we are training teachers on why, this is not a good practice, because we can influence the trajectory of how these kids' teachers teach. So that was more like a research based decision on like. There's a lot of research on rules that expire, and how just having kids associate, an alligator is not*

helping prepare for algebra readiness, but that was the only that was like the one like major shift that we all made, and we like had to back it up with evidence, and then went through some of those things.

Smaller refinements (38%) included tweaks to wording and symbols presented in teacher scripts to align better with their instructional vocabulary and help teachers anticipate when to emphasize certain phrases. Moreover, adaptations were made to shorten (31%) and lengthen the duration (23%) of a program, as participant 12 remarked an intervention designed for kindergarten students with difficulty in math

The only instruction they're getting in math is in their core. So, an intervention on top of that you're probably pulling away from the highest priority area which is reading. We were going to do like 30 min for an intervention block, and we just doubted down to 20, based on sort of feedback, from teachers.

Fewer studies integrated a new treatment to the program (23%), such as P12 stated: "Now, as part of that, we've gotten really interested in ways that we can support language development within math programs and math vocabulary. So, we're going back. we're adding on to the program." Other changes that were less frequent included changes to the packaging of the program (31%) as explained by participant 3 who was assessing the efficacy of an intervention delivered by teachers, in addition to shortening the sessions she noted:

We also audio record the sessions. So, we look at things like timing… Often we try to jam too much into a lesson, or, you know, maybe we should slow down. And also we learned on the PowerPoints that if we put the script um, not that we want teachers to

be completely scripted, but we put the script on the bottom of the PowerPoint … where teachers have it right there and then like ideas for discussion, and so on. And again if you're trying to have fidelity of implementation. You do kind of have things have to have things relatively scripted.

Participant 1 confronted major changes that drifted away from the original theory of action. She explains the events occurring during a developmental study conducted in partnership with a school district:

The chief also said they no longer needed to use the same curricula across schools. And so, CMP basically went away as well over the course of that next four years there was a lot of shifts. The coaching that was probably the biggest ,um loss, and that was the other thing is like. We also had data on mathematical knowledge for teaching of those coaches, their visions their qualitative data around teachers. They were such an expert people, and so losing that source of expertise in a district like that, was, it's just devastating.

This experience highlights the extent to which the context of the programs can vary and the challenges this may present for adaptation. The types of changes made to the program context are shown in Figure 3.3. All cases demonstrated some form of adaptations at the level of the program context, specifically in the re-formatting of the programs from one-to-one delivery systems to group focused implementation. Another example is provided by participant 9 whose program is focused on teacher professional development, he remarked

We pivoted from the professional development because, we want to try to create an experience for students to learn this content, not periodically throughout the year. but, after all, of the high-stake testing is finished.

In addition, the personnel that implement the programs often transitioned from researchers to practitioners already embedded within the schools (43%).

Participant 3 explained her rationale in doing this,

We originally had the researchers carry out the intervention in small groups and in larger groups, and then we train teachers, and because we wanted the intervention to be more authentic and see if teachers could use it and use it effectively. We are getting data, but it's slow.

As projects scaled-up there were also changes to the populations involved, such as when a program designed for students struggling with math was implemented in a classroom to benefit all students (36%), and changes such as altering the setting (43%) in which the program was delivered. Participant 2 explained this adaptation,

It's a changed program in order to be um, so that teachers can use it in an efficient way. All you know, with all children. And so what we're doing is the lessons um are the same… it's not exactly the same program, but it's the same content and the same lessons. Just deliver through a different platform.

Least frequent were modifications to the training (57%) and implementation and evaluation protocols (29%) that researchers originally designed (show in Figure 3.1). However, the changes made to professional development and implementation strategies reveal important lessons for researchers developing and accessing math programs. Participant 7 offers one of these approaches that worked in their external evaluation:

But we also now incorporated coaching and modeling in the classroom throughout the period. So, if we notice that there is a slacking off software usage. Then we arrange for coaching and modeling session in the classroom with those teachers who are not using it effectively. Very early on I mean there's no point in rushing around, you know, a month before the post-test, but we do it like one month at the beginning. 2 months, 3 months. So, 3 coaching, modeling sessions in order to alleviate any chance of this kind of like this, you know, kind of like a fading effect.

Most adaptations maintained fidelity to the core components of the original theory of action or research plan (93%), yet almost half of the adaptations reported did not preserve fidelity(47%), as shown in Figure 4. In the case of participant 1, the program implementation did have to stray away from integral assumptions of the theory of change due to regulation changes at the district level. In a separate case participant 8 shares an important deviation from fidelity:

there was less professional development in the second year and that may have accounted for why teachers changed their behavior. And maybe that was positive for the program. and things things changed, for example, it was very clear that in the second year the the the curriculum itself needed to be more aligned with the State curriculum…So they [teachers] decided like, okay, I'm, they're not gonna be implementing it with the same fidelity that the developer wanted them. And so they adjusted it but they still did it more than the comparison group before. And we believe that that's really what one of the reasons why and students did it better, which is my theory would be. And I guess maybe maybe I think and and I say that probably in the article is, is fidelity really that important?

Researcher teams (86%) and practitioners (79%) such as teachers, school interventionists, and curriculum specialists within the schools often made the decisions to make the adaptations, see Figure 3.5. In some cases the programs were implemented by external evaluators– independent researchers contracted to evaluate the effectiveness of the program to preclude bias or conflict of interest (n=3). In these cases, two external evaluators contracted via research firms differed in that they only permitted adaptations that were pre-specified in the routine implementation protocol from the original study or reported adaptations made by the practitioners outside of their control as researchers. In one case, the external evaluator was also a intervention developer in the same field; in this case the evaluators adaptations encompassed the full spectrum of content to implementation changes as they saw best suited the new scale-up context.

Other prominent decision makes included the school district leadership such as principals, principal supervisors, and the math department (36%) and the grant funders or program officers (29%). School students (14%) and project coordinators (7%) were rarely included in the decision making and when the math programs were being evaluated by external evaluators, the primary investigators and developers of the program were not able to participate very much (14%).

Most (93%) case studies reported making at least one adaptation in reaction to constraints or challenges occurring during the process of implementation and often in alignment with the core components of the original intervention plan. About 50% of the cases reported also making changes proactively, prior to the implementation of the program, in some cases to test the impacts of these changes. Lastly, 43% of the cases faced changes that were made to the program that were not planned, such as a change in the

personnel implementing a program, reductions in the dosage of the program, and teacher's tendencies to return to their traditional instruction as opposed to the practices taught during the math program implementation.

Across the cases there were regularities in the types of adaptations made, with most adaptations focusing on making changes to the program content and context. The majority of the 76 adaptations reported in the interviews preserved fidelity and integrity to the key components of the math program design, the decisions were informed primarily by researchers and practitioners, and most changes were planned in reaction to events occurring during the implementation of the math program. All together the substantial presence of these regularities may reveal important phenomenon that frequently arises in the design and implementation of math programs. This is of important consideration when the majority of the changes were not anticipated previous to implementation and can pose additional challenges to the planned theory of change and implementation model.

Goal and Reasons

The predominant goal of the adaptations made was to improve the program fit with its recipients (85%) such as making items age appropriate, catering to different cognitive abilities, or designing for teachers with different levels of training. A prominent study goal and subsequent goal of these adaptations was to cater the content and format to students who were struggling with mathematics or designed specifically for students with learning difficulties (n cases = 6). However, some adaptations also occurred to address constraints in disadvantaged districts (n cases = 3), or to better suit the general student population as it was impacted by broader societal forced like the COVID-19 pandemic or other forms of educational reform (n cases =7). Indeed, many adaptations aligned to the broader stated

aims of the research studies many of which focused on ensuring equitable high-quality math instruction through different mechanisms including professional development, curriculum, additional school programming, and educational tools.

All coded goals and reasons are shown in Figure 3.6 Panel A followed by the sources of influence on these goals in Panel B. Although all cases focused on improving student outcomes, a large share of the changes made focused on ensuring the program fit with the teacher's implementation could be feasible. Participant 6 details some of this process in her work as a curriculum & instructional strategies developer & evaluator:

> "The school consultant is a specialist in teaching mathematics in the elementary school, but her job is to consult other teachers, but she's, however specialist in teaching mathematics, and she, Her knowledge is really crucial because she knows how teachers react. Yes, it is. Teach us every day, every day, every day it's. It's her job to work with teachers. Teachers are different in different circumstances, in different modes, and and she has a lot of experience. and when she looks through our proposal, and she can comment on like mathematical with the didactical ideas for kids, if it will work with kids, but also how we can approach teachers if the teachers will be able to do? If not, what should we do? How how should we present it to teaches? How should we work through with teachers, and what is missing, etc. So it was her role and on the second project was 2 of them."

Next, most frequent goal was to improve the effectiveness and outcomes of the program, with most of the decisions towards this goal coming from the researchers yet arising from the constraints teachers faced in schools. Participant 11 demonstrated this in the following excerpt:

"The other thing that we did is again with that 30 min constraints like these teachers that teach intervention groups all day genuinely like they teach for 30 min they have a 5 min break to drop off and pick up kids, and they teach for 30 more minutes, and they're doing that all day long. It was 30 min, and we're done so. We told them that a non-negotiable was that they couldn't cut the training because the training is the teacher directed component of the lesson. but the on the relay which does the kind of like in the "I do", "We do", "You do" it's the "We do" section. And then, in the independent practice that you do that they could predetermine which problems they were going to cut so they could like. cut the left side of the page, or cut the right side of the page, or cut the evens, or cut the odds. But then they needed to think systematically about which problems they were going to cut and not just like haphazardly. Okay, just do the first 3, you know, and that is their only rule. So, we felt like we had to make that, because we didn't want them to cut the training because we wanted that direct instruction there. We really didn't want to cut the independent practice of the relay, either. But we had to give them some options when they didn't have enough time".

Relatedly, the goals for adaptations also focused on improving recipient satisfaction (38%), feasibility (38%), fidelity (46%), sustainability (46%), addressing inequities or disparities observed (23%), increasing reach and engagement (31%), addressing inequities in the school system (19%), addressing cultural factors (31%), increasing retention(8%), and reducing cost were least frequently reported (23%). Many of these goals overlapped, as shown in the excerpt before the changes not only address student learning and a better

fit to the interventionists schedule, but in prioritizing these factors the fidelity and feasibility of implementing the program was also improved. The factors that often influenced these goals arose from different ecological levels, as shown in the bottom plot of Figure 6. For example, there was overlap across adaptations made that were influenced by the student level (learning progressions), the teacher level (personal preferences and professional judgements), the organizational level (school service structures and resource constraints), and the broader sociopolitical level (educational reform, state testing mandates).

Most adaptations have two predominant goals in improving fit with recipients and addressing multiple sources of influence from the individual differences in outcomes at the student level, the feedback from teachers, the constraints of school resources and service structures, and the sociopolitical norms and mandates. All levels of influence are associated with each other and influence specific modifications revealing the patterns across the regularities researchers can anticipate facing when doing educational research in dynamic systems with changing counterfactual conditions across organizations and across time.

Are there regularities in these adaptations across different research goals?

To build evidence towards math programs that can be beneficial across the complex educational systems the National Center for Education Research (NCER), the National Center for Special Education Research (NCSER), and the Institute of Education Sciences (IES) have developed reports and recommendations detailing the progression through different phases that studies must make to build incremental evidence (Taylor & Doolittle, 2017). Among these stages there is an important point of transition from developing a program and demonstrating its efficacy (goal 3), to then evaluating that program in a

different setting, sometimes by an independent evaluator under the routine conditions the developers designed (goal 4). If the program demonstrated success in goal 4 then a scale-up study is conducted with a much larger and diverse sample (goal 5) to assess generalizability, however making it to goal 5 is rare.

Across all case studies only one project had reached this stage, and all cases included in this study fall along one or multiple points in this spectrum. To better understand how adaptations play a role in these phases of development I compared differences in processes and goals of adaptations detailed below. I hypothesized that programs in development would often make changes to adapt to the practitioners' requests in contrast to programs in evaluation phases wherein the goal is to implement core design components with high fidelity. Technically, projects in the developmental stages should have less specified theories of change and plans of implementation as they are in the process of developing this through iterative testing and discussions with stakeholders. In contrast studies in the evaluation stages should have robust specifications for exactly what the evidence based theory of change is and a stricter implementation plan based what worked in previous studies.

Contrary to the technical specifications, adaptations took place across all phases of a program's development and evaluation (see table 3.1) providing several regularities for deeper examination. Since many cases reported multiple projects at different stages of development the following analysis details proportions of coded segments between cases that were categorized as either in the development phase or the evaluation phase. Figure 3.7 presents all the previously shown figures now broken down by the different study goals of developmental and evaluation focused studies in different panels. First, there was very

little difference between the number of changes occurring in either study, developmental studies reported 99% changes made and 1% changes rejected and evaluation studies reported 93% changes made and 7% changes rejected (proportions are proportions of segments coded within each study, between study goals (Figure 3.7 Panel A). Across both study goals the adaptations made were mostly reactive (development 64%, evaluation 70%). However, they largely differed in the number of proactive changes (development 9%, evaluation 25%), and the number of changes that were not planned (development 25%, evaluation 5%), see Figure 3.7 Panel B.

The points in time in which these adaptations took place where before implementation (development 22%, evaluation 20%), during implementation (development 68%, evaluation 56%), during scale-up (development 2%, evaluation 11%), during maintenance and sustainment (development 8%, evaluation 3%), and after the full implementation or after the grant funding is completed (development 0%, evaluation 13%). This breakdown is shown in Figure 3.7 Panel C.

The decision makers varied more across the project goals with researchers making most of the decisions (Figure 3.7 Panel D; development 61%, evaluation 82%), followed by teachers (development 41%, evaluation 33%). There were more district level and school based practitioners involved in developmental work (development 4-41%, evaluation 0-4%), and more student involvement (development 4%, evaluation 0%). Across both projects there was similar involvement by grant funders and program developers (development 6%, evaluation 2%).

Regarding the goals of these adaptations, there is still a larger share of focus on improving fit with recipients (Figure 3.7 Panel E; development 41%, evaluation 32%) and

improving effectiveness outcomes (development 38%, evaluation 36%). However, ensuring feasibility (development 0%, evaluation 28%), increasing retention (development 0%, evaluation 3%), and reducing cost (development 0%, evaluation 3%) were present in evaluation studies as opposed to absent in the developmental studies (development 0%, evaluation 28%). Across both studies the importance of fidelity, sustainability, satisfaction, promoting equity, and increasing engagement were almost equal in proportion. Lastly, the influential factors or reasons for these changes were similarly distributed between the different ecological levels: student (Figure 3.7 Panel F; development 18%, evaluation 33%), teacher (development 44%, evaluation 41%), school (development 39%, evaluation 53%), and social political (development 36%, evaluation 15%). However, the evaluation studies had greater influence arising from the student factors than the sociopolitical factors, and the developmental studies had greater influence from social political factors than student factors.

There are apparent differences across the project goals with patterns demonstrating greater proportions of adaptations taking place in developmental studies (total n= 29, see Table 3.1) that are influenced by practitioners and focused on addressing cultural factors, feasibility, and satisfaction. In contrast, evaluation studies (total n= 38, see Table 3.1) had more goals towards improving feasibility and fidelity of implementation with decisions predominantly coming from the research teams. These differences are not unexpected based on the kinds of grant funding requirements of these studies. A surprisingly large proportion of adaptations took place even when the program or intervention had been previously proven as efficacious, demonstrating that evaluation studies have greater adaptations to better fit the opportunities and constraints at the local level of

implementation. The amount of reactive and unplanned adaptations across both study types also demonstrates that there is plenty of changes that researchers can learn from understanding the experiences of the researchers interviewed to understand what adaptations to anticipate. A key takeaway is that among studies engaged in adaptation, there is mostly reactive change in the design of developmental (68%) and evaluation studies(70%) , but and more proactive change in the implementation of evaluation studies (25%) in contrast to developmental studies (9%).

How might regularities benefit program planning and design?

The multiplicity of regularities across different math programs, research teams, and research goals demonstrates important insights that could be widely beneficial to researchers, grant funders, and school partners. To make sense of these regularities I present them as systemic challenges and opportunities thematically organized at different ecological levels across the different ecological levels of influential factors (Bronfenbrenner, 1995). I also provide guiding principles that summarize advice and lessons learned directly from the researchers interviewed.

Individual / Student Level

Multiple systemic challenges to program design and implementation arose from the variation of individual differences that can be observed in classrooms and across settings. Most challenges included issues with adapting program design to reduce student disengagement, implementation designs that accounted for increased absenteeism, and content delivery that could account for the comorbidity of students with multiple learning difficulties. Researchers often remarked these changes when recounting transitions from one context to another, or when designing for dramatically different geographical contexts

which revealed important variation in the kinds of lives students lived and the challenges they faced. Many of these challenges revealed opportunities to test which designs would best adapt to students who had below grade level content knowledge, required more time to learn, and when explicit language instruction would be most generative to their learning. Most of these adaptations targeted the students but only two cases obtained student feedback on adaptations revealing opportunities to engage in more student feedback as researchers attempt to design programs that fit students in their local context. An important guiding principle from this analysis is to build in feedback systems for teachers and student to determine program and implementation quality. One researcher modeled this principle in the following excerpt:

> "They have an exit interview each year, and although some of our teachers, it's not a true exit interview, because they'll be with us next year as well. But that exit interview is comprehensive. Takes about an hour to do, and probably has at least 20 questions about. You know, "How has this worked for your students?" "What improvements could we provide to the materials that we're giving to you?" You know what other things you see that are absent from [program], and so, like [program] is in its second year, and in year one we did a lot of focus on fractions and word problem solving." (P10)

Practitioner / Teacher Level

Teacher's professional judgement and preferences provided innovative adaptations and challenges to some researchers. Adaptations such as adding and removing elements to programs were made with the understanding that either way teachers will add or remove

things at their will during implementations. With this knowledge researchers opted-to build in guidance for what is best to keep or remove, and often shortened the duration of lessons to better fit the time constraints teacher had. This revealed one important guiding principle across multiple studies, programs had to be between 20-30 minutes to fit in teachers and school interventionists schedules.

A separate guiding principle arose from the data in researchers' descriptions for how they built adaptive professional development to enhance implementation. As shown in the following excerpt:

> "So how we pivoted from the professional development is, we want to try to create an experience for students to learn this content, not periodically throughout the year. but, after all, of the high-stake testing is finished." (P09)

Researchers additionally recommended that to optimize PD implementation they needed to ensure: (1) continuous professional development & support such as through the design of Peer learning communities; (2) Combining coaching & modeling, even when coaching cannot be separately implemented incorporating it into training sessions with large groups was still beneficial; (3) Check-ins with teachers early on during implementation will ensure higher fidelity than rushing to increase engagement in the last few months; (4)
Developing scripts (with specific symbols to prime behaviors like vocal emphasis) for implementers is important to ensure fidelity of implementation no matter what the context of the school might be.

Organizational / School Level & Sociopolitical / National Level

Multiple challenges arose at the organizational level fueled by the broader changes in the sociopolitical context. One research termed this phenomenon as the changing counterfactual as national educational reform was rolled out including No Child Left Behind and the Common Core State Standards. The intervention developer and evaluator explained:

> "We understood as a function of uh, the data that we collected over the course of the center that fractions instruction and intervention really needed to change um and reflect the education reform. And so that's why we um now think that super solvers uh is really the more efficacious and effective um program. And it's kind of interesting because um, you know, it's kind of like what happened with Reading First um in the early two thousand, where um, in kindergarten classrooms, for example, they really began teaching reading for really the first time." (P02)

The policy roll-out and the school level adoption of these policies resulted in the following challenges: added pressure from state test and accountability, competing demands, teacher turnover, and limited funding. The changing counterfactual at these time scales is not obsolete and an important phenomenon to plan and design for. A guiding principle here is the design with systems change in mind. One researcher modeled how to account for organizational challenges such as teacher turnover rates and the changing proportions of students in need in the following excerpt:

"We design our stuff definitely upfront with thinking about the end user when we're not. There is probably going to be an instructional assistant. They probably aren't going to have a math background for sure they may or may not have. Like a teaching certificate or

background, it can really vary. So we sort of like. we try to say ahead of time. who's going to deliver this in schools? And what does that mean for our design of these things? We have to sort of make them be able to be picked up by almost anybody in the school and delivered." (P12)

Another researcher modeled this in the context of low teacher turnover:

"It creates for teachers some kind of support, moral support also like working support…If the school team decide that it will, they will work with you it. It means that the people they are different. It's the people who work together, the people who are understand the need. It's a different case when you just pick some benevolent volunteers from different school, or even to persons from the same school. Maybe 2 persons work together, and they likely to to do a good job. But if they are not understood by the school team. especially in our case, because it's it's a huge difference that we propose It's a huge difference." (P06)

DISCUSSION

This study demonstrates a substantial extent of adaptational change that takes place in research on math programs regardless of whether they are in the developmental phase or the evaluation phase. Adaptations were present in every single case observed regardless of the goals or constraints of a study, highlighting the ongoing cyclical nature of iterating through program development within dynamic contexts. Most cases made adaptations with the goal of enhancing program fit with students and teachers and the goal of improving their outcomes. Although not all the cases explicitly set out to do design based research the iterative improvement, mixed methodologies, and localized design principles that the interviewees described in their research processes aligned closely to the approaches of

design-based research (Anderson & Shattuck, 2012). The current study provides concrete examples of factors influencing common adaptations, in some cases the success of these adaptations is also described which can inform mental models (Lichand, Serdeira, & Rizardi, 2023) of the types of adaptations researchers should anticipate, whether they arise from challenges or opportunities, and whether they be in the theory of change or the implementations models. These key insights are summarized as guiding principles and the modified FRAME-IS in Appendix B, is provided as a concrete tool to get researchers started on documenting their adaptations.

The curse of knowledge (Newton, 1990) plays an important role in the work of developing, improving, implementing, and assessing math programs. Often researchers build a set of knowledge and skills throughout the years of development that are not easily transferrable to researchers wishing to replicate or apply their work. This is apparent in the results whereby seven of the cases have reported the adaptations in published reports and six of cases have not or plan to in the future. Among the cases observed 5 cases have developed dedicated publications detailing specific adaptations made and the impact of these changes, however adaptations that could not be empirically assessed were not described. The adaptations that are not reported across publications include useful observations and challenges face. The differences in reporting and the extent to which adaptations are reported indicate potential areas of improvement.

Transferring the knowledge on what adaptations were productive is complicated by the curse of knowledge in which the minutia of the key components, decisions, and strategies of a successful math program are known it is difficult to understand what it is like for someone to not know these details. Although, researchers create thorough reports

and manuals on these programs an area that is often underreported is the numerous adaptations that were made across different contexts to ensure the program worked well. Such as, in what school structures was it best to provided 1:1 training to teachers and in what contexts were group trainings better? This task is difficult as programs cannot always statistically or qualitatively assess which method works best, instead this information is stored in the experiential knowledge of the previous developers and implementers. One way to address the Curse of Knowledge is to provide concrete examples and stories that can inform developers, implementers, and evaluators about the contextual complexities of math programs (Heath & Heath, 2006). To facilitate this communication of knowledge this paper highlights the important role of quantifying and qualifying the adaptations made across populations and contexts as math programs are developed to reveal and capture important sources of program effect variation.

Limitations

An important limitation in this work is the representation of the researchers involved in developing math programs, although saturation was reached with the 13 participants the recruitment methods precluded very early programs from being included in this work. Therefore, there may be additional challenges that arise for math program developers that have not yet received funding from IES or NSF grants that have not been captured. Additionally, the changes in context across all projects are only representative of changes occurring when a program is applied across contexts within the same countries. Different forms of adaptation may be more prevalent (such as the who gets to make decisions and how the program context varies when applying programs across locations within greater geographical distance from one another. For this reason, the current paper

brings forth the utilitarian importance of using a tool to capture adaptations to honor their value and highlight the importance of change and difference across groups.

Perspectives from graduate students are not captured and sometimes these can include the most on the ground kind of information. However, they may not necessarily reveal changes made but potential changes to make that are not taken up by the researchers. This was not explored further to not violate the trust of interviewees and focus on the changes actually made.

Future Directions

In future work I would like to obtain grant funding to conduct a mixed-methods evaluation of project teams scaling effective math programs using different approaches to understand the moderation of adaptations across different contexts of national importance such as when designing math programs for highly economically constrained school districts and marginalized teacher and student populations such as recent immigrants and impoverished communities. An important addition to future research is the intentional mapping of the development of core ingredients of each math program as adaptations take place. This approach can follow the use of conjecture maps to identify key changes in design and highlight the key adaptations that transform the overall design of a math program (Sandoval, 2014). By integrating the use of the Conceptual Framework for Studying Variation in Program Effects, Treatment Contrasts, and Implementation (Weiss et al, 2014), the Framework for Reporting Adaptations and Modifications to Evidence-Based Implementation Strategies (FRAME-IS; Miller et al., 2021), and the approaches of design based research we can develop more refined theory about what math program designs work for whom, why, and under what conditions.

GENERAL DISCUSSION

Researchers are tasked to develop and test solutions to wicked problems in education, such as social problems with multiple interacting factors and systems which obscure operationalized distinctions of the causes and symptoms of problems, leaving multiple problems that can have multiple solutions (Rittel & Weber, 1973; Buchanan, 1992). The issues of social inequities in human development and education are a wicked problem that the researcher's described in this dissertation have attempted to solve to. All have faced barriers some withstanding from the limited reach of their methodological approaches in the face of numerous challenges. In this dissertation I have demonstrated three separate solutions to common challenges that arise for educational researchers attempting to develop programs, practices, and interventions that can inform policy making and educational practice.

In Study 1, I address the challenges to making accurate forecasts about the medium- to long-term outcomes of successful math interventions are omitted variables bias, measurement over-alignment, and measurement under-alignment. By including multiple pretest measures of student's mathematical competencies, the accuracy of impact measurement and forecasting is greatly improved. Although, controlling for multiple confounding variables is necessary in program evaluation it is not sufficient to also develop accurate forecasts of the longer-run outcomes of educational programs. In addition to removing bias using longitudinal randomized trials another potential solution to this challenge is to combine the use of both conceptually proximal and conceptually distal

measures of student's mathematical competencies. By averaging the estimated impacts of the two measures on long-term math skills researchers can develop stronger tests of their theories of mathematical development. In addition, these calculations can be used in power-calculations to demonstrate to funding agencies the appropriate sample sizes that should be funded to determine what programs work best for math education.

Methodological challenges also arise when well-funded longitudinal projects are implemented. In Study 2, I use the case of the Baby's First Years project, a well implemented multi-site RCT that evaluated the impact of income on various child and maternal health outcomes. Although, the project was well-powered to detect effect sizes of .20 SD there were multiple mixed effects across all the outcomes with some effects being too small to measure with precision. The BFY study is representative of many cases in educational research whereby nonexperimental data is used to determine the factors that are the strongest predictors of educational outcomes to design potential interventions that improve those factors. As observed in this case, the factors selected may not always yield the anticipated results. The strength of the BFY study is that it allowed me to assess the correspondence of experimental and nonexperimental estimates to determine under what conditions nonexperimental data can be modeled to predict the outcomes of child allowance policies and to test theories of how income influences child development.

In doing so, I found little correspondence between individuals experimental income impacts and nonexperimental estimates of income gradients. However, when comparing the average impact of all experimental estimates to the average of all the nonexperimental estimates they had much better correspondence. This is informative for researchers and policy makers who use nonexperimental data to develop potential interventions; rather

than using a single strong association it is recommended to average across multiple qualitatively similar measures to get a better estimate of a potential causal effect. Moreover, another guiding principle for researchers is to be cautious of experimental estimates that are particularly large in magnitude relative to their corresponding non-experimental estimates, because they are estimated with far more error.

In Study 3, I provide a tool that can be useful to better understand cases in which successful small-scale educational programs fail to yield the same benefits when they are scaled-up and tested for effectiveness across different contexts (Means & Harris, 2013; List, Suskind, & Supplee, 2021; Bryk, 2015; Snow, 2015). Currently, there is not a norm for reporting how educational programs adapt to local contextual opportunities and constraints unless the adaptations are intentionally tested in the evaluation or considered to be useful for generalizable knowledge. I posit the use of the FRAME-IS Codebook and Interview Protocol (Miller et al., 2021) to capture the complex and numerous adaptations that arise in the theories of change and the implementation models of researchers. The combination of this tool with the *Conceptual Framework for Studying Variation in Program Effects, Treatment Contrasts, and Implementation* (Weiss et al., 2018) to guide research on variation of program impacts and to determine the sources of this variation for analysis and program improvement. I conducted 13 semi-structured interviews that demonstrated how the adaptation variation informed important guiding principles for the development and evaluation of math programs.

As a result, regardless of the math program goal and the phase of evaluation that it is in, it is important for researchers to consider.

1. Build in feedback systems in the data collection and analytic repertoire of the research team. This facilitates for teachers and student feedback about what works well and what doesn't, in addition this can be a useful measure of implementation quality.
2. When planning and implementing programs they often had to be extended based on new additions teachers needed to make, in addition they needed to be shortened to fit between the 20-30 minute windows that most school based teachers and interventionists have with the students. Rather than ignoring this constraint designing in anticipation of this constraint can save money and time in later implementations.
3. Design adaptive professional development sessions to enhance implementation and bypass common school service structure and time constraints. Researchers that had to adapt these kinds of sessions recommended the following design components: (1) continuous professional development & support such as through the design of peer learning communities; (2) Combine direct training, coaching, and modeling there may not be enough time for each, even when coaching cannot be separately implemented incorporating it into training sessions with large groups was still beneficial; (3) Incorporate early check-ins with teachers during implementation to ensure higher fidelity rather than rushing increase engagement in the last few months; (4) developing scripts for program implementation, schools often have high turnover and diverse personnel so it's important to design without assumptions that the eventual implementer will have sufficient training in mathematics. In addition, including

specific symbols to prime behaviors like vocal emphasis in the script is helpful for implementers and important to ensure fidelity of implementation no matter what the context of the school might be.

4. Design theories of change and implementation models with systems change in mind, such as focusing on how to ensure a program can transform the norms of a department rather than the norms of a single teacher. Many researchers opted for providing treatment at the school level to enhance the authenticity of the program, improve feasibility, and enhance ongoing morale and working support that teachers receive to sustain the program. Otherwise, single teacher implementers may face several barriers if the broader departmental and school culture does not understand the new program.
5. Researchers and funders should prepare for the changing counterfactuals in the educational landscape. Changes such as educational reform and school district policies have the potential to disrupt the effectiveness of a program or its core components. Building in feedback systems as shown in (guiding principles 1) can keep researchers involved in these discussions with practitioners and district leaders. Without preparation for changing counterfactual conditions the research field may be underprepared for selecting the optimal programs for the changing times.
6. Challenges to scaling the use of math software programs were experienced in soliciting commercial partnerships due to a reluctance by some businesses to partner due to concerns this new product would disrupt other programs they had already sold to schools (despite their own excitement for the innovative

capabilities of the program). This example reveals a potential area where funding agencies can find ways to support researchers in commercializing their math programs. Current overreliance on market forces to support innovation have proven insufficient. Furthermore, there are additional needs for better means to protect copyrights, as some publishers and test makers will freely take the content of without providing adequate acknowledgement and compensation. Program funders and researchers should discuss and consider options for commercializing and protecting the intellectual property of math programs, especially in a changing world where math program delivery can take many different modalities (i.e., teachers, computers, books).

This research is important in two ways, it provides a new tool for researchers to use, and it highlights guiding principles that researchers can use as they begin to design and evaluate new math programs. It is already difficult to conduct research in schools, and the difficulty is amplified when the design is not ready to adapt to the differences across contexts. To supplement the assumptions that researchers may already have about conducted research in the field the insights from this work can further prepare research teams for constraints they had not anticipated. Most importantly, researchers can use these guiding principles to better design programs for different communities and further investigate the factors and processes that contribute to the diversity and variation of program implementation and effectiveness.

Future Work

My future work will focus on validating the methods I have recommended in this dissertation to form specific guiding principles that researchers and funding agencies can

use in the future to ensure that they collect data that can be used to inform forecasts of program impacts, collective within-study designs for theory testing, and be transparent about the adaptations that took place so that others may replicate and test these changes.

To validate the use of conceptually proximal and conceptually distal measures I plan to gather data from different randomized controlled trials focusing on the mathematical development of children. I will test the different forms of forecasting to determine if averaging effects across the proximal and distal measures consistently reveals more accurate forecasts than using a single measure alone. By including additional studies, I can also estimate if there are any differences in long-term outcomes based on potential peer-effect studies, such as contrasting forecast accuracy for studies that treated multiple peers in the same classroom as opposed to studies that used few children in the classroom and pulled them out for the intervention. In addition, I will test the hypothesis that omitted variable bias is more difficult to account for in preschool aged children because of the difficulty of giving a comprehensive battery of pretest assessments and more measurement error. I will assess this by comparing the forecast accuracy of studies implementing a wider range of pre-test measures. This work has the potential to provide important guidance for researchers who need to reduce omitted variables bias but are constrained by the capacities and context of assessing very young children.

My second goal is to use the forecasting method to design multi-intervention randomized controlled trials where we use the forecasted long-term effects of two different interventions to see if the delivery of both interventions has multiplicative effects in contrast to the delivery of a single intervention. I plan to design this as a four-arm randomized control trial in the form of a preregistered persistent collaboration (Makel et al

., 2019) with two different research teams that develop and evaluate math programs for children with mathematical difficulties at different developmental stages. For example, let's say intervention A is designed to help students in preschool to develop important mathematical competencies for school entry. In contrast, intervention B is designed to help students in kindergarten with commonly observed math difficulties. I would be interested in coordinating the data collection of these two projects to organize them as one large scale RCT where students are randomly assigned to (1) Intervention A only, (2) Intervention B only, (3) Both Intervention A & B , or (4) no intervention at either time point. Importantly students will be randomized at kindergarten and again at 1st grade to ensure that there is no omitted variables bias influencing their exposure to the intervention. By conducting this study, I plan to advance the field in the practice of building large scale collaborative and mutually reinforcing designs of intervention delivery and data collection to answer bigger questions that what a single research team could be capable of alone.

My third goal is to continue to leverage my experience with large scale collaborative science trials to organize a preregistered adversarial collaboration (Makel et al., 2019) between research teams that wish to target the same problems in math education with different methodological approaches. In contrast to goal 2, I will seek to answer questions about how researchers that approach a problem leveraging co-design as a method to adapt a math program may have different impacts than researchers that approach this from more interventionist perspective therefore seeking to replicate previous impacts with minimal adaptation. There are many challenges to making fair comparisons between teams, to facilitate this goal I will recruit research teams that have clearly designed development and implementation models that they intend to follow to operationalize the key differences

between them. In addition, I plan to identify two to three different measures of research-practitioner quality, implementation quality, and qualitative and quantitative math performance outcomes that can be collected in the same way across both studies to facilitate comparison. Lastly, I will ensure that the multiple teams involved have personal investment in the research questions and genuine interest in the outcomes for mutual benefit. As is required by adversarial collaborations both teams will have a clear understanding of the goals of this research and must have agreed to change their minds or views on the subject matter based on the results.

TABLES

TABLE 3.1

Interviewed Participant Characteristics

				Adaptations			
Participant	**Research Approaches**	**Study Phases Discussed**	**Contact Source**	**Total**	**Developmental Phase**	**Evaluation Phase**	**Where were adaptations reported?**
P01	Partnership	Development & Efficacy	Continuous Improvement	7	7	0	Discussion, Dedicated Publication
P02	Interventionist	Efficacy & Effectiveness	IES-WWC	12	2	10	Introduction summary of changes, developmental process not reported
P03	Interventionist	Efficacy & Effectiveness	IES-WWC	7	1	6	NR
P04	Partnership	Development	Continuous Improvement	2	2	0	NR
P05 (N = 2)	Interventionist & Partnership (DBIR)	Development & Efficacy & Effectiveness	IES-WWC	6	6	0	Dedicated Publication
P06	Partnership	Development & Efficacy	NSF Exploratory	3	3	0	Dedicated Publication
P07	External Evaluator	Efficacy & Effectiveness	IES-WWC	2	0	2	NR
P08	External Evaluator & Partnership	Efficacy & Effectiveness	IES-WWC	1	0	1	Dedicated Publication
P09	Partnership	Development & Efficacy	NSF Exploratory	3	3	0	NR

P10	Interventionist	Development & Efficacy & Evaluation& Replication & Scale-Up	IES-WWC	15	8	7	NR
P11	Interventionist	Efficacy & Effectiveness	IES-WWC	9	0	9	Technical reports
P12	Interventionist	Efficacy & Effectiveness	IES-WWC	6	4	2	Dedicated Publication
P13	Interventionist	Efficacy & Effectiveness	IES-WWC	1	0	1	NR
			Totals	67	29	38	

Note. NR= Adaptations not yet reported in publication

Figures

Figure 3.1

Parent Level Codes for What was Changed in the Math Program

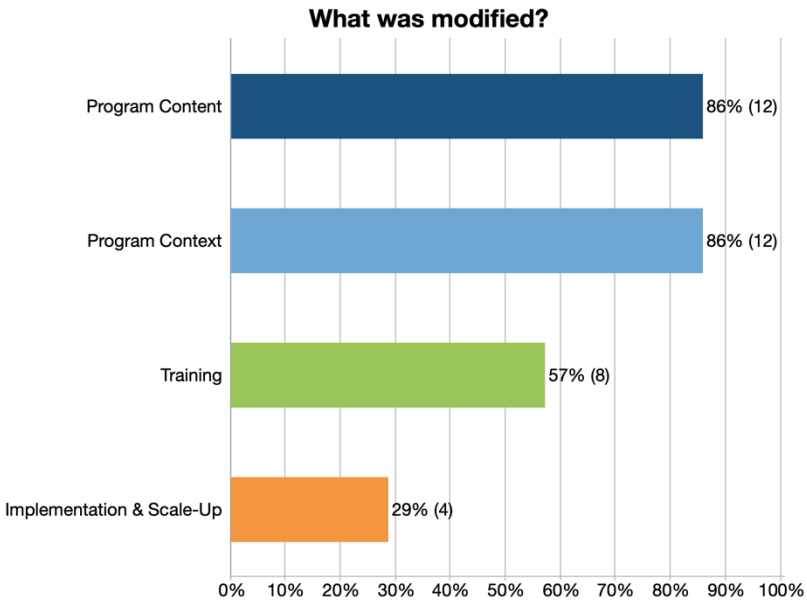


Figure 3.2

Subcodes for Content Level Changes

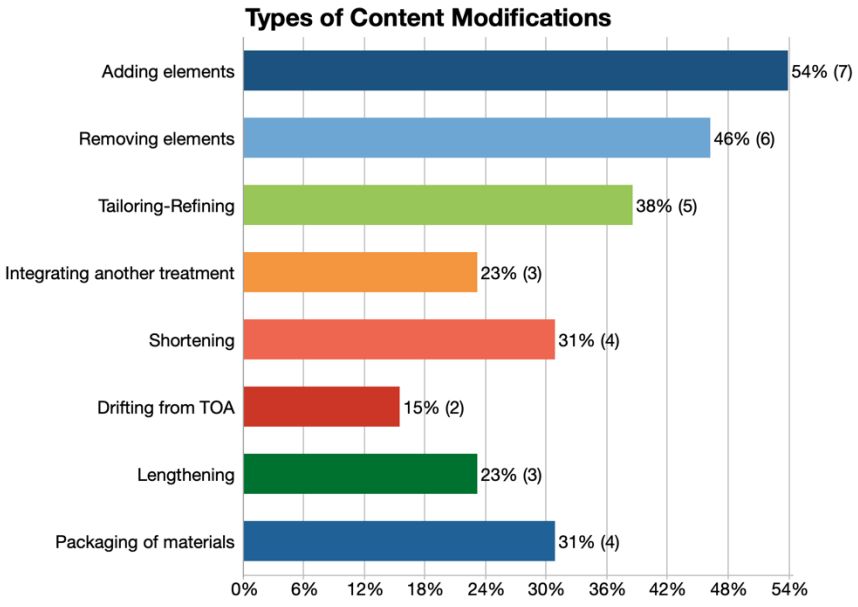


Figure 3.3

Subcodes for Program Context Level Changes

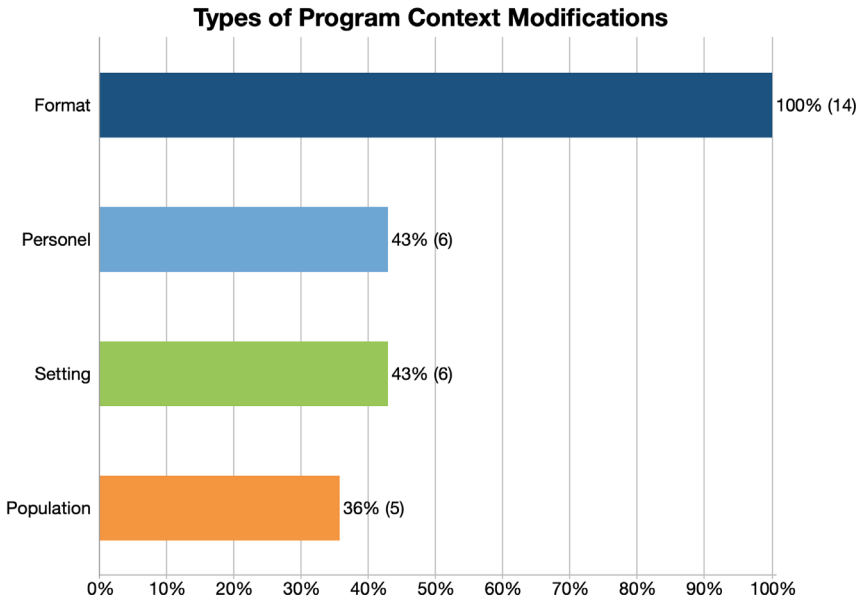


Figure 3.4

Proportion of Changes that maintained Fidelity Original Theory of Change or Core Components

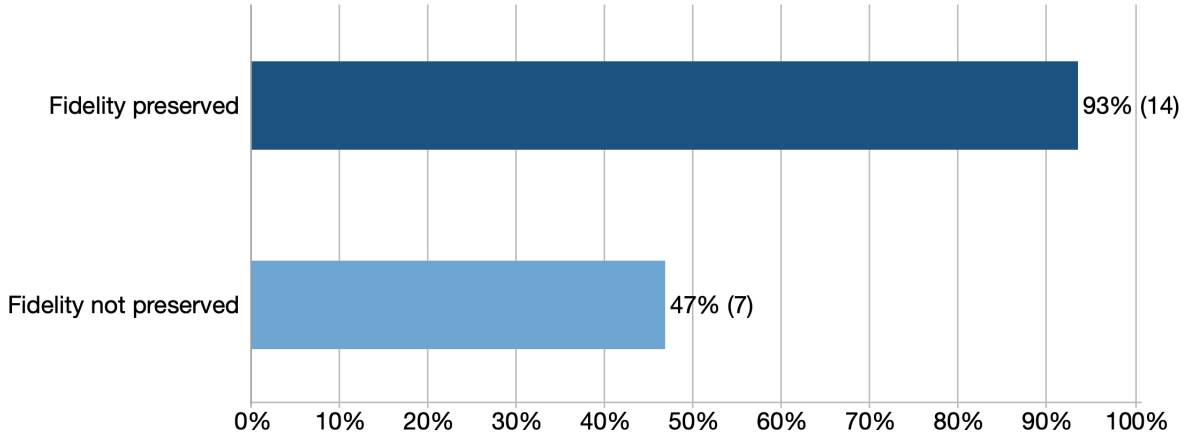


Figure 3.5

Who Participates in the Decision to Modify?

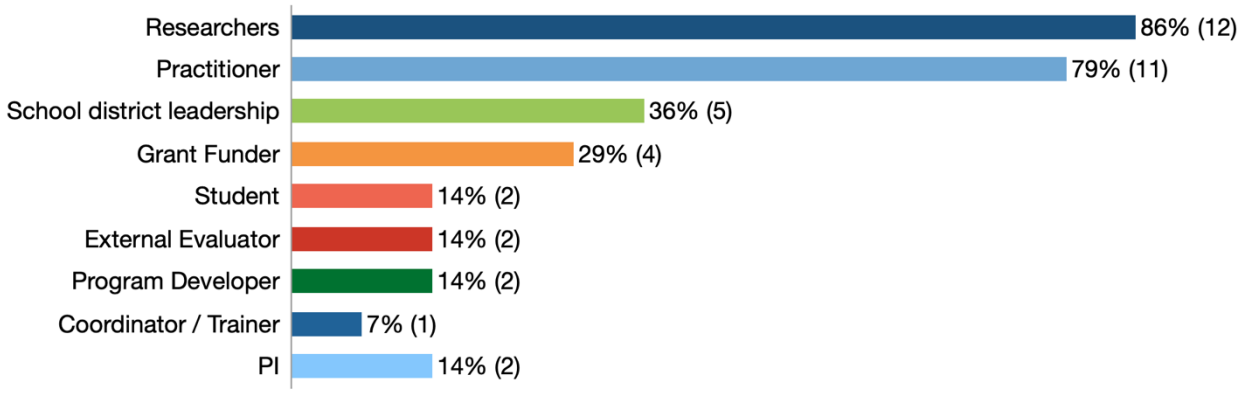
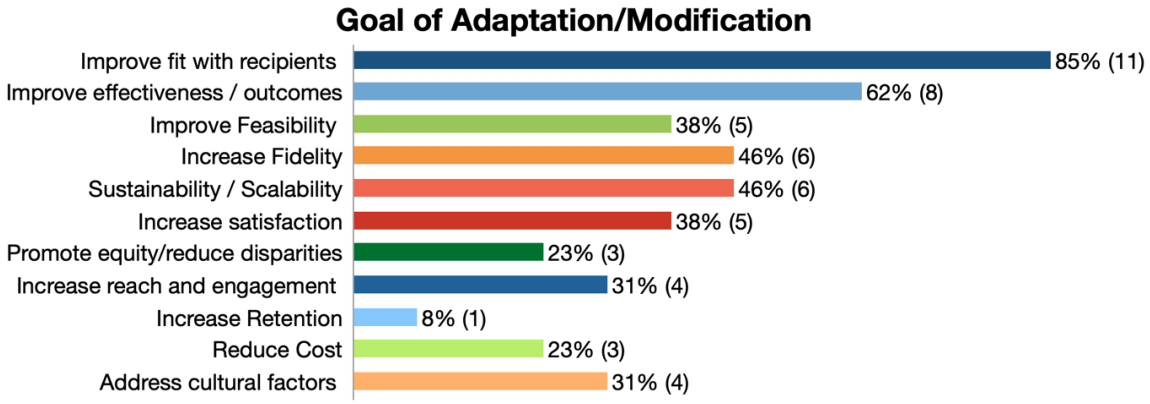


Figure 3.6

Goals and Reasons for Modifications / Adaptations

Panel A



Panel B

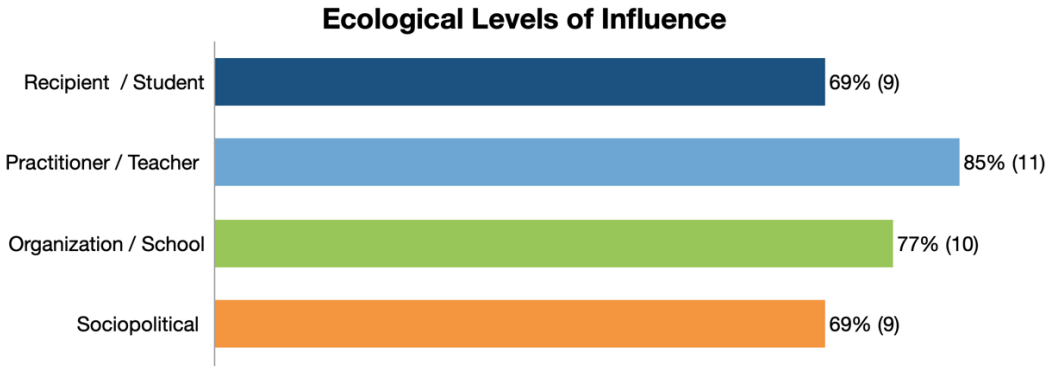
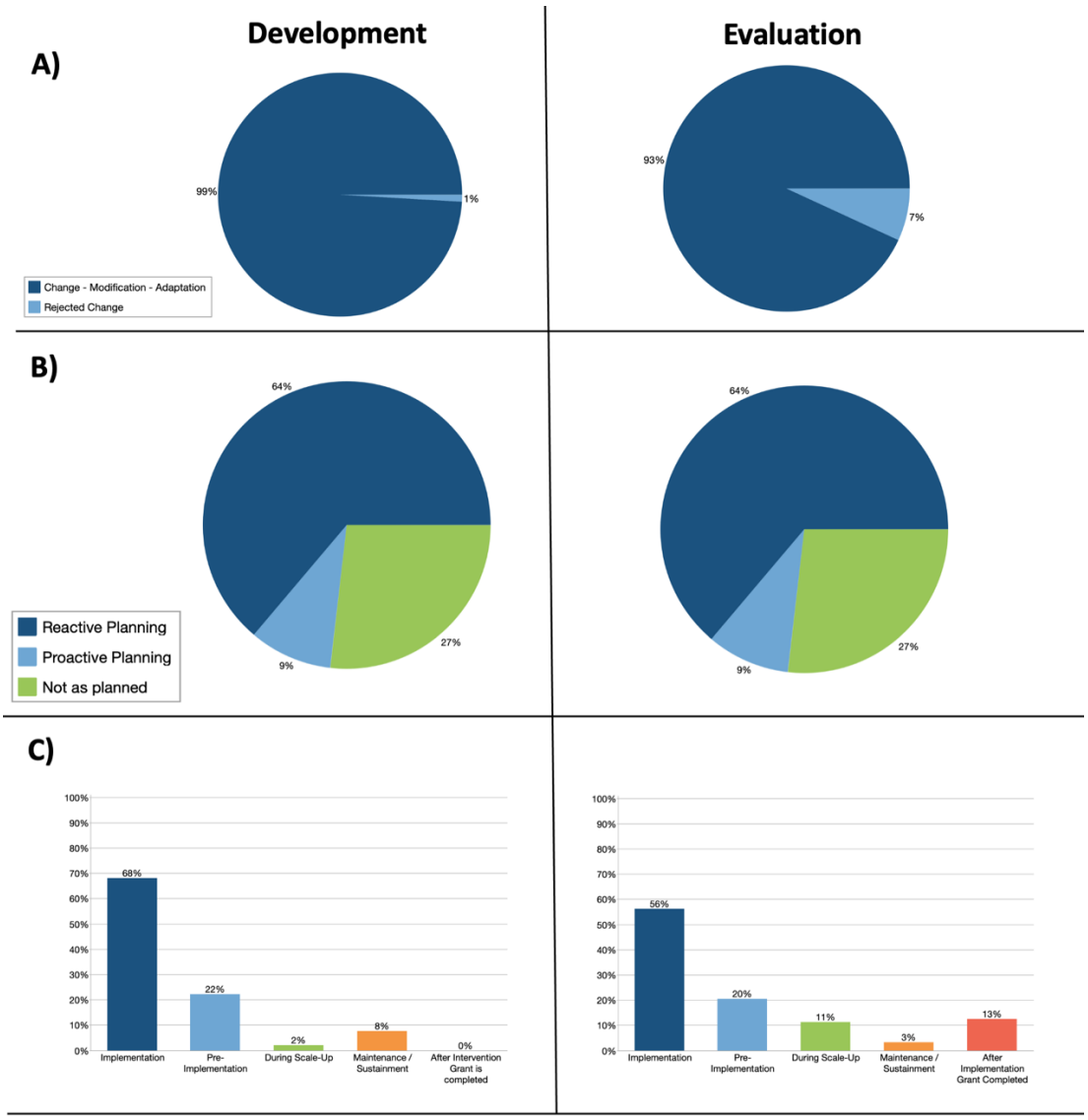


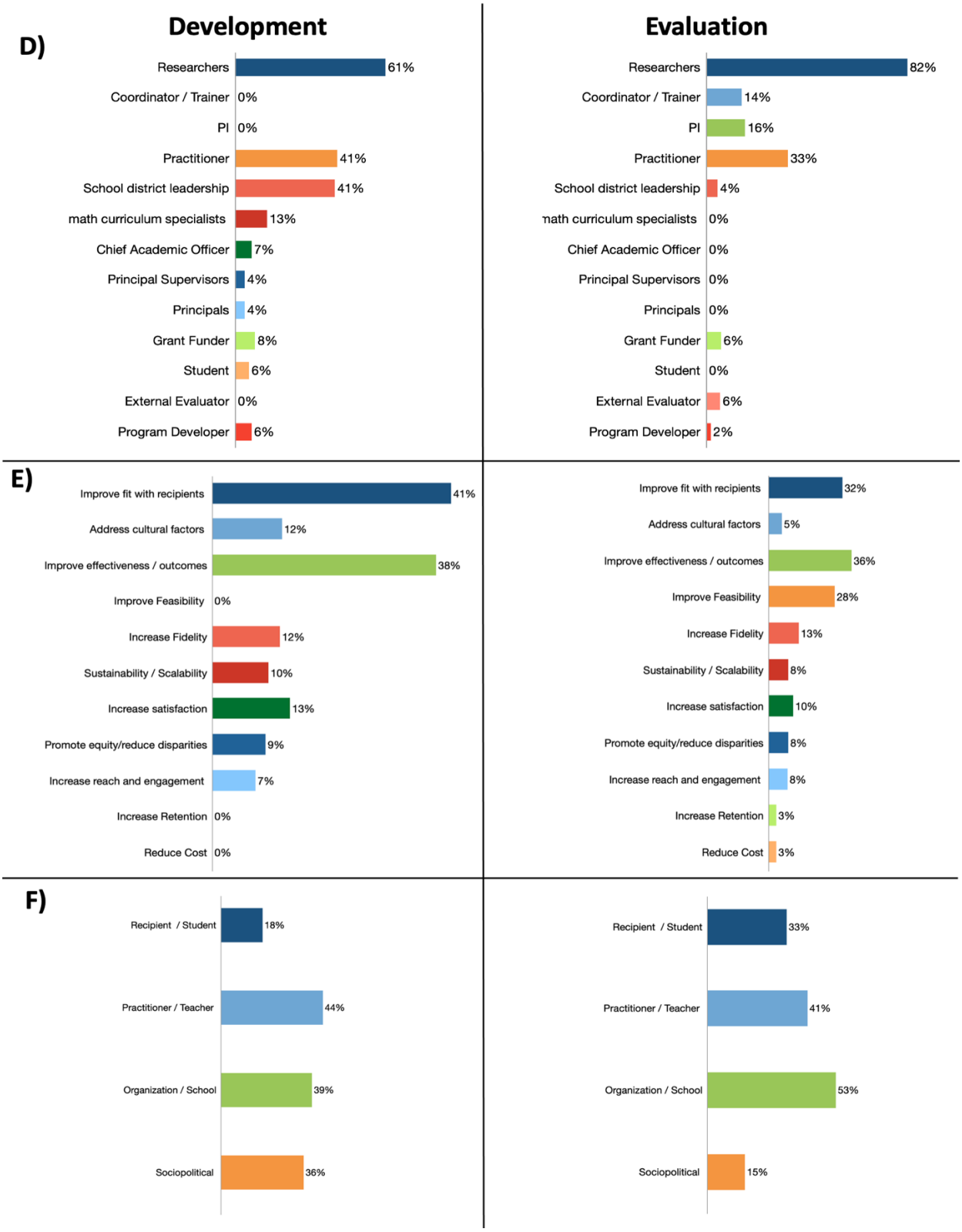
Figure 3.7

Process and Goal Codes Split by Study Goal



(continued on next page*)*

Figure 3.7 *continued*



REFERENCES

Abadie, A., Athey, S., Imbens, G. W., & Wooldridge, J. (2017). When should you adjust standard errors for clustering? (No. w24003). National Bureau of Economic Research.

Akee RKQ, Copeland WE, Keeler G, Angold A, Costello EJ. 2010. Parents' incomes and children's outcomes: a quasi-experiment using transfer payments from casino profits. Am. Econ. J. Appl. Econ. 2(1):86–115

Akee, R. K., Copeland, W. E., Keeler, G., Angold, A., & Costello, E. J. (2010). Parents' incomes and children's outcomes: a quasi-experiment using transfer payments from casino profits. American Economic Journal: Applied Economics, 2(1), 86-115.

Anyon Y, Bender K, Kennedy H, Dechants J. (2018) A Systematic Review of Youth Participatory Action Research (YPAR) in the United States: Methodologies, Youth Outcomes, and Future Directions. Health Educ Behav.45(6):865-878. doi: 10.1177/1090198118769357. Epub 2018 May 11. PMID: 29749267.

Arce-Trigatti, P., Chukhray, I., & Turley, R. N. L. (2018). Research–practice partnerships in education. In Handbook of the sociology of education in the 21st century (pp. 561-579). Springer, Cham.

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.

Autobiographical book recollecting the events that took place in MDRC as social experiments became more common for policy decision making.

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

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

Bandelj, N., Wherry, F.F., Zelizer, V.A. (2017). Money Talks: Explaining How Money Really Works. Princeton, N.J: Princeton University Press.

Bang, M., & Vossoughi, S. (2016). Participatory design research and educational justice: Studying learn- ing and relations within social change making. Cognition and Instruction, 34(3), 173–93.

Bassok, D., Markowitz, A., & Morris, P. (2021). Introducing the Issue. The Future of Children, 31(1), 3-19.

Becker GS. 2009. A Treatise on the Family. Cambridge, MA: Harvard University Press.

Beeber, L. S., & Miles, M. S. (2003). Maternal mental health and parenting in poverty. In M. S. Miles & D. Holditch-Davis (Eds.), Annual review of nursing research, Vol. 21, 2003: Research on child health and pediatric issues (pp. 303–331). Springer Publishing Company.

Benish SG, Quintana S, Wampold BE. (2011) Culturally adapted psychotherapy and the legitimacy of myth: a direct-comparison meta-analysis. J Couns Psychol. 58(3):279-89. doi: 10.1037/a0023626. PMID: 21604860.

Berman, P., & McLaughlin, M. W. (1976, March). Implementation of educational innovation. In The educational forum (Vol. 40, No. 3, pp. 345-370). Taylor & Francis Group.

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.

Bronfenbrenner, U. 1995. Developmental ecology through space and time: A future perspective. In Examining lives in context: Perspectives on the ecology of human development, ed. P. Moen, G.H. Elder Jr., and K. Lüscher, 619–647. Washington, DC: American Psychological Association.

Bryan, C. J., Tipton, E., & Yeager, D. S. (2021). Behavioural science is unlikely to change the world without a heterogeneity revolution. Nature human behaviour, 5(8), 980-989.

Bryk, A. S. (2015). 2014 AERA distinguished lecture: Accelerating how we learn to improve. Educational researcher, 44(9), 467-477.

Bryk, A. S., Gomez, L. M., Grunow, A., & LeMahieu, P. G. (2015). Learning to improve: How America's schools can get better at getting better. Harvard Education Press.

Bryk, Anthony S. "Fidelity of Implementation: Is It the Right Concept?" Carnegie Foundation for the Advancement of Teaching, Carnegie Foundation, 2 Feb. 2017, https://www.carnegiefoundation.org/blog/fidelity-of-implementation-is-it-the-right-concept/.

Buchanan, R. (1992). Wicked problems in design thinking. Design issues, 8(2), 5-21.

Cabassa LJ, Baumann AA. (2013) A two-way street: bridging implementation science and cultural adaptations of mental health treatments. Implement Sci. 8:90. doi: 10.1186/1748-5908-8-90. PMID: 23958445; PMCID: PMC3765289.

Cammarota, J., & Fine, M. (2010). Revolutionizing education: Youth participatory action research in motion. Routledge.

Caspi, A., Houts, R.M., Belsky, D.W., Harrington, H., Hogan, S., Ramrakha, S., … Moffitt, T. (2016). Childhood forecasting of a small segment of the population with large economic burden. Nature Human Behaviour, 1, 0005.

Chambers, D. A., Glasgow, R. E., & Stange, K. C. (2013). The dynamic sustainability framework: addressing the paradox of sustainment amid ongoing change. Implementation Science, 8(1), 1-11.

Chen, E., Martin, A. D., & Matthews, K. A. (2006). Socioeconomic status and health: do gradients differ within childhood and adolescence?. Social science & medicine, 62(9), 2161-2170.

Chetty, R., Friedman, J. N., Hilger, N., Saez, E., Schanzenbach, D. W., & Yagan, D. (2011). How does your kindergarten classroom affect your earnings? Evidence from Project STAR. The Quarterly journal of economics, 126(4), 1593-1660.

Coburn, C. E., Penuel, W. R., & Geil, K. E. (2013). Research-practice partnerships: A strategy for leveraging research for educational improvement in school districts. William T. Grant Foundation.

Coburn, C. E., Touré, J., & Yamashita, M. (2009). Evidence, interpretation, and persuasion: Instructional decision making at the district central office. Teachers College Record, 111(4), 1115-1161.

Connolly, A. J. (2007). KeyMath 3: diagnostic assessment. San Antonio, TX: Pearson.

Cooper, A., MacGregor, S., & Shewchuk, S. (2020). A Developmental Evaluation of Research-Practice-Partnerships and Their Impacts. International Journal of Education Policy and Leadership, 16(9), n9.

Cox, N. (2006) MLOWESS: Stata module for lowess smoothing with multiple predictors, Statistical Software Components S456777, Boston College Department of Economics. https://ideas.repec.org/c/boc/bocode/s456777.htm

Cunningham, S. (2021). In Causal inference the mixtape (pp. 3.1–3.3). essay, Yale University Press. https://mixtape.scunning.com/dag.html#dag.

Dahl, G.B. & Lochner, L. 2012. The impact of family income on child achievement: evidence from the earned income tax credit. American Economic Review, 102(5):1927–56

Datnow, A., Hubbard, L., & Mehan, H. (2002). Extending educational reform. Taylor & Francis.

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. Science, 243(4899), 1668-1674.

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.

Domenech Rodríguez M, Wieling E. Developing culturally appropriate evidence based treatments for interventions with ethnic minority populations. In: Rastogi M, Wieling E, editors. Voices of color: First person accounts of ethnic minority therapists. Thousand Oaks, CA: Sage Publications; 2004. pp. 313–333.

Domenech Rodríguez MM, Bernal G. Bridging the gap between research and practice in a multicultural world. In: Bernal G, Domenech Rodríguez MM, editors. Cultural Adaptations: Tools for Evidence-Based Practice with Diverse Populations. Washington, DC: American Psychological Association Press; 2012. pp. 265–287.

Donovan, M. S. (2013). Generating improvement through research and development in education systems. Science, 340(6130), 317-319.

Donovan, M. S., Snow, C. E., & Daro, P. (2013). The SERP approach to problem-solving research, development, and implementation. National Society for the Study of Education Yearbook, 112, 400–425.

Drahota, A. M. Y., Meza, R. D., Brikho, B., Naaf, M., Estabillo, J. A., Gomez, E. D., ... & Aarons, G. A. (2016). Community-academic partnerships: A systematic review of the state of the literature and recommendations for future research. The Milbank Quarterly, 94(1), 163-214.

Duncan, G. J., & Murnane, R. J. (Eds.). (2011). Whither opportunity? Rising inequality, schools, and children's life chances. Russell Sage Foundation.

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.

Duncan, G. J., Magnuson, K., & Votruba-Drzal, E. (2017). Moving beyond correlations in assessing the consequences of poverty. Annual review of psychology, 68, 413-434.

Duncan, G. J., Magnuson, K., Kalil, A., & Ziol-Guest, K. (2012). The importance of early childhood poverty. Social Indicators Research, 108(1), 87-98.

Duncan, G., & Petersen, E. (2001). The long and short of asking questions about income, wealth, and labor supply. Social Science Research, 30, 248–263.

Duncan, G., Magnuson, K., Murnane, R., & Votruba-Drzal, E. (2019). Income inequality and the well-being of American families. Family Relations, 68(3), 313-325.

Duncan, G., Morris, P., & Rodrigues, C. (2011). Does money matter? Estimating impacts of family income on young children's achievement with data from random-assignment experiments. Developmental Psychology, 47(5), 1263-1279. doi: 10.1037/a0023875

Evans GW. (2004) The environment of childhood poverty. Am Psychol;59(2):77–92

Evans, W., Wolfe, B., & Adler, N. (2012). The SES and health gradient: a brief review of the literature. The biological consequences of socioeconomic inequalities, 1-37.

Farrell, C. C., Davidson, K. L., Repko-Erwin, M. E., Penuel, W. R., Quantz, M., Wong, H., Riedy, R., & Brink, Z. (2018). A descriptive study of the IES Researcher–Practitioner Partnerships in Education Research Program: Final report (Technical Report No. 3). Boulder, CO: National Center for Research in Policy and Practice.

Farrell, C.C., Penuel, W.R., Coburn, C., Daniel, J., & Steup, L. (2021). Research-practice partnerships in education: The state of the field. William T. Grant Foundation.

Ferrer-Wreder, L., Sundell, K., & Mansoory, S. (2012). Tinkering with perfection: Theory development in the intervention cultural adaptation field. In Child & Youth Care Forum (Vol. 41, No. 2, pp. 149-171). Springer US.

Finch, B. K. (2003). Socioeconomic gradients and low birthweight: Empirical and policy considerations. Health Services Research, 38(6 Pt 2), 1819–1841. doi:10.1111/j.1475-6773.2003.00204.x.

Fishman, B. J., Penuel, W. R., Allen, A.-R., Cheng, B. H., & Sabelli, N. (2013). Design-based implementation research: An emerging model for transforming the relationship of research and practice. National Society for the Study of Education Yearbook, 112, 136–156

Fishman, B., & Penuel, W. (2018). Design-based implementation research. International handbook of the learning sciences (pp. 393-400). Routledge.

Fuchs L. S., Geary D.C., et al., (2013) Effects of First-Grade Number Knowledge Tutoring with Contrasting Forms of Practice. Journal of Educational Psychology. 105 (1) 58-77. doi:10.1037/a0030127

Geary 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

Geary, D. C., Bailey, D. H., & Hoard, M. K. (2009). Predicting mathematical achievement and mathematical learning disability with a simple screening tool: The number sets test. Journal of Psychoeducational Assessment, 27(3), 265-279.

Geary, D. C., Hoard, M. K., Byrd-Craven, J., Nugent, L., & Numtee, C. (2007). Cognitive mechanisms underlying achievement deficits in children with mathematical learning disability. Child development, 78(4), 1343-1359.

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

Griner D, Smith TB. Culturally adapted mental health intervention: A meta-analytic review. Psychotherapy (Chic). 2006 Winter;43(4):531-48. doi: 10.1037/0033-3204.43.4.531. PMID: 22122142.

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.

Gutiérrez, K. D. & Jurow, A. S. (2016). Social design experiments: Toward equity by design. Journal of the Learning Sciences, 25(4), 565-598.

Gutiérrez, K. D., Jurow, A. S., & Vakil, S. (2020). Social design-based experiments: A utopian methodology for understanding new possibilities for learning. In N. i. S. Nasir, C. D. Lee, R. Pea, & M. McKinney de Royston (Eds.), Handbook of the cultural founda- tions of learning (pp. 330-347). Routledge.

Hackman, D. A., & Farah, M. J. (2009). Socioeconomic status and the developing brain. Trends in cognitive sciences, 13(2), 65-73.

Halpern-Meekin, Sarah, Edin, Kathryn, Tach, Laura and Sykes, Jennifer. It's Not Like I'm Poor: How Working Families Make Ends Meet in a Post-Welfare World, Berkeley: University of California Press, 2015. https://doi.org/10.1525/9780520959224

Hansson, Sven Ove. (2021) "Science and Pseudo-Science", The Stanford Encyclopedia of Philosophy. Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/fall2021/entries/pseudo-science/>.

Harding, J. F., Morris, P. A., & Hughes, D. (2015). The relationship between maternal education and children's academic outcomes: A theoretical framework. Journal of Marriage and Family, 77(1), 60-76.

Haveman, R., & Wolfe, B. (1995). The determinants of children's attainments: A review of methods and findings. Journal of economic literature, 33(4), 1829-1878.

Heckman, J. J., Stixrud, J., & Urzua, S. (2006). The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior. Journal of Labor Economics, 24(3), 411-482.

Hedges, L. V., & Schauer, J. M. (2019). More than one replication study is needed for unambiguous tests of replication. Journal of Educational and Behavioral Statistics, 44(5), 543-570.

Henrick, E. C., Cobb, P., Penuel, W. R., Jackson, K., & Clark, T. (2017). Assessing Research-Practice Partnerships: Five Dimensions of Effectiveness. William T. Grant Foundation.

Huston, A. C., McLoyd, V. C., & Coll, C. G. (1994). Children and poverty: Issues in contemporary research.

Jäntti, M. (2009) "Mobility in the United States in Comparative Perspective," in Changing Poverty. Retrieved from: https://www.irp.wisc.edu/publications/focus/pdfs/foc262g.pdf

Jenkins, J. M., Watts, T. W., Magnuson, K., Gershoff, E. T., Clements, D. H., Sarama, J., & Duncan, G. J. (2018). Do high-quality kindergarten and first-grade classrooms mitigate preschool fadeout?. Journal of Research on Educational Effectiveness, 11(3), 339-374

Jordan, N. C., & Hanich, L. B. (2000). Mathematical thinking in second-grade children with different forms of LD. Journal of learning disabilities, 33(6), 567-578.

Kalil, A., Mayer, S., & Shah, R. (2020). Impact of the COVID-19 crisis on family dynamics in economically vulnerable households. University of Chicago, Becker Friedman Institute for Economics Working Paper, (2020-139).

Kelly, A. (2004). Design research in education: Yes, but is it methodological?. The journal of the learning sciences, 13(1), 115-128.

Kirshner, B. (2010). Productive tensions in youth participatory action research. Yearbook of the National Society for the Study of Education, 109(1), 238–251.

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

Kling, J. R., Liebman, J. B., & Katz, L. F. (2007). Experimental analysis of neighborhood effects. Econometrica, 75(1), 83-119.

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

LaLonde, R. J. (1986). Evaluating the Econometric Evaluations of Training Programs with Experimental Data. The American Economic Review, 76(4), 604–620. http://www.jstor.org/stable/1806062

Larson, K., & Halfon, N. (2010). Family income gradients in the health and health care access of US children. Maternal and child health journal, 14(3), 332-342.

List, J. A., Suskind, D., & Supplee, L. H. (Eds.). (2021). The Scale-up Effect in Early Childhood and Public Policy: Why Interventions Lose Impact at Scale and what We Can Do about it. Routledge.

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

Magnuson, K., Sexton, H. R., Davis-Kean, P. E., & Huston, A. C. (2009). Increases in maternal education and young children's language skills. Merrill–Palmer Quarterly, 55, 319–350.

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"

Mayer, S. (1997). What money can't buy: The effect of parental income on children's outcomes. Cambridge, MA: Harvard University Press.

Maynard, R., & Murnane, R. (1979). The effects of a negative income tax on school performance: Results of an experiment. Journal of Human Resources, 14(4), 463–476.

Means, B., & Harris, C. J. (2013). Towards an Evidence Framework for Design-Based Implementation Research. Yearbook of the National Society for the Study of Education, 112(2), 350-371.

Michalopoulos, C., Bloom, H. S., & Hill, C. J. (2004). Can propensity-score methods match the findings from a random assignment evaluation of mandatory welfare-to-work programs?. Review of Economics and Statistics, 86(1), 156-179.

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

Miller, C., Barnett, M.L., Baumann, A.A. et al. The FRAME-IS: a framework for documenting modifications to implementation strategies in healthcare. Implementation Sci 16, 36 (2021). https://doi.org/10.1186/s13012-021-01105-3

Milligan, K., & Stabile, M. (2011). Do child tax benefits affect the well-being of children? Evidence from Canadian child benefit expansions. American Economic Journal: Economic Policy, 3(3), 175-205.

Minkler, M., Wallerstein, N., & Wilson, N. (2008). Improving health through community organization and community building. K. Glanz, BK Rimer, K. Viswanath, Health behavior and health education. theory, research, and practice, 37-58.

Mistry RS, Vandewater EA, Huston AC, McLoyd VC. (2002) Economic well-being and children's social adjustment: the role of family process in an ethnically diverse low-income sample. Child Dev.73(3):935–951

Morris, P., Duncan, G. J., & Clark-Kauffman, E. (2005). Child well-being in an era of welfare reform: the sensitivity of transitions in development to policy change. Developmental psychology, 41(6), 919.

National Academies of Sciences, Engineering, and Medicine. 2017. Communities in Action: Pathways to Health Equity. Washington, DC: The National Academies Press. https://doi.org/10.17226/24624.

National Academies of Sciences, Engineering, and Medicine., et al. 2019. "Consequences of Child Poverty." A Roadmap to Reducing Child Poverty, The National Academies Press., Washington, DC, 2019, pp. 67–91.

Noble KG, Magnuson K, Gennetian LA, et al. Baby's First Years: Design of a Randomized Controlled Trial of Poverty Reduction in the United States. Pediatrics. 2021;148(4):e2020049702

Noble, K. G., Houston, S. M., Brito, N. H., Bartsch, H., Kan, E., Kuperman, J. M., ... & Sowell, E. R. (2015). Family income, parental education and brain structure in children and adolescents. Nature neuroscience, 18(5), 773-778.

Parr, J. M., & Timperley, H. S. (2015). Exemplifying a continuum of collaborative engagement: Raising literacy achievement of at-risk students in New Zealand. Journal of Education for Students placed at Risk (JESPAR), 20(1-2), 29-41.

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

Penuel, W. R., & Gallagher, D. J. (2017). Creating Research Practice Partnerships in Education. Harvard Education Press. 8 Story Street First Floor, Cambridge, MA 02138.

Penuel, W. R., & Hill, H. C. (2019). Building a Knowledge Base on Research-Practice Partnerships: Introduction to the Special Topic Collection. AERA Open. https://doi.org/10.1177/2332858419891950

Penuel, W. R., Roschelle, J., & Shechtman, N. (2007). Designing formative assessment software with teachers: An analysis of the co-design process. Research and Practice in Technology Enhanced Learning, 2(1), 51–74.

Petterson, S. M., & Albers, A. B. (2001). Effects of poverty and maternal depression on early child development. Child development, 72(6), 1794-1813.

"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"

Reardon, S. F., & Bischoff, K. (2011). Income inequality and income segregation. American journal of sociology, 116(4), 1092-1153.

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.

Ridley, M., Rao, G., Schilbach, F., & Patel, V. (2020). Poverty, depression, and anxiety: Causal evidence and mechanisms. Science, 370(6522).

Rittel H. W. J., Webber M. M. (1973). Dilemmas in a general theory of planning. Policy Sciences, 4, 155-169.

Robinson, D. H., Levin, J. R., Schraw, G., Patall, E. A., & Hunt, E. B. (2013). On going (way) beyond one's data: A proposal to restrict recommendations for practice in primary educational research journals.

Rueda, R.M. & Conejero, A. (2020) Effects of Poverty on Early Neurocognitive Development. In Stevens, C., Pakulak, E., Segretin, M.s., & Lipina, S.J. (Eds.) Neuroscientific perspectives on poverty (pp.63-65). ISBN: 978-987-86-6736-2

Sabol, T. J., McCoy, D., Gonzalez, K., Miratrix, L., Hedges, L., Spybrook, J. K., & Weiland, C. (2022). Exploring treatment impact heterogeneity across sites: Challenges and opportunities for early childhood researchers. Early Childhood Research Quarterly, 58, 14-26.

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.

Schneider, Mark. (2018) “Developing an Evidence Base for Researcher-Practitioner Partnerships.” IES Directors Blog, IES, 30 July 2018, https://ies.ed.gov/blogs/director/post/rpps.

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.

Sirin, S. R. (2005). Socioeconomic status and academic achievement: A meta-analytic review of research. Review of Educational Research, 75, 417–453.

Smith, L. T. (2021). Decolonizing methodologies: Research and indigenous peoples. Zed Books Ltd.

Soman, D., & Ahn, H.-K. (2011). Mental accounting and individual welfare. In Perspectives on framing (pp. 65–92). Psychology Press.

Stahmer, A. C., Dababnah, S., & Rieth, S. R. (2019). Considerations in implementing evidence-based early autism spectrum disorder interventions in community settings. Pediatric medicine (Hong Kong, China), 2.

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.

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.

Steiner, P. M., & Wong, V. C. (2018). Assessing correspondence between experimental and nonexperimental estimates in within-study comparisons. Evaluation review, 42(2), 214-247.

Steiner, P. M., Wong, V. C., & Anglin, K. (2019). A causal replication framework for designing and assessing replication efforts. Zeitschrift für Psychologie, 227(4), 280-292. http://dx.doi.org/10.1027/2151-2604/a000385

Stirman SW, Gamarra JM, Bartlett BA, Calloway A, Gutner CA. Empirical examinations of modifications and adaptations to evidence-based psychotherapies: methodologies, impact, and future directions. Clin Psychol Sci Pract. 2017;24(4):396–420.

Sykes, J., Križ, K., Edin, K., & Halpern-Meekin, S. (2015). Dignity and dreams: What the Earned Income Tax Credit (EITC) means to low-income families. American Sociological Review, 80(2), 243-267.

Tseng V. (2017). The next big leap for research-practice partnerships: Building and testing theories to improve research use. New York, NY: William T. Grant Foundation.

Tseng, V. (2012). Partnerships: Shifting the dynamics between research and practice. New York, NY: William T. Grant Foundation.

U.S Census Bureau, American Community Survey 2011-2015, Current Population Survey 1960 to 2017 Annual Social and Economic Supplements. Retrieved from: https://www.census.gov/library/visualizations/2017/demo/poverty_measure-how.html

Waller, N. G., & Meehl, P. E. (2002). Risky tests, verisimilitude, and path analysis.

Wallerstein N (2006). What is the evidence on effectiveness of empowerment to improve health? Copenhagen, WHO Regional Office for Europe (Health Evidence Network

report;http://www.euro.who.int/Document/E88086.pdf, accessed 03 November 2021).

Wallerstein, N. (1999). Power between evaluator and community: research relationships within New Mexico's healthier communities. Social Science & Medicine, 49(1), 39-53.

Wallerstein, N. B., & Duran, B. (2006). Using community-based participatory research to address health disparities. Health promotion practice, 7(3), 312-323.

Wandersman, A.H., & Florin, P. (2003). Community interventions and effective prevention. The American psychologist, 58 6-7, 441-448.

Watts, T. W. (2020). Academic achievement and economic attainment: Reexamining associations between test scores and long-run earnings. AERA Open, 6(2), 2332858420928985.

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.

Weidmann, B., & Miratrix, L. (2021). Lurking inferential monsters? Quantifying selection bias in evaluations of school programs. Journal of Policy Analysis and Management, 40(3), 964-986.

Weiss, M. J., Bloom, H. S., & Brock, T. (2014). A conceptual framework for studying the sources of variation in program effects. Journal of Policy Analysis and Management, 33(3), 778-808.

Welsh, R. O. (2021). Assessing the quality of education research through its relevance to practice: An integrative review of research-practice partnerships. Review of Research in Education, 45(1), 170-194.

Wentworth, L., Mazzeo, C., & Connolly, F. (2017). Research practice partnerships: A strategy for promoting evidence-based decision-making in education. Educational Research, 59(2), 241-255.

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

Whyte, W. F., Greenwood, D. J., & Lazes, P. (1991). Participatory action research: Through practice to science in social research. Participatory action research, 32(5), 19-55.

Wilkinson, G. S. (1993). WRAT-3: Wide range achievement test administration manual. Wide Range, Incorporated.

Woodcock R, McGrew KS, Mather N. Woodcock-Johnson tests of achievement. 3rd ed. Itasca, IL: Riverside; 2001.

Wright, S. (1921). Correlation and Causation. Journal of Agricultural Research, 20, 557-585.

Xu, D., Zhang, Q., & Zhou, X. (2020). The impact of low-ability peers on cognitive and non-cognitive outcomes: random assignment evidence on the effects and operating channels. Journal of Human Resources.

"York, A., Valladares, S., Valladares, M.R., Snyder, J., & Garcia, M. (2020).

Community Research Collaboratives. Boulder, CO: National Education Policy Center. Retrieved [date]

from http://nepc.colorado.edu/publication/crc."

Zavala, M. (2013). What do we mean by decolonizing research strategies? Lessons from decolonizing, Indigenous research projects in New Zealand and Latin America. Decolonization: Indigeneity, Education & Society, 2(1), 55–71.

Zavala, M. (2016). Design, participation, and social change: What design in grassroots spaces can teach learning scientists. Cognition & Instruction, 34(3).

Zayas, L. H., Bellamy, J. L., & Proctor, E. K. (2012). Considering the multiple service contexts in cultural adaptations of evidence-based practice. In R. C. Brownson, G. A. Colditz, & E.K. Proctor (Eds.), Dissemination and implementation research in health: Translating science to practice. Oxford: University Press.

Ziol-Guest, K. M., Duncan, G. J., Kalil, A., & Boyce, W. T. (2012). Early childhood poverty, immune-mediated disease processes, and adult productivity. Proceedings of the National Academy of Sciences, 109(Supplement 2), 17289-17293.

APPENDIX A

Interview Protocol

Hello, [name] thank you so much for meeting with me today. I am currently a doctoral candidate, and I am interested in talking with you about [your work on / Project] called [name]. For my dissertation, I am interested in mapping the process of the changes that take place when research teams develop and implement new math educational [C: programs | P: interventions] .

I've planned this meeting out as a semi-structured interview because I really want this to be a conversation between us to develop understanding of the complexity that really goes into developing these programs and ensuring their success. Since we will probably get into a lot of details, may I have your permission to record this session? If you wish to stop recording anytime, please let me know and I will pause or end the recording.

For the following questions, I'd like for you to consider the version of the [C: programs | P: interventions] from [study name – year] .

[if possible, send a link]

In your own words could you tell me your story in doing this work, for example what brought you to doing this work?

- [Optional] Do you refer to [study name] as an intervention, a practice, or a program?

- [Optional] If you made any changes to the design of [the work] or the implementation of [the work] do you refer to those as adaptations, modifications, refinements, or something else?

§ *As they talk, note the language they use to connect to or align with their work below:*

Get them into a conversation about incubation activity / doing the work

Q1: Who did you work on this project with?

Q2: Did it work out as you planned?

Q3: Let's say this project is working and you want to take what you've learned to scale to a new setting that could benefit as well? How would you imagine this would be made to fit in a new place?

§ [Alternative]Do you remember making any changes the design or delivery, modifications, or refinements to the [C: program | P: intervention]?

IF YES: Q1.1: Can you share more details about how the [content/ evaluation/ training / context] changed?

(Continue to Q2)

IF NO: Mention example from their literature to guide conversation. If that doesn't work use alternative approach, à Could you tell me about your program development and implementation timeline?

Then ask the following questions about important key moments in the development.

(Alternative questions shown in red)

Modification Loop

Questions below are specific to each [adaptation/modification/refinement]. Iterate through questions until reaching saturation.

Q2: Why was this [adaptation/modifications/refinement] made?

Q2.1: If not mentioned- what was the goal of this change?

(skip)

Q3: I'm wondering if there were any moments that you felt the work may be straying away from incorporating the active ingredients in your theory of change? For example, was your implementation consistent with maintaining the [P: fidelity | C: Integrity] of the original [P: core components | C: design conjectures]?

Q3.a: I'm wondering if throughout your work you ever felt like you were straying away from the key ingredients that were theorized to make your study work? For example, was fidelity

maintained across your intervention implementation. If not what deviations from fidelity occurred?

Q4: Who or What influenced the decision to make the change(s)?

§ Examples: national mandates, available materials, administrators, teachers, or students?

Q4.1: Did anyone else participate in the final decision to make this change?

Q4.2: What proportion of the decision was made by researchers or other community stakeholders? For example, did the researchers get more say in the decision, did the teachers, or was there always a specific group for this with like 2 researchers and 4 teachers? Anything like that?

Q4.a.1: Who participated in the decision making around the development and implementation of this intervention?

Q4.a.2: What proportion of the decision was made by researchers or other community stakeholders? For example, did the researchers get more say in the decision, did the teachers, or was there always a specific group for this with like 2 researchers and 4 teachers? Anything like that?

Q5: Did the decisions to make changes occur more in development, implementation, or as you scaled up to new contexts?

Q5.1. Did you have a structure set up for collective decisions making? Such as specific meetings, a polling strategy, or even email threads?

Q5.1: If not yet stated, were the changes made planned or unplanned? Such that the decision made were in reaction to a specific situation?

Q5.a: Did you make any changes to the intervention at different phases of implementation? As it was scaled-up? Or to sustain it over time without your active involvement?

Q5.a.1: Did you have a structure set up for collective decision making? Such as specific meetings, a polling strategy, or even email threads?

Q6: Did this adaptation/modification work or not work for your intended goals? Or were any conjectures non-productive?

If something didn't work, ask:

Q6.1: Was there any way you could have known sooner that it didn't work?

IF YES:

Q6.1.2: How would this experience change what you would do in the future?

[continue to Q6. still ask Q's below]

IF NO:

Q6.2: How did you determine that adaptations did or did not work? What sources of evidence did you look for or gather?

Q6. 3: Looking back at this now, do you wish you would have measured anything else at this time or thereafter to see if the changes you made did work for the intended purpose? [try to give an example]

Q6.a.1: Did the work overall turn out as you intended? OR was there anything that occurred that was no productive or in line with your theory of change?

Q7: Were there any additional challenges in making these changes that we haven't spoken about? For example: did you need more time to do it or did you wish you had a different process?

Q7.a.1: Were there any additional challenges in conducting this work that we haven't spoken about? Could you share a story with me about some of these challenges and how they affected you or your work?

If there is time, ask modification loop questions about additional modifications that they mention that did or did not work.

General Intervention Questions

The following questions are asked about the specific project/program/intervention under investigation, these questions are for all interviewees unless otherwise specified in the question sequence.

Q7: Throughout your work did you [theorizing or conjectures] about what would work, change?

IF NO:

Q.7.1: Did you change your choice of tools or measures at any point because you weren't arriving at the desired outcomes?

Q8: One thing I am interested in is how researchers describe [adaptations/modifications/refinements]. Because authors have limited space in their papers, some of these descriptions are understandably omitted when studies are written up as [blogs/ memos/ internal reports/ or as scientific journal articles]. Have you added descriptions of [adaptations/modifications/refinements] to any kinds of publications/memos/ blogs/ meeting notes or even emails?

Q8.1: Were these [adaptations/modifications/refinements] written about anywhere or where they may be circulated internally?

IF YES: Would it be possible to access these descriptions? (If so, coordinate how to access these descriptions as artefacts)

Q9: Have you worked on other math interventions/programs before?

Q9.1: Did any of those projects make more adaptations than the work we talked about today?

Q9.2: Were your previous projects done in partnership with any other stakeholders in the community?

Q9.3: Was the research process that we talked about today always your research process in previous projects or has this evolved over time?

IF THEY HAVE DONE BOTH IN-PARTNERSHIP AND OUT-OF-PARTNERSHIP WORK:

Q9.4: What were the differences between your work on math programs that were done in a partnership without a partnership?

Q9.5: (IF GOOD RAPPORT: did you find one approach to be more effective than another? "effective" as they frame it)

Q9.6: How were changes made to these programs different when you worked within partnerships?

Q10.1: Will there be a future study related to this work, and if so, what kinds of changes would you make in the future?

Q10.2: Will you continue to use the same approach or method [as in working in a partnership or not working in a partnership] to do this kind of work?

Thank you so much for your time, we have reached the end of the questions. If possible, I would like to reach back out to you to make sure that what I write about our interview is represented correctly.

Q11: May I send you a written sample to make sure that I am representing you and your work appropriately?

Q12: Would you like me to use a pseudo name for you and your intervention?

IF pseudonym: what would your *preferred* pseudo-name be? For you and your work?

APPENDIX B

The following framework was directly copied with authorization, from the Framework For Reporting Adaptations and Modifications-Expanded Codebook (Stirman et al., 2013; Wiltsey-Stirman et al., 2019; Miller et al., 2021). The original manual is freely available, with permission from the authors I replicated this manual to demonstrate its usage and modifications as shown below (retrieved from chrome extension: //efaidnbmnnnibpcajpcglclefindmkaj/https://med.stanford.edu/content/dam/sm/fastlab/documents/FRAMECodebook2010.30.19.pdf). Modifications made to this coding manual to fit the educational research context were tracked in red font.

When Did the Modification Occur?

1. **Pre-implementation/ planning/pilot** - prior to the formal beginning of the planned implementation
2. **Implementation**
3. **Scale up** - efforts to scale up/spread the intervention beyond the initial implementation site (e.g., to other regions, communities, or organizations)
4. **Maintenance/Sustainment** - beginning after initial supports or funding are withdrawn, rousing the "2 years after implementation" rule of thumb.

Were adaptations planned?

1. **Proactive -** a process of planned adaptation ideally as early as possible in the implementation process. Occurs through a planning process that identifies ways to maximize fit and implementation success while minimizing disruption of the intervention.
2. **Reactive** - less systematic, occur during program implementation, often due to unanticipated obstacles. In an impromptu manner, in reaction to constraints or challenges that are encountered; may or may not be aligned with the elements of the intervention that make it effective. Note that iteration can accommodate unanticipated challenges, e.g., during the "Act" portion of a "Plan-do-study-act"

cycle, an adaptation would not be considered reactive, because it was determined through a systematic process rather than through improvisation.

WHO made the decision to modify?

1. Team – healthcare team, organizational unit
2. Individual practitioner/ facilitator — the individual who delivers the intervention.

   Included teachers or interventionists in schools in this code
3. Non-program staff — e.g., front desk staff schedule fewer sessions than indicated in protocol, contractors.
4. Administration — Leadership within the organization

   Renamed this as School District Leadership with additional subcodes for principal supervisors, principals, chief academic officers, superintendents, and math curriculum specialists
5. Program developer/ purveyor — intervention/treatment developer or expert
6. Researcher — researcher or team that leads research effort.

   Primary investigator of the research grant discussed.
7. Treatment/Intervention team — treatment team (often smaller than team listed above, which may comprise all providers within an organizational unit). This focuses on the smaller team that delivers care.

   In my coding I treated the team as the "research team" that worked together to design and/or implement the program, if an individual in the team was not noted as having made the primary decision then this code was used.
8. Community members - individual stakeholders in the community who may or may not ultimately receive the intervention.
9. Recipients - individuals with the identified problem or risk factor who are the planned recipients of the intervention.

   Students receiving the intervention, even if the intervention targeted the teachers, the students are the primary unit of analysis
10. Coalition of stakeholders — implementation team or advisory board that includes stakeholders from multiple stakeholders.
11. Unknown/unspecified
12. Optional: Include a specifier to indicate who made the ultimate decision

WHAT is modified?

1. Content - Modifications made to content itself, or that impact how aspects of the treatment are delivered.
2. Contextual - Modifications made to the way the overall treatment is delivered.
3. Training and Evaluation - Modifications made to the way that staff are trained in or how the intervention is evaluated.

   Separated training and evaluation as training in the form of Professional Development was often referred to as a key active ingredient in many research projects.

4. Implementation and scale-up activities - Modifications to the strategies used to implement or spread the intervention.

At what LEVEL OF DELIVERY (for whom/what is the modification made?)

Individual-individual patient or recipient: use this code if the clinician states that they modify the EBP for a particular patient (e.g. simplifying language if a patient has cognitive issues or if language barriers exist; cultural modifications for an individual consumer)

1. Target Intervention Group (e.g., all individuals with the problem that is being targeted)
2. Cohort/individuals that share a particular characteristic (e.g., individuals who do not speak the language in which the intervention was originally developed), all individuals with the target problem plus a specific comorbidity, individuals with lower levels of literacy, new mothers with the target problem)
3. Individual practitioner — an individual makes the adaptation/modification for all individuals with whom they work.
4. Clinic/unit level — an entire unit or clinic makes a modification (e.g., limiting the number of meetings/sessions, changing the format of an intervention)
5. Organization — the full organization makes the modification/adaptation.
6. Network System/Community (e.g., VA healthcare system, County or Community that is implementing)

Contextual modifications are made to which of the following?

1. **Format:** use this rating if changes are made to the format of treatment delivery (e.g. a treatment originally designed to be used one-on-one that is now delivered in a group format or via technology)

2. **Setting**: use this rating if the treatment is being delivered in a different setting (e.g. a treatment originally designed to be used in a mental health setting that is now delivered in primary care)
3. **Personnel:** use this rating if the treatment is being delivered by different personnel (e.g. A treatment originally designed to be administered by a psychologist that is now delivered by a psychiatric nurse or clergy)
4. **Population:** use this rating if the treatment that was SPECIFICALLY DEVELOPED to target population is being delivered to a different population than originally intended (e.g., if an intervention developed for adults is now being delivered to older adults or teens; or if an intervention for borderline personality disorder is being delivered to individuals with PTSD).

Note the following examples for clarification:

- Delivering a treatment to a Hispanic/Latino individual that wasn't originally designed with a particular ethnic group in mind: Context, population
- Delivering a treatment to a Hispanic/Latino population that was originally specifically designed for African Americans: context, population.
- Delivering a treatment that wasn't originally designed with a particular ethnic grouping mind, but modifying it to accommodate cultural or language differences: content, population, at the group level of delivery (content may also be addressed though tailoring, see below).
- Delivering a treatment to a Hispanic/Latino population that was originally designed for African Americans, and also modifying the treatment itself to accommodate cultural or language differences: context, population, delivered at the group level of delivery, AND content, group (tailoring, see below). Note—if a context-level modification is made, it is also possible that a content-level modification was also made, but that's not always the case. So sometimes 2 modifications would be made—such as:
- Delivering a treatment in Spanish to some (but not all) Hispanic/Latino clients/patients that was originally specifically designed for African Americans: 1) context, population, cohort (individuals that share the same characteristics) 2) Content, tailoring (language)]; if other adaptations are made to address cultural differences, the goal of improving fit—to address culture would apply.

What is the NATURE of the content modification? (Examples can be changed to fit the intervention being evaluated)

1. **Tailoring/tweaking/refining:** the clinician describes a change to the intervention that leaves all of the major intervention principles and techniques intact (e.g.

modifying language, creating somewhat different versions of handouts or homework assignments, cultural adaptations) If you would like to specify that the tailoring was a cultural adaptation, a separate "1C" code can be used to differentiate it from other forms of tailoring.

2. **Integrating intervention into another framework:** the clinician indicates, or it is apparent, that another treatment approach is the starting point, but elements of the intervention are brought into the treatment (e.g. selecting particular intervention elements or modules to use in the context of another treatment
3. **Integrating another treatment into EBP:** the clinician indicates, or it is apparent, that the intervention is the starting point, but that they are also using aspects of different therapeutic approaches or EBP's in their treatment (e.g. integrating an empty chair exercise into a standard "CBT for Depression" treatment protocol). To use this code for interview data, the strategy or treatment should be specifically named, and should not be the use of general therapeutic skills (e.g., validation, listening would not be used, but if someone says, "I integrate a more client-centered approach into the EBP", this code could be assigned).Integration of Motivational Interviewing (MI) into a protocol that does not specify MI principles is another common example.
4. **Removing/skipping core modules or components:** the clinician indicates that their baseline or standard treatment is based on the EBP, but notes that they are dropping elements of the EBP. Note that this code may be used if interventions (e.g., agenda setting) or modules (e.g., the Cognitive Processing Therapy safety module) are intentionally left out.
5. **Lengthening/extending (pacing/timing):** the clinician reports spending a longer amount of time than prescribed by the manual to complete the intervention or intervention sessions(whether due to changed spacing between sessions, or longer sessions, more sessions, or spending more time on one or more modules or concepts)
6. **Shortening/condensing (pacing/timing):** the clinician reports spending a shorter amount of time than normal to complete the intervention or intervention sessions (whether due to changed spacing between sessions, or shortening sessions, offering fewer sessions, ongoing through modules or concepts more quickly without skipping material)

    If material is skipped in the context of shortened or abbreviated sessions, then this would qualify as two modifications (both "Removing/skipping" and "Shortening/condensing," e.g. shortening a protocol from 12 to 8 sessions by both condensing material and skipping some materials/interventions entirely).

7. **Adjusting the order of intervention modules or segments**: the clinician indicates that they have presented intervention modules or concepts in a different order than originally described in the manual, regardless of the reason (e.g. if the clinician deemed the patient not ready for a particular module, or if the clinician wanted to cover other material that seemed especially relevant to the patient at that time). If the intervention provides flexibility around the order of modules, then this code would not apply.
8. **Adding modules:** the clinician indicates that they inserted additional distinct materials or areas of focus consistent with the fundamentals of the intervention(e.g. a therapist doing CBT for depression who adds on a few sessions of CBT for insomnia would be coded here, but adding DBT or mindfulness modules to CBT would be "Integrating another treatment into EBP" above); or modules that are in some way complimentary (e.g., adding psycho-education about parenting to an anger management protocol). This differs from integration 6 in that this is adding a distinct/discrete element/focus rather than weaving in other interventions or techniques.
9. **Loosening structure:** If a clinician indicates that they don't always structure a session as prescribed in the manual but still believe that the intervention is the starting point from which they work, this code is appropriate (e.g., if they say they don't use the formal structure, but still endorse the use of Cognitive Therapy throughout the session; or if they say they allow a brief period of off-topic discussion or processing prior to the start of the CT session/agenda setting). If they also name specific elements that they do not use, a separate code would also be used, namely, "Removing/skipping". This code should not be used if they endorse something more along the lines of weaving CT into another framework(in which case, use Integrating intervention into another framework). Note that saying something like "it's not as formal" is not specific enough (as this could mean they just change the language)—they need to indicate in some way that they emphasize structureless in some way.
10. **Repeating:** If a module or intervention that is normally prescribed once during a protocol done more than once, this code should be applied. For example, if one session of breathing re-training is prescribed, but a clinician later repeats this intervention, "repeating" would be coded. If no mention is made regarding implications for the length of the session or protocol, no assumptions should be made about length. However, if it is mentioned that repeating resulted in lengthening of the session/protocol, both codes should be applied as separate modifications.

11. **Substituting:** A module or activity is replaced with something that is different in substance (e.g., replacing a module on condom use with one on abstinence in an HIV prevention program).
12. **Spreading-breaking up** session content over multiple sessions, e.g., a 1-sessionintervention gets broken up into 2 sessions.
13. **Departing from the intervention:** ("drift") followed by a return to protocol within the encounter, e.g., moving from CBT to supportive therapy for 10 minutes or more, then getting back to the protocol
14. **Drift from protocol without returning:** (e.g., start using another intervention); e.g., stop using CBT, do supportive therapy or another approach for the rest of the session; stop discussing lifestyle changes in a diabetes prevention intervention before module is complete and discussing contraception for the remainder of the meeting.
15. **Not a modification**: If activities are consistent with the intervention, even if the clinician does not think they are, it should not be coded as a modification (unless it meets the definition of tailoring/tweaking above). This code can also be used if clinicians endorse making referrals for adjunct services unless this is inconsistent with the intervention.
16. **Not enough information to code**---use sparingly!

Relationship to fidelity

Look to manual and fidelity tools for guidance about proscribed and essential elements. In their absence, determinations for coding should be in conjunction with someone who knows the protocol/literature/theory well, or after a review of theory and research.

1. **Fidelity-consistent modifications**: preserve core elements/functions of a treatment that are needed for the intervention to be effective.
2. **Fidelity-inconsistent modifications**: alter the intervention in a manner that fails to preserve its core elements/functions.
3. **Unknown**: use when there is no theory or evidence to inform a decision about whether an element is core vs. peripheral.

What was the goal of the Modification?

1. **Increase reach or engagement**: changes intended to increase the # of people that are willing to engage in the intervention.

2. **Increase retention**: changes intended to increase the # of people that are willing to engage in a full dose of the intervention.
3. **Improve feasibility**: accommodate time or space constraints, etc

   I added separate codes for

   (1) Increasing sustainability: modifying program to fit into the structure of the educational system such that it can continue to be delivered under the routine conditions when the grant funding is over, and the researchers are no longer there.

   (2) Scalability: modifying the program to ensure that it can be used in separate kinds of contexts, researcher must note intent to bring program to different contexts not just one.

   (3) Fidelity: modify components of the program that facilitate the implementers adherence to specific requirements such as time constraints or word usage. This was mostly discussed in the context of participants still being involved in a program implementation cycle.

   These were separate goals that researchers reported having.
4. **Improve fit with recipients**: this can include factors such as preferences, needs, abilities. Note that cultural adaptation is a sub-category of fit.
   a. **To address cultural factors**: factors specifically identified to be unique to a particular group, that require a change from the original intervention. Consider language and meanings of words and terms, culture, and context in such a way that it is compatible with the client's cultural patterns meanings and values.

      Note that culture is also listed under reasons (see below) and can be endorsed for cultural adaptations (creating some redundancy), but by stakeholder request, this specifier is a way to identify an adaptation specifically as a cultural adaptation. Note that some adaptations may be made in conjunction with cultural adaptations that are not in and of themselves cultural adaptation—these adaptations should be coded separately and not use the cultural specifier. For example, adaptations such as tailoring may be for the goal of improving fit with a specific culture, and culture may be checked off as a reason. Additionally, services may be delivered to this population in home rather than in a clinic with the same population, to improve engagement or fit, with the reason being the recipient's access to resources such as transportation or the clinic's location/accessibility, but NOT due to culture.

5. **Improve effectiveness/outcomes**: health outcomes, as opposed to satisfaction, engagement, etc.
6. **Reduce cost:**
7. **Increase satisfaction:**
8. **Promote equity/reduce disparities**-for use when there are identified inequities in the availability, quality, or provision of services and the adaptation is intended to address those disparities or promote more equitable care/service delivery.

Reasons—What factors influenced the Decisions?

Multiple items across levels can be coded, but stick to the most salient, clear reasons for the actual adaptations. For example, if the recipients were of a different ethnicity than the original populations, but decisions to adapt were really made due to funding constraints, and no cultural adaptations or other elements to address the differences in populations were made, only code funding. Note that an "Other" code is included on the coding sheet, but it should be used sparingly, and efforts should be made to identify a code (or combination of codes) that fits.

Sociopolitical/Outer Context

1. **Existing Laws**: Mandates, Policies, and Regulations that might place constraints or requirements on an implementation.
2. **Political climate:** e.g., if some aspects of the intervention that are controversial are altered, or new elements are included due to significant political attention (e.g., integration of suicide or violence prevention or screening due to political attention)
3. **Funding Policies**: requirements for funding, constraints placed on funding or reimbursement (e.g., if telehealth isn't reimbursable, etc)
4. **Socio-historical context**: e.g., if aspects of the intervention raise concerns or adaptations are requested due to the history or social context of a community.
5. **Societal/Cultural Norms**: e.g., if there are norms regarding where mental health support is received (e.g., through clergy or spiritual advisors) or who provides it; norms that may necessitation alteration of intervention aspects or terminology used; stigma may also be considered here.
6. **Funding and Resource Allocation/Availability**: adaptations made because more or fewer resources are available (e.g., shortening or expanding; changing what materials are distributed, etc.)

Organizational level

1. **Available resources:** (funds, staffing, technology, space)

2. **Competing demands or mandates**:Competing demands, de-prioritization of an intervention
3. **Time constraints**
4. **Service structure** (e.g., a clinic only provides group or time-limited interventions)
5. **Location/accessibility**
6. **Regulatory/compliance** – e.g., legal concerns may lead to certain aspects of an intervention not being delivered (e.g., limits to the types of physical activity or activities that may occur off-premises).
7. ~~**Billing constraints** – e.g., only certain providers can bill for certain intervention, limits to frequency or amounts of services that can be billed.~~

   This was removed because billing is not a common practice in education research.
8. **Social context** (culture, climate, leadership support)—organizational climate and context
9. **Mission** –goal and purpose of the organization
10. **Cultural or religious norms**—E.g., norms that care providers need to be the same gender as the patient, or that a family member would remain present during care; religious norms that don't fit with an intervention; cultural norms shared by members of the organization that are at odds with aspects of the intervention (e.g., assertiveness training). These could also impact org culture or mission but may not necessarily 11.Identified disparities or inequities at the organizational level (identified disparities in quality, access to pr provision of services

Provider Level

For the four below: if the provider is of a different race, ethnicity, sexual orientation or gender identity than the recipient and components are added to facilitate cultural competence and shared understandings, or to acknowledge different experiences that the provider and recipient may have had.

1. **Race/ Ethnicity**
2. **Sexual/gender identity**
3. **Cultural competency**
4. **First/spoken languages**—e.g., if training needs to include translation of concepts and terminologies; if intervention may need to include use of multiple languages to facilitate understanding.
5. **Previous Training and Skills**—knowledge and familiarity with an intervention

6. **Preferences**—comfort and interest in providing aspects of the intervention.
7. ~~**Clinical**~~ **Expert/Professional Judgement**—decisions based on clinical presentation and judgement about the needs of the individual.
8. **Perception of intervention**—beliefs about the intervention and it's fit, complexity, and effectiveness.
9. **Comfort with technology**

Recipient level

For the four below: if the recipient (s) is/are of a different race, ethnicity, sexual orientation or gender identity than the recipient and components are added to facilitate cultural competence and shared understandings, or to acknowledge different experiences that the provider and recipient may have had. Decisions on whether to code these factors at the provider, recipient level, or both may depend on who identifies the need, or whether the adaptation applies to a single or few recipients or provider…

1. **Race; Ethnicity**
2. **Sexual/gender identity**
3. **Sexual Orientation**
4. **Cognitive capacity / Physical capacity:** ability to engage in the intervention due to physical or cognitive abilities.
5. **Access to resources**: e.g., resources that make engagement in certain aspects of an intervention feasible.
6. **Literacy and education level**: ability to engage with written materials or complex content.
7. **First/spoken languages**.
8. **Legal status**: e.g., individuals involved in the criminal justice system may not be able to engage in certain aspects of an intervention.
9. **Cultural or religious norms**
10. **Comorbidity/Multimorbidity**—presence of other conditions that require intervention.
11. **Immigration Status**—e.g., undocumented individuals may require changes to reduce.
12. **Risk of legal problems**: some concepts may require additional attention or tailoring, such as concepts of autonomy and control may need to be addressed differently due to uncertainties related to immigration status.

13. **Crisis or emergent circumstances**: emergencies (e.g., health risks, suicide risks);significant life events that require intervention or attention.
13. **Motivation and Readiness**—willingness to engage in the intervention.
14. **Comfort with Technology**

Potential additions

What was the result of making this change?

APPENDIX C

Brief Case Description of Interview with P01

P01 has worked in a continuous improvement project with practitioners representing 4 districts across several years. The project goal was to focus on providing professional improvement sessions to math teachers that were co-designed with the chief academic officer. Multiple modifications were made to the program during its planning and development across the districts. The first phase represented the planning and piloting period where teachers, instructional coaches, and principals across four districts were observed, and feedback was given to the broader district. Then, in the second phase there was ongoing development as instructional improvement institutes that were implemented in two school districts. During phase two the team collaboratively worked on refining the institutes with the chief academic officers, principals, and instructional coaches.

During the implementation and iterative improvement cycle there were unplanned modifications that took place to address the growing pressures from the national educational reform "No Child Left Behind" which manifested through added pressures from the principal supervisors. The changes occurred at the sociopolitical and organizational levels where new curriculum and training for teachers was endorsed in fact memorization to increase short term gains on students standardized tests. The program developers aimed to inform the principal supervisors about the negative consequences of focusing on short term gains while addressing concerns about performance on state testing which lead to joint work to address the broader socio political changes in NCLB and the common core standards as well as the organizational pressures of state testing. This change in joint work helped to maintain the integrity of the theory of action, such as the focus on instructional improvement that prioritized problem solving, reasoning, and understanding.

Change #1 Comparing New State Test to Old State Test

To maintain the integrity to the theory of change researchers worked with the guidance of the math department to present a comparison of the kinds of questions that would be asked on the new state test. Because the items on the new test benefited from the instructional approaches that were originally emphasized in the instructional institutes this helped bring the districts principal supervisors, the math department, and the researchers to an agreement about the maintenance of the in the instructional institutes focus on instruction that prioritizes problem solving, reasoning, and understanding rather than rote memorization.

Change #2 From Part Time Coaches to District B Coaches

The goal of this change was to improve the reach and feasibility of instructional improvement the math department practitioners and researchers changed the strategy of having one part-time teacher / part-time coach in each middle school, to having three full-time coaches for the district that got additional professional development on how to work with groups of teachers as opposed to individual teachers. This shift was consistent with the group's commitment to the idea that any sort of shift in practice required intentional professional learning. This change was based on the effectiveness of instruction observed at the practitioner level; however, it influenced the whole organizational structure of the school districts and maintained the focus on the district coaches.

Change #3 Adapting to Changes in Leadership in District D

A new chief academic officer disbanded the cadre of math coaches in this district, and the institutes could no longer be held. In their best efforts to maintain the continuous instructional improvement researchers continued to partner with the math leads to design and co-design and co-lead professional development. However, the instructional improvement institutes were disbanded. This was an unanticipated district level change that was not aligned with the shared assumptions laid out in the theory of action nor the integral values of the research teams and their partners.

Brief Case Description of Interview with P02

P02 has worked on developing a fraction intervention for students with a history of mathematics learning difficulty. The intervention's development process spanned a period of about 10 years. It evolved from *Fractions Challenge* to *Fraction Face-Off*, to *Super Solvers*. The changes were formulated on the basis of a series of randomized controlled trials contrasting the added value of alternative instructional components. Changes were also made to reflect the intensified and upgraded fractions instruction at grades 3 – 5 due to College- and Career-Readiness Standards replacing No Child Left Behind standards and to address different objectives and learning needs at grade 3 versus grades 4 and 5. Most recently, the Fuchs research team has assessed the program's efficacy when implemented as whole-class fraction intervention, with strong effects demonstrated for students with and without histories of mathematics learning difficulties.

Change #1: Fine-Tuned Instruction

Across multiple randomized controlled trials, each with three several arms, the researchers fine-tuned the instructional content and procedures by selecting design components that yielded the better learning outcomes for the students. One example is a randomized controlled trial that contrasted business-as-usual intervention versus *Super Solvers* with blocked instructional design for the program's calculations component, versus *Super Solvers* with interleaved instructional design for the program's calculations component. (Blocked instruction involves teaching each fraction calculation topic in sequence. Interleaved instruction involves teaching the fraction calculation topics together.) Results revealed that *Super Solvers* with interleaved calculations instruction produced superior effects 1 year after intervention ended, compared to that *Super Solvers* with blocked calculations instruction and compared to control. For this reason, the team incorporated interleaved calculations instruction within *Super Solvers*.

Change #2: Change in Content to Address the Changing Counterfactual

The landscape of fractions at grades 3 – 5 instruction was affected by the roll out of Common Core (i.e., College- and Career-Readiness Standards). This reform upgraded and deepened regular classroom fraction instruction, which benefited average and high achievers. The reform did not, however, benefit struggling students, so achievement gaps grew. In response, the content of *Fractions Face Off* was refined within a revised intervention – referred to as *Super Solvers*, that better addresses the reformed educational landscape.

Change #3: Small Group → Whole Class Instruction

To increase impact by addressing students with and without mathematics learning difficulties, the researchers re-engineered *Super Solvers* for whole-class implementation. Each session, classroom teachers (1) conduct a class wide lesson, (2) then divide the class into dyads for guided rehearsal and peer explanations, and (3) then oversee independent practice to ensure individual student accountability. *Inclusive Super Solvers* relies on the same lessons as the *Super Solvers* small-group intervention, but it is delivered by classroom teachers in a different context (regular classrooms).

Brief Case Description of Interview with P03

P03 focuses on developing interventions for children with math learning difficulties. The Fraction Sense Intervention (FSI) discussed during the interview was built from a focus on providing children with math learning difficulties more opportunities to learn fractions in 6th grade so they can move on to higher math. This work included a team of university researchers, a local school district lead that formed a conduit between researchers and the teachers in the schools. To assess the intervention there was a preliminary development study, which experienced multiple changes over time, and a larger efficacy study that is currently ongoing. The intervention has been iteratively developed across the years and improved based on the researchers' experiences with teacher observations, teacher feedback, and evidence from students' math learning. The changes that were made to the intervention took place across phases of piloting, implementation, and scaling-up to new schools and classrooms. Changes we highlighted during the interview are described in detail in the following sections. In general, changes were made to the content, training, and evaluation, and to the implementation; these changes encompass the format of delivery, the setting in which the intervention is delivered, and the personnel delivering the intervention. Changes were designed to proactively address differences at the cohort level (i.e., for students who are low-income) and unit level (i.e., for schools with intervention teachers that do not have formal training in math instruction). The nature of these changes includes the refinement of intervention materials such as the math problems used and the timing of the lessons to adapt to classroom instructional time. All changes preserved fidelity to the core elements of the initial design and plan.

Change #1: Researcher implemented → Teacher Implemented

The intervention changed from being delivered by researchers, who would pull students out of the classroom for small group instruction, to being delivered by the teachers. This was a proactively planned change to the setting of treatment delivery to understand how teachers would implement the intervention without the direct involvement of researchers. In addition, the professional development of the teachers changed to adapt to the teacher's time constraints. The broader goal of these changes is to improve feasibility and address factors of the local context such as organizational resources, competing demands, time constraints, and the service structure of each school.

Change #2: Small group instruction → Whole class instruction

The variation of available resources across school contexts motivated the team to proactively plan to change the format of the intervention from being delivered to small groups to whole intervention classrooms. Especially if the intervention had now shifted to teacher-based implementation. The goal of this change was to align with organizational level constraints as in the previous change.

Chage #3: Focus more on arithmetic and go back to basics

Across the data collected researchers realized that to reinforce conceptual understanding, better fit the intervention with student needs, and improve effectiveness they should include simpler problems that were more meaningful. Such as showing addition and multiplication together (½+½ = and ½ * ½ =) using the principle of interleaved practice. This is a change that retained the fidelity of the intervention with the 4 original design components that were drawn from the large-scale longitudinal study that formed the basis of this ongoing work. As the intervention was implemented the research team reactively made changes as they learned how the cognitive principles were best implemented to help children better learn.

Change #4 Intervention timing per session + adding script

To improve the feasibility, fidelity, and reach of the intervention the researchers designed open sourced and low-cost materials. These changes included PowerPoint lessons that contain the scripts in the notes section and lessons that are manageable within a grading period and 30 minute class periods. These changes broadly address the variation between teachers' practice and norms; the researchers also provide video tapes so that teachers may observe the lesson. Some of these changes were made in reaction to the constraints observed across different schools and teachers.

Brief Case Description of Interview with P04

P04 has been developing and validating rubrics for measuring instructional practices in math classrooms. As a former high school math teacher and researcher of the MIST project at Vanderbilt University, she discussed her collaborative work using the Instructional Quality Assessment (IQA) tool for understanding changes in teachers' instructional practices over time. Her interest in rubric development stemmed from a desire to capture additional dimensions of practice not measured by the IQA. Her research team received funding from the National Science Foundation to validate and refine the rubrics. P04 mentioned the challenges and iterative nature of the rubric development process, including the need for changes based on cognitive interviews, generalizability studies, and iterative cycles of refinement. P04 engages in research-practice partnerships and continuous improvement approaches in her work. One important challenge she has encountered is navigating where to publish articles that detail the iterative process of developing rubrics, interrater reliability, and surveys for authentic educational contexts. There are no clear pathways for these kinds of publications, which further obscures the transparency of the adaptive and iterative process of developing math programs, measurements, and observational tools.

Change #1: Rubric Content Changes to Observational Rubrics

One significant change they made was merging the math expectations and social expectations rubrics into one, as it was challenging to disentangle the two in practice. The interviewee acknowledges that not all rubrics functioned well, especially in elementary classrooms, and adjustments were made based on evidence and feedback from the coders. This was an unplanned change that emerged during the implementation of the rubric validity study.

The interviewee highlights the importance of iteration and design-based approaches in their work. They emphasize the need for multiple iterations of generalizability studies and making changes based on the data collected. They also mention the challenge of using rubrics to measure things that occur frequently in classrooms and the complexity of assessing non-standard English and cultural dialects. Although a specific change to the rubrics based on this matter was discussed. It was evident that multiple refinements were made to the content of the rubrics in order to better fit the cultural factors of the classroom, such as modifications that would better fit the observable behaviors of teachers and students that would be observed with the rubrics. Specifically, there was less non-standard English used by the students than the researchers expected, which can be attributed in-part to the established norms of the classroom.

Change #2: Survey Content Validation in the Local Context

In a separate project where researchers were developing a student survey the interviewee also brought to light the importance of validating any measure with individuals embedded in the local context. This validation work ensures that the research tool fits the recipient's needs, preferences, and abilities. For example: the researchers in this team conducted cognitive interviews to ascertain that student understood a survey as the researchers intended. By doing this, the cognitive interview responses revealed that the student's did not all understand words like "value" and that they had developed different meanings for words like "race", "culture" , and even in some cases how they felt towards doing math. This led to proactive modifications for the survey prior to implementation in a large school district.

Brief Case Description of Interview with P05

P05 has been leading a team of researchers and software engineers to create Math-Mapper, a software composed of an interactive map of nine big ideas for middle school mathematics. The ideas are subdivided into clusters and constructs, each of which has an underlying learning trajectory based on empirical research from the learning sciences. These learning trajectories are associated with the Common Core standards. Multiple diagnostic assessment items are written for each of the levels of the trajectories. Teachers assign students short diagnostic assessments throughout the year and the data is returned immediately to both teachers and students showing student areas of strengths and areas for improvement by level. Retesting with equated assessments is made available.

As the team developed and refined the design of Math Mapper, P05 collaborated with teachers across two different schools located very geographically and economically diverse school systems. The complex differences between these groups created design challenges and opportunities which supported an iterative design process for Math Mapper and influenced the speed of its technological development. Both student data and teacher feedback influenced changes and new features. Many modifications involved adaptations to content organization based on practices in schools such curricular chunking or sequencing needs/choices. All changes maintained fidelity to the key components of the original plan of this design.

Adaptations to Theory of Change
Content and Context Modifications

The teachers engaging in iterative designs with the research teams suggested that:
1. misconceptions associated with levels be added into the map together with diagnostic items and reporting,
2. the big ideas of the program be extended to include Algebra 1, and
3. the assessments be offered at finer grain size by assessing at the construct level as well as at the cluster level of the big ideas. These requests were based on teachers' instructional practices and were carefully designed into Math Mapper by the researchers.

Another challenge addressed by the team was how to overcome teacher's automatic questioning or rejection of the value of multiple choice items. Simultaneously, teachers would raise concerns that the conceptual items were too difficult, before giving their students adequate time to learn how to approach them. It took time to gain credibility with the teachers on the validity of the measurements. In other cases, important topics assessed by the program were skipped by the teachers. For example, unaware of the research on the importance of students understanding the shape of univariate data before the study of measures of central tendency, teachers would rush to teach mean, median, and mode without a foundation in variability. These challenges ramified into challenges and innovations to the ongoing validation process concerning the psychometric qualities of the measures and led to strong collaborations among learning scientists, psychometricians and practitioners.

However, teachers also created compelling innovations and opportunities in their own practices. For instance, one teacher orchestrated the use of student grouping based on like performance patterns and assigned students to work collaboratively using the software's practice mode and later return to work individually to revise and resubmit correct answers to missed questions. These adaptations modified the format of the implementation but in a way that was still consistent and even enhanced within the core theory of change.

Adaptations to Implementation Model
Paucity in Scale-Up Due to Partnerships Ending
An important disruption occurred in the development of this project when a prominent foundation terminated their funding stream that had supported a partnership with an after school math literacy program which was building a means to use Math Mapper program as an assessment tool in their certification program. Other challenges to scaling the use of the program were experienced in soliciting commercial

partnerships due to a reluctance by some businesses to partner due to concerns this new product would disrupt other programs they had already sold to schools (despite their own excitement for the innovative capabilities of the program). Another tendency which limited the uptake of the program was that many teachers relied excessively on wanting to assess a single isolated standard. Since standards are of widely divergent grain size and when approached individually can limit coherence of how students develop mathematical reasoning, this commitment to a single standard was viewed by the developers as restrictive.

These examples reveal a potential area where funding agencies can find ways to support researchers in commercializing their math programs. Current overreliance on market forces to support innovation have proven insufficient. Furthermore, there are additional needs for better means to protect copyrights, as some publishers and test makers will freely take the content of, for example, the learning trajectories without providing adequate acknowledgement and compensation.'

Professional Learning Communities

The use of Math Mapper in diverse school settings raised an important consideration about how to operate Professional Learning communities around the use of diagnostic data within formative assessment practices. Two different experiences occurred in working with math supervisors and yielded clear evidence of the need for their support and expertise in using data for ongoing improvement. In one case, the school supervisor was a strong advocate for the use of the program and did so in ways that maintained fidelity to the theory of change and enhanced methods of implementation. She focused on teachers' understanding of student reasoning in relation to specific content and paid close attention to how to obtain improvements for all performance levels. In contrast, the other supervisor was sporadic in attending training sessions and often advocated for a huge variety of programs with little attention to their coherence or contradictions. Furthermore, the researchers/designers noted that teachers also needed an appropriate combination of support, and accountability to shift to the ongoing and robust use of data to improve instruction and focus on student thinking. The implementation process was continuously refined as the researchers developed and tested conjectures of how to operate those groups and gain the teachers trust as teachers went into about their experiences and developed their own conjectures depth on student reasoning .

Brief Case Description of Interview with P06

P06 has conducted multiple research projects focusing on changing instructional practices in math education in direct partnership with math teachers, math curriculum specialists, and local schools. Her their initial project, funded for three years, aimed to create teaching materials and tasks to enhance the teaching of additive word problem solving. They collaborated with teachers, proposing activities, testing them in classrooms (by teachers) and discussing their implementation in classrooms. The project received positive feedback from teachers and the Ministry, leading to a follow-up project on multiplicative structures for higher grade levels. Throughout our discussion P05 highlighted a cycle of feedback from teachers and through qualitative studies that was used to refine the content of the classroom lessons and the professional development provided. Experimentation with students within the classroom paralleled the cycles of feedback to inform design. Important challenges to doing this work were discussed.

Challenges

P05 highlights the challenge of maintaining instructional changes when students transition to new grade levels with different teachers who are unaware of the project and its approach. This issue led them to change their approach in the next project by working with teachers of all levels within a school instead of specific grade levels. This informed an important guiding principle that rather than approaching program implementation with individual teachers, researchers should consider implementation at a systems level, such as to cohorts of teachers or entire schools so that teachers can build community and support around the new instructional practices. She discusses strategies such as publishing books and presenting at venues for teachers to share their findings and approaches. They also mention that teachers have started applying some ideas from their approach, albeit partially, which reinforces the need for working with schools as a collective effort to provide support and ensure continuity.

Although the project achieved success within the two schools, the interviewee expresses uncertainty about replicating the project on a larger scale due to financial limitations and the lack of a continuous education system for teachers in their context. Funding is not available for the longitudinal work that is needed to test how the instructional approaches would benefit students across multiple grades. This is important for funders to consider because these institutional structures may prevent effective math programs from being able to scale. Importantly, another guiding principle emerged when P05 emphasizes the importance of scaling knowledge rather than the research itself, tying this back to the previous challenge of implementing new programs with systems in mind.

Adaptations

Some of the changes that were made across the different implementations include content changes to the wording of the lessons. The school consultant who was a partner in the research held a lot of power in deciding which activities and didactic approaches would work for students and how to best approach the teachers. Another change was to the personnel who participated in the project. Additional support from school consultants and specialists was requested by the teachers to help with the students with the most challenging math difficulties across several classrooms.

Overall, the interview sheds light on the challenges faced in scaling instructional changes in math education, given financial constraints and the absence of a continuous education system for teachers. The interviewee emphasizes the significance of working with schools as a whole to create a supportive environment and shares strategies for sharing knowledge with practitioners.

Brief Case Description of Interview with P07

P06 is a research professor who has conducted an external evaluation of the Odyssey Compass Learning program. In collaboration with a team of researchers and intervention implementers, she conducted a randomized control trial of the math software as part of a contract with the Regional Education Lab. The evaluation focused on fourth-grade students and involved a team of researchers from several academic and for-profit research organizations. The evaluation aimed to assess the effectiveness of implementing the software as instructed by the developers in comparison to a control group. Although variations in software usage were observed across different sites, the study maintained its strong fidelity to the original research design.

Changes

P07 highlights an interesting aspect of the evaluation, where the developers suggested changing the study to implement the software as an after-school program, citing previous research that showed positive outcomes in such contexts. However, this change was rejected as a precaution against introducing confounds that would disrupt the causal design. This decision ensured the integrity and validity of the study and showcases a situation in which making adaptations would not be desirable or in line with the research goals.

Adaptations were made to the professional development that teachers received as part of the implementation; this is planned reactively as teachers provide feedback on aspects of the software that they were not familiar with. This provided feedback for the researchers as they refined the checklist of what components were necessary for the proper implementation. This marked an important aspect of the evaluation that can be overlooked by other researchers yet is essential to the proper scaling of this intervention.

Guiding Principles

During our conversation, a practical guiding principle emerged when the researcher described an important adaptation that was made to increase the retention of participants was their recruitment and consenting protocols. When working with marginalized populations the experimenters experience very low consenting rates when using the opt-in approach as opposed to the opt-out approach. This created significant challenges to the researchers and based on the data collected the organization IRB changed their protocols to allow for opt-out consenting processes. This situation was not formally published; however it was reported to the funding agency which highlights an opportunity to share this knowledge more widely as other researchers may be facing similar challenges.

In addition, P07 discusses the importance of maintaining fidelity in efficacy studies and shares her experiences in conducting other interventions. As she explained, researchers benefit from a strategy to monitor and support fidelity, including regular check-ins, usage reports, coaching, and modeling sessions with teachers early on during implementation.

Overall, the interview provides insights into the evaluation of the Odyssey Compass Learning program and the challenges of maintaining fidelity in research studies. The interviewee emphasizes the importance of preserving the integrity of the study design to ensure accurate and reliable results, while also acknowledging the importance of calibrating research quality by making adaptations in changing conditions.

Brief Case Description of Interview with P08

P08 is a researcher at a for profit research corporation, in this interview she describes an external evaluation of the effectiveness of the Cognitive Tutor for Algebra for middle and high-school students. Her broader research interests include providing pathways for STEM education for women of color that come from low-income backgrounds. This intervention was evaluated as a potential tool for shaping students' academic paths and motivating them towards STEM careers. The interviewee collaborated with a team of researchers to lead the evaluation study, combining rigorous evaluation methodologies, including a strong Randomized Controlled Trial (RCT) design. The U.S. Department of Education's Institution of Education Sciences (IES) supported the evaluation, aligning with their emphasis on rigorous causal models. Below I summarize the adaptations that were made during the evaluation process which lasted approximately two years, these changes have also been reported in scientific publications.

Change #1 Adaptations in Implementation Variation

P08 collaborated closely with the developers, attending professional development sessions to operationalize the program's components and understand the desired classroom practices. By developing quantitative measures, such as surveys and interviews, the interviewee aimed to capture variations in implementation across schools and classrooms. This approach facilitated a comprehensive assessment of implementation fidelity, considering the program's theory and impact on outcomes. The interviewee discusses the outcomes of the evaluation, highlighting that the initial hypothesis assumed improved implementation fidelity in the second year, leading to stronger effects on student outcomes. However, the analysis revealed that teachers in the second year implemented the Cognitive Tutor program with less fidelity compared to the first year. Surprisingly, these teachers adjusted their approach, deviating from the developer's instructions, while still incorporating the program's unique features to a higher extent than the teachers in the control group. The changes the teachers made appeared to be aimed at aligning the intervention better with the state standards.

P08 Discovered this unexpected variation through quantitative data on implementation and inquiries into classroom practices. The changes made by teachers in implementing the program, indicate that fidelity to specific components varied. However, teachers adapted the program to suit their classrooms and students' needs, resulting in positive outcomes. The interviewee emphasizes the importance of measuring dosage and quality rather than solely focusing on fidelity, highlighting the complexity of quantifying implementation quality, which often requires observations and rubrics.

In addition, the careful measurement approaches used by P08 serve as a guiding principle for researchers to capture components such as implementation fidelity, dosage, and the quality of program delivery to be able to determine when adaptations and other forms of variation of implementation may explain differences in intervention impacts. The unexpected finding underscores the complexity of program implementation and the need for nuanced evaluations. Overall, P08's perspective underscores the importance of understanding implementation nuances and the dynamic nature of program delivery, providing guidance for evaluators and researchers in considering comprehensive impact analysis.

Brief Case Description of Interview with P09

P09 has been involved with the Young People's Project since tenth grade, he has graduated with a Bachelors in Computer Information Systems, a Masters in Science and Technology Management, and has received his own grants to fund ongoing projects in graduate school. We discussed his involvement in YPP and his current project which is a new approach that builds from the roots that YPP has established. The Young People's Project, founded in 1996, aimed to address the issue of Jim Crow segregation in education, particularly in mathematics education. The project involved young people teaching math to their peers in the community. P09's participation in the project exposed him to travel, new experiences, and the realization that there was more to life beyond high school. His involvement in computer science and coding within the project influenced his decision to major in computer science and pursue a career in technology. The founder of the Algebra Project Dr. Bob Moses approached him about a funded research project that required someone who could bridge the gap between computer science, math, and young people. Since then P09 has been involved in developing and teaching a computer science course called Exploring STEM Literacy.

Change #1 Personnel implementing the STEM literacy course

The exploring STEM literacy courses were iteratively developed in collaboration with the math literacy workers and school teachers. The approach to build this project is not so much focused on scaling, instead P09 describes a process of replication utilizing the key components of YPP across new contexts while adapting to the opportunities and challenges that each context offers. P09 reflects on how different cultures can make a difference during implementation at the school and classroom context. The key components of the program include involving youth from the community in the research and program implementation, this is referred to as a model of distributed excellence originating from the African proverb that says "it takes a village to raise a child". One adaptation that deviated from this component arose from a single teacher's decision to take on the role that would originally be allotted for a near-peer college math literacy worker. In this case, the change was not planned and instead a reactive adjustment to the local context. As P09 explains "teachers aren't real people…if you do see them in public, you're going to run away. But when kids see young people who are math literacy workers. They run towards them. So that's totally different".

Change #2 Adapting to the changing context

Another adaptation was planned in response to the constraints that teachers had throughout the school year in order to meet the adequate performance levels on the state standardized tests. Teachers were under so much pressure to adhere to the mandates that the plan to deliver the program changed. Professional development was delivered over a longer period of time throughout the school year, so that by May teacher's built up sufficient experiences to implement the new activities with their students after testing was finished.

As the project has developed using a design-based approach, now into a Summer based professional development, the overall goal has been to develop a set of professional development experiences that incorporate the previous adaptations to help math teachers become more like the math literacy workers to understand where students need to be mentally and empathetically, to better understand students and serve them better. The goal of the design is to model YPP's peer-to-peer approach to replicate the components of YPP that make its informal learning experiences successful during the school day. This project engages in a cycle of feedback where the teachers learn new content, then they go try it out with new students, and then the students give feedback back to the teacher's as they refine. It will be insightful to understand how the feedback from the students and the teachers shapes the design and implementation to the project moving forward.

Brief Case Description of Interview with P10

P10 develops math intervention materials that focus on math skills such as word problem solving, the interpretation of the equal sign, and fractions for students struggling with math. She has conducted efficacy grants to determine the impacts of these math interventions on third and fourth grade math learning and developmental grants focusing on models of professional development for teachers. Across both grant types, the adaptations made are largely based on practitioner feedback, however the extent to which changes can be implemented vary by the grant specifications where efficacy grants are more restrictive and developmental grants allow for more frequent changes. A routine of feedback is built into each project to improve intervention effectiveness and continuous improvement. Feedback is provided by various practitioners including tutors and teachers and the team of researchers decide which adjustment can be made within the capacity constraints prior to the next intervention implementation. Across the projects several changes have been made prior to implementation, during implementation, and at scale-up. Adaptations were proactively planned by the research team however most of the adaptations were content and treatment group focused. The nature of content modifications was to tailor or refine the materials, add new elements, lengthening the lessons or breaking them up, and reordering the lessons to better fit the students receiving the intervention. All changes maintained fidelity to the core elements of the original intervention design.

Change #1 Content Changes to Improve Efficacy

For the efficacy grants that focused on tutoring students, multiple changes were made to the intervention lessons content such as explaining regrouping in terms of money and analogies of going to the bank to change money to regroup 10 ones into one dollar bills. The goal of these changes was to improve effectiveness and the rationale behind it arose from conversations with teachers where they brought in their experiences and skillsets. Similarly, feedback from the tutors providing the intervention or teachers observing the interventions also inspired modifications like increasing time on difference problems in contrast to total problems, creating more lessons for reviewing multiplication and division facts, and even improving the clarity and relevance of practice word problems by incorporating names in the word problems that are similar to students peers and references that are culturally relevant to students such as the popular virtual game called *Roblox*.

Change #2 Content Changes to Address Changing Counterfactual

Content changes also occurred considering the sociopolitical changes caused by COVID-19, although the intervention program previously demonstrated effectiveness the new scale-up to a different school district following the pandemic did not elicit significant impacts on the equal sign test during the first year of implementation. The grant program officer allowed intervention changes to the equal sign activities to address differences in students' mathematical performance as this group had been affected by school closures in first and second grade. This is the only unplanned reactive change that was made due to the unexpected outcomes of the first year of implementation.

Change # 3 Recurring Adaptations in Developmental Projects

The granting structure of the developmental grants focusing on developing a professional learning and coaching model for math teachers allows for more flexibility in program design. The goals of these adaptations were to improve fit with the recipients of the program since they encompassed teachers across different schools, districts, and states. An additional goal is to improve the effectiveness of the program while addressing differences in teachers' personal skills and experiences as well as differences across their schools' resources like curriculum and textbooks, and competing work demands. Some of the changes made directly influenced by the teacher's exit interviews was to include a focus on place value and math language in addition to the focus on fractions and word problem solving. Researchers also created additional sessions to discuss strategies to transfer word problem strategies to online standardized testing since the students mostly practice on paper pencil but then complete their state standardized tests online.

Brief Case Description of Interview with P11

P11 is an implementation manager for a randomized controlled trial of an efficacious intervention being implemented naturalistically with fourth graders in across three different sites. She brings her expertise as a middle school math teacher and an elementary school math specialist to her current research. As the implementation manager, she has managed three site coordinators at three different universities and implemented the training for all the interventionists who implemented the intervention at her site. Across this work she has balanced the importance of maintaining fidelity of the intervention while making planned and unplanned adaptations to its implementation by based on interventionists response to the expectations. Adaptations to the program were made by the implementation team and the teacher implementers. Although some modifications were planned out during the pre-intervention planning, most adaptations occurred during implementation of this study. Minor changes were made to the to the context in which the intervention was delivered, the intervention design, how the implementers were trained, and to the evaluation of the implementation. Adaptations were made for groups of implementing teachers that shared similar constraints and to address differences in settings, personnel, and target population. The nature of the modifications was to tailor the intervention to better fit the shared constraints and the differences across contexts, mostly the changes encompassed removing elements of- and shortening the- curriculum, breaking up training sessions, and loosening the structure of the intervention a bit. All changes were consistent with maintaining fidelity to the original intervention design.

Change # 1 Contextual Changes in Group Sizes

Across the scale-up study there are many changes that occurred due to differences in context. The original plans were to recruit four teachers at each of ten schools, making it 40 teachers total. However, at one of the districts the organizational structure provided one Tier 3 interventionists at each of 30 schools, and so researchers adapted to this to ensure that sufficient teachers were recruited for the project. Teacher training previously had been done all at one time across 3 days. However, with this study the training was instead done "just-in-time" for a set of six lessons, which also allowed for the incorporation of coaching, and emailed feedback to teachers. Similarly, the organizational structures of different schools created variation in the student group sizes that teachers taught. To improve the feasibility of the intervention with teachers the original treatment group size was changed from one group of 3 to larger groups or multiple groups when teachers had larger class sizes.

Change # 2 Intervention Design Changes

At the end of the intervention, students play a game for review. Teachers became concerned about challenges the students would have and chose to make modifications to increase student engagement, satisfaction, and overall fit with their student groups. For example, one teacher with a group of 10 kids and did not want 1 child solving a problem, as 9 children watched, so she broke the students up into groups. As training facilitator P11 encouraged teacher agency but ensured that key components the intervention were met including the requirements that each student solves 3 or 4 problems, use the intervention cards, explain their answers, and play the game to facilitate competition and fun. Further curricular adaptations arose that were handled in a similar manner and in some cases, modifications were suggested by teachers were not accepted and rationale was provided to the teachers. Additional adaptations that were accepted included: (1) teacher created anchor charts to increase reach and address cultural factors in bilingual districts and (2) cutting the word problem warm up to decrease time for intervention, as it covered fifth grade standards and students were being intervened at in fourth grade and (3) cutting independent practice problems to also improve feasibility for teachers to deliver intervention within the 30 minutes. All these changes were made while intentionally maintaining intervention effectiveness with the key components of initial instruction, repetition and practice, and support.

Change # 3 Content Changes with Greater Gator

Prior to implementation, the research team decided to make one modification to the curriculum that was based on recent research showing that some mathematical rules that are taught in elementary expire in middle school. Specifically, the research team removed all references to the "Greater Gator" in the curriculum and explicitly addressed this change with the teachers in the training sessions.

Brief Case Description of Interview with P12

P12 has an academic background in school psychology and assessment, he has worked on the development and evaluation of school curricula for several years. His primary focus has been on student needs in early numeracy, and he has designed and evaluated multiple interventions including: (1) developing a core program called ELM for kindergarten, (2) the subsequent development and evaluation of the intervention program called Roots, (3) and further feasibility, replication, and scale-up wherein iterative improvements are made to the original designs. The researcher describes their iterative development process, which involves feasibility and usability trials, pilot studies, and efficacy trials. They also highlight the collaborative nature of their research team, which includes math content experts, research coordinators, and graduate students. The team members bring different areas of expertise and interests, such as math cognition and English language learners.

Change #1 Adaptation to Systems of Delivery

Throughout the extensive research conducted P12 reflected on the importance of building a "*system that is more closely aligned with how interventions are delivered in school settings versus how they're delivered in research studies where you're trying to ask and answer very specific questions*" (49). This aim is the focus of a new grant that is an intervention designed with greater flexibility so that teachers can adapt to changing contexts such as when they have different proportions of students in the classroom that qualify as being *at-risk*. The key guiding principle P12 acknowledges is that researchers should ask themselves "*What next steps do we need to do to sort of build better math delivery systems?*" (51). This can be at the classroom level, when considering how many children may need intervention in a classroom, as well as at the district level when assessing needs across schools.

Change #2 Systematic Development of Content Adaptations

Multiple modifications and adaptations have been made to the curriculum programs based on feedback and their iterative development process. P12 gives examples of changes such as adding visual cues for intervention delivery, including lesson previews, and underlining key vocabulary. These modifications aim to enhance the effectiveness and delivery of the programs, considering the variation in the local school contexts. For example, in some schools the teachers delivering intervention could be trained in a completely different field, such as history, and so the design of the intervention has this adaptation built in by providing very specific scripts and cues so that anyone with any level of training can deliver the intervention with fidelity to the key components. P12 mentions that most of the modifications have worked well and are informed by their extensive experience in schools. While these modifications are primarily maintained within the research team, they have written articles about their design process and how they applied it to their interventions. They also highlight a second guiding principle, and that is the importance of having a systematic development process and how it can benefit grant proposals and review panels.

Brief Case Description of Interview with P13

P13 is a former math teacher with extensive experience. We discussed his work on a fourth-grade math intervention project. The intervention is research-based and aims to address gaps in students' math understanding. The project trains interventionists at schools to implement the intervention themselves, allowing for potential dissemination of the program within the school district. The researcher mentions that the math content of the intervention is solid and avoids misleading information.

During the interview we discuss some challenges and hiccups they have encountered in their role as an observer and coach, providing feedback to interventionists, who are the current math teachers already embedded within the schools. During the implementation the protocols and the additional requirements have varied across teams based on the differences in their local contexts. The researcher also shares their observations of differences in consistency when implementing the program's procedures, which has led to confusion among students. However, they mention that overall, the intervention has been effective and well-received by the students.

Challenges to Adaptations

Some important challenges that arose from the implementation training, manual, and protocols were differences in the underlying assumptions of the researchers. P13 highlights the importance of addressing tacit assumptions in elementary math education and the need for a deeper understanding of the reproduction of knowledge in teaching practices. They share their observations during a training session with the intervention designers and how their feedback brought a new perspective to the program. P13 also mentions the editing process of the manual and the removal of a specific mathematical representation for teaching early mathematics. They express their view that this representation can be helpful for students and criticize the hostility towards them in some educational contexts. Further, assumptions were incongruent in the intervention design such as the emphasis on procedure rather than conceptual understanding. As well as incongruent mathematical representations that deviate from their definition.

The majority of these content modifications were suggested by P13. However, their objections were met with a dismissive response, implying that they were being pedantic. This is the only case that demonstrates the various challenges that researchers face in making adaptive changes to programs when there is unequal distribution of power across the research teams. A guiding principle from this example is for researchers to consider that they may not be aware of the underlying assumptions of their designs. An important exercise to make these explicit is to speak to practitioners and request feedback on this specific topic to ensure that assumptions are being explicitly stated and challenged.