UNIVERSITY OF CALIFORNIA,
IRVINE

Academic Skills and Long-Run Outcomes

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Education

by

Tyler W. Watts

Dissertation Committee:
Distinguished Professor Greg Duncan, Chair
Assistant Professor Drew Bailey
Associate Professor Damon Clark
Chancellor's Professor Carol Connor

2017

DEDICATION

To

Mr. Hollis for first showing me the value of education.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF APPENDIX TABLES

ACKNOWLEDGMENTS

I would first like to recognize the invaluable contribution of Greg Duncan to this work and to my career as a researcher and scholar. Greg has been a constant source of inspiration, and he has been an unparalleled role-model throughout my time in graduate school. I will always be grateful for the opportunity to work alongside him for five years. There is no doubt that his mentorship left a profound impression on me both professionally and personally.

I would also like to recognize the contribution of each of my committee members. Drew Bailey and Damon Clark have been fantastic mentors throughout graduate school, and I would not be in the position I am today without their help. They both influenced me intellectually, and their impression can be seen throughout this dissertation and on my research program in general. I am also grateful to Carol Connor for openly sharing her project with me, and I hope we have an opportunity to begin more collaborations in the future.

I would also like to thank other mentors who have been instrumental in my success in graduate school. I have appreciated the advice and mentorship of Jacque Eccles, who challenged me to think in new ways and has given me countless bits of advice for my research and career. I would also like to thank Mimi Engel and Amy Claessens for extending their branch of the "Greg Duncan family tree." Doug Clements and Julie Sarama were also instrumental in my success, as they openly shared their evaluation data with me, and encouraged me to pursue most of the projects that carried me through graduate school. I have also appreciated the help of George Farkas, Deborah Vandell, Peg Burchinal, Marianne Bitler, Emily Penner, Andrew Penner, Thad Domina, Bob Siegler, Pamela Davis-Kean, Dale Farran, Josh Lawrence and members of the P01 grant who influenced many of the research projects that led to this dissertation.

I must also recognize the contributions of the great friends that I have had during graduate school. I would not be at UCI without Ana Auger, who knew before I did that I belonged in the Education program. Her advice and friendship has been invaluable. I should also thank Jade Jenkins and Kim Pierce, who have become lifelong friends and true sources of professional help. I also owe thanks to Neil Young, Brian Jenkins, Tutrang Nguyen, Ryan Lewis, Chris Stillwell, NaYoung Hwang, Maureen Spanier, Robert Garcia, Kat Schenke and many other UCI students for their contributions to my work and for their personal friendship.

Finally, I would like to thank my parents for their unwavering support.

This research was supported by the Eunice Kennedy Shriver National Institute of Child Health & Human Development of the National Institutes of Health under Award Number P01HD065704. This work was also supported by the Institute of Education Sciences, U.S. Department of Education through Grants R305K05157 and R305A120813. The content is solely the responsibility of the author and does not necessarily represent the official views of the National Institutes of Health nor the U.S. Department of Education.

The text of Chapter 2 is a reprint of the material as it appears in *Child Development.* Greg Duncan, Douglas Clements and Julie Sarama were co-authors of this manuscript. It has been reprinted by permission of John Wiley & Sons, Inc.

CURRICULUM VITAE

Tyler W. Watts
School of Education
University of California - Irvine
3200 Education
Irvine, CA 92697-5500
Phone: 254/744.2008
E-mail: twatts@uci.edu
Website: tylerwwatts.wordpress.com

POSITIONS

2017 Assistant Professor of Research
New York University
Steinhardt School of Culture, Education, and Human Development
Applied Psychology (start date: June 2017)

EDUCATION

2017 Ph.D. in Education
University of California, Irvine

2015 M.A., Education
University of California, Irvine

2011 B.A., double major in Psychology and Religious Studies with high honors
University of Texas

PEER-REVIEWED PUBLICATIONS

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

Watts, T. W., Duncan, G. J., Clements, D. H., Sarama, J., Wolfe, C. B., & Spitler, M. E. (2017). What is the long-run impact of learning mathematics during preschool? *Child Development.* Advance online publication. doi:10.1111/cdev.12713

Schenke, K., Nguyen, T., **Watts, T.W.,** Clements, D. H., Sarama, J., Wolfe, C. B., & Spitler, M. E. (2017). Differential effects of the classroom on African American and non-African American's mathematics achievement. *Journal of Educational Psychology.* Advance online publication. doi: 10.1037/edu0000165

Watts, T. W., Clements, D. H., Sarama, J., Wolfe, C. B., Spitler, M. E., & Bailey, D. H. (2016). Does early mathematics intervention make lower-achieving children learn like higher-achieving children? *Journal of Research on Educational Effectiveness, 10*(1), 95-115. doi: 10.1080/19345747.2016.1204640

Engel, M., Claessens, A., **Watts, T. W.,** & Stone, S. (2016). Socioeconomic inequality at school entry: A cross-cohort comparison of families and schools. *Children and Youth Services Review, 71,* 227-232. doi: 10.1016/j.childyouth.2016.10.036

Engel, M., Claessens, A., **Watts, T. W.,** & Farkas, G (2016). Mathematics content coverage and student learning in kindergarten. *Educational Researcher, 45*(5), 293-300. doi: 10.3102/0013189X16656841

Nguyen, T., **Watts, T. W.,** Duncan, G. J., Clements, D. H., Sarama, J. S., Wolfe, C., & Spitler, M. E. (2016). Which preschool mathematics competencies are most predictive of fifth grade achievement? *Early Childhood Research Quarterly, 36*, 550-560. doi: 10.1016/j.ecresq.2016.02.003

Watts, T. W., Duncan, G. J., Chen, M., Claessens, A., Davis-Kean, P. E., Duckworth, K., Engel, M., Siegler, R. S., Susperreguy, M. I. (2015). The role of mediators in the development of longitudinal mathematics achievement associations. *Child Development, 86*(6), 1892-1907. doi: 10.1111/cdev.12416

Watts, T. W., Duncan, G. J., Siegler, R. S., & Davis-Kean, P. E. (2014). What's past is prologue: Relations between early mathematics knowledge and high school achievement. *Educational Researcher, 43*(7), 352-360. doi: 10.3102/0013189X14553660

Bailey, D. H., **Watts, T. W.,** Littlefield, A. K., & Geary, D. C. (2014). State and trait effects on individual differences in children's mathematical development. *Psychological Science, 25*(11), 2017-2026. doi: 10.1177/0956797614547539

Harte, C. B., **Watts, T. W.**, & Meston, C. M. (2013). Predictors of 1-, 6-and 12-month smoking cessation among a community-recruited sample of adult smokers in the United States. *Journal of Substance Use, 18*(5), 405-416.

INVITED CHAPTERS

Vandell, D.L., & **Watts, T.W.** (forthcoming). Self care. In M.H. Bornstein (Editor-in-Chief) M. Arterberry, J. E. Lansford, & K. L. Fingerman (Eds.), *The SAGE encyclopedia of lifespan human development (*pp. xx-xx). Thousand Oaks, CA: SAGE.

Vandell, D.L., Larson, R., Mahoney, J.L., & **Watts, T.W.** (2015). Children's Organized Activities. In R.M. Lerner (Series Ed.), M.H. Bornstein & T. Leventhal (Vol. Eds.), *Handbook of child psychology: Vol. 4. Ecological Settings and processes in developmental systems* (7th ed.). New York: Wiley

MANUSCRIPTS IN PROGRESS

Jenkins, J. M., **Watts, T. W.,** Magnuson, K., Gershoff, E., Clements, D. H., Sarama, J., Wolfe, C. B., & Spitler, M. E. *Preventing preschool fadeout through instructional intervention in kindergarten and first grade.* Manuscript under review.

Watts, T. W. *Will boosting test scores improve labor market outcomes?* Manuscript in progress.

Watts, T. W., *Evaluating the effects of a two-year individualized instruction intervention in mathematics.* Manuscript in progress.

CONFERENCE PRESENTATIONS

Watts, T. W., (April, 2017). *Revisiting the correlation between test scores and adult earnings.* Paper presented at the 2017 biennial meeting for the Society for Research in Child Development.

Bailey, D. H., Duncan, G., **Watts, T. W**., Clements, D. H., & Sarama, J. (September, 2016). *Risky Business: Correlation and Causation in Longitudinal Studies of Skill Development.* Invited talk, Conference of the International Mind, Behavior, and Education Society, Toronto, Canada.

Watts, T. W., Clements, D. H., Sarama, J., Wolfe, C. B., Spitler, M. E.., Bailey, D. (April, 2016). *Effects of an early mathematics intervention on stable and time-varying components of mathematics achievement.* Paper presented at the 2016 annual meeting for the American Educational Research Association.

Watts, T. W., Duncan, G. J, Clements, D. H., Sarama, J. (November, 2016). *What is the long-run impact of learning math during preschool?.* Paper presented at the 2015 annual meeting for the Association for Public Policy Analysis and Management.

Engel, M., Claessens, A., **Watts, T. W.,** & Farkas, G. (April, 2015). *The misalignment of kindergarten mathematics content.* Paper presented at the 2015 annual meeting for the American Educational Research Association.

Engel, M., Claessens, A., & **Watts, T. W.** (March, 2015). *Rising inequality at school entry: A cross cohort comparison.* Paper presented at the 2015 annual meeting for the American Educational Research Association.

Engel, M., Claessens, A., **Watts, T. W.,** & Farkas, G. (March, 2015). *The misalignment of kindergarten mathematics content.* Paper presented at the 2015 biennial meeting for the Society for Research in Child Development.

Engel, M., Claessens, A., & **Watts, T. W.** (March, 2015). *Rising inequality at school entry: A cross cohort comparison.* Paper presented at the 2015 biennial meeting for the Society for Research in Child Development.

Watts, T. W., Duncan, G. J., Clements, D. H., Sarama, J., Wolfe, C. B., & Spitler, M. E.. (March, 2015). *Preschool growth in mathematics and long-run achievement: An instrumental variables approach.* Paper presented at the 2015 annual meeting for the Society for Research on Educational Effectiveness.

Watts, T. W., Nguyen, T., Schenke, K., Duncan, G. J., Clements, D. H., Sarama, J., Wolfe, C. B., & Spitler, M. E.. (March, 2015). *Great expectations: The effect of teacher expectations on the mathematics achievement of African American students in a preschool mathematics intervention.* Paper presented at the 2015 annual meeting for the Society for Research on Educational Effectiveness.

Jenkins, J. M., **Watts, T. W.,** Magnuson, K., Gershoff, E., Clements, D. H., Sarama, J., Wolfe, C. B., & Spitler, M. E.. (March, 2015). *Preventing preschool fadeout through instructional intervention in kindergarten and first grade.* Paper presented at the 2015 annual meeting for the Society for Research on Educational Effectiveness.

Engel, M., Claessens, A., **Watts, T. W.,** & Farkas, G. (November, 2014). *The misalignment of kindergarten mathematics content.* Paper presented at the 2014 annual meeting for the Association for Public Policy Analysis and Management.

Engel, M., Claessens, A., & **Watts, T. W.** (November, 2014). *Rising inequality at school entry: A cross cohort comparison.* Paper presented at the 2014 annual meeting for the Association for Public Policy Analysis and Management.

Watts, T.W., Duncan, G. J. *The Groove of Growth: How Early Gains in Math Ability Influence Adolescent Achievement.* (March, 2014). Poster presented at the 2014 Spring meeting for the Society for Research on Educational Effectiveness.

Watts, T.W., Spanier, M., Duncan, G.J. *Predicting adolescent math achievement with preschool math skills.* (April, 2014). Paper presented at the 2014 annual meeting for the American Educational Research Association.

Duncan, G.J., Chen, M., Claessens, A., Davis-Kean, P.E., Duckworth, K., Engel, M., Siegler, R., Susperreguy, M.I., **Watts, T.W.,** (2013, April). *Self-concepts, school placements and executive functioning as mediators of links between early and later school achievement*. Paper presented at the biennial meeting of the Society for Research in Child Development. Seattle, WA.

Meston, C.M., **Watts, T.W.,** Stephenson, K.R, Lorenz, T.A., Pujols, Y.K., Pulverman, C.A., (2013, January)**.** *Mediators of the long term impact of childhood sexual abuse on sexual function: Stigmatization, powerlessness, and traumatic sexualization.* Poster session presented at the annual meeting of the International Society for the Study of Women's Sexual Health, San Diego, CA**.**

Watts, T.W., Harte, C.B., & Meston, C.M (2010, May). *Predictors of relapse following an 8-week smoking cessation intervention program.* Poster session presented at the annual meeting of the Association for Psychological Science, Boston, MA.

Harte, C.B., Perlman, **Watts, T.W.**, H., Bentkowski, O.Z., Whitaker, A., & Meston, C.M (2010, May). *Quitting smoking improves sexual health in men.* Poster session presented at the annual meeting of the Association for Psychological Science, Boston, MA.

Harte, C.B., **Watts, T.W.**, Perlman, H., Bentkowski, O.Z., Whitaker, A., & Meston, C.M (2010, April). *Effects of smoking cessation on sexual health in men.* Poster session presented at the annual meeting of the Society for Behavioral Medicine, Seattle, WA.

RESEARCH POSITIONS

NICHD PO1 Grant- Projects I and IV
School of Education
University of California, Irvine

Supervisors: Greg Duncan and George Farkas

2012~ present Title: Graduate Student Researcher

Duties: Oversaw numerous research projects investigating various hypotheses regarding long run processes in child development. Worked with a variety of datasets, including survey and panel data, and built new datasets from administrative records. Quantitative analyses have been published in academic journals and presented at multiple conferences.

NSF and IES Funded TRIAD Evaluation Study
University of Denver

Supervisors: Doug Clements

2014~ 2017 Title: Data Analyst

Duties: Conducted analyses focused on understanding why treatment impacts faded. Work included building usable data from variety of sources (e.g., surveys, school records, state test information, etc.). Implemented quantitative analyses using multiple regression, structural equation modeling, and instrumental variables. Presented work at multiple public policy and educational conferences.

Out of School Time Laboratory
School of Education
University of California, Irvine

Supervisor: Deborah Vandell, Ph.D.

2012~2013 Title: Graduate Student Researcher

Duties: Helped conduct literature review and wrote text regarding afterschool programs and out of school time for invited chapter in Handbook of Child Psychology.

TEACHING EXPERIENCE

2017 **Primary Instructor**, Introduction to Statistics

Average score of 4.65 (5-point scale) across all items on end of term evaluation

2016 **Teaching Assistant**, Multiple Regression (graduate course), Professor Greg Duncan

2012 **Teaching Assistant,** Adolescent Development, Professor Joseph Mahoney

INVITED LECTURES

2015 **"Common Core and Obama's Education Policy,"** Guest lecture in EDUC 50, Issues in K-12 Education.

SERVICE WORK

Journal referee	Psychological Science; Developmental Psychology; Applied Developmental Science; Learning and Individual Differences; Campbell Collaboration
Grant referee	Administration for Children and Families, US Department of Health and Human Services
UC, Irvine	Student Representative- Associated Doctoral Students of Education
	Student Mentor for the School of Education DECADE Program

RESEARCH SKILLS

- **Analytic Techniques**
  - Multiple Regression
  - Structural Equation Modeling
  - Experimental and Quasi-Experimental Methods
  - Hierarchical Models
  - Descriptive Statistics
- **Statistics Software Expertise**
  - Stata
  - Mplus
  - SPSS
  - HLM
- **Courses Taken in Research Methods**
  - Social and Behavioral Statistics
  - Qualitative Research Methods I and II
  - Structural Equation Modeling
  - Multiple Regression Analyses
  - Applied Econometrics I and II

- Hierarchical Linear Modeling

PROFESSIONAL MEMBERSHIP

American Educational Research Association
Society for Research on Educational Effectiveness
Association for Public Policy Analysis and Management
Society for Research in Child Development
Association for Psychological Science

ABSTRACT OF THE DISSERTATION

Academic Skills and Long-Run Outcomes

By

Tyler W. Watts

Doctor of Philosophy in Education

University of California, Irvine

Distinguished Professor Greg J. Duncan, Chair

Mathematics and reading skills are targeted by wide-ranging educational policies in hopes that boosting academic achievement will improve adult attainments. Relying on theories of skill building, researchers and policy-makers have pursued the idea that early gains in skills will lead to the acquisition of later skills, and this skill-building trajectory should lead to adult economic success. In this dissertation, I examined this long-run academic skill acquisition process in mathematics by investigating several approaches to promoting mathematics achievement during the preschool and elementary school years. I then turned to the hypothesis that boosting academic achievement during the schooling years should lead to greater economic success in adulthood.

In the first study, I investigated whether early math learning impacted later math achievement in a sample of children recruited for participation in a preschool mathematics intervention program. To generate causal estimates of the impact of early learning on later achievement, I leveraged random assignment to the preschool mathematics program as an instrument for gains in early math skills. I found some indication that instrumented gains in early math skills affected math achievement measured 6 to 7 years later, but estimates were smaller than had been reported in previous correlational studies. These findings suggested that

theories of skill building may have over predicted the long-run returns to early investments in mathematical skill development.

In the second study, I evaluated the effects of a 2-year intervention that encouraged second- and third-grade teachers to individualize instruction in mathematics. Individualized instruction is thought to help students at all achievement levels gain skills through tailoring instruction to the individual needs of each student. Results suggested that the intervention had little impact on math achievement at both grades assessed, and I found no impact of spending two consecutive years in individualized math instruction. However, teacher implementation was poor, suggesting that teachers may be resistant to programs that encourage them to differentiate instruction in mathematics.

The final study examined the link between adolescent achievement test scores and adult earnings. Although many studies have reported links between test scores and earnings (e.g.., Currie & Thomas, 2001; Murnane et al., 2000), most studies have only controlled for simple demographic characteristics (e.g., race, gender), leaving concerns that reported estimates might contain substantial bias. Using nationally representative data from the United Kingdom, I found that adolescent math and reading scores predicted adult earnings through age 50, but results were highly sensitive to the inclusion of a large set of controls (e.g., IQ, personality, parenting characteristics). Although fully-controlled estimates were still positive and significant, my results suggest that using the correlation between test scores and earnings to project educational program impacts may lead to biased predictions.

In final chapter, I discuss the implications of these findings for educational theory and policy. In particular, I suggest that theories of skill building need revision and more work is needed to understand the mechanisms that connect academic skills to important life outcomes.

Chapter 1

Introduction

The mathematics and reading achievement of the K-12 student population in the U.S. has received a great deal of research and policy attention in recent decades. Much of this attention stems from concerns that the U.S. has fallen behind its industrialized peers in academic skill levels (Hanushek, 2009; National Mathematics Advisory Panel, 2008; OECD, 2013), and many fear that this could put the U.S. at a competitive disadvantage economically as the global economy shifts toward sectors that require high levels of technical ability (Deming, 2015; Levy & Murnane, 2005). Moreover, policy-makers have recently become especially concerned with the mathematics achievement of children in K-12 schools, as the Obama administration made mathematics achievement a central piece of its educational policy initiatives.[1] Further, the most recent federal education law, which largely dismantled the No Child Left Behind Act, still required the yearly testing of mathematics and reading achievement in grades 3 through 8, and at least once in high school (Every Student Succeeds Act, 2015).

Longitudinal research suggests that efforts to promote academic skills, particularly during the early schooling years, should lead to long-run improvements in cognitive ability. Many well-controlled studies have shown that early measures of mathematics and reading ability strongly predict later measures of school achievement, even in the presence of a host of child and environmental control variables (e.g., Aunola, Leskinen, Lerkkanem, & Nurmi, 2004; Byrnes & Wasik, 2009; Claessens, Duncan & Engel, 2009; Duncan et al., 2007). These studies imply that if educational programs can raise the academic skills of children during the early grade years,

[1] For example, see the "Educate to Innovate" initiative, which focuses on fostering science, technology, engineering, and mathematics (STEM) skills, primarily in minority youth: https://www.whitehouse.gov/issues/education/k-12/educate-innovate

then such improvements should last throughout school. The path from early skill development to later achievement is thought to unfold through a skill-building process in which the acquisition of skills in an earlier time period increase the chance of further skill acquisition later (e.g., Cunha & Heckman, 2008).

Programs that raise math and reading skills are also hypothesized to have long-run impacts on children's eventual economic attainment. Correlational studies relying on large longitudinal datasets have reported that adolescent academic test scores strongly predict later earnings measured during adulthood (e.g., Currie & Thomas, 2001; Lin, Lutter and Ruhm 2016; Murnane et al., 2000). These studies have been viewed as evidence that if an academic program raises achievement scores, then through this impact on math and reading skills, the intervention should also affect economic outcomes (e.g., Krueger, 2003).

Thus, taken together, the research on skill-building (e.g., Cunha & Heckman, 2008; Duncan et al., 2007) and the literature detailing associations between achievement test scores and adult earnings (e.g., Murnane et al., 2000) paint a cohesive picture of the importance of promoting academic skills during childhood and adolescence. If we can identify interventions that successfully boost academic skills, particularly during the early years, then this investment should pay-off by enhancing the future economic attainment of affected children.

However, recent research has raised questions regarding which methods are most effective for boosting academic skills, and other questions remain about whether such programs will reliably improve later attainment. Although many studies have reported strong correlations between academic skills measured several years apart (e.g., Duncan et al., 2007), recent meta-analytic intervention evidence suggests that most early academic interventions produce effects on test scores that fade out within a few years of the intervention ending (McCoy et al., 2015).

This intervention evidence implies that correlational research reporting strong correlations between early academic skills and later skill attainment may contain substantial bias.

Perhaps the most illuminating example of the discrepancy between the correlational work on skill building and the more sobering intervention evidence comes from Clements and Sarama's *Building Blocks* scale-up evaluation (see Clements, Sarama, Spitler, Lange, & Wolfe, 2011; Clements, Sarama, Wolfe, & Spitler, 2013). They developed a highly successful preschool mathematics curriculum that was evaluated in a random assignment study implemented across 42 low-income elementary schools. Students assigned to the intervention group had much higher mathematics achievement scores at the end of preschool when compared with students in the control condition (Hedges $g = 0.71$). Using this same sample of children, Bailey, Duncan, Watts, Clements and Sarama (in press) reported that the end-of-preschool math score strongly predicted math achievement measured through grade 5. However, the end-of-preschool treatment impact fell by 60% by the end of first grade, and faded to 0 by the beginning of fourth grade. Thus, the correlation between early math achievement and later achievement suggested that an early math intervention should strongly boost achievement measured through grade 5. Yet, the intervention evidence suggested otherwise.

The discrepancy between the predictions of correlational models and the evidence from interventions has led some to argue that the instructional environment following an early intervention may be largely responsible for the observed fadeout (e.g., Stipek, 2017). This argument contends that early-grade teachers spend most of their time teaching content that is targeted below the needs of children who received high quality preschool instruction (e.g., Engel, Claessens, & Finch, 2013), allowing children from the control group to "catch-up" to treated children. One way to remedy this problem would be to alter the "one-size-fits-all" model of

instruction found in most early-grade classrooms (see Gregory & Chapman, 2013). Instead of teaching the same curriculum to all students, teachers could differentiate their content in order to meet the diverse needs of every student in the class. Indeed, Connor and colleagues (2013) found that an individualized instruction program in reading had large positive effects on the reading achievement of children in grades 1 through 3. However, questions remain as to whether this approach could work in other subjects, like mathematics, and other studies have found evidence that even high-quality subsequent instructional environments may not be able to abate fadeout effects (e.g., Claessens, Engel, & Curran, 2014; Jenkins et al., under review).

Nevertheless, even if we do find instructional methods that boost academic achievement throughout K-12 schooling, the path between academic skill attainment and adult outcomes also remains unsettled. Much of the literature that has reported associations between academic test scores and adult earnings has failed to attend to possible sources of omitted variables bias (e.g., Currie & Thomas, 2001; Murnane et al., 2000), leaving questions as to whether academic programs that impact achievement test scores will also boost adult earnings. If unobserved factors account for much of the reported association between test scores and earnings, then any intervention that affected test scores but failed to affect these unobserved factors would likely fail to have an impact on earnings.

In this dissertation, I attempt to address some of the gaps in the literature on the promotion of academic skills and the influence of these skills on long-run outcomes. In particular, the three studies described here address three distinct, yet related issues regarding academic achievement and long-run outcomes:

1. Do early skill gains in mathematics lead to further math skill acquisition during later periods?

2. Does differentiating instruction in mathematics boost student math achievement in early elementary school?
3. If an academic program boosts mathematics and reading skills, what impact might the program have on adult earnings?

Overview of Studies

Study 1: What is the long-run impact of learning mathematics during preschool? Mathematics is understood to be a particularly hierarchical subject, as basic competencies are needed to advance to more difficult concepts and procedures (Siegler, Thompson, & Schneider, 2011). This implies that early skill gains should translate to later skill development by giving children an early advantage in mathematical knowledge of foundational concepts and procedures, and correlational studies have found that early gains in math achievement strongly predict later achievement (e.g., Watts, Duncan, Siegler, & Davis-Kean, 2014).

In this study, I leveraged random assignment to a preschool mathematics program to test whether exogenously-produced variation in school-entry math test scores predicted math achievement measured in late elementary school. I found some indication that instrumented test score gains led to later achievement, but effects were much smaller than what had been reported in previous correlational studies, and the relation between early gains and later achievement was not consistently observed at every time-point tested. These findings imply that although early math skill gains may lead to later achievement, the relation between early skills and later achievement may be weaker than was previously thought.

Study 2: Evaluating the Effects of a Two-Year Individualized Instruction Intervention in Mathematics. In the second study, I investigated the effects of a mathematics intervention program that took a particularly promising approach to boosting early-grade math

achievement. This intervention attempted to individualize instruction by training second- and third-grade teachers to tailor their math teaching to the individual needs of each student. Many have hypothesized that such instructional efforts would greatly improve the achievement of children with varying ability levels (e.g., Gregory & Chapman, 2013), and individualized instruction has also been targeted as a possible approach for curbing fadeout effects after early academic intervention (e.g., Clements et al., 2013).

I found little indication that the program positively boosted math scores at either second or third grade. Across most models tested, results were largely null, though I found positive treatment effects on one of the math subtests given at the end of second grade ($\beta = 0.16$), and I found a negative treatment impact on one of the subtests given in third grade ($\beta = -0.11$). Finally, I found no effect of spending two consecutive years in the individualized math intervention program. Although these results suggest that individualizing instruction in mathematics may be an ineffective means of boosting math achievement, I found that program implementation by study teachers was poor. Thus, questions remain as to whether such approaches could be more successful with stronger buy-in from teachers.

Study 3: Will boosting test scores improve labor market outcomes? Many papers have found that adolescent achievement test scores predict early-career earnings (e.g., Currie & Thomas, 2001; Lin et al., 2016; Murnane et al., 2000). This correlation is often used to project what educational program impacts on earnings might be when researchers observe impacts on test scores but lack measures of earnings (e.g., Krueger, 2003). However, this projection may lead to erroneous predictions if the correlation between test scores and earnings contains bias due to unobserved factors that are unlikely to be affected by educational programs.

In this study, I evaluated whether estimates of the correlation between test scores and earnings were robust to the addition of a substantial number of control variables for important child and environmental characteristics that were left unconsidered by previous studies (e.g., personality, IQ, parenting characteristics). I found that math and reading scores measured at age 16 positively predicted earnings measured through age 50. However, these correlations were substantially reduced with the addition of control variables, and correlations from fully-controlled models were smaller than what had been reported in the previous literature. These results suggest that researchers should apply caution when projecting program impacts on earnings by using the test score and earnings correlation.

References

Aunola, K., Leskinen, E., Lerkkanen, M. K., & Nurmi, J. E. (2004). Developmental dynamics of math performance from preschool to grade 2. *Journal of Educational Psychology*, *96*(4), 699.

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

Byrnes, J. P., & Wasik, B. A. (2009). Factors predictive of mathematics achievement in kindergarten, first and third grades: An opportunity–propensity analysis. *Contemporary Educational Psychology*, *34*(2), 167-183.

Claessens, A., Duncan, G., & Engel, M. (2009). Kindergarten skills and fifth-grade achievement: Evidence from the ECLS-K. *Economics of Education Review*, *28*, 415–427. doi:10.1016/j.econedurev.2008.09.00

Claessens, A., Engel, M., & Curran, F.C. (2014). Academic content, student learning, and the persistence of preschool effects. *American Educational Research Journal*, *51*(2), 403-34.

Clements, D. H., Sarama, J., Spitler, M. E., Lange, A. A., & Wolfe, C. B. (2011). Mathematics learned by young children in an intervention based on learning trajectories: A large-scale cluster randomized trial. *Journal for Research in Mathematics Education*, *42*, 127-166.

Clements, D. H., Sarama, J., Wolfe, C. B., & Spitler, M. E. (2013). Longitudinal evaluation of a scale-up model for teaching mathematics with trajectories and technologies: Persistence of effects in the third year. *American Educational Research Journal*, *50*(4), 812-850.

Cunha, F., & Heckman, J. J. (2008). Formulating, identifying and estimating the technology of cognitive and noncognitive skill formation. *Journal of Human Resources, 43*(4), 738-782.

Currie, J. & Thomas, D. (2001). Early test scores, school quality and SES: Longrun effects on wage and employment outcomes. In S. W. Polachek (Ed.), *Worker wellbeing in a changing labor market*. Amsterdam: Elsevier Science.

Deming, D. J. (2015). *The growing importance of social skills in the labor market* (No. w21473). National Bureau of Economic Research.

Duncan, G. J., Dowsett, C. J., Claessens, A., Magnuson, K., Huston, A. C., Klebanov, P., et al. (2007). School readiness and later achievement. *Developmental Psychology, 43*, 1428-1446.

Engel, M., Claessens, A., & Finch, M. A. (2013). Teaching students what they already know? The (Mis) Alignment between mathematics instructional content and student knowledge in kindergarten. *Educational Evaluation and Policy Analysis, 35*, 157-178. doi:10.3102/0162373712461850

Every Student Succeeds Act, 20 U.S.C. §§ 1111 (2015).

Gregory, G. H., & Chapman, C. (2013). *Differentiated instructional strategies: One size doesn't fit all.* Thousand Oaks: Corwin.

Hanushek E A. (2009). The economic value of education and cognitive skills. *In: Handbook of Education Policy Research.* New York, NY: Routledge; 2009:39-56.

Jenkins, J. M., Watts, T. W., Magnuson, K., Gershoff, E., Clements, D. H., Sarama, J., Wolfe, C. B., & Spitler, M. E. (under review). *Preventing preschool fadeout through instructional intervention in kindergarten and first grade.*

Krueger, A. B. (2003). Economic considerations and class size. *The Economic Journal, 113*(485), F34-F63.

Levy, F., & Murnane, R. J. (2005). *The new division of labor: How computers are creating the next job market.* New York, NY: Russell Sage Foundation.

Lin, D., Lutter, R., & Ruhm, C. J. (2016). *Cognitive Performance and Labor Market Outcomes* (No. w22470). National Bureau of Economic Research.

McCoy, D. C., Yoshikawa, H., Ziol-Guest, K. M., Duncan, G., J., Schindler, H. S., Magnuson, K. A., & Yang, R. (2015). *Long-term impacts of early childhood education programs on graduation, special education placement, and grade retention: A meta-analysis.* Manuscript in preparation.

Murnane, R. J., Willett, J. B., Duhaldeborde, Y., & Tyler, J. H. (2000). How important are the cognitive skills of teenagers in predicting subsequent earnings? *Journal of Policy Analysis and Management*, *19*(4), 547-568.

National Mathematics Advisory Panel (2008). *Foundations for success: The final report of the National Mathematics Advisory Panel.* Washington D.C.: U.S. Department of Education.

OECD. (2013). *Survey of Adult Skills, First Results: United States.* Retrieved from http://www.oecd.org/site/piaac/Country%20note%20-%20United%20States.pdf

Siegler, R. S., Thompson, C. A., & Schneider, M. (2011). An integrated theory of whole number and fractions development. *Cognitive Psychology, 62*(4), 273-296.

Stipek, D. (2017, March 17). The preschool fade-out effect is not inevitable. *Education Week.* Retrieved from http://www.edweek.org/ew/articles/2017/03/17/the-preschool-fade-out-effect-is-not-inevitable.html?cmp=soc-edit-tw&intc=es

Watts, T. W., Duncan, G. J., Siegler, R. S., & Davis-Kean, P. E. (2014). What's past is prologue: Relations between early mathematics knowledge and high school achievement. *Educational Researcher*, *43*(7), 352-360.

Chapter 2

What is the long-run impact of learning mathematics during preschool?

Abstract

The current study estimated the causal links between preschool mathematics learning and late elementary school mathematics achievement, using variation in treatment assignment to an early mathematics intervention as an instrument for preschool mathematics change. Estimates indicate (n= 410) that a standard-deviation of intervention-produced change at age 4 is associated with a 0.24 standard deviation gain in achievement in late elementary school. This impact is approximately half the size of the association produced by correlational models relating later achievement to preschool math change, and is approximately 35% smaller than the effect reported by highly-controlled OLS regression models (Claessens et al., 2009; Watts et al., 2014) using national datasets. Implications for developmental theory and practice are discussed.

What is the long-run impact of learning mathematics during preschool?

An accumulating body of research suggests that early mathematical skills are critical to developing long-run success in school (Aunola, Leskinen, Lerkkanem, & Nurmi, 2004; Byrnes & Wasik, 2009; Claessens & Engel, 2013; Duncan et al., 2007; Geary, Hoard, Nugent, & Bailey, 2013; Jordan, Kaplan, Ramineni, & Locuniak, 2009; Stevenson & Newman, 1986; Watts, Duncan, Siegler & Davis-Kean, 2014). Among these studies, Duncan and colleagues' (2007) analysis of six longitudinal datasets provides the most robust evidence of strong associations between early and later mathematics achievement. Their investigation of school readiness skills asked a seemingly straight-forward question: if one examined a broad range of child skills and behaviors at school entry, and controlled for a host of child and family background characteristics, which characteristics would emerge as the strongest predictors of the child's eventual school achievement? Among the candidates investigated were academic competencies, attention problems, and internal and externalizing problem behaviors. Across the datasets, a consistent pattern emerged: mathematics achievement at school entry was the strongest predictor of later success in mathematics, and in some cases reading, even when all other characteristics tested were controlled. Since the publication of this study, other correlational studies have found similar results (Claessens, Duncan, & Engel, 2009; Claessens & Engel, 2013; Foster, 2010), including one that extended the outcome measurement into high school (Watts et al., 2014).

Developmental and cognitive theories predict that early mathematics knowledge is associated with later achievement because early numerical skills facilitate students' future mathematical skill acquisition (e.g. Aunola et al., 2004; Entwisle & Alexander, 1990; Gersten et al., 2009; Jordan et al., 2009). This skill-building framework rests on the idea that mathematics is a particularly hierarchical subject, in which mastery of simple concepts and procedures is

required for understanding more difficult mathematics. For example, solving even a simple algebraic equation would be impossible without knowledge of operations such as division and multiplication, and this operational knowledge depends on understanding the basic principles of counting. Relatedly, Siegler, Thompson, and Schneider (2011) describe how students gradually broaden the class of numbers that they understand as they progress through mathematics, with successful students moving from mastery of whole numbers in early grades to fractions in later elementary and middle school. Indeed, a well-developed body of empirical work documents the carefully-sequenced cognitive steps students take as they expand their understanding of numbers and mathematics (e.g. Booth & Siegler, 2006; Gilmore, McCarthy, & Spelke, 2007; Laski & Siegler, 2007; Opfer & Thompson, 2008; Sarama & Clements, 2009).

Beyond the cognitive skill-building framework lie other developmental reasons to expect that early success in mathematics would set children on a successful trajectory throughout school. Complex interactions between the child and her environment in the early schooling years are likely to leave long-lasting influences on the child's developmental trajectory (Bronfenbrenner & Morris, 2006). For example, high-achieving children in kindergarten are more likely to receive positive feedback regarding their academic proficiency from teachers, parents, and peers, which in turn may boost their perception of their own math competence (Bong & Skaalvik, 2003; Meisels, 1998). Relatedly, early mathematics achievement could be a gateway to higher-ability tracking in school, which would also support further academic development. Indeed, these pathways from early to later mathematics achievement have received empirical support, as evidence suggests that self-concepts and placement into gifted and talented programs both mediate the association between early and later mathematics (Watts et al., 2015).

From Level to Change in Early Mathematics

Much of the correlational evidence linking early and later mathematics ability is based on measures of early levels of math skills. Other studies show strong associations between early *gains* in mathematical ability and later success in school. For example, using longitudinal data, Watts and colleagues (2014) found that gains in mathematical skills during the first 2 years of school were more predictive of later achievement than were level-measures of school-entry skills. Moreover, early math gains were just as predictive of high school achievement as grade-3 math achievement, even after controlling for concurrent gains in other cognitive skills, such as working memory and reading achievement. Using nationally-representative data, Claessens et al. (2009) found that change in mathematics achievement across kindergarten was highly predictive of both fifth grade mathematics and reading achievement. Finally, using a growth-curve modeling approach, Jordan and colleagues (2009) found that change in number competence, measured six times in kindergarten and first grade, strongly predicted third-grade mathematics achievement.

Taken together, these studies suggest that the process of *learning* mathematics during the early-grade years may set students on a higher-achievement trajectory throughout their time in school. If the associations between early change and later achievement reported by these correlational studies approximate causal effects, then such long-run impacts could be expected from educational interventions that successfully promote early mathematics learning. Although past studies of early math change controlled for a host of child characteristics, including initial level of mathematics achievement, it is still unclear whether the regression-adjusted association between early change in mathematics and later achievement represents a causal effect. Here we ask: Do early mathematics gains produced by random assignment to an intervention predict later math achievement as strongly as the naturally-occurring gains used in past studies? If the

associations reported in past studies are driven by unobserved characteristics, such as interest, motivation, parental support for mathematics, or cognitive aptitude, then even highly successful early mathematics interventions may have no detectable impact on later achievement.

Indeed, experimental and observational studies suggest that the regression-adjusted associations reported by correlational research overstate the potential long-run impacts of early mathematics intervention. Bailey, Watts, Littlefield and Geary (2014) hypothesized that the stable correlation observed between measures of early mathematical ability and the sequence of later mathematics measures may be due to stable but unobserved factors that heavily influence mathematics achievement throughout development. Using a latent-factor state-trait model, they separated the variance in longitudinal measures of mathematics achievement into time-variant (state) and time-invariant (trait) components. They found that most of the variation in repeated measures of mathematics achievement was trait-like, as variation in individual differences in mathematics achievement were highly stable over time. Conversely, changes in any single measure of mathematics ability had relatively small effects on subsequent achievement scores once the stable variance was partitioned into a single, latent, factor. They concluded that correlational studies investigating the association between early and later measures of achievement fail to take into account the multitude of stable environmental and individual factors that likely influence achievement over time, and this omission leads to an overstatement of the importance of early measures of achievement on later measures.

Further, experimental evidence from intervention studies also suggests that long-run correlational models may not accurately represent causal impacts. *Building Blocks*, a preschool mathematics curriculum designed by Clements and Sarama (2008), was evaluated as part of a multi-site scale-up evaluation of an intervention model called TRIAD (Technology-enhanced,

Research-based, Instruction, Assessment, and professional Development; see Clements, Sarama, Spitler, Lange, & Wolfe, 2011; Clements, Sarama, Wolfe, & Spitler, 2013). In the TRIAD evaluation study, state-preschool programs were randomly assigned to either a curriculum implementation condition or a business-as-usual control condition. Although the intervention produced a large impact on mathematics achievement at the end of preschool (Hedge's $g = 0.72$), this effect faded by over 60% by the end of first grade (Clements et al., 2011; Clements et al., 2013). The fade-out pattern reported by Clements and colleagues resembles the results of a meta-analysis of early childhood education interventions (Leak et al., 2010), which found that most early interventions faded substantially in the years immediately following the end of treatment. Moreover, recent evidence from a large-scale middle childhood mathematics intervention has shown similar fadeout effects (Taylor, 2014).

Although these intervention findings dim hopes that producing gains in early mathematical skills might transform long-run academic trajectories, analysis of intervention effects do not directly test the causal returns of early skill gains. Even if an early intervention such as TRIAD produced a large boost in skills during the treatment period, estimates of the intervention's impact on later-grade math achievement would merely test the effect of being assigned to the treatment group on later achievement, not the effect of students' math skills gains across the treatment period. Further, traditional "treatment on the treated" analyses in such contexts test the effect of actually participating in the program on later outcomes, but this analysis still falls short of directly examining the long-run effects of growth in early skills.

If we want to understand how long-run developmental trajectories might be altered as a result of spurring early gains in academic skills, a different analytic approach is needed. To be effective, this approach would need to separate variation in early mathematics change from

sources of unobserved characteristics (e.g. child IQ, parental investment, interest) that might induce an upward bias in the estimated relationship between early skill gains and later achievement. Yet, unlike long-run analyses of intervention effects, this approach should also directly test the effect of early mathematics change on later achievement, not the effect of program participation, or assignment to a program, on later measures of math ability.

Current Study: Instrumental Variables

To obtain a causal estimate of the association between early mathematical skill change and later achievement, the current study employs instrumental variables (IV) techniques, which are widely used in applied econometric studies (see Angrist & Pischke, 2008; Murnane & Willett, 2010). IV methods have recently garnered considerable attention from developmental scientists; Gennetian, Magnuson, and Morris (2008) demonstrated the potential utility of the method for answering questions of theoretical importance in developmental psychology. Auger, Farkas, Burchinal, Duncan and Vandell (2014) employed IV for estimating the causal impact of childcare quality on later academic outcomes, and Crosby, Dowsett, Gennetian and Huston (2010) used IV to examine the impact of childcare type on child behavioral problems.

The intuition behind an IV approach is relatively simple: if the variation in a theoretically-interesting predictor variable can be purged of the portion of its variation that stems from unobserved factors (i.e. selection bias), then the "clean" variation left can be used to estimate a causal effect. To generate this clean variation, the observational dataset must contain a variable (i.e. instrument) that satisfies two conditions. First, the instrument must have a strong effect on the predictor of interest (in our case early math gains). Second, the instrument can only affect the eventual dependent variable of interest (in our case later-grade achievement) through the main predictor. In other words, the effect of the instrument on the dependent variable should

be completely mediated by the key endogenous predictor. Both requirements are essential to the success of the IV analysis, and finding instruments that satisfy these criteria in developmental research can be difficult (Gennetian et al., 2008).

In the current study, we seek to identify the causal impact of early mathematical skill change on later mathematics achievement. We test this causal relation by leveraging random assignment within the TRIAD scale-up evaluation as an instrument for preschool mathematics change. We then relate this "exogenously-produced" change (i.e., the change in mathematics learning that is only due to random assignment to the intervention, not other personal or environmental factors such as cognitive ability or parenting) to mathematics achievement measured in fourth and fifth grade. We chose the fourth and fifth grade outcome measures because they closely align with the time at which outcomes were measured in previous correlational work (e.g., Duncan et al., 2007) and because they were the most distal measures of mathematics achievement available in the data.

To produce exogenous variation in preschool math change, we take advantage of the fact that the Building Blocks intervention randomly assigned treatment to classrooms within clusters of preschools (called "blocking groups" and described below). The intuition behind our IV approach is that, to the extent that the relationship between early math change and later math achievement is causal, preschool clusters showing particularly large treatment impacts on math gains across the preschool year should also show larger-than-average impacts on later-grade achievement. The IV estimate is essentially the ratio of the later-grade impacts to early-gain impacts – both of which are produced by random assignment to treatment status. Mechanically, we use blocking group and treatment status interactions as instruments for early mathematics

change in a two-stage least squares (2SLS) model (e.g., Duncan, Morris, & Rodruigues, 2011). The 2SLS estimator is a common technique for IV analyses (see Murnane & Willett, 2010).

If the instrumental variable criteria mentioned above are satisfied (i.e., the instrument strongly predicts preschool math gains, and the instrument only affects later achievement through its effect on preschool math learning), then the 2SLS model should provide an unbiased estimate of the causal effect of preschool mathematics change on later mathematics achievement. Prior research leads us to hypothesize that early mathematics change will have a causal effect on later achievement, because early success in mathematics is likely to improve the chances of later mathematics achievement through both skill acquisition and other related personal and environmental processes (e.g. boosting positive self-concepts, placement in higher-achievement tracks in school). However, we expect that the causal impact will be smaller than the relations reported by correlational studies, as recent evidence suggests that omitted factors probably bias estimates of the association between early and later measures of mathematics achievement.

Method

Study Design

The design of the TRIAD scale-up evaluation is crucial to our analytic model. The intervention evaluation study researchers recruited 42 elementary schools with state-funded preschool programs serving low-income communities in New York and Massachusetts to participate in the evaluation, and they then grouped these schools into 8 blocks. The blocking groups were determined based on fourth-grade state-collected achievement test scores alone, and were not linked to district or other shared characteristics. This process was done to help ensure that schools in the treatment and control condition were balanced on unobserved characteristics (see Clements et al., 2011).

Within each block, schools were randomly assigned to one of three conditions: 1) control condition (business as usual); 2) *Building Blocks* curriculum during preschool only; 3) *Building Blocks* curriculum during preschool with extended pedagogical development (PD) in kindergarten and first grade. Schools assigned to either treatment condition (i.e., conditions 2 and 3) implemented the *Building Blocks* curriculum along with aspects of the TRIAD model that included PD and extensive instructional support (described below). Thus, the TRIAD evaluation study tested the success of the *Building Blocks* preschool curriculum in comparison with other preschool approaches to teaching mathematics, as students in the control condition still received mathematical instruction in their preschools (see Clements et al., 2011). As explained below, our analysis focuses just on the first and second groups.

The *Building Blocks* curriculum (Clements & Sarama, 2013), implemented during preschool, was based on theory and research on early childhood learning and teaching. The basic approach was finding the mathematics in, and developing mathematics from, children's activities by helping children extend and mathematize these activities. All components were based on learning trajectories for each core topic. First, empirically based models of children's thinking and learning were synthesized to create a developmental progression of levels of thinking in the goal domain that emphasized conceptual understanding, procedural skill, and problem solving competencies. Second, sets of activities were designed to engender those mental processes or actions hypothesized to move children through a developmental progression.

Preschool teachers working in schools assigned to either treatment condition attended 13 pedagogical development sessions over the course of two years. The PD sessions were designed to help teachers understand the developmentally-sequenced learning trajectories that form the basis of the *Building Blocks* curriculum, and teachers also learned the core mathematics

procedures and concepts for each topic. Teachers were also trained to use formative assessment and the *Building Blocks* software, called *Building Blocks Learning Trajectories* (BBLT). BBLT was an individually-paced program for students that was aligned with the curriculum and intended to provide additional instructional support. Finally, throughout the preschool year, teachers interacted with program mentors who offered instructional guidance and also asssessed the fidelity of implementation. Analyses showed that teachers taught the curriculum with adequate fidelity (mode and mean of 1, "agree" on -2 to +2 Likert scale) (see Clements & Sarama, 2011; Clements et al., 2011). On an observational instrument focused on mathematics, *Building Blocks*, compared to control, teachers had significantly higher scores on the classroom culture scale, total number of mathematics activities observed, and the number of computers on and working for students to use. However, there were no observed statistically significant differences in the number of minutes mathematics was taught (Clements et al, 2011).

The current study only considers children attending schools assigned to the preschool only treatment condition or control (school N= 30). Unfortunately, we were not able to use the alternative treatment condition in our current analyses as the requirements for a viable instrument (described below) were not met by this third condition. We describe our attempts to use the third, follow-on treatment arm in more detail in the Appendix.

The key component of our analyses, the instrumental variable, is derived by generating treatment by block interactions, which we then relate to preschool mathematics change. We use these interactions because we expect that some blocks were more successful at producing preschool mathematics change than others, and these block differences should produce more variation in intervention-caused preschool math learning. As explained above, our IV procedure

assesses whether blocks with the largest treatment-induced gains in early math also produced the largest impacts on measures of fourth and fifth grade achievement.

Data

We use data drawn from the TRIAD evaluation study, which randomly selected 880 students from the preschool classrooms of the schools assigned to either the preschool curriculum intervention or the control condition. Students' mathematical knowledge was assessed at the beginning and end of preschool, spring of kindergarten and first grade, fall and spring of grade four, and the spring of grade five. The current study relies on data collected during preschool and grades four and five. As described below, we employ two separate model specifications. The first group of models uses a balanced panel, which only includes students with non-missing test score data during preschool and grades four and five (subsequently referred to as the "grade-pooled" sample; n= 410). The second group of models considers students that had data on any of the respective follow-up measures (fall of fourth grade n= 469; spring of fourth grade n= 543; spring of fifth grade n= 502). The missing cases in the grade-pooled sample are missing due to study-attrition. Of the baseline characteristics assessed, only free or reduced price lunch (FRPL) status contains any non-response (approximately 20%), and non-response was not related to treatment status (p= 0.30). In the regression models that follow, FRPL was included as a covariate, and missing cases were set to 0. A dummy variable was then included in each regression indicating whether an observation had missing data on the FRPL indicator.

Table 2.1 presents sample characteristics for participants in the full sample, grade-pooled sample, treatment, and control. As Table 2.1 reflects, half of the students recruited for participation in preschool were African American, 23% were Hispanic, and 21% were White.

Further, 85% of the sample qualified for FRPL (only the 773 non-missing cases were considered here). Students who are included in the pooled sample, and thus, did not leave the study in the later rounds of data collection, were more likely to attend a New York school ($p < 0.001$). They were also more likely to be Hispanic ($p= 0.063$) and less likely to be White ($p = 0.027$). However, students in the pooled sample did not statistically significantly differ on the preschool entry test, and were not more or less likely to be in the treatment or control group.

[Insert Table 2.1]

A comparison of columns 4 and 5 from Table 2.1 shows that treatment and control groups were balanced on baseline observable characteristics, as no statistically significant differences were detected between the two groups.

Measures

Mathematics achievement. During preschool, mathematics achievement was assessed at the beginning and end of the preschool year with the Research-based Early Math Assessment (REMA; Clements, Sarama, & Liu, 2008; Clements, Sarama, & Wolfe, 2011). The REMA was designed specifically for use with children ages 3 through 8, and it was administered through two one-on-one interviews with a trained administrator. The test was administered in two sections: number and geometry. Topics found on the number portion of the exam included counting, subitizing, number sequencing, cardinality, number composition and decomposition, place value and adding and subtracting. Topics on the geometry part of the exam included shape recognition, congruence, measurement, patterning, and shape composition and decomposition.

The REMA included 225 items that were ordered according to difficulty. The study administrator stopped the exam once a student incorrectly answered 4 consecutive items. The testing process was videotaped and subsequently coded for correctness and strategy use.

Approximately 10% of the assessments were double coded, and assessors and coders were blind to study condition. The REMA scores were then converted to Rasch-IRT scores to account for random guessing and item difficulty. The measure was validated in three diverse samples of young children, and it has been shown to have a 0.86 correlation with the Child Math Assessment: Preschool (see Clements et al., 2008), a .74 correlation with the *Applied Problems* subtest of the Woodcock Johnson III (see Weiland et al., 2012), and strong internal reliability (Cronbach's $\alpha = 0.94$; see Clements et al., 2008). The REMA was also administered in the spring of kindergarten and first grade. The current study employs both the standardized Rasch-IRT scores and simple raw-counts of the number of items correctly answered (subsequently referred to as "raw scores").

During the fall and spring of grade 4 and spring of grade 5, an extension of the REMA, called the TEAM 3-5, was administered (Clements, Sarama, Khasanova, & Van Dine, 2012). The TEAM 3-5 is a paper-and-pencil assessment that can be administered in a group setting. It is aligned with the developmental progressions as the REMA although some topics are "retired" (e.g., simple counting, subitizing, shape recognition) while others, similarly drawn from research-based developmental progressions (see Maloney, Confrey, & Nguyen, in press; Wilson, Mojica, & Confrey, 2013) are introduced or receive greater emphasis (e.g., multiplication and division, fractions and decimals, measurement of area and volume, coordinate systems, and more sophisticated analysis of geometric shapes). In the current sample, the TEAM 3-5 was found to have good internal reliability (Cronbach's $\alpha = 0.91$). Further, correlations between the assessment and state grade-5 achievement tests in New York ($r(351) = 0.82$, $p < 0.001$) and Massachusetts ($r(110) = 0.76$, $p < 0.001$) were high for the subset of students for which state tests

were available (approximately 40% of the full sample). As with the REMA, the TEAM 3-5 was also converted to a standardized Rasch-IRT score.

The key measure in the study, mathematics change, was constructed by taking the simple difference between the standardized post-preschool IRT-scored REMA and the standardized preschool entry IRT-scored REMA. Thus, model coefficients should be interpreted as "a standard deviation of change," which makes the effects most comparable to effect sizes reported in both intervention and correlational literature. However, because IRT scores can be difficult to interpret, we have also calculated a simple measure of the change in the raw number of items correctly answered on the pre- and posttests. When considering this measure in comparison with the IRT scores, recall that the IRT score takes into account correctness, as well as strategy use and item difficulty. Thus, the raw scores reflect a much simpler, and less comprehensive, measure of mathematics knowledge that do not have the characteristics of measurement that the IRT scores possess.

Table 2.2 presents descriptive statistics for both IRT-scaled and raw score measures of the pretest, posttest, and change measure for both the treatment and control groups. On average, students in the treatment group correctly answered approximately 11 items on the pretest, and students in the control group answered 12 items, a statistically non-significant difference ($p = 0.526$). By the end of preschool, students in the treatment group correctly answered approximately 21 more questions than on the pretest measure, and students in the control group correctly answered roughly 16 more items than on the pretest ($p < 0.01$). Thus, both groups grew substantially in their mathematics knowledge. The standardized IRT scores also reflect the substantial change students made in both the treatment and control groups. The REMA IRT scores were standardized to have a mean of zero at approximately first grade, thus the change

from an average score of -3.25 for the treatment group at pretest to a score of -1.87 at the posttest reflects positive growth toward the normed first grade mean.

[Insert Table 2.2]

Covariates. Information regarding child ethnicity, gender, age, limited English proficiency, special education status, and FRPL status were collected at baseline from the study schools' administrative data. The measures are included as controls in the following analyses.

IV Model

We used a two-stage least-squares (2SLS) modeling procedure in Stata 13.0 to estimate the causal effect of preschool mathematical skill change on later mathematics achievement. In the first stage regression, we regressed our key predictor, preschool mathematics change, on treatment status, blocking group, preschool-entry mathematics achievement, baseline measures of student characteristics, and, most importantly, the interaction between treatment status and blocking group. The resulting equation for the i^{th} child in the j^{th} block is as follows:

1. $$MathChange_{ij} = a_1 + \beta_1 Tx_{ij} + \sum_{j=1}^{8} \beta_{2j} Block_i + \beta_3 Block_i * Tx_{ij} + \beta_4 EntryMath_{ij} + \beta_5 Covariates_{ij} + e_{ij}$$

where $MathChange_{ij}$ is the post-test math score subtracted from the pre-test math score of the i^{th} student in the j^{th} block, and the instruments are represented by the treatment dummy variable (Tx_{ij}) and the treatment and block interactions ($Block * Tx_{ij}$). The use of interactions between random-assignment design characteristics (such as site) and treatment status as instruments has been used in other quasi-experimental studies of educational settings (Auger et al. 2014, Duncan et al., 2011; Taylor, 2014). The second stage regression, which estimated the impact of preschool math change on later achievement, then used the predicted values for preschool math change generated in the first equation:

2. $MathAchievement_{ijt} = a_1 + \theta_1 PredictedMathChange_{ij} + \sum_{j=1}^{8} \theta_{2j} Block_i + \theta_3 EntryMath_{ij} + \theta_4 Covariates_{ij} + z_{ij}$

where $MathAchievement_{ijt}$ represents the math achievement test score for the i^{th} child, in blocking group *j*, at time *t* (either fall or spring of fourth grade, or spring of fifth grade). In this equation, the instruments from the first equation (treatment status, and treatment and block interactions) do not appear, and θ_1 represents the causal impact of preschool mathematical skill change on later achievement. If the key IV assumptions described below are satisfied, then z_{ij}, the error term, should only represent random shocks, and should not include the sources of omitted variable bias that typically plague correlational models.

Whenever IV methods are employed, the instrumented parameter of interest (θ_1 in *Equation 2*) should be interpreted as the local average treatment effect (LATE), where "local" describes compliant students (see Angrist & Pischke, 2008; Murnane & Willett, 2010). In other words, IV methods only identify the effect for participants who were compelled to participate in the treatment based on random assignment. In our setting, this means that we identify the effect of preschool mathematics change for students who grew in mathematics only as a result of random assignment to the treatment.

As described in more detail below, we estimated separate 2SLS models for fall and spring of fourth grade, and spring of fifth grade measures of mathematics achievement, respectively. However, we also estimated models in which we pooled mathematics achievement scores across these three grades. All models presented included robust standard errors that were adjusted for clustering at the school level.

Correlations between instruments and mathematics change. To be effective in an IV analysis, an instrument must have a strong effect on the endogenous predictor variable. In this

case, the treatment by block interactions need to produce enough variation to reliably predict mathematics change in *Equation 1*. Indeed, in the intervention considered here, the treatment was specifically designed to affect mathematics change during the preschool year. However, some blocks may have been more successful at this goal than other blocks. To assess the correlation between the instruments and preschool mathematics change, we ran a regression predicting our key measure of preschool mathematics change on baseline characteristics (including preschool-entry mathematics score), block and treatment dummies, and interactions between treatment and block. With standard errors adjusted for clustering at the school level, the joint-test for the set of treatment and block interactions produced a large-enough F-statistic ($F(8) = 41.46$, $p < 0.001$) to confidently conduct 2SLS analyses, as an F-statistic of 10 is usually considered the threshold for an effective instrument (e.g. Angrist & Pischke, 2008). Column 1 of Table 2.3 displays the coefficients produced by this model, including the block and treatment interactions. Block 5 was omitted from the regression as the comparison group, as this was the block with the most students (n=162). In this model, the treatment had a large main effect ($\beta = 0.699$, $SE = 0.138$), and some blocks produced positive interactions with treatment status, while others produced negative coefficients. This indicates considerable variability between blocks on the effect of the treatment on mathematics change.

Exclusion restriction. To produce only exogenous variation in the endogenous predictor, the instrument should not be correlated with the error term in *Equation 2*. In other words, the instrument should not have an effect on the dependent variable (late elementary school mathematics achievement) except through the endogenous predictor (preschool mathematics change).

Theoretically, this should be the case in the current analysis. The model by which the intervention was designed conceptualizes the impact of the intervention on elementary school mathematics achievement through a skill-building framework that hinges upon gains made in preschool mathematics achievement (Clements et al., 2011; 2013). Thus, future mathematical skill production relies on the mathematics skills children carry at the end of preschool, as the preschool mathematical competencies allow them to learn and master new, more difficult, material. Further, we found no differences in baseline observables between the treatment and control groups (see Table 2.1), indicating that at baseline the treatment group was not advantaged in a way that would have improved their chances of becoming high achievers later on.

However, it is possible that the intervention could have affected later elementary school mathematics achievement through other mechanisms, such as boosts in language skills, motivation or executive functioning. Further, treatment students could have been sorted into higher quality classrooms after preschool, which could have, in turn, boosted their later mathematics achievement. Our data include observational measures of classroom instructional quality from the children's kindergarten and first grade classrooms (observations were recorded for approximately 73% of the current analysis sample; see Clements et al., 2013 for full description of the observational measure). We found no indication that treatment status was correlated with kindergarten or first grade instructional quality. We also found that treatment status was not related to the likelihood of staying in the same school through kindergarten, first grade, or fifth grade.

Unfortunately, we lack the broad measures of child characteristics needed to rule out unexpected changes in child functioning due to the preschool mathematics intervention. However, language skills were measured at the beginning of the kindergarten year, and Sarama

and colleagues (2012) reported a standardized statistically significant treatment impact of approximately 0.10 on the measure of language achievement (measure included the ability to recall key words, use of complex utterances, willingness to reproduce narratives independently, and inferential reasoning). We tested whether this boost in language skills could bias our models by running our primary OLS and IV models with, and without, the kindergarten entry language score. Including the language measure did not change our estimates (results shown in the Appendix), indicating that although the treatment impacted language functioning, this boost in language did not affect later mathematics achievement.

Given that the intervention was only the implementation of a preschool mathematics curriculum (that ran for approximately 15 minutes per day; Clements et al., 2011), not a global program targeted at a wide array of socio-emotional and cognitive skills, it seems most plausible that the primary mechanism through which the intervention affected students was through preschool mathematical skill development. Still, we cannot rule out whether the treatment might have caused changes in unobserved child characteristics, such as motivation or executive functioning. In both cases, changes in these unobserved skills could bias our estimates if boosts in these skills also impacted later mathematics achievement. Previous correlational studies that have examined relations between mathematics achievement and various socio-emotional and cognitive skills suggest that any likely bias-causing candidate would probably have a small effect on our model (e.g. Claessens & Engel, 2013; Duncan et al., 2007; Jordan et al., 2009; Watts et al., 2014). Nevertheless, if such biases were present in our models, they would likely have positive correlations with later mathematics achievement and preschool change, and would then bias our key estimate in an upward direction. Because we lack the measures to totally rule out this potential threat, our findings should be considered upper-bound estimates of the causal

relation between preschool mathematical skill change and later mathematics achievement.

Grade-Pooled Estimates

In the analyses that follow, we rely primarily on estimates generated from a grade-pooled dataset. In these models, we pooled observations across the fall and spring of fourth grade and the spring of fifth grade, such that each student was observed three times, and students were only included in this sample if they had non-missing data on both fourth grade measures and the fifth grade test (n= 410). We chose this path for two reasons. First, IV models typically generate relatively large standard errors, because IV models depend only on variation produced by the instruments, and thus have less variation with which to produce estimates (Angrist & Pischke, 2008). Thus, to generate precise estimates, more statistical power is required.

Second, this model is justified by the high correlations between the fourth and fifth grade test scores, as these measures each had an average correlation of 0.84. Further, after pooling across grades, we regressed fall of fourth grade, spring of fourth grade, and spring of fifth grade mathematics achievement on preschool mathematics change and covariates. In this model, we included dummies for grade level and interactions between grade and change. This set of interactions, which test whether the relation between change and later achievement differs between grade levels, were jointly not statistically significantly different from 0 ($F(2)= 0.50$, $p = 0.610$).

However, because the impact of preschool change on achievement at different grade levels is of theoretical interest, we also present models that were estimated using non-pooled data. In these models, fall and spring of fourth grade and spring of fifth grade achievement were each regressed independently on instrumented-preschool mathematics change.

Results

We begin with results from OLS models in which we regressed our later measures of mathematics achievement (fall and spring of fourth grade and spring of fifth grade) on preschool mathematics change, preschool entry mathematics achievement, and other baseline characteristics. Columns 2 through 4 of Table 2.3 presents results from non-pooled, OLS, models in which we examined the relation between preschool mathematics change and fourth and fifth grade mathematics achievement, respectively. Key independent and dependent variables were standardized, and all models presented included the full list of control variables (correlations for all predictor variables are shown in the Appendix). Columns 2 through 4 show the relatively stable predictive relation between preschool mathematics change and later achievement, as a standard deviation of change had approximately a one-half standard deviation effect on fall and spring fourth- and spring of fifth-grade achievement. The effects reported in columns 2 through 4 are larger than the OLS-adjusted effects of early mathematical skill change reported by Claessens et al. (2009) and Watts et al. (2014), as their studies produced standardized effects of approximately 0.35. This discrepancy probably reflects the greater availability of cognitive control measures available in the datasets employed by those studies.

[Insert Table 2.3]

Grade-Pooled IV Estimates

Next, we turn to estimates generated from pooled models that used block and treatment interactions as instruments for preschool mathematics change. Recall that in the pooled models, each student's fourth and fifth grade tests were considered as separate observations in one model. In each model, standard errors were adjusted for school-level clustering, but we also tested models that adjusted for student-level clustering to account for the panel structure of the dataset, and results did not qualitatively differ.

In column 1 of Table 2.4, we begin with the reduced form estimates, which show the effect of the instrument on the eventual outcome variable of interest. In our study, the reduced form model can be interpreted as a basic treatment impact model, as we show the average treatment impact of random assignment to the TRIAD intervention on mathematics achievement in fall and spring of 4th grade and spring of 5th grade. Across the grades, the average treatment impact was positive, but not significant (β= 0.094, *SE*= 0.064, p = 0.154). However, our IV results suggest that the simple treatment impact estimate masks the effect of treatment-induced change in mathematics on later achievement.

[Insert Table 2.4]

For purposes of comparison, column 2 of Table 2.4 presents grade-pooled OLS results comparable with the estimates displayed in columns 2 through 4 of Table 2.3, as a standard deviation of preschool mathematics change was related to a 0.535 standard deviation gain in later mathematics achievement (*SE*= 0.044, p < 0.001). Column 3 displays the 2SLS-estimated (instrumental variables) impact of standardized mathematics change on later achievement with only site, blocking group, and preschool entry math score controlled. In this model, the effect fell by over 50% when compared with the OLS models, though the estimate was still substantively and statistically significant (β= 0.236, *SE*= 0.113, p = 0.037). In column 4, we added the full list of background characteristics, and the coefficient was nearly unchanged, though the standard error fell, reflecting the control variables' added utility for increasing precision (β= 0.242, *SE*= 0.081, p = 0.003). The lack of change in the coefficient on preschool change after the addition of these control variables provides some degree of confidence that the exclusion restriction assumption is fairly safe in our models, as this indicates that the relation

between instrument-produced change and later achievement was not correlated with baseline observables.

Additional Models

Column 5 through 7 present 2SLS estimates generated from non-pooled models in which every student was only observed one time, and the fall of fourth grade, spring of fourth grade, and spring of fifth grade scores were considered individually. We present these models because they can provide theoretically interesting information regarding whether the relation between exogenously-produced mathematics change and later achievement may differ by grade. However, we hesitate to draw strong inferences based on these models because our sample sizes drop considerably in each of them, and this limits our ability to generate precise estimates when using IV (see Angrist & Pischke, 2008). Thus, these models merely inform the primary estimates presented in columns 3 and 4 of Table 2.4, but drawing strong conclusions based solely on these models would be inadvisable.

As columns 5 and 7 demonstrate, the significant and positive effect detected in the pooled models was not found in models relating change to either measure of fourth grade achievement. Although the fall of fourth grade model presented a positive coefficient with a large standard error (β= 0.132, *SE*= 0.109, p = 0.223), the spring of fourth grade model produced a coefficient of nearly zero (β= 0.039, *SE*= 0.096, p = 0.683). However, we were surprised to find that preschool math change strongly predicted fifth grade mathematics achievement in our disaggregated IV model (β= 0.257, *SE*= 0.079, p = 0.001). It would seem that the fifth grade effect was largely driving the positive grade-pooled estimate, as the grade-pooled estimate roughly represents an average of the three disaggregated effects.

In the Appendix, we present results from additional analyses in which we estimated our grade-pooled IV model in only key subgroups (i.e., African Americans, Limited English Proficient students, high and low achieving students, FRPL students). Across all groups we found positive effects within the confidence interval of our key estimate shown in column 3 of Table 2.4. We found the largest effect for African American children ($\beta = 0.379$, $SE = 0.104$, $p < 0.001$), but we did not find that this effect was statistically significantly different from the effect for Non-African American students ($p = 0.150$).

In analyses presented in the Appendix, we also tested the sensitivity of our primary findings to various model specifications. As mentioned above, we tested whether controlling for kindergarten measures of language and literacy skills changed our results, and we found no indication that our models were affected by these measures. Further, we examined models that did not control for baseline mathematics achievement, and found that this did not substantively change our estimates. Next, we tested whether controlling for grade level changed the grade-pooled IV estimates, and again found that our results were robust to this specification. Finally, we tested whether changing our IV estimation procedures affected our results. We found that using only the single treatment status indicator as an instrument produced a positive, marginally statistically significant, coefficient of 0.154 ($SE = 0.094$, $p = 0.104$), and using "limited information maximum likelihood" IV estimator instead of the 2SLS estimator produced a coefficient quite similar to the one reported in column 3 of Table 2.4.

Discussion

The current study tested the extent to which learning mathematics during preschool improves mathematics achievement in late elementary school. We leveraged variation in preschool learning produced by a preschool mathematics intervention to generate causal

estimates of the impact of gains in preschool mathematics knowledge. In our main models, we found that a 1-SD boost in preschool math learning produced approximately a quarter-SD gain in late elementary school achievement. However, we were surprised that this relation was only detected between preschool math learning and fifth grade achievement, and we found no such association between preschool gains and fourth grade achievement.

Taken together, these results lead us to make two primary conclusions. First, correlational approaches to questions regarding longitudinal achievement patterns should be approached with great caution. Second, early learning does not appear to be an "inoculation" that necessarily produces later achievement gains, and consequently, theories regarding skill-building processes probably require some amount of revision.

Comparisons with Correlational Literature

Our results suggest that the correlational literature, based primarily of OLS models that controlled for a host of family and child background characteristics, probably overstated the long-run effects of preschool mathematics achievement. When compared with OLS models estimated in the current study, the IV models reduced the effect of preschool change on later mathematics achievement by nearly 50%. When considered alongside the intervention literature, perhaps this finding should not be surprising, as preschool interventions often show steady fadeout patterns as time after the end of treatment elapses. Yet, why did the correlational literature fail to predict the modesty of the causal relation between early math skill gains and later achievement?

The answer could simply be that it is nearly impossible to control for all of the potential confounds between early and later test scores. Indeed, previous correlational investigations (Claessens et al., 2009; Watts et al., 2014) included a wide array of cognitive, academic, and

socio-emotional skills not included in our study, but these controls apparently failed to account for all of the underlying sources of bias. Watts and colleagues (2014) even controlled for *gains* in reading achievement and domain-general cognitive skills, and still found a 1-SD gain in early math achievement was associated with a 0.37 SD boost in late elementary school achievement. When compared with our grade-pooled models, these estimates are approximately 35% larger than the 0.24-SD effect that we found using instrumental variables (it should be noted that the 95% confidence interval for our primary grade-pooled model ranged from 0.08 to 0.39).

Further, compared with previous examinations, we did not find the IV-produced effect of preschool change to be consistent across grades, as we found no evidence of a strong relation between change and achievement in fourth grade, but we detected a substantial link between change and achievement in fifth grade. Certainly, the developmental period over which change was measured should be considered when drawing such comparisons, as Claessens and colleagues measured mathematical skill change during kindergarten, and Watts et al. measured mathematics change from preschool through the end of first grade. It is possible that change during kindergarten or first grade could be a stronger predictor of later achievement than change during preschool. Yet, given that we found a comparably large, OLS-adjusted, relation between preschool change and later achievement, we find it unlikely that this difference accounts entirely for the discrepancies between our IV estimates and the associations reported in previous correlational research.

If previous correlational models simply lacked the necessary set of controls, what factors might need to be controlled if correlational models stand a chance of replicating causal estimates? Indeed, future work should seek to find the set of measures that can fully reduce bias in analyses of longitudinal academic achievement data, and it is likely that such measures would

need to include indicators of a wide variety of environmental and personal characteristics that could influence the development of math achievement over time. However, a few recent investigations also demonstrate that alternative approaches to modeling correlational data may provide a more productive path forward. Bailey and colleagues (2014) found that a state/trait model, which accounted for omitted-variables bias by modeling the stable variation present in repeated measures of mathematics achievement as a single, latent factor, substantially reduced the predictive relation between gains in an early measure of math ability and later measures of achievement.

Alternatively, the current paper provides another possible approach for generating more accurate causal predictions. If researchers can find instruments that satisfy the criteria described above, then such analyses could better improve our understanding of many developmental processes, as this approach is not necessarily limited to investigations of cognitive and academic development. Finding viable instruments is no easy task, but other quasi-experimental approaches can also provide more robust causal estimates (see Murnane and Willett (2010) for an approachable review of a variety of quasi-experimental methods). For example, Cortes and Goodman (2014) found that students who were approximately randomly assigned to an extra mathematics course in high school (generated from a regression discontinuity in assignment based on prior-year math scores) had higher graduation rates and were more likely to attend college. Such findings provide robust causal evidence of the possible benefits of mathematics education, and offer an important test of developmental theories that would predict better outcomes for students with enhanced math learning opportunities. Thus, although quasi-experimental methods may be difficult to pursue, the benefits of generating more accurate causal estimates should make such efforts worthwhile.

Implications for Developmental Theory and Practice

Our most surprising result, perhaps, was that we found a strong impact of instrument-produced change on fifth grade mathematics achievement, but we found no impact on achievement in our two fourth grade measures of math ability. We did not hypothesize this pattern of results, and because these models were less precisely estimated than our grade-pooled models, we do not wish to overstate these findings. Nevertheless, when considering what processes might have given rise to these results, recall that the same test was administered at both fourth grade measurement points and at the spring of fifth grade measurement point. Thus, changes in the measure should not account for differences in the pattern of findings. However, it is likely that the curriculum students encountered in school changed substantially between the fourth and fifth grade years. During the fifth grade year, the schools in Massachusetts and New York both switched to the Common Core Standards, which emphasizes conceptual understanding of mathematics (Common Core Standards Initiative, 2010). Further, it has been argued that this shift toward conceptually-focused math would especially alter the way math was taught in low-income schools (Schmidt & Burroughs, 2013).

It is quite possible that the knowledge gained from the intervention during preschool only benefited students once the more conceptually-rich content was emphasized in fifth grade. Certainly, this finding warrants further investigation and replication before major conclusions can be drawn. Yet, it should be noted that even if preschool math change only positively impacted mathematics achievement in fifth grade, but not fourth grade, then this finding strongly contradicts the predictions made by correlational models. Previous studies (e.g., Duncan et al., 2007; Claessens & Engel, 2013; Watts et al., 2014) have all reported stable relations between early mathematics achievement and later measures of achievement, no matter when the

dependent variable was measured. Indeed, these findings led previous studies to predict that early intervention efforts would have stable long-run effects (Duncan et al., 2007; Watts et al., 2014). Our findings suggest that this is not likely to be the case.

Our pattern of results has implications for developmental theory. If our fifth grade finding is found to be robust to replication, then this would suggest that skill-building processes do not necessarily unfold in a monotonic manner. In other words, early math skills might not reliably lead to the development of later mathematical knowledge across all settings. Rather, early mathematical knowledge may only lead to the production of later knowledge when this early knowledge base is paired with the correct mix of content and teaching. This suggests that subsequent environments play a critical role in sustaining cognitive development in the wake of early investments in cognitive skills. This also suggests that skill-building theories that predict that early knowledge gains will necessarily lead to advantages in later achievement (e.g., Cunha & Heckman, 2008) may need some revision, as our results imply that skill development may be a more complex process that relies on many factors other than the mere possession of early skill advantages.

However, we also wish to underscore that our preferred estimates, the grade-pooled models, suggested that intervention-spurred early gains in mathematics led to approximately a fifth of a SD gain in mathematics across fourth and fifth grade. This implies that early skill gains do matter for developing long-run achievement trajectories. Although the effect was not as large as was previously predicted by correlational work (e.g., Duncan et al., 2007), our results do demonstrate the long-run utility of early skills advantages. When considering what these results imply for developmental theory and practice, we should recall the "LATE" interpretation of instrumental variables results (see Angrist & Pischke, 2009). Instrumental variables techniques

identify effects for the "complier" population within the sample. In our study, compliers are students who responded to the intervention, and gained in mathematics knowledge as a result of participation in the program. This is perhaps intuitive, as this means that we identified the effect of early math gains for students that, for whatever reason, were able to particularly benefit from participation in *Building Blocks.* Understanding what types of students respond best to early academic programs, like *Building Blocks,* presents a promising avenue for further research, as it opens the door for targeting programs toward students that might stand to benefit the most from early cognitive investments.

Although our results imply that early gains in mathematics ability should lead to moderate advantages in math achievement later in elementary school, for interventions, it is important to consider the amount of change that would be required of a program to replicate the effect reported here. For an intervention effect to produce a 1-sd end-of-treatment effect on mathematics gains, students in the treatment group would need to gain a full standard deviation *more* in mathematics achievement than students in the control group. Although our raw score measure compares imperfectly to the standardized Rasch-IRT scores (recall that IRT scores take into account strategy use and item difficulty), the raw scores presented in Table 2.2 show that students in the control group still learned a considerable amount of mathematics during preschool. If we trace the raw score means back to the test items, our results suggest that students would need to move from simple number recognition to addition and subtraction by the end of preschool to produce a full standard deviation of change beyond the control group. Although such a progression in average mathematical ability during preschool may not be impossible, current data from nationally representative samples indicates that addition and subtraction is taught far less than more simple mathematics topics in even kindergarten, and only

5% of students have mastered adding and subtracting at kindergarten entry (Engel, Claessens, & Finch, 2013). Thus, our results likely reflect an upper-bound of the probable long-run effects of successful early math interventions.

Limitations and Conclusion

The results should also be considered against the limitations of the study. As was discussed previously, the exclusion restriction assumption could be violated if the intervention affected later mathematics achievement through unknown pathways unaccounted for by the present models. Unfortunately, we lack the data to extensively test for extraneous treatment-effect pathways. Yet, we found no evidence that boosts in language skills might have also affected later mathematics achievement, and our results did not change with the inclusion of background control variables. We also tested whether students in the treatment group were more likely to remain in the same school throughout the elementary school years, and whether they entered into higher quality kindergarten and first grade classrooms. In both cases, we found no evidence that treatment students' schooling environments changed after the treatment year. This also suggests that peer effects should not bias our results, as students in the treatment group were not more likely to remain in school with the same peers than students in the control condition.

Further, although we employed fairly comprehensive measures of mathematics achievement, it is likely that these measures still failed to capture all dimensions of children's mathematics knowledge. Thus, it remains possible that the benefits of gains in early math skills were not fully detected by the later mathematics measures. Finally, when interpreting our results, one should recall that our models were only tested within a relatively low-income sample of children. Thus, it is unclear how our results might relate to students from different

socioeconomic backgrounds. This further implies the need for replicating our results in diverse settings and samples.

Nevertheless, the threat of omitted variable bias was not completely eradicated, meaning the current estimates produced by the 2SLS models likely reflect upper-bound estimates of the effect of intervention-caused mathematics change on later math achievement. Thus, although we found some indication that a standard deviation of change during preschool might lead to approximately a quarter of a standard deviation gain in later mathematics achievement, intervention fade out is likely to be substantial even in the years following a treatment successful enough to produce an average treatment effect of a full standard deviation. As a result, if educational practitioners and policy-makers wish to produce early childhood interventions that sustain effects in the years following the end of preschool, time and attention might be better placed on developing methods designed to build upon preschool gains during the early elementary school years (see Clements and colleagues (2013) for description of a follow-through treatment that abated early intervention fadeout effects to a degree).

In sum, the current paper demonstrated the use of a quasi-experimental method for better understanding how mathematics skills develop during the early and middle childhood years. Our results illustrate that previous correlational approaches overstated the long-run benefits of early math intervention, and that more robust approaches are necessary for generating better causal estimates. Further, such approaches are also fundamental to our ability to test developmental theories, as the current findings imply that early math skills do not automatically lead to future academic success.

References

Angrist, J. D., & Pischke, J. S. (2008). *Mostly harmless econometrics: An empiricist's companion*. Princeton University Press.

Aunola, K., Leskinen, E., Lerkkanen, M. K., & Nurmi, J. E. (2004). Developmental Dynamics of Math Performance From Preschool to Grade 2. *Journal of Educational Psychology*, *96*, 699. doi:10.1037/0022-0663.96.4.699

Auger, A., Farkas, G., Burchinal, M. R., Duncan, G. J., & Vandell, D. L. (2014). Preschool center care quality effects on academic achievement: An instrumental variables analysis. *Developmental Psychology*, *50*, 2559. doi:10.1037/a0037995

Bailey, D. H., Watts, T. W., Littlefield, A. K., & Geary, D. C. (2014). State and trait effects on individual differences in children's mathematical development. *Psychological Science, 25*, 2017-2026. doi: 10.1177/0956797614547539

Bong, M., & Skaalvik, E. M. (2003). Academic self-concept and self-efficacy: How different are they really? *Educational Psychology Review*, *15*, 1-40. doi:10.1023/A:1021302408382

Booth, J. L., & Siegler, R. S. (2006). Developmental and individual differences in pure numerical estimation. *Developmental Psychology*, *42*, 189. doi: 10.1037/0012-1649.41.6.189

Bronfenbrenner, U., & Morris, P. (2006). The bioecological model of human development. In W. Damon & R. M. Lerner (Eds.), *Handbook of child psychology: Vol 1. Theoretical models of human development* (6th ed., pp. 793-828). New York: Wiley.

Byrnes, J. P., & Wasik, B. A. (2009). Factors predictive of mathematics achievement in kindergarten, first and third grades: An opportunity–propensity analysis. *Contemporary Educational Psychology*, *34*, 167-183. doi:10.1016/j.cedpsych.2009.01.002

Claessens, A., Duncan, G., & Engel, M. (2009). Kindergarten skills and fifth-grade achievement: Evidence from the ECLS-K. *Economics of Education Review*, *28*, 415–427. doi:10.1016/j.econedurev.2008.09.00

Claessens, A., & Engel, M. (2013). How important is where you start? Early mathematics knowledge and later school success. *Teachers College Record*, *115*, 060306.

Clements, D. H., & Sarama, J. (2008). Experimental evaluation of the effects of a research-based preschool mathematics curriculum. *American Educational Research Journal*, *45*, 443-494. doi:10.3102/0002831207312908

Clements, D. H., & Sarama, J. (2011). Early childhood mathematics intervention. *Science*, *333*(6045), 968-970. doi: 10.1126/science.1204537

Clements, D. H., & Sarama, J. (2013). *Building Blocks, Volumes 1 and 2*. Columbus, OH: McGraw-Hill Education.

Clements, D. H., Sarama, J., Khasanova, E., & Van Dine, D. W. (2012). *TEAM 3-5—Tools for elementary assessment in mathematics*. Denver, CO: University of Denver.

Clements, D. H., Sarama, J. H., & Liu, X. H. (2008). Development of a measure of early mathematics achievement using the Rasch model: the Research-Based Early Maths Assessment. *Educational Psychology*, *28*(4), 457-482. doi:10.1080/01443410701777272

Clements, D. H., Sarama, J., Spitler, M. E., Lange, A. A., & Wolfe, C. B. (2011). Mathematics learned by young children in an intervention based on learning trajectories: A large-scale cluster randomized trial. *Journal for Research in Mathematics Education*, *42*, 127-166. Retrieved from: http://www.jstor.org/stable/10.5951/jresematheduc.42.2.0127

Clements, D. H., Sarama, J., & Wolfe, C. B. (2011). *TEAM—Tools for early assessment in mathematics.* Columbus, OH: McGraw-Hill Education.

Clements, D. H., Sarama, J., Wolfe, C. B., & Spitler, M. E. (2013). Longitudinal evaluation of a scale-up model for teaching mathematics with trajectories and technologies Persistence of effects in the third year. *American Educational Research Journal*, *50*, 812-850. doi:10.3102/0002831212469270

Cortes, K. E., & Goodman, J. S. (2014). Ability-tracking, instructional time, and better pedagogy: The effect of Double-Dose Algebra on student achievement. *The American Economic Review*, *104*, 400-405. doi:http://dx.doi.org/10.1257/aer.104.5.400

Crosby, D. A., Dowsett, C. J., Gennetian, L. A., & Huston, A. C. (2010). A tale of two methods: comparing regression and instrumental variables estimates of the effects of preschool child care type on the subsequent externalizing behavior of children in low-income families. *Developmental Psychology*, *46*, 1030. doi:10.1037/a0020384

Duncan, G. J., Dowsett, C. J., Claessens, A., Magnuson, K., Huston, A. C., Klebanov, P., et al. (2007). School readiness and later achievement. *Developmental Psychology*, *43*, 1428-1446. doi: 10.1037/0012-1649.43.6.1428

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

Engel, M., Claessens, A., & Finch, M. A. (2013). Teaching students what they already know? The (Mis) Alignment between mathematics instructional content and student knowledge in kindergarten. *Educational Evaluation and Policy Analysis*, *35*, 157-178. doi:10.3102/0162373712461850

Entwisle, D. R., & Alexander, K. L. (1990). Beginning school math competence: Minority and majority comparisons. *Child Development, 61*, 454-471. doi:10.1111/j.1467-8624.1990.tb02792.x

Foster, E. M. (2010). The value of reanalysis and replication: Introduction to special section. *Developmental Psychology, 46*, 973. doi:10.1037/a0020183

Geary, D. C., Hoard, M. K., Nugent, L., & Bailey, D. H. (2013). Adolescents' functional numeracy is predicted by their school entry number system knowledge. *PLoS ONE, 8*, e54651. doi:10.1371/journal.pone.0054651

Gennetian, L. A., Magnuson, K., & Morris, P. A. (2008). From statistical associations to causation: what developmentalists can learn from instrumental variables techniques coupled with experimental data. *Developmental Psychology, 44*, 381. doi:10.1037/0012-1649.44.2.381

Gersten, R., Chard, D. J., Jayanthi, M., Baker, S. K., Morphy, P., & Flojo, J. (2009). Mathematics instruction for students with learning disabilities: A meta-analysis of instructional components. *Review of Educational Research, 79*, 1202-1242. doi:10.3102/0034654309334431

Gilmore, C. K., McCarthy, S. E., & Spelke, E. S. (2007). Symbolic arithmetic knowledge without instruction. *Nature, 447*, 589-591. doi:10.1038/nature05850

Jordan, N. C., Kaplan, D., Ramineni, C., & Locuniak, M. N. (2009). Early math matters: kindergarten number competence and later mathematics outcomes. *Developmental Psychology, 45*, 850–867. doi:10.1037/a0014939

Laski, E. V., & Siegler, R. S. (2007). Is 27 a big number? Correlational and causal connections among numerical categorization, number line estimation, and numerical magnitude comparison. *Child Development, 78*, 1723-1743. doi:10.1111/j.1467-8624.2007.01087.x

Leak, J., Duncan, G. J., Weilin L., Magnuson, K., Schindler, H., & Yoshikawa, H. (2010). *Is timing everything? How early childhood education program impacts vary by starting age, program duration, and time since the end of the program.* UC-Irvine working paper, presented at the fall 2010 meetings of the Association for Public Policy Analysis and Management, Boston, MA. Retrieved from: http://education.uci.edu/docs/Leak_Duncan_Li_Timing_Paper_APPAM_102810.pdf

Maloney, A. P., Confrey, J., & Nguyen, K. H. (Eds.). (in press). *Learning over time: Learning trajectories in mathematics education*. New York, NY: Information Age Publishing.

Meisels, S. J. (1998). *Assessing readiness* (Report no. 3-002). Ann Arbor, MI: Center for the Improvement of Early Reading Achievement. Retrieved from http://www.ciera.org/products/meisels-1998/reports32.html

Murnane, R. J., & Willett, J. B. (2010). *Methods matter: Improving causal inference in educational and social science research*. Oxford University Press.

Opfer, J. E., & Thompson, C. A. (2008). The trouble with transfer: Insights from microgenetic changes in the representation of numerical magnitude. *Child Development, 79*, 788-804. doi:10.1111/j.1467-8624.2008.01158.x

Sarama, J., Lange, A. A., Clements, D. H., & Wolfe, C. B. (2012). The impacts of an early mathematics curriculum on oral language and literacy. *Early Childhood Research Quarterly, 27*, 489-502. doi:10.1016/j.ecresq.2011.12.002

Sarama, J., & Clements, D. H. (2009). *Early childhood mathematics education research: Learning trajectories for young children.* New York, NY: Routledge.

Schmidt, W. H., & Burroughs, N. A. (2013). How the common core boosts quality and equality. *Educational Leadership, 70*, 54-58.

Siegler, R. S., Thompson, C. A., & Schneider, M. (2011). An integrated theory of whole number and fractions development. *Cognitive Psychology, 62*, 273-296. doi:10.1016/j.cogpsych.2011.03.001

Stevenson, H. W., & Newman, R. S. (1986). Long-term prediction of achievement and attitudes in mathematics and reading. *Child Development, 57*, 646-659. Retrieved from: http://www.jstor.org/stable/1130343

Taylor, E. (2014). Spending more of the school day in math class: Evidence from a regression discontinuity in middle school. *Journal of Public Economics*, 117, 162-181. doi:10.1016/j.jpubeco.2014.06.002

Watts, T. W., Duncan, G. J., Chen, M., Claessens, A., Davis-Kean, P. E., Duckworth, K., Engel, M., Siegler, R. S., Susperreguy, M. I. (2015). The role of mediators in the development of longitudinal mathematics achievement associations. *Child Development.* Advance online publication. doi:10.1111/cdev.12416

Watts, T. W., Duncan, G. J., Siegler, R. S., & Davis-Kean, P. E. (2014). What's past is prologue: Relations between early mathematics knowledge and high school achievement. *Educational Researcher, 43*, 352-360. doi:10.3102/0013189X14553660

Weiland, C., Wolfe, C. B., Hurwitz, M. D., Clements, D. H., Sarama, J. H., & Yoshikawa, H. (2012). Early mathematics assessment: Validation of the short form of a prekindergarten

and kindergarten mathematics measure. Educational Psychology, 32, 311–333. doi:10.1080/01443410.2011.654190

Wilson, P. H., Mojica, G. F., & Confrey, J. (2013). Learning trajectories in teacher education: Supporting teachers' understandings of students' mathematical thinking. *The Journal of Mathematical Behavior, 32*, 103-121. doi: 10.1016/j.jmathb.2012.12.003

Table 2.1
Sample Characteristics

	Full Sample	Pooled Sample	p-values	Treatment	Control	p-values
	(1)	(2)	(3)	(4)	(5)	(6)
PreK Entry Math	-3.210	-3.210	0.984	-3.249	-3.164	0.467
	(0.830)	(0.808)		(0.856)	(0.795)	
Site						
New York	0.725	0.815	0.001	0.702	0.753	0.756
Massachusetts	0.275	0.185	0.001	0.298	0.247	0.756
Ethnicity						
African American	0.502	0.488	0.275	0.519	0.482	0.814
Hispanic	0.231	0.198	0.063	0.198	0.270	0.523
White- Non-Hispanic	0.211	0.249	0.027	0.246	0.169	0.506
Other	0.0557	0.0659	0.055	0.0372	0.0783	0.237
Female	0.497	0.556	0.003	0.496	0.497	0.893
Age at PreK Entry	4.359	4.339	0.302	4.331	4.392	0.382
	(0.352)	(0.352)		(0.353)	(0.348)	
Special Education	0.167	0.156	0.476	0.173	0.159	0.678
Free/Reduced Lunch	0.849	0.850	0.951	0.824	0.881	0.25
Limited Eng Prof.	0.167	0.163	0.417	0.124	0.220	0.279
Observations	880	410	-	484	396	-

Note. For each variable, mean values are displayed. Standard deviations are in parentheses. Column 2 displays mean characteristics for students included in the primary analysis model, in which only participants who had non-missing test score data in fall or spring of fourth grade and spring fifth grade were considered. Column 3 displays p-values from regressions comparing students who were included in the pooled sample with those who were not. P-values listed in column 6 indicate the extent to which treatment participants differed from controls. In each regression, standard errors were adjusted for clustering at the school level (30 schools). F-test results indicate whether the set of baseline characteristics were jointly-significantly different from 0 in a model in which treatment status was regressed on all the baseline covariates simultaneously.

Table 2.2
Math Change Descriptives

	Treatment	Control	P-Values
PreK Entry Math			
IRT Score	-3.249	-3.164	0.467
	(0.856)	(0.795)	
Number Correct	11.46	12.10	0.526
	(7.493)	(7.781)	
PreK Post Math			
IRT Score	-1.872	-2.245	0.004
	(0.672)	(0.749)	
Number Correct	32.70	28.02	0.022
	(12.11)	(12.07)	
PreK Change			
IRT Score	1.376	0.919	0.001
	(0.705)	(0.650)	
Number Correct	21.25	15.92	0.001
	(8.647)	(8.053)	
Observations	456	378	

Note. Entries show means and standard deviations are shown in parentheses. The IRT scores were scaled such that a score of "0" approximates the achievement level of a student in first grade. The p-value column lists p-values from regressions in which each variable listed was regressed on treatment status. P-values less than 0.001 were rounded to 0.001.

Table 2.3

OLS Models Predicting Preschool Change and Late-Elementary School Math Achievement

		Later Achievement		
	Math Change	Fall- 4th Grade	Spring- 4th Grade	Spring- 5th Grade
	(1)	(2)	(3)	(4)
Math Change		0.568***	0.582***	0.529***
		(0.044)	(0.042)	(0.043)
Treatment	0.699***	-0.313***	-0.371***	-0.234***
	(0.138)	(0.047)	(0.041)	(0.061)
Controls	Inc.	Inc.	Inc.	Inc.
Blocking Group	Inc.	Inc.	Inc.	Inc.
*Block * Treatment*				
1	-0.127			
	(0.262)			
2	-0.320*			
	(0.135)			
3	-0.281			
	(0.165)			
4	0.102			
	(0.162)			
6	-0.262			
	(0.186)			
7	-0.189			
	(0.187)			
8	0.045			
	(0.182)			
Observations	834	469	543	502
R-squared	0.425	0.499	0.496	0.448

Note. Robust standard errors were adjusted for clustering at the school level, and are displayed in parentheses. In each model, the dependent variable was standardized, as was math change and age. Column 1 displays coefficients produced by treatment and block and treatment group interaction (the main component of the IV analysis) predicting math change during preschool. Columns 2 through 4 display the results of OLS models predicting standardized math achievement in grades 3 through 5, respectively, with baseline characteristics and preschool math change. Coefficients produced by control variables (prek entry math, gender, race, whether limited English proficient, age, whether designated for special education, whether FRPL, site and blocking group) can be found in the Appendix. * $p<0.05$ ** $p<0.01$ *** $p<0.001$

Table 2.4

IV Estimates Relating Preschool Change to Late-Elementary School Achievement

	Reduced Form	OLS	IV- Reduced Control	IV- Full Controls	IV- Fall 4th Grade	IV- Spring 4th Grade	IV- Spring 5th Grade
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Math Change		0.535***	0.236*	0.242**	0.132	0.039	0.257**
		(0.041)	(0.113)	(0.081)	(0.109)	(0.096)	(0.079)
Treatment	0.094						
	(0.064)						
Controls							
Entry Math Score	Inc.	Inc.	Inc.	Inc.	Inc.	Inc.	Inc.
Site	Inc.	Inc.	Inc.	Inc.	Inc.	Inc.	Inc.
Block	Inc.	Inc.	Inc.	Inc.	Inc.	Inc.	Inc.
Background Characteristics	Inc.	Inc.		Inc.	Inc.	Inc.	Inc.
Observations	1230	1230	1230	1230	469	543	502

Note. Robust standard errors, shown in parentheses, were adjusted for clustering at the school level. IV estimates were generated using 2SLS. In the models presented in columns 1 through 4, students were observed three times (fall and spring of fourth grade and spring of fifth grade). In the models presented in columns 5 through 7, students were observed only once, and the measurement point of the dependent variable varies in each model. "Inc" denotes the inclusion of various sets of control variables. The dependent variable, late-elementary school math achievement, was within-grade standardized, and the main independent variable, preschool math change, was also standardized. For full list of background controls, see Table 2.3 note. Coefficients produced by background controls can be found in the Appendix. * $p<0.05$ ** $p<0.01$ *** $p<0.001$

Appendix

Chapter 2: What is the long-run impact of learning mathematics during preschool?

Exclusion of Third Treatment Condition

The TRIAD study included three study conditions to which schools were randomly assigned: 1) control condition (business as usual); 2) *Building Blocks* curriculum during preschool only; 3) *Building Blocks* curriculum during preschool with extended pedagogical development in kindergarten and first grade. In the current paper, we only analyzed data from the control condition and the "*Building Blocks* curriculum during preschool only" condition. We hoped to incorporate the third treatment arm into our analyses, but a few analytic and conceptual limitations prevented us from doing so.

First, the follow-through condition lasted until the end of first grade. So, instead of preschool math gains, our key variable for this condition would be gains between preschool and the end of first grade. When we tested if our instrument (block and treatment group interactions) produced enough variation in preschool through first grade gains, we found that the F-test fell far short of the critical threshold of a score of 10 (F= 5.26). Thus, the instruments were weak predictors of gains in math from preschool through first grade and this suggests that IV methods based on them are not warranted.

Further, we were even less comfortable with the exclusion restriction with this condition, as we found that results changed with the addition of baseline controls. This indicated that other processes correlated with student characteristics might have accounted for the association between instrumented gains and later achievement.

Additional Results

Appendix Table 2.1 presents correlations between all predictor variables used in the primary models.

Appendix Table 2.2 presents results from models shown in Table 2.3 of the main text with all control variable coefficients displayed.

Model 1 of Appendix Table 2.3 presents the results from models shown in Table 2.4 of the main text with coefficients produced by all control variables displayed.

Model 2 of Appendix Table 2.3 displays results from a model that does not control for preschool entry math score. In this model, the coefficient produced by our change measure dropped to 0.148 (SE = 0.106, p = 0.165). Although IV methods typically purge any measurement error in the endogenous predictor from the model, the results presented in Model 2 suggest that measurement construction may still affect our IV results. As is often found in longitudinal data, the observed correlation between preschool entry math score and preschool change is negative (r (410) = -0.450, $p < .001$), which means that students that were high achieving at the beginning of preschool grow *less* throughout preschool than lower-achieving students. In a typical correlational OLS model, this negative correlation between change and entry-level necessitates controlling for the level score. Although the IV model only depends on change caused by the instruments (random assignment and block interactions), we would still expect there to be a ceiling to the amount of change the intervention could actually spur among initially higher-achieving students. Because the intervention provided teachers with a prescribed curriculum, high-achieving students that had already mastered many of concepts and procedures contained in the curriculum before preschool would not experience any instrumented change. Consequently, some high achieving students would be near 0 on the measure of instrumented change, but would have high math scores in later elementary school.

To further assess whether measure constraints for high achieving students restricted the prediction reported in Model 2, we tested a model that included the interaction between the entry-level score and change, but we did not include the main effect of entry level math in the model. Model 3 of Appendix Table 2.3 shows that when this interaction is controlled, the change measure coefficient grows to a value much closer to the effect reported in Model 1 (β= 0.210, SE= 0.110, $p < 0.10$). This result gives further indication that the negative correlation between change and entry-level still affects the IV results to some degree.

In Model 4 of Appendix Table 2.3, we tested another model that did not control for the preschool entry math score, and we also used the preschool posttest score as our key measure instead of preschool change. When modeling the relation between change and later achievement, two different model specifications provide the same result: 1) regress later achievement on a measure of change calculated by subtracting the pretest from the posttest and control for the pretest; 2) regress later achievement on the posttest and control for the pretest. A full explanation of why both models produce the same result can be found in Watts, Duncan, Siegler, and Davis-Kean (2014). As Model 4 shows, removing the pretest and instrumenting for the posttest score provides a similar result to the preferred model specification shown in column 1, further indicating that unobserved ability does not substantially bias our key estimate.

Model 5 of Appendix Table 2.3 presents results from a model that controls for language skills measured in kindergarten. The minimal change in the coefficient for preschool change between Models 1 and 5 provides further evidence that we cannot detect violations to the exclusion restriction with available measures. Although Sarama and colleagues (2012) reported that the intervention did impact kindergarten entry language skills, including this measure does not affect our key coefficient.

In Models 6 and 7, we present results from models that took two different approaches to controlling for grade level, as to adjust for differences between the time in which the student took the various follow-up tests (fall of fourth grade, spring of fourth grade, or spring of fifth grade). In Model 6, we included a continuous measure for grade (coded: 1=4th grade fall, 2= 4th grade spring, 3= 5th grade spring), and in Model 7, we included grade fixed effects (i.e. grade dummy variables). Across both models, results did not differ from our key estimates shown in column 1.

In Models 8 and 9, we tried two different approaches to specifying our IV model. The results presented in Model 8 were estimated by only using the treatment status variable as the instrument for preschool mathematics change. This model differs from our preferred model, as the single treatment dummy instrument provides far less variation in preschool math change than the treatment and block interactions. When only treatment assignment is used, the estimate falls to a marginally significant 0.154 (SE = 0.094, p = 0.102). Finally, we tested a model that used limited information maximum likelihood as the IV estimator instead of 2SLS. This is a common alternative estimator for IV (see Taylor, 2014). This model again produced a similar result (β= 0.211, SE= 0.095, $p < 0.05$).

Finally, Appendix Table 2.4 presents regressions from grade-pooled models that restricted the sample to key subgroups. Much like the non-pooled estimates, these estimates were also hampered by relatively small sample sizes, yet they provide some indication that unanticipated heterogeneity in mathematics change may exist. In Columns 1 and 2, we present models that were run with only African American and Non-African American students (majority Hispanic or White), respectively. We found some indication that African Americans especially benefitted from preschool math change, as instrumented preschool math change produced a large

coefficient in the model restricted to only African American children (β= 0.379, SE= 0.104, p < 0.001). To test whether this difference between African American and non-African American students was statistically significant, we also tested a model that included an interaction term between preschool change and whether African American (model not shown). In this model, the interaction term produced a positive, but statistically non-significant, coefficient (β= 0.271, SE= 0.199, p = 0.15). The large standard error produced by the interaction term suggests that we simply lacked the statistical power to test for such differences with the necessary level of precision.

Further, we also tested for subgroup-specific effects for students identified as Limited English Proficient (LEP) and Non-LEP students (columns 3 and 4, respectively), students who scored below and above the median on the preschool entry math test (columns 5 and 6, respectively), and students who did and did not qualify for FRPL (columns 7 and 8, respectively). Across these models, we only found significant effects for "Non-LEP" students (β= 0.271, SE= 0.081, p < 0.001) and FRPL students (β= 0.311, SE= 0.093, p < 0.001). However, these effects did not substantively differ much from the effects reported for the other subgroups, but because these two subgroups had relatively large sample sizes, their effects were statistically significant. All of the tested subgroup effects were positive and ranged from 0.143 to 0.311, and none of these effects were much larger or smaller than the main effects reported in grade-pooled estimates of Table 2.4 in the main text.

Appendix Table 2.1

Correlations Among Key Independent and Dependent Variables

		1	2	3	4	5	6	7	8	9	10	11	12
1	Math Change												
2	PreK Entry Math	-0.450***	1.000										
3	Treatment Group	0.356***	-0.044	1.000									
4	Female	0.019	0.072	-0.032	1.000								
5	African American	-0.109*	-0.099*	0.065	-0.032	1.000							
6	Hispanic	0.125*	-0.110*	-0.111*	0.024	-0.484***	1.000						
7	Other	0.008	0.001	-0.106*	-0.060	-0.259***	-0.132**	1.000					
8	Limited Engl. Prof	0.153**	-0.163***	-0.127**	-0.030	-0.365***	0.543***	0.255***	1.000				
9	Age at PreK Entry	0.033	0.253***	-0.110*	-0.029	-0.145**	0.282***	-0.002	0.264***	1.000			
10	Special Education	0.046	-0.064	-0.000	-0.008	-0.111*	0.090	-0.114*	0.028	0.055	1.000		
11	FRPL	0.078	-0.235***	-0.034	0.096	0.179***	0.118*	0.048	0.132**	0.021	-0.167***	1.000	
12	Site- New York	-0.157**	-0.049	0.005	0.016	0.177***	-0.331***	-0.076	-0.366***	-0.426***	-0.037	-0.060	1.000
13	Missing- FRPL	-0.064	0.164***	-0.068	-0.059	-0.090	-0.072	-0.048	-0.118*	-0.039	0.103*	-0.661***	0.025

Note. N= 410 (grade pooled sample). * p<0.05 ** p<0.01 *** p<0.001

Appendix Table 2.2

OLS Models Predicting Preschool Change and Late-Elementary School Math Achievement

		Later Achievement		
	Math Change	Fall- 4th Grade	Spring- 4th Grade	Spring- 5th Grade
	(1)	(2)	(3)	(4)
Math Change		0.568***	0.582***	0.529***
		(0.044)	(0.042)	(0.043)
Treatment	0.699***	-0.313***	-0.371***	-0.234***
	(0.138)	(0.047)	(0.041)	(0.061)
Controls				
PreK Entry Math	-0.483***	0.672***	0.662***	0.616***
	(0.049)	(0.043)	(0.050)	(0.045)
Female	0.058	-0.015	0.022	0.035
	(0.042)	(0.070)	(0.092)	(0.084)
Afrian American	-0.313***	-0.019	-0.244**	-0.215***
	(0.085)	(0.089)	(0.078)	(0.050)
Hispanic	-0.212*	-0.006	-0.171	-0.171
	(0.092)	(0.157)	(0.152)	(0.144)
Ethnicity- Other	-0.081	0.270	0.061	0.387*
	(0.167)	(0.150)	(0.193)	(0.168)
Limited Eng Prof.	0.159	0.304***	0.233**	0.180
	(0.111)	(0.072)	(0.080)	(0.159)
Age	0.135***	-0.077	-0.118***	-0.077
	(0.025)	(0.042)	(0.032)	(0.038)
Special Education	-0.108	-0.079	-0.125	0.001
	(0.071)	(0.100)	(0.107)	(0.119)
Free/Reduced Lunch	0.059	-0.141	-0.142	-0.167
	(0.104)	(0.098)	(0.085)	(0.113)
Site- New York	-0.274**	-0.395***	-0.182*	-0.019
	(0.078)	(0.089)	(0.075)	(0.113)
Missing- FRPL	0.239*	0.228	0.096	0.072
	(0.095)	(0.163)	(0.115)	(0.119)
Block Group				
1	-0.202	-0.042	0.121	-0.178**
	(0.236)	(0.095)	(0.065)	(0.064)
2	0.367**	-0.134	0.057	-0.109
	(0.121)	(0.076)	(0.050)	(0.087)
3	0.330*	0.004	0.035	-0.294**
	(0.120)	(0.113)	(0.079)	(0.097)
4	0.073	-0.108	0.181**	-0.104
	(0.134)	(0.105)	(0.055)	(0.107)
6	0.138	0.103	0.225*	0.035
	(0.140)	(0.135)	(0.106)	(0.180)
7	0.299*	-0.140	-0.149**	-0.311**
	(0.123)	(0.091)	(0.053)	(0.086)
8	0.143	-0.265*	-0.206	-0.391***
	(0.157)	(0.114)	(0.134)	(0.090)
*Block * Treatment*				
1	-0.127			

	(0.262)			
2	-0.320*			
	(0.135)			
3	-0.281			
	(0.165)			
4	0.102			
	(0.162)			
6	-0.262			
	(0.186)			
7	-0.189			
	(0.187)			
8	0.045			
	(0.182)			
Constant	-0.139	0.600***	0.522***	0.449*
	(0.188)	(0.161)	(0.139)	(0.187)
Observations	834	469	543	502
R-squared	0.425	0.499	0.496	0.448

Note. Robust standard errors were adjusted for clustering at the school level, and are displayed in parentheses. In each model, the dependent variable was standardized, as was math change and age. Column 1 displays coefficients produced by treatment and block and treatment group interaction (the main component of the IV analysis) predicting math change during preschool. Columns 2 through 4 display the results of OLS models predicting standardized math achievement in grades 3 through 5, respectively, with baseline characteristics and preschool math change. * $p<0.05$ ** $p<0.01$ *** $p<0.001$

IV Estimates Generated from Grade-Pooled Models Predicting Fourth and Fifth Grade Math Achievement- Additional Model Specifications

	IV-Full Controls (1)	IV- Change No Pretest (2)	IV- Change No Pretest (3)	IV- Posttest (4)	IV- Language Control (5)	IV- Grade Control (6)	IV- Grade Control (7)	IV- Single Instrument (8)	IV- LIML Estimator (9)
Math Change	0.242**	0.148	0.210+		0.227***	0.242**	0.242**	0.154	0.211*
	(0.081)	(0.106)	(0.110)		(0.064)	(0.081)	(0.081)	(0.094)	(0.095)
PreK Posttest Math				0.289***					
				(0.080)					
PreK Entry Math	0.489***				0.392***	0.489***	0.489***	0.446***	0.474***
	(0.068)				(0.072)	(0.068)	(0.068)	(0.077)	(0.075)
Entry Math * Change			0.036						
			(0.026)						
Kindergarten Language					0.226***				
					(0.055)				
Grade Level (Continuous)						0.006			
						(0.023)			
Grade Dummies									
4th Grade Fall							-0.012		
							(0.046)		
4th Grade Spring							-0.006		
							(0.039)		
Background Controls									
African American	-0.224**	-0.381***	-0.376***	-0.273**	-0.171**	-0.224**	-0.224**	-0.252**	-0.234**
	(0.075)	(0.099)	(0.096)	(0.084)	(0.065)	(0.075)	(0.075)	(0.081)	(0.078)
Hispanic	-0.116	-0.313+	-0.312+	-0.190	-0.054	-0.116	-0.116	-0.139	-0.124
	(0.148)	(0.169)	(0.170)	(0.158)	(0.113)	(0.148)	(0.148)	(0.150)	(0.150)
Ethnicity- Other	0.261	0.243	0.253	0.271	0.427*	0.261	0.261	0.244	0.255
	(0.185)	(0.217)	(0.211)	(0.187)	(0.188)	(0.185)	(0.185)	(0.198)	(0.189)
Female	0.092	0.175+	0.172+	0.125	0.065	0.092	0.092	0.100	0.095
	(0.080)	(0.092)	(0.091)	(0.079)	(0.090)	(0.080)	(0.080)	(0.084)	(0.082)
Age	-0.027	0.130**	0.129**	0.042	-0.053	-0.027	-0.027	-0.017	-0.023
	(0.038)	(0.043)	(0.044)	(0.046)	(0.045)	(0.038)	(0.038)	(0.039)	(0.039)
Special Education	-0.130	-0.277*	-0.274*	-0.200*	-0.085	-0.130	-0.130	-0.135	-0.132
	(0.101)	(0.108)	(0.110)	(0.095)	(0.116)	(0.101)	(0.101)	(0.102)	(0.101)
Limited Eng Prof.	0.219*	0.016	-0.009	0.120	0.261**	0.219*	0.219*	0.216*	0.218*
	(0.096)	(0.135)	(0.140)	(0.112)	(0.099)	(0.096)	(0.096)	(0.094)	(0.095)

Free/Reduced Lunch	-0.126	-0.288**	-0.301**	-0.209*	-0.138	-0.126	-0.126	-0.125	-0.126
	(0.092)	(0.105)	(0.101)	(0.084)	(0.112)	(0.092)	(0.092)	(0.097)	(0.094)
Site- New York	-0.275**	-0.364***	-0.363**	-0.286**	-0.305***	-0.275**	-0.275**	-0.306**	-0.286**
	(0.096)	(0.107)	(0.111)	(0.094)	(0.074)	(0.096)	(0.096)	(0.097)	(0.097)
Missing- FRPL	0.224+	0.256+	0.248+	0.242+	0.152	0.224+	0.224+	0.222+	0.223+
	(0.122)	(0.140)	(0.140)	(0.127)	(0.184)	(0.122)	(0.122)	(0.124)	(0.123)
Blocking Group									
1	-0.277**	-0.379*	-0.337*	-0.284**	-0.258***	-0.277**	-0.277**	-0.319*	-0.292*
	(0.104)	(0.153)	(0.150)	(0.099)	(0.075)	(0.104)	(0.104)	(0.135)	(0.118)
2	-0.110	-0.178	-0.160	-0.133	-0.210**	-0.110	-0.110	-0.120	-0.113
	(0.072)	(0.111)	(0.107)	(0.084)	(0.064)	(0.072)	(0.072)	(0.082)	(0.075)
3	-0.075	-0.231+	-0.220+	-0.160	-0.118	-0.075	-0.075	-0.070	-0.073
	(0.096)	(0.139)	(0.131)	(0.106)	(0.102)	(0.096)	(0.096)	(0.106)	(0.099)
4	-0.029	-0.045	-0.042	-0.034	-0.003	-0.029	-0.029	-0.031	-0.029
	(0.075)	(0.130)	(0.119)	(0.080)	(0.061)	(0.075)	(0.075)	(0.093)	(0.081)
6	0.085	-0.131	-0.113	-0.009	0.076	0.085	0.085	0.072	0.081
	(0.183)	(0.185)	(0.186)	(0.186)	(0.081)	(0.183)	(0.183)	(0.184)	(0.184)
7	-0.184*	-0.196	-0.196	-0.197*	-0.156+	-0.184*	-0.184*	-0.177*	-0.181*
	(0.076)	(0.133)	(0.129)	(0.096)	(0.087)	(0.076)	(0.076)	(0.088)	(0.079)
8	-0.337**	-0.437*	-0.429*	-0.364**	-0.289*	-0.337**	-0.337**	-0.358*	-0.344**
	(0.121)	(0.205)	(0.192)	(0.129)	(0.123)	(0.121)	(0.121)	(0.150)	(0.132)
Constant	0.420*	0.805***	0.823***	0.563***	0.603***	0.385+	0.426*	0.468**	0.437*
	(0.166)	(0.170)	(0.167)	(0.160)	(0.169)	(0.208)	(0.169)	(0.178)	(0.173)
Observations	1230	1230	1230	1230	1014	1230	1230	1230	1230
R-squared	0.415	0.233	0.240	0.391	0.449	0.415	0.415	0.383	0.405

Note. Robust standard errors were adjusted for clustering at the school level and are shown in parentheses. In each model, students were observed three times (fall and spring of fourth grade, and spring of fifth grade). The dependent variable, late-elementary school math achievement, was within-grade standardized, and the main independent variable, preschool math change, was also standardized. + $p < 0.10$ * $p<0.05$ ** $p<0.01$ *** $p<0.001$

Appendix Table 2.4
Pooled IV Estimates- Subgroup Effects

	African American	Non-African American	LEP	Non-LEP	Low Math	High Math	FRPL	Non-FRPL
Math Change	0.379***	0.162	0.232	0.271***	0.143	0.214	0.311***	0.220
	(0.104)	(0.143)	(0.143)	(0.081)	(0.112)	(0.116)	(0.093)	(0.129)
Entry Math Score	Inc.	Inc.	Inc.	Inc.	Inc.	Inc.	Inc.	Inc.
Site	Inc.	Inc.	Inc.	Inc.	Inc.	Inc.	Inc.	Inc.
Block	Inc.	Inc.	Inc.	Inc.	Inc.	Inc.	Inc.	Inc.
Background Characteristics	Inc.	Inc.	Inc.	Inc.	Inc.	Inc.	Inc.	Inc.
Observations (pooled)	600	630	201	1029	618	612	903	327

Note. Robust standard errors were adjusted for clustering at the school level. "LEP" stands for limited English proficient and "FRPL" stands for free or reduced price lunch. In each model, students were observed three times (fall and spring of fourth grade, and spring of fifth grade). In each model, the dependent variable, late-elementary school math achievement, was within-grade standardized. "Inc" denotes the inclusion of various sets of control variables * $p<0.05$ ** $p<0.01$ *** $p<0.001$

Chapter 3

Evaluating the Effects of a Two-Year Individualized Instruction Intervention in Mathematics

Abstract

Individualized instruction has been identified as a way to target instruction to the specific needs of students in order to maximize the learning gains of students across the achievement distribution. The current study examined the effects of a two-year individualized instruction intervention in mathematics for a sample of second grade students (n= 519) attending elementary schools in northwest Florida. Classrooms were randomly assigned to one of two conditions: 1) individualized instruction in mathematics; 2) individualized instruction in reading. These students were then followed into third grade (n=472), and classrooms were again randomly assigned to receive individualized instruction in either mathematics or reading. In second grade, I found largely null impacts of the mathematics program on math test scores, though a positive and statistically significant effect (β = 0.16) was detected for one of the math subtests given at the end of the school year. In third grade, math program effects were again largely null, though a negative impact (β = -0.11) was detected on one of the math subtests administered. I found no impact of assignment to two consecutive years of individualized math instruction on end of third grade math scores. Analyses of classroom observations suggested that teachers did not make great efforts to individualize instruction, and correlational models suggested that individualized instruction in mathematics did not strongly predict math achievement. Implications for educational theory and policy are discussed.

Evaluating the Effects of a Two-Year Individualized Instruction Intervention in Mathematics

Individualizing instruction in elementary school has been identified as a promising approach to boosting the achievement of all students in a classroom, regardless of their achievement levels at the beginning of the year. Programs that support individualizing instruction have garnered more attention recently as proponents have criticized traditional elementary school academic instruction as a "one size fits all" model that largely ignores the heterogeneity that exists across students (e.g., Engel, Claessens, Watts, & Farkas, 2016). Indeed, Connor and colleagues (2013) recently demonstrated the utility of individualizing instruction during the early-grade years in reading achievement, as a randomized control trial of individualized instruction in reading produced substantial impacts on achievement tests in language and reading measured across grades 1 through 3.

In the current study, I tested whether random assignment to an individualizing student instruction (ISI) program in mathematics boosted the math achievement of second and third graders enrolled in 8 elementary schools in northern Florida. The program ran over two years, and in both years, classrooms were randomly assigned to implement ISI in either math or reading. I tested whether the program affected test scores in both second and third grade, and I tested whether assignment to the ISI group in math across *both* second and third grade affected scores measured at the end of grade 3. Further, because ISI programs are designed to positively affect students across all levels of the ability distribution, I also tested for treatment heterogeneity by interacting treatment status with baseline measures of math achievement.

Across most models tested, I found no impact of the math ISI program on end-of-year measures of math achievement. For the second-grade year of the study, the standardized program impact on the average of all the math subtests given in the spring was 0.06 (SE = 0.06),

and at third grade the effect was also null (β = -0.06, SE = 0.06). I found some statistically significant effects on certain math subtests at both grades, as the program improved Woodcock-Johnson (WJ-III) *Math Fluency* scores in grade 2 (β = 0.16, SE = 0.09; $p < 0.10$) but appeared to lower WJ-III *Applied Problems* scores in grade 3 (β = -0.11, SE = 0.05; $p < 0.05$).

Tests for treatment heterogeneity produced statistically significant results only in grade 3, as students in the top third of the math achievement distribution at baseline were found to have a stronger program impact (β = .26, SE = 0.11; $p < 0.05$), yet the main effect for treatment was negative in this model (β = -0.17, SE = 0.07; $p < 0.05$), which largely negated the positive impact found for this subgroup. I also found no indication that any combination of treatments over the course of both grades 2 and 3 produced impacts over being assigned to the control group during both years. Finally, analyses of the classroom observational measures suggested that teachers did not spend much time individualizing instruction in mathematics despite participating in an intervention program that encouraged them to do so. Further, correlational models suggested that individualization efforts did not strongly predict higher achievement scores at any measurement point during the two study years.

Background

Many studies have identified the importance of early-grade mathematical ability in determining long-run academic success (e.g., Aunola, Leskinen, Lerkkanen, & Nurmi, 2004; Bailey, Siegler, & Geary, 2014; Duncan et al., 2007; Geary, Hoard, Nugent, & Bailey, 2013; Jordan, Kaplan, Ramineni, & Locuniak, 2009; Siegler et al., 2012; Watts, Duncan, Siegler, & Davis-Kean, 2014), as early math achievement is often found to be a strong predictor of later math achievement. However, long-run results from early math intervention studies have been largely disappointing, as intervention designers have found it difficult to make long-lasting

impacts on children's math achievement (for review see Bailey, Duncan, Watts, Clements, & Sarama, in press). These results have proven puzzling, as mathematics is thought to be a hierarchical subject in which early skill gains should lead to learning advantages in later periods. For example, children who master counting in preschool should be better prepared to learn adding and subtracting in kindergarten (Baroody, 1987). Yet, long-run intervention results suggest that even if early math intervention studies provide students with higher levels of skills at the beginning of kindergarten or first grade, these skill advantages are not readily transferred into further skill gains.

In trying to explain the discrepancy between correlational models that predict strong long-run returns to early gains in math skills and early math intervention studies that show substantial impact fadeout, some have pointed to the lack of curricular differentiation in early-grade math instruction (Bailey, Duncan, Odgers, & Yu, 2017; Claessens, Engel, & Curran, 2014; Stipek, 2017). This argument contends that early interventions raise students' skill levels beyond the level taught in students' subsequent classroom environments. So, students who mastered counting in preschool might spend the better part of kindergarten learning to count again. Thus, even successful early intervention efforts could be doomed to fail in the long-run because teachers do not differentiate instruction to the needs to students with varying levels of ability.

Indeed, recent descriptive research suggests that early elementary school teachers largely fail to deliver content that meets the developmental needs of the children in their classrooms, instead preferring to target curriculum at only the lowest-achieving children. Using nationally-sampled data from the late 1990's, Engel, Claessens and Finch (2013) found that during the kindergarten year, teachers spent most of their time teaching math concepts that children had already mastered prior to kindergarten entry. They also found that students learned more math

over the course of the school year when they were paired with teachers who taught more challenging content (i.e., content better aligned with student ability levels). More recently, Engel, Claessens, Watts, and Farkas (2016) found that even after the accountability reforms of the early 2000's, kindergarten teachers in 2010 still reported spending most of their instructional time on the same basic content that most students had already learned prior to kindergarten. Thus, the early-grade mathematics content taught to students appears to be largely divorced from the mathematics knowledge that students possess when they enter the classroom.

For proponents of early-grade mathematics intervention, such findings are particularly troubling, as early intervention may raise ability levels only to see affected students enter into subsequent classrooms that fail to build upon the gains students previously made. Further, the standard approach of teaching the same content to all students in a given class (often referred to as the "one-size-fits-all" model for instruction) appears to contradict many developmental theories of cognitive growth (e.g., Brofenbrenner & Morris, 2006; Vygotsky, 1978). Vygotsky's zone of proximal development (ZPD) predicts that teaching will be most effective if instruction is targeted at the point where a child transitions from knowledge they can learn without help to knowledge they can learn only with the help of others. Given that studies using both nationally-representative samples (e.g., Fryer & Levitt, 2006) and smaller-scale samples (e.g., Aunola et al., 2004) have reported wide variation in children's early mathematical skills, it seems unlikely that teachers would be able to target instruction to each child's ZPD without radically differentiating the content taught within their classroom.

Unfortunately, little work has closely investigated how tailoring instruction to student ability levels might influence mathematics achievement over time, and even less research has investigated such approaches through experimental designs. A meta-analysis of early-grade

math interventions found slightly higher average effect sizes for interventions that included formative assessment as a key component compared with interventions that did not, but this difference was not statistically significant (p = 0.15; Burns, Codding, Boyce, & Lukito, 2010). However, most studies included in the meta-analysis did not report how the formative assessment was actually used, so it remains unclear whether teachers included in the studies used the formative assessment to provide content better targeted to students' specific needs.

Further indirect evidence is available from the broad literature on cognitive tutoring programs, many of which have been designed to improve mathematics achievement (e.g., Anderson, Corbett, Koedinger, & Pelletier, 1995; Ritter, Kulikowich, Lei, McGuire, & Morgan, 2007). Cognitive tutor programs are computer-based interventions in which students interact with an instructive software that adapts content to the student's specific learning needs over time. A recent meta-analysis of 26 K-12 math cognitive tutor intervention studies found small effect sizes ranging from 0.01 to 0.09 depending on the model specification (Steenbergen-Hu & Cooper, 2013). Yet, these programs were typically administered in a short period of time in a setting removed from the child's typical classroom instruction, so a curriculum intervention that encouraged teachers to differentiate instruction might produce quite different effects.

Some other relevant experimental evidence is provided by Clements, Sarama, Wolfe, and Spitler (2013), who evaluated the long-run effects of a preschool math intervention that substantially boosted math achievement at the end of preschool (Hedge's g = 0.71). Some of the students who received the intervention were randomly assigned to a separate condition designed to abate fadeout effects. This "follow-through" intervention condition included professional development (PD) sessions for kindergarten and first-grade teachers where study-designers encouraged teachers to differentiate their instruction to further challenge the students who

already mastered basic math skills during the preschool intervention. Indeed, results indicated that students in this "follow-through" treatment condition outperformed students that only received the preschool treatment at the end of first grade. However, it remains unclear if teachers in the "follow-through" condition truly differentiated their instruction, or if the PD sessions merely convinced them to teach more advanced content to all students in the class.

The most thorough, and promising, experimental evidence on differentiated instruction comes from a reading intervention designed by Connor and colleagues (2013). In this study, classrooms were randomly assigned to an Individualizing Student Instruction (ISI) intervention in reading. As part of the intervention, student reading achievement was regularly tested throughout the school year, and teachers attended PD sessions in which they discussed strategies for grouping students based on their reading assessment results. The groups were organized so that students with similar skills would work together, and the teacher would deliver specialized content to each group to help them improve in specific areas. The intervention was tested on a single cohort of students over the course of first, second, and third grade, and classrooms were randomly assigned to the treatment or comparison condition each year. Connor and colleagues reported fairly large treatment effects (0.25 – 0.44 SD's) on reading measures at all three grades assessed, and they also found that receiving individualized instruction in reading over the course of three consecutive years boosted reading achievement measured at the end of third grade. Thus, a well-designed individualized instruction program may be highly effective at raising early academic skills.

Questions remain as to whether such approaches could be effective in early-grade mathematics, and the success of any individualization program would also depend upon the buy-in of teachers. If teachers prefer teaching a single curriculum, they may be resistant to dropping

the “one size fits all” model in order to adapt an ISI program that may require more effort and attention than they are accustomed to providing. Further, elementary school teachers have been identified as having high levels of discomfort and anxiety in teaching mathematics (e.g., Bursal & Paznokas, 2006), and an individualization program would require a high level of flexibility and proficiency in math in order to adapt curriculum to meet the specific needs of students with different skill sets. However, if ISI was shown to be effective at boosting early-grade math achievement, such instructional models could become a key component of any effort to raise mathematics achievement throughout K-12 schooling, and ISI may also provide an important tool for abating fadeout in the years following early intervention.

Current Study

In the current study, I test the effects of an ISI program in mathematics on the math achievement of students in second and third grade. I also test the effects of participating in the ISI math intervention over the course of two consecutive years. Based on the success of a similar program targeted at reading reported by Connor and colleagues (2013), and the limited success of tailored cognitive tutor programs in math (Steenbergen-Hu & Cooper, 2013), I expect that ISI will positively affect student math achievement.

Method

Study Design

Data were drawn from an early elementary school intervention designed to test the effectiveness of ISI for children from low-income communities. Study designers recruited second grade classrooms (n=44) in 6 schools serving low-income families in northern Florida. Immediately following the start of the school year, classrooms were randomly assigned to one of two conditions: 1) ISI in math; 2) ISI in reading. Students enrolled in these classes (classroom

n= 44) were recruited for study participation (total second grade *n* = 646), and their mathematics and reading achievement was assessed throughout the year.

During the following school year, students were tracked into third grade, and third-grade teachers (classroom *n*= 40) were again randomly assigned to implement ISI in either math or reading. Many students from the second-grade sample remained with the study through third grade (*n*=439), and their new classmates were also recruited for study participation (total third grade *n* = 626). As with the previous study year, students' mathematics and reading achievement was repeatedly assessed throughout the school year.

Thus, for both study years, the reading arm of the intervention provides an active control group for the math arm of the intervention, as teachers in the reading condition attended the same number of PD sessions as teachers in the math condition. This should help ensure that the motivation of teachers was similar across both groups and mitigate Hawthorne effects.

It should also be noted that the study began in first grade, where participating first-grade teachers in study schools were assigned to implement either ISI in reading or a math curriculum called Math Pals (Fuchs et al., 1997). Because the math arm of the study in first grade was not an ISI program, I limit the focus of the current paper to the second and third grade study years. However, in the Appendix, I present OLS models in which I estimate treatment impacts for the first-grade intervention, and I found no treatment effects.

Intervention. During second and third grade, teachers in both the math and reading conditions attended PD sessions in August and January, and PD sessions were focused on individualizing instruction in the respective subjects. Teachers were also encouraged to attend monthly meetings with other teachers implementing the same curriculum in their school (i.e.,

other teachers in the same treatment condition), and they received bi-weekly instructional coaching in either math or reading.

The key component of the math intervention was the regular use of the ISI-math assessment, the *Concepts and Applied Skills Assessment* (CASA), which was used to group children into skill-based learning groups. The CASA items ranged in difficulty from content typical of kindergarten through fifth grade, and items covered topics such as numeracy, geometry, measurement and data, and word problems (alpha = .92). The CASA was administered three times over the school year, and teachers discussed score reports at their monthly teacher meetings. At these meetings, study researchers met with teachers to assist in devising strategies to help students make achievement gains in areas identified as weaknesses on the CASA.

Teachers were directed to create small student groups within their classrooms based on CASA results (a typical class would have students sorted into 4 or 5 small groups), and they would restructure these groups throughout the year depending on how students performed. Groups were created so that students with similar learning needs were grouped together, and study researchers attending the monthly teacher meetings would provide math activities designed to supplement the curriculum already taught in class. These math activities were adapted from the Math Pals curriculum (Fuchs et al., 1997), and teachers were instructed to target the activities to the respective levels of their small groups.

The intervention was designed such that teachers would conduct their standard whole-class curriculum in mathematics as usual, then follow their standard math content with small group time (i.e., ISI-math program) for about 20 minutes. This small group time was called "station time," as one small group would meet with the teacher, and the other groups would work

independently or together on targeted activities. During station time, students would work on activities designed to reinforce skills identified by the CASA as needing improvement. Thus, station-time content may deviate considerably from the whole-class topic taught in class depending on the specific needs of a particular small group. Teachers were instructed to meet with each group at least twice per week, and they were encouraged to meet most frequently with students who had weaker scores on the CASA.

Previously reported findings. Connor and colleagues (2013) reported positive effects of assignment to the reading arm of the intervention on WJ-III *Letter-Word Identification* scores in grades 1 (Cohen's d = 0.32), 2 (Cohen's d = 0.44), and 3 (Cohen's d = 0.25). Further, they found positive effects on *Passage Comprehension* scores in grades 1 (Cohen's d = 0.36) and 2 (Cohen's d = 0.43), but not grade 3 (Cohen's d = 0.06, not statistically significant). Connor and colleagues (in press) also recently reported effects for the second-grade year of the math treatment arm on math achievement. Using a hierarchical linear modeling (HLM) approach with baseline math scores mean centered at the classroom level, they reported positive effects for two of the math subtests given at the end of second grade (0.41 to 0.60). In the current paper, I replicate these second-grade math intervention treatment impact models using alternative modeling approaches (described in detail below), and I also extend this analysis to the third-grade year of the study.

Analytic Approach

The current study seeks to estimate the effect of assignment to the individualized mathematics instruction condition on the end-of-year math achievement of students in second and third grade, and I also investigate the impact of spending two years in individualized instruction in mathematics on spring grade 3 math scores. In comparison with the analyses

reported by Connor and colleagues (2013; in press), I take a more traditional econometric approach by simply modeling student outcomes as a function of treatment status and baseline characteristics using OLS models:

$$Math_{ijs} = \alpha_0 + \beta_1 Tx_{ijs} + ØBaseline_{ijs} + \delta School_s + \varepsilon_{ijs}$$

where $Math_{ijs}$ is math achievement measured in the spring of a given school year for the *ith* student in classroom *j* in school *s.* Treatment status (coded "1" if assigned to the math intervention; coded "0" if assigned to the reading intervention), determined by random assignment prior to the school year, is indicated by Tx_{ijs}, and β_1 represents the treatment impact for a given school year. In models presented, I control for a series of baseline covariates (denoted by $Baseline_{ijs}$) to account for any random assignment imbalance that might bias results. This vector of baseline characteristics also contains fall measures of student achievement, which should correlate highly with spring outcome measures and consequently improve the precision of the model's estimates. Because random assignment was conducted at the classroom level, I also test models that control for a school fixed effect ($\delta School_s$), which should adjust for any unmeasured school differences and any selection factors that led students to enrolling in different schools. [2] Finally, the error term, ε_{ijs}, should be uncorrelated with β_1 if random assignment produced groups that were equal on all observable and unobservable characteristics.

In most models presented, I adjust standard errors for classroom-level clustering (i.e., the level of random assignment) using the Huber-White standard error estimator in Stata 13.0

[2] Unlike Connor and colleagues (in press), I do not group-mean center baseline test scores at the classroom level. If random assignment did not produce balanced groups, then group-mean centering the scores may mask differences between classrooms in the treatment and control group. This could bias results when using the group-mean centered scores as control variables in treatment impact models. However, in some models I do include school fixed effects, which essentially estimates models within each respective school. This is akin to adding a dummy variable for each school to the model. Because random assignment occurred within schools, this should not mask any baseline imbalance issues.

(Rogers, 1994). In models with school fixed effects, standard errors were clustered at the school level. However, as there were only 6 schools in the study, it is likely that these standard errors were biased because approximately 30 clustering units are typically needed to produce reliable clustered standard errors (see Cameron & Miller, 2015). Therefore, the preferred estimates will come from fully-controlled models (i.e., student-level baseline characteristics) with standard errors clustered at the classroom level, and the school fixed effect models will provide a check for whether school selection factors may have influenced treatment effects.

Because the intervention was designed to help teachers individualize instruction for children with varying levels of mathematics achievement, I also investigate whether the treatment worked better for high- or low-achieving students by interacting treatment status with baseline math test scores. Also, recall that students were followed longitudinally over the course of second and third grade, allowing me to test whether random assignment to the intervention during both grades 2 and 3 produced positive long-run impacts on math achievement measured at the end of grade 3. In effect, this model is the same as the single-year treatment impact models described above, except that treatment status is measured by a series of 4 dummy variables indicating different treatment combinations across the two years: 1) treated both years; 2) treated second grade, control (i.e., reading intervention) third grade; 3) control second grade, treated third grade; 4) control both years. In this model, I omit students that were in the control condition both years as the comparison group, though I also test for comparisons between the three groups that received treatment in at least one year. Finally, in this model, I control for baseline characteristics measured at the fall of grade 2.

For both the second and third grade year, respectively, I limit the analytic sample to students who had valid random assignment data in the fall and at least one non-missing

mathematics achievement test in the spring (second grade n = 519, classroom n = 33; third grade n = 472, classroom n = 32). For the models testing combinations of treatments over both study years, I limit the sample to children who had at least one valid spring of third grade math test score and valid random assignment data in grades 2 and 3 (n = 379).

Measures

Mathematics achievement. Student mathematics achievement was assessed using the Woodcock-Johnson III Tests of Achievement (WJ-III; Woodcock, McGrew, & Mather, 2001). Two subscales of the WJ-III were used to measure mathematics achievement: *Applied Problems* and *Math Fluency*. Both tests were administered at five individual time-points over the two study years: during the fall (i.e., baseline), and spring (i.e., post-treatment) of second grade; during the fall (i.e., baseline), winter (i.e., mid-treatment) and spring (i.e., post-treatment) of third grade.

The *Applied Problems* subtest is a commonly used measure of mathematics ability that asks children to work through a series of mathematical problems that cover a range of mathematical concepts and procedures. Over the grades tested here, the *Applied Problems* test covers topics such as counting, adding and subtracting, and multiplication and division. The subtest usually takes approximately 15 minutes to complete, and it is administered by a trained examiner who presents the child with questions in a one-on-one setting. The child answers either verbally or through pointing at the test materials. The psychometric properties of the *Applied Problems* subtest have been widely reported, and it has been shown to have strong internal reliability (alpha= 0.88; Woodcock et al. 2001).

The *Math Fluency* subtest provided a second measure of mathematics achievement. Like the *Applied Problems* test, the *Math Fluency* subtest asks students to work through a series

questions involving addition, subtraction, multiplication and division. However, the *Math Fluency* test heavily involves making simple calculations, and it is administered via paper and pencil. The test is timed and students are asked to complete as many problems as possible within the time constraint. As with the *Applied Problems* subtest, the psychometric properties of the *Math Fluency* test are widely available, and it has been shown to have good internal reliability (alpha = 0.90; Woodcock et al., 2001).

Finally, study designers also administered *The KeyMath- Third Edition* (Connolly, 2007) as a measure of basic math concepts (e.g., numeration, geometry, measurement), operations (e.g., addition, subtraction, multiplication, etc.), and problem applications. Questions on the *KeyMath* are presented orally, and students respond orally. The test is not timed, and the exam is stopped when students hit a ceiling of 3 incorrect consecutive answers. The test was administered in the fall and spring of second grade, and the fall of third grade. I use the total score across all subtests administered at each time-point, and the test has been shown to have good internal reliability (alpha=0.87; Connolly, 2007).

For all of the analyses that follow, I used the standard scores available for all three math exams. These standard scores have been nationally normed, which allows for easy comparisons between the study sample and a nationally representative sample. For each time point, I also created a composite measure of math achievement, which was the average of each students' non-missing math tests. For the fall and spring of second grade, and the fall of third grade, this included both WJ-III subtests and the *KeyMath* test. For the winter and spring of third grade, this only included the two WJ-III tests.

Reading achievement. Reading achievement was also measured using the WJ-III. Three reading subtests were administered: *Letter-Word Identification, Picture Vocabulary,* and

Passage Comprehension. For young children, the *Letter-Word Identification* subtest asks students to name basic letters, and the test progresses to reading more challenging words as students grow older. In the *Picture Vocabulary* test, students are presented with a series of drawings and they are asked to identify the object or action depicted in the picture. Finally, the *Passage Comprehension* subtest asks students to read passages and answer questions designed to measure their comprehension of what they read. I use the reading subtests as baseline control measures in the fall of both the second- and third-grade study years.

Intervention Fidelity. During both years, classroom observations were videotaped and conducted live during mathematics and reading instruction. For teachers assigned to the math arm of the intervention, observers scheduled observations at the teachers' convenience, and they attended class during math instruction. Observations were conducted in the fall, winter and spring of both treatment years, and each observation lasted approximately 45 to 60 minutes. For classes assigned to the mathematics intervention, observers rated math instruction using a math-focused version of the *Individualizing Student Instruction Fidelity Scale* (short form; see Connor et al., 2013; Connor et al., in press), which included items regarding teachers' ability to individualize instruction in mathematics. The observational measure also included items designed to measure teacher organization and warmth/responsiveness to student needs. Inter-rater agreement was calculated for 10% of the sample, and kappa values ranged from .73 to .82. In supplementary analyses, I use the observational assessment for classrooms assigned to the math treatment condition to examine whether individualization efforts correlated with student achievement measures.

Other student characteristics. Information regarding student ethnicity, gender, and date of birth was obtained via administrative data from study schools. Age has been scaled as age in years at first-grade entry.

Results

Baseline Equivalence

Table 3.1 presents descriptive characteristics for students assigned to the math and reading conditions during both study years. For ease of exposition, I refer to the math intervention as the "treatment group" and the reading intervention as the "control group."

In order to assess whether randomization produced groups equivalent on observable measures, I ran two different analytic checks. First, I assessed whether each individual observable characteristic differed between treatment and control by running a series of bivariate regressions where a given baseline characteristic for either second or third grade was regressed on a dummy indicator for treatment status in that corresponding year. Next, I tested whether the entire set of baseline characteristics jointly differed between the two groups by running a single regression in which treatment status was regressed on the entire set of baseline observables. For the test scores, I used the fall tests for a particular year as the baseline measures of achievement (i.e., for the second-grade study year, fall of second grade tests serve as baseline measures; for the third-grade study year, fall of third grade tests serve as baseline measures).

[Insert Table 3.1]

As Table 3.1 reflects, I found few indications of baseline differences between the two groups when each baseline characteristic was examined individually. Across both years, no single characteristic was found to be statistically significantly different between the two groups, though African Americans were slightly more likely to be assigned to the control group in grade

3 (p = 0.08). However, when looking at the joint set of baseline characteristics, there was some cause for concern for the second-grade study year. I regressed second grade treatment status on the set of baseline covariates listed in Table 3.1, and found that the *Math Fluency* test positively predicted placement in the treatment group (β = 0.15, SE= 0.05), as did age (β = 0.22, SE= 0.07). Further, a joint F-test produced a statistically significant result [$F(11,31) = 6.32$, $p < 0.001$], indicating that the set of baseline covariates jointly differed between the treatment and control groups.[3] This suggests that students in the second-grade treatment group may have been higher achieving in mathematics before the intervention began, but no other baseline test, including the two other tests of math achievement, were found to differ between the two groups. However, in all models presented, I will control for all baseline tests in order to adjust for possible differences, observed and unobserved, in baseline levels of achievement.

Sample Characteristics

Table 3.1 also shows descriptive characteristics of the study sample during each year. Across both years of the intervention, approximately 80% of study children were White, and 7% identified as Black. The sample was evenly split between boys and girls, and the average child was a little over six and a half years old during the fall of first grade.

The test score data across all three years shows that at various time-points the sample did differ from a nationally representative sample on specific subtests. Recall that the scores were standardized to national norms to have a mean of 100 and SD of 15, and I converted the scores to z-scores using these national parameters. For each score displayed, a nationally representative

[3] In this OLS model, list-wise deletion was used to drop students who were missing baseline test scores (see Table 3.2). As a comparison, I also tested a structural equation model (SEM) with full information maximum likelihood (FIML) to account for missing data. In this model, the baseline *Math Fluency* test was again a significant predictor of placement in treatment (β = 0.15, SE= 0.04), though age was not (β = 0.12, SE= 0.07). However, a Chi-square test of model fit also indicated that the set of baseline covariates jointly differed between treatment and control (p = 0.034).

sample would have a z-score mean of 0, so the averages displayed in Table 3.1 indicate how far the study sample was from the national average in SD units. Most of the average scores are close to zero, with no consistent pattern discernable across either math or reading. The *Letter Word Identification* task may be an exception, as students consistently scored approximately one half of a SD higher than the national norm across both years, but this was not seen across any of the other reading tests.

Attrition and Missing Data

In Table 3.2, I present descriptive statistics detailing patterns of missingness and attrition across both study years. Attrition for each year was defined as having valid random assignment data in the fall and no valid math score data in the spring. In grade 2, approximately 20% of the sample was lost due to attrition, and this did not differ between the treatment and control groups ($p = 0.749$). In grade 3, 21% of the treatment group was lost due to attrition and 28% of the control group was lost. Again, this attrition rate was not significantly different between the two groups ($p = 0.147$).

[Insert Table 3.2]

I also present rates of missingness on various baseline test measures. Because there were some indications of baseline imbalance, I must control for baseline achievement tests to adjust for possible differences in baseline ability between the treatment and control groups. However, controlling for these tests could introduce bias if one group has more missing baseline test scores than the other group. In general, although I found some high rates of missing baseline tests, these rates were not different between the treatment and control groups.

In grade 2, 25% of the treatment group was missing a baseline WJ-III math test, as opposed to 32% of the control group, but this rate difference was not statistically significant ($p =$

.279). Missing rates were much better for the *KeyMath* test and the WJ-III reading tests, as few students from either group (2-4%) were missing on these baseline measures.

In grade 3, few students (2-3%) were missing any WJ-III baseline test, but missing rates were higher for the *KeyMath* test. Approximately 37% of the treatment group was missing a *KeyMath* test at baseline, and 46% of the control group was missing a baseline test, though this rate was not statistically significantly different ($p = .237$).

In the regressions that follow, I imputed missing data on baseline test scores by giving a student the standardized average of all non-missing cognitive subtests from that same measurement period (e.g., if a student was missing the fall *Math Fluency* test in second grade, they were given the average of their scores across their non-missing math and reading tests from the fall of second grade). I then included dummy indicators for missing baseline subtests in the regression models to adjust for the imputation. I also ran models that used listwise deletion (i.e., students with any missing test score data were dropped from the model), and results were nearly identical. Finally, in the appendix, I present results that used SEM with FIML to adjust for missing data, and treatment impact results were again nearly identical. Here, I present results from OLS models to allow for more flexible modeling specifications (i.e., standard error adjustments; school fixed effects).

Treatment Impact Results

I begin with results for each individual year of the math intervention, with results for grade 2 shown Table 3.3 and results for grade 3 shown in Table 3.4. In Table 3.5, I present results from models that tested the impact of various combinations of treatments across grades 2 and 3 on third grade math outcomes. In most models, robust standard errors were adjusted for classroom-level clustering, and in school fixed effect models, standard errors were adjusted for

school-level clustering. Test scores were converted to nationally normed z-scores, so coefficients can be interpreted as effect sizes. Within each set of results, I begin with the bivariate association between treatment status and the spring test score. I then add baseline covariates (preferred estimates) and finally, school fixed effects.

[Insert Table 3.3]

Single-Grade Findings. Table 3.3 presents results for math achievement in second grade, and results were largely null. In columns 1 through 3, I present models that tested the impact of assignment to the treatment on the math composite score (i.e., the average of every students' non-missing *Applied Problems*, *Math Fluency*, and *KeyMath* subtests). The unadjusted treatment effect was 0.13 (*SE*= 0.13), and this fell to 0.06 (*SE*= 0.06) when baseline controls were added.

The results for the specific subtests are shown in columns 4 through 12. I detected some positive program impacts on the *Math Fluency* subtest (columns 7 through 9). Unfortunately, this same subtest was found to be significantly related to treatment assignment at baseline in the aforementioned model that tested the association between second grade treatment assignment and all baseline characteristics. Thus, it is difficult to draw strong causal conclusions regarding the treatment impacts in second grade. Indeed, the treatment impact estimates on the *Math Fluency* subtest also show how baseline imbalance appears to affect results. The bivariate effect was 0.25 (*SE* = 0.19), yet this fell by approximately 35% when baseline controls were added to the model (β = 0.16, *SE* = 0.09), and it fell again when school fixed effects were added (β = .13, *SE* = 0.10). The difference in the point estimate between the bivariate and fully-controlled models suggests that baseline imbalance may be responsible for some of the treatment effect. Although the fully-controlled model should adjust for much of this baseline imbalance, it is

unclear if unmeasured characteristics might also drive the difference between the treatment and control groups.

In Table 3.4, I present results from grade 3 of the study. Much like second grade, I found little indication that the treatment positively affected grade 3 test scores. For the composite math score (recall that no *KeyMath* test was administered in the spring of grade 3), I found a bivariate treatment effect of 0.03 (SE = 0.12) and this fell to a negative, though not-statistically significant, effect when controls were added (β = -0.06, SE = 0.06). Surprisingly, I found a negative and statistically significant effect for the *Applied Problems* subtest when all controls were included in the model (β = -0.11, SE= 0.05). Yet, this negative effect was not detected on the *Math Fluency* subtest, so it remains unlikely that the treatment substantially hindered mathematics achievement during third grade.

Heterogeneity and Impact of Treatments Across Grades. In Table 3.5, I present results from analyses that tested for treatment heterogeneity based on baseline measures of math ability, and I also present results from models that tested for impacts of various combinations of second and third grade treatments on spring of third grade test scores.

To investigate treatment heterogeneity, I tested for interactions between baseline math achievement and treatment status. For these models, the dependent variable was the composite math achievement score, and I included the full set of controls and school fixed effects. For both years of the treatment, I averaged the three baseline achievement tests together, and split the distribution on both measures into thirds. I then interacted treatment with the bottom and top third of the distribution (the middle third acted as the comparison group). For second grade, I found no indication of treatment heterogeneity as neither interaction term produced a statistically significant coefficient. Further, a joint F-test testing whether the interaction terms jointly

contributed to the model was not statistically significant (p = 0.229). For third grade, I found some indication that the treatment may have produced a positive effect for high-achieving students, as I found a positive treatment interaction for students in the top third of the distribution (β = 0.26, SE= 0.11). However, the main effect for treatment was negative and statistically significant (β = -0.17, SE= 0.07), and the joint F-test produced a marginally statistically significant result (p = 0.090). Although the treatment may have worked better for students at the top of the distribution, this certainly was not the stated goal of ISI, and the overall negative main effect of treatment status in this model dims any strong positive interpretations for even this group of students. In models not shown, I also tested treatment interactions with demographic variables, and found no consistent pattern of results.

Finally, I tested for treatment effects across both years of the intervention. For these models, the sample was limited to students that were present in both the second and third grade year of the study (n= 379), and the outcome measure was the composite achievement test for third grade. Recall that the design of the study produced four unique groups over the course of these two years: 1) treated in both second and third grade; 2) treated in second grade only; 3) treated in third grade only; 4) control group (i.e., reading intervention) in both grades. In the models presented in Table 3.5, students that were in the control group during both years serve as the comparison group.

Again, across the bivariate and controlled models, I found no treatment impacts for any of the treatment combinations over the two years. The F-test presented in column 5 tested for significant differences between the 3 groups that included at least one year of treatment status, and this test did not produce a statistically significant result (p = 0.504).

Supplemental Models

Across most of the models tested, I found little impact of the ISI math program on math achievement measures. It is difficult to know why the math arm of the intervention largely failed to move math achievement, yet two possibilities seem most likely: 1) the program was not implemented by teachers; 2) the program itself was incapable of affecting student achievement. In order to provide some sense as to why the program might have failed to produce large positive effects, I turned to the classroom observations. However, my ability to make strong claims based on the classroom observations is severely limited due to small sample sizes and considerable missing data. It should also be noted that analyses with the classroom observational measure are purely post-hoc and exploratory in nature.

Classroom observations were conducted in the fall, winter and spring of both the second and third grade years of the study, and observers looked for evidence of individualization in mathematics instruction in the classrooms assigned to the math arm of the intervention. For second grade, 17 of the 18 teachers had at least one non-missing observational measure, and for third grade, all 17 treatment teachers had at least one non-missing observational measure.

Observers rated teachers across 15 different criteria that assessed how often the teacher made efforts to individualize instruction. Across these 15 items, observers rated teachers along two dimensions of individualization: 1) the use of small groups in general; 2) the use of individualized instruction within small groups. Each item was rated using a Likert-scale that ranged from "0" (i.e., teacher does not implement small groups or individualized instruction) to "3" (i.e., teacher always implements small groups or individualized instruction).

Table 3.6 presents the averages for both the small group and individualized instruction dimensions at each time point. Across both study years, I found few indications that teachers made strong efforts to individualize instruction or use small groups. However, observers saw

more evidence of individualization efforts in second grade than during third grade. In third grade, the small group averages were below 1.5 at each measurement point (a score of "1" indicated that the teacher "rarely" used small groups) and individualization averages fell just above 1.0 at all three points. During second grade, small group and individualization scores grew over the course of the year, but averages were still fairly low considering that teachers were explicitly encouraged to use small groups during math instruction. In the fall of second grade, observers rated teachers below 1 on both small group and individualization. By the spring, the average small group score was 2.18 (a score of "2" indicated that teachers often used small groups), and the individualization average was 1.87.

Thus, based on the classroom observations available, it appears that teachers made limited effort to individualize instruction in mathematics, especially in grade 3. In Table 3.7, I present results from correlational models testing concurrent associations between math achievement scores and individualization efforts as measured by the classroom observational assessment. In effect, these models examine whether individualization efforts correlate with higher achievement scores. In each of these models, I averaged together the scores on the "small group" and "individualization" dimensions of the observational assessment because the correlation between these two dimensions was over 0.90 at each measurement point. Further, because sample sizes were low due to missing test scores (see Table 3.2) and the fact that the math observational assessment was only administered in treatment classes, I pooled the data and treated each observational measure as independent observations. Students with at least one non-missing test score and matching observational assessment were included in the pooled model, which led to the inclusion of 282 children in 31 different classrooms. However, it should be noted that for this sample, variation on the measure of "individualization" was limited. Only

15% of valid cases ever had a measurement rating over "2," with the vast majority of the variation occurring between "0" (i.e., teacher never uses individualized instruction) and "1" (i.e., teacher rarely uses individualized instruction).

In each of the models presented in Table 3.7, standard errors were adjusted for classroom-level clustering and all variables were standardized. I also tested models that adjusted standard errors for non-independence at the student level, and results were nearly identical. In column 1 of Table 3.7, I present results from a bivariate model that tested the simple correlation between individualized instruction efforts and achievement, and this association was nearly zero ($\beta = 0.01$, $SE = 0.04$). In column 2, I added controls, including a lagged measure of average math achievement from the previous measurement wave, and I again found no association between individualization and math achievement ($\beta = 0.04$, $SE = 0.05$). In column 3, I present results from the same model tested using SEM with FIML to adjust for missing data, and results were nearly identical.

Finally, in columns 4 and 5, I present results from models that predicted time spent individualizing instruction using the average math score from the previous wave and demographic variables. These models were tested to investigate whether observable student characteristics predicted time spent on individualizing instruction in mathematics. Across both models, I found no evidence that achievement in the previous wave significantly predicted individualization by the teacher in the next wave. Further, no other student characteristics tested was found to significantly predict time spent individualizing instruction in math.

Discussion

In the current study, I found little indication that the ISI program for mathematics positively affected student math achievement. In second grade, I found a non-significant effect

of 0.06 on the composite score of all math subtests given in the spring; the analogous third grade effect was a non-statistically significant -0.06. I also found no indication that spending two years in the math intervention affected math scores measured at the end of grade 3. Although I found some significant results on certain math subtests given at both grade 2 (effect of 0.16 on *Math Fluency*) and grade 3 (effect of -0.11 on *Applied Problems*), the general pattern of results across most models tested was null.

Poor implementation appears to be a possible culprit for the null findings in both years, though implementation looked somewhat better in grade 2. Classroom observers rated teachers as surprisingly low on measures of "small group time" and "individualization" across both study years, with observational scores being especially low in grade 3. This was somewhat surprising, given that the intervention explicitly asked teachers to organize students in small groups, and the PD sessions administered throughout the year provided teachers with strategies for managing their small groups in order to individualize instruction. Further, it is possible that the observational averages could be biased upwards, because observers scheduled each observational assessment with the study teachers. Thus, demand characteristics may have played a role if study teachers knew that observers were looking for evidence of small group time. Even if they did not substantially alter their instruction on observational days, it is hard to imagine that teachers would have demonstrated a lower proclivity for small group and individualized instruction than was normal during the observational periods. However, it is possible that better implementation scores could have led to similar results, as I did not find a significant correlation between the measure of individualized instruction and student achievement scores.

Interestingly, Connor and colleagues (2013) did not find similar problems with implementation when analyzing the effects of the reading arm of the intervention, and the large

effects reported on reading achievement for the reading intervention also imply a higher rate of implementation. Why might teachers have been so resistant to implementing the intervention in mathematics? Between grades 2 and 3, study schools switched math curricula, changing from Saxon Math to a new curriculum called Everyday Math. Saxon Math is a commonly used curriculum that emphasizes regular procedural practice (see Larson, 2004). Everyday Math also includes procedural practice, but it was designed to be aligned with the Common Core Standards, which emphasizes conceptual content (Common Core Standards Initiative, 2010). Switching to the Common Core Standards aligned curriculum may have negatively affected implementation during grade 3, as teachers likely had a substantial amount of new mathematics content to learn in order to implement Everyday Math. Adding on the ISI program at the same time may have simply been too much.

Indeed, observers found more evidence of implementation in grade 2, especially during the spring. I did find some evidence of positive treatment effects in grade 2, as treatment students scored higher on the spring *Math Fluency* test ($\beta = 0.16$). Unfortunately, I found some indication of baseline imbalance on this same test in the fall, and adding covariates to the model substantially reduced the treatment effect when compared with the uncontrolled model. However, given that the control group for this study was not a traditional control group, but instead a group that received an ISI program in reading, it is not surprising that I only detected positive effects on the *Math Fluency* test. The other two tests administered, *Applied Problems* and *KeyMath*, were both language intensive, as both tests required students to respond orally to study examiners. In contrast, the *Math Fluency* test is a timed test administered via paper and pencil, and students are simply asked to perform as many calculations as possible. Connor and colleagues (2013) reported positive effects of the reading intervention on language skills, and it

is possible that these language skill boosts also helped the students assigned to the reading condition perform better on the two language-heavy math tests. Many studies have reported correlations between language and math skills in early childhood (e.g., Fuchs et al., 2010; Purpura & Ganley, 2014), but experimental evidence of transfer remains limited. Nevertheless, although the "active" control group helped ensure balance between intervention and control teachers in terms of PD sessions attended, it made it difficult to rule out transfer effects as boosting language skills may have also boosted math skills for at least some students.

The active control group can be counted as one of a few limitations to the current study. It should also be noted that the original designers of the study (e.g., Connor and colleagues, 2013) are primarily reading researchers, who originally designed the intervention as an ISI program for reading achievement. The math arm of the intervention was originally intended to be the "active control" group for the reading arm. Thus, the math arm of the intervention could have simply been less developed than the reading arm. Indeed, the ISI reading program had been pilot tested in multiple studies before study designers brought it to scale in the current data (Connor et al., 2007; Connor, Morrison, Schatschneider, et al., 2011; Connor, Piasta, et al., 2009), but the math intervention had not received this earlier treatment. Consequently, the current study represents a pilot testing of the math program, and future efforts to implement the ISI program in mathematics may prove more successful. Further, small sample sizes and a high prevalence of missing data prevented me from fully exploring correlations between the observational assessment and concurrent achievement. I also could not test whether intervention-driven impacts on the observational assessment affected student achievement because the math observational assessment was only administered in classrooms assigned to the math intervention.

Given these limitations, and given the positive reading effects reported by Connor and colleagues (2013), it seems premature to write off ISI programs in mathematics as unworthy of further pursuit. If future intervention efforts gained more traction with study teachers, then student achievement effects could be quite different from the effects reported here. Yet, it is concerning that the individualized measure from the observational assessment had no apparent correlation with student achievement, even if most of the variation was centered between "never" individualizing instruction and "rarely" individualizing instruction. Thus, it seems plausible that the null effects were due to some combination of the fact that the program was newly developed and that the study teachers were apparently resistant to implementing it. The apparent resistance from teachers to implement the program analyzed here should not be overlooked. Individualizing instruction in mathematics may simply be a difficult task for elementary school teachers who are typically much less comfortable with mathematics than reading (e.g., Bursal & Paznokas, 2006). These results should be seen as only a first step toward understanding how ISI programs might work in elementary school mathematics, but future efforts may need to substantially revise the approach to differentiating instruction in mathematics from the approach explored here.

References

Anderson, J. R., Corbett, A. T., Koedinger, K. R., & Pelletier, R. (1995). Cognitive tutors: Lessons learned. *The journal of the learning sciences*, *4*(2), 167-207.

Aunola, K., Leskinen, E., Lerkkanen, M. K., & Nurmi, J. E. (2004). Developmental dynamics of math performance from preschool to grade 2. *Journal of Educational Psychology*, *96*(4), 699.

Bailey, D., Duncan, G. J., Odgers, C. L., & Yu, W. (2017). Persistence and fadeout in the impacts of child and adolescent interventions. *Journal of Research on Educational Effectiveness*, *10*(1), 7-39.

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

Bailey, D. H., Siegler, R. S., & Geary, D. C. (2014). Early predictors of middle school fraction knowledge. *Developmental Science, 17*(5), 775-785. doi: 10.1111/desc.12155

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

Bronfenbrenner, U., & Morris, P. (2006). The bioecological model of human development. In W. Damon & R. M. Lerner (Eds.), *Handbook of child psychology: Vol 1. Theoretical models of human development* (6th ed., pp. 793-828). New York: Wiley.

Burns, M. K., Codding, R. S., Boice, C. H., & Lukito, G. (2010). Meta-analysis of acquisition and fluency math interventions with instructional and frustration level skills: Evidence for a skill-by-treatment interaction. *School Psychology Review*, *39*(1), 69.

Bursal, M., & Paznokas, L. (2006). Mathematics anxiety and preservice elementary teachers' confidence to teach mathematics and science. *School Science and Mathematics*, *106*(4), 173-180.

Cameron, A. C., & Miller, D. L. (2015). A practitioner's guide to cluster-robust inference. *Journal of Human Resources*, *50*(2), 317-372.

Claessens, A., Engel, M., & Curran, F.C. (2014). Academic content, student learning, and the persistence of preschool effects. *American Educational Research Journal*, *51*(2), 403-34.

Clements, D. H., Sarama, J., Wolfe, C. B., & Spitler, M. E. (2013). Longitudinal evaluation of a scale-up model for teaching mathematics with trajectories and technologies: Persistence of effects in the third year. *American Educational Research Journal*, *50*(4), 812-850.

Common Core State Standards Initiative: About the Standards. (2010). Retrieved from: http://www.corestandards.org/about-the-standards

Connolly, A. J. (2007). *KeyMath diagnostic assessment (3rd ed)*. Minneapolis: Pearson, Inc.

Connor, C. M., Morrison, F. J., Fishman, B. J., Crowe, E. C., Al Otaiba, S., & Schatschneider, C. (2013). A longitudinal cluster-randomized controlled study on the accumulating effects of individualized literacy instruction on students' reading from first through third grade. *Psychological Science*, *24*(8), 1408-1419. doi:10.1177/0956797612472204

Connor, C. M., Morrison, F. J., Fishman, B., Giuliani, S., Luck, M., Underwood, P. S., . . . Schatschneider, C. (2011). Testing the impact of Child Characteristics × Instruction interactions on third graders' reading comprehension by differentiating literacy instruction. *Reading Research Quarterly*, *46*, 189–221.

Connor, C. M., Morrison, F. J., Fishman, B. J., Schatschneider, C., & Underwood, P. (2007). The early years: Algorithm-guided individualized reading instruction. *Science*, *315*, 464–465.

doi:10.1126/science.1134513

Connor, C. M., Mazzocco, M. M., Kurz, T., Crowe, E. C., Tighe, E. L., Wood, T. S., & Morrison, F. J. (in press). Using assessment to individualize early mathematics instruction. *Journal of School Psychology*.

Connor, C. M., Piasta, S. B., Fishman, B., Glasney, S., Schatschneider, C., Crowe, E., . . . Morrison, F. J. (2009). Individualizing student instruction precisely: Effects of child x instruction interactions on first graders' literacy development. *Child Development, 80,* 77–100.

Duncan, G. J., Dowsett, C. J., Claessens, A., Magnuson, K., Huston, A. C., Klebanov, P., et al. (2007). School readiness and later achievement. *Developmental Psychology, 43,* 1428-1446. doi: 10.1037/0012-1649.43.6.1428

Engel, M., Claessens, A., & Finch, M. A. (2013). Teaching students what they already know? The (Mis) Alignment between mathematics instructional content and student knowledge in kindergarten. *Educational Evaluation and Policy Analysis, 35*(2), 157-178.

Engel, M., Claessens, A., Watts, T., & Farkas, G. (2016). Mathematics content coverage and student learning in kindergarten. *Educational Researcher*, *45*(5), 293-300.

Fryer, R. G., & Levitt, S. D. (2006). The black-white test score gap through third grade. *American Law and Economics Review, 8*(2), 249-281.

Fuchs, L. S., Fuchs, D., Hamlett, C. L., Phillips, N. B., Karns, K., & Dutka, S. (1997). Enhancing students helping behavior during peer-mediated instruction with conceptual mathematical explanations. *The Elementary School Journal, 97*, 223–249.

Fuchs, L. S., Geary, D. C., Compton, D. L., Fuchs, D., Hamlett, C. L., Seethaler, P. M., … & Schatschneider, C. (2010). Do different types of school mathematics development depend

on different constellations of numerical versus general cognitive abilities? *Developmental Psychology, 46*(6), 1731-1746.

Geary, D. C., Hoard, M. K., Nugent, L., & Bailey, D. H. (2013). Adolescents' functional numeracy is predicted by their school entry number system knowledge. *PLoS ONE, 8*, e54651. doi:10.1371/journal.pone.0054651

Jordan, N. C., Kaplan, D., Oláh, L. N., & Locuniak, M. N. (2006). Number sense growth in kindergarten: A longitudinal investigation of children at risk for mathematics difficulties. *Child Development*, *77*(1), 153-175.

Larson, N. (2004). *Saxon math*. Norman, OK: Saxon Publishers.

Purpura, D. J., & Ganley, C. M. (2014). Working memory and language: Skill-specific or domain-general relations to mathematics? *Journal of Experimental Child Psychology, 122*, 104-121.

Ritter, S., Kulikowich, J., Lei, P. W., McGuire, C. L., & Morgan, P. (2007). What evidence matters? A randomized field trial of Cognitive Tutor Algebra I. *Frontiers in Artificial Intelligence and Applications, 162*, 13.

Rogers, W. (1994). Regression standard errors in clustered samples. *Stata technical bulletin, 3*(13).

Siegler, R. S., Duncan, G. J., Davis-Kean, P. E., Duckworth, K., Claessens, A., Engel, M., … Chen, M. (2012). Early predictors of high school mathematics achievement. *Psychological Science*, *23*(7), 691–697. doi:10.1177/0956797612440101

Steenbergen-Hu, S. & Cooper, H. (2013) A meta-analysis of the effectiveness of intelligent tutoring systems on K-12 students' mathematical learning. *Educational Psychology, 105*(4), 970-987. doi:10.1037/a0032447

Vygotsky, L. (1978). *Mind in Society*. Cambridge: Cambridge University Press.

Watts, T. W., Duncan, G. J., Siegler, R. S., & Davis-Kean, P. E. (2014). What's past is prologue: Relations between early mathematics knowledge and high school achievement. *Educational Researcher, 43*(7), 352-360. doi: 10.3102/0013189X14553660

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock-Johnson tests of achievement*. Itasca, IL: Riverside Publishing.

Table 3.1

Average Baseline Characteristics for Grade 2 and Grade 3 Interventions

	Grade 2			Grade 3		
	Treatment	Control	p-values	Treatment	Control	p-values
Female	0.55	0.54	0.738	0.52	0.55	0.485
Ethnicity						
White	0.84	0.83	0.811	0.84	0.77	0.133
Asian	0.01	0.01	0.732	0.02	0.01	0.912
African American	0.06	0.07	0.800	0.05	0.10	0.077+
Hispanic	0.03	0.04	0.502	0.05	0.05	0.923
Other	0.06	0.05	0.806	0.05	0.07	0.355
Age (years) at G1 fall test	6.64	6.60	0.327	6.58	6.62	0.482
	(0.45)	(0.49)		(0.42)	(0.50)	
Cognitive Baseline Tests						
Applied Problems	0.28	0.25	0.858	0.19	0.07	0.347
	(0.91)	(0.88)		(0.85)	(0.87)	
Math Fluency	0.47	0.29	0.193	0.03	-0.05	0.442
	(0.84)	(0.83)		(0.86)	(0.81)	
KeyMath	-0.34	-0.37	0.704	0.31	-0.04	0.265
	(0.94)	(0.83)		(0.86)	(1.45)	
Picture Vocabulary	0.20	0.13	0.375	0.25	0.16	0.382
	(0.64)	(0.61)		(0.68)	(0.64)	
Letter-Word ID	0.52	0.56	0.751	0.50	0.42	0.509
	(0.85)	(0.87)		(0.67)	(0.78)	
Passage Comp.	-0.01	-0.07	0.461	-0.13	-0.17	0.753
	(0.72)	(0.74)		(0.62)	(0.68)	
Joint F-Test	F(12,31)= 5.80		0.001***	F(12,31)=1.23		0.304
Observations	274	245		261	211	

Note. Standard deviations are presented in parentheses. All cognitive baseline tests were standardized to national norms at each time point. For every test, a nationally representative sample would have a mean of 0 and SD of 1. Thus, the averages displayed here indicate the distance between the study sample and the national average in standard deviation units at each grade. Baseline test scores were measured in the fall of each school year. All baseline tests were taken from the Woodcock-Johnson III (WJ-III) Tests of Achievement except for *KeyMath*. P-values were generated from a series of regressions in which each baseline characteristic was regressed on treatment status for that given year, with standard errors adjusted for clustering at the classroom level (i.e., the level of random assignment). F-test values were generated from regression models that regressed treatment assignment on the entire set of baseline characteristics for the corresponding year. The joint F-tests evaluates whether the entire set of baseline covariates jointly differ between the treatment and control groups.

+ $p<0.10$ * $p < 0.05$ *** $p < 0.001$

Table 3.2

Proportions of Students Attriting and Missing Baseline Achievement Measures

	Treatment	Control	p-values
Grade 2			
Attrition	0.20	0.19	0.749
Missing WJ-III Math	0.25	0.32	0.279
Missing WJ-III Reading	0.04	0.04	0.578
Missing *KeyMath*	0.02	0.02	0.900
Missing Any Test	0.25	0.32	0.279
Grade 3			
Attrition	0.21	0.28	0.147
Missing WJ-III Subtest	0.03	0.02	0.611
Missing *KeyMath*	0.37	0.46	0.218
Missing Any Test	0.38	0.46	0.237

Note. P-values were generated from a series of regressions in which indicators for attrition and missing baseline test scores were regressed, respectively, on treatment status. Attrition was coded "1" if a student was present at random assignment but was missing on all math test scores in the spring of a given school year. Descriptive statistics for students missing on baseline tests were generated off of the analysis sample (i.e., the non-attriting sample in each school year: G2 n = 519; G3 n= 472).

Table 3.3

Grade 2 Treatment Impacts on Spring Math Test Scores

	Math Composite			Applied Problems			Math Fluency			Key Math		
	Bivariate	Controls	School FE	Bivariate	Controls	School FE	Bivariate	Controls	School FE	Bivariate	Controls	School FE
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Treatment	0.128	0.059	0.055	0.097	0.055	0.053	0.245	0.163+	0.131	0.071	0.004	0.020
	(0.126)	(0.057)	(0.047)	(0.147)	(0.061)	(0.066)	(0.186)	(0.090)	(0.101)	(0.115)	(0.070)	(0.046)
Female		-0.012	-0.018		0.005	-0.001		0.027	0.013		-0.048	-0.043
		(0.044)	(0.026)		(0.058)	(0.032)		(0.055)	(0.023)		(0.049)	(0.042)
Asian		0.066	0.046		-0.093	-0.109		0.424	0.439		-0.085	-0.122
		(0.240)	(0.169)		(0.294)	(0.284)		(0.383)	(0.375)		(0.168)	(0.174)
Black		-0.012	-0.006		0.025	0.023		0.060	0.054		0.102	0.103***
		(0.082)	(0.048)		(0.059)	(0.073)		(0.127)	(0.086)		(0.068)	(0.017)
Hispanic		0.038	0.035		0.061	0.055		0.178	0.190		0.107	0.085
		(0.090)	(0.080)		(0.171)	(0.060)		(0.174)	(0.151)		(0.106)	(0.090)
Other		-0.079	-0.080		-0.018	-0.018		-0.060	-0.064		-0.093	-0.100
		(0.055)	(0.062)		(0.096)	(0.085)		(0.142)	(0.169)		(0.065)	(0.079)
Age (years)		-0.086	-0.078		0.007	0.017		-0.047	-0.053		-0.053	-0.042
		(0.054)	(0.068)		(0.082)	(0.049)		(0.071)	(0.061)		(0.059)	(0.084)
Baseline Tests												
Applied Problems		0.208**	0.191**		0.268***	0.254**		0.196**	0.172**		0.225***	0.214**
		(0.060)	(0.051)		(0.059)	(0.061)		(0.062)	(0.043)		(0.049)	(0.037)
Math Fluency		0.349***	0.349***		0.238***	0.234**		0.765***	0.766***		0.073+	0.069*
		(0.032)	(0.020)		(0.047)	(0.052)		(0.060)	(0.044)		(0.037)	(0.027)
Key Math		0.327**	0.336+		0.343**	0.355*		-0.093	-0.076		0.609***	0.621***
		(0.108)	(0.143)		(0.096)	(0.113)		(0.066)	(0.090)		(0.038)	(0.048)
Picture Vocab.		0.046	0.050		0.176**	0.183*		-0.050	-0.063		0.060	0.073
		(0.047)	(0.031)		(0.058)	(0.071)		(0.068)	(0.040)		(0.050)	(0.038)
Letter Word ID.		0.141*	0.132**		0.144+	0.142+		0.213**	0.189**		0.061	0.062
		(0.054)	(0.033)		(0.083)	(0.060)		(0.077)	(0.044)		(0.063)	(0.052)
Passage Comp.		-0.107+	-0.092*		-0.090	-0.083		-0.101	-0.070		0.029	0.026
		(0.056)	(0.034)		(0.079)	(0.056)		(0.097)	(0.066)		(0.056)	(0.033)
N	519	519	519	376	376	376	376	376	376	514	514	514
R-squared	0.005	0.794	0.789	0.003	0.701	0.695	0.015	0.679	0.663	0.002	0.793	0.794

Note. Robust standard errors were adjusted for classroom-level clustering and are presented in parentheses. In school fixed effects models, standard errors were clustered at the school level. All test scores were standardized to national norms, so coefficients can be interpreted as effect sizes. The Math Composite is the average of the Applied Problems, Math Fluency, and Key Math subscores. For gender, males are the comparison group, and for ethnicity, Whites are the comparison group. Missing dummies for missing baseline test scores were used in all models with control variables.
+ p<0.10 * p< 0.05 ** p < 0.01 *** p < 0.001

Table 3.4

Grade 3 Treatment Impacts on Spring Math Test Scores

	Math Composite			Applied Problems			Math Fluency		
	Bivariate	Controls	School FE	Bivariate	Controls	School FE	Bivariate	Controls	School FE
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Treatment	0.032	-0.063	-0.049	0.002	-0.111*	-0.093	0.062	-0.016	-0.004
	(0.121)	(0.058)	(0.050)	(0.127)	(0.049)	(0.051)	(0.133)	(0.084)	(0.095)
Female		-0.017	-0.011		-0.072	-0.069		0.039	0.046
		(0.046)	(0.024)		(0.053)	(0.051)		(0.061)	(0.042)
Asian		0.095	0.159		0.100	0.138		0.091	0.180
		(0.166)	(0.143)		(0.217)	(0.153)		(0.168)	(0.176)
Black		-0.019	-0.017		-0.151+	-0.163+		0.112	0.129
		(0.087)	(0.101)		(0.084)	(0.079)		(0.112)	(0.140)
Hispanic		-0.046	-0.053		-0.083	-0.104		-0.009	-0.002
		(0.087)	(0.092)		(0.133)	(0.151)		(0.105)	(0.078)
Other		0.020	0.009		0.025	0.004		0.015	0.014
		(0.063)	(0.070)		(0.089)	(0.153)		(0.095)	(0.074)
Age (years)		-0.065	-0.070*		-0.215**	-0.220**		0.086	0.079
		(0.061)	(0.024)		(0.067)	(0.035)		(0.075)	(0.042)
Baseline Tests									
Applied Problems		0.272***	0.263***		0.475***	0.468***		0.069	0.059
		(0.033)	(0.036)		(0.050)	(0.046)		(0.042)	(0.039)
Math Fluency		0.348***	0.355***		0.020	0.021		0.676***	0.689***
		(0.029)	(0.030)		(0.034)	(0.040)		(0.037)	(0.024)
Key Math		0.100*	0.092*		0.107	0.107		0.093***	0.078**
		(0.046)	(0.031)		(0.078)	(0.067)		(0.021)	(0.014)
Picture Vocab.		-0.013	-0.013		0.028	0.028		-0.054	-0.054
		(0.035)	(0.034)		(0.051)	(0.045)		(0.044)	(0.044)
Letter Word ID.		0.107*	0.128**		0.043	0.048		0.171**	0.208**
		(0.048)	(0.029)		(0.056)	(0.028)		(0.061)	(0.041)
Passage Comp.		0.046	0.043		0.129*	0.134*		-0.036	-0.048
		(0.040)	(0.027)		(0.054)	(0.052)		(0.057)	(0.034)
N	472	472	472	472	472	472	472	472	472
R-squared	0.001	0.709	0.718	0.000	0.629	0.626	0.001	0.631	0.647

Note. Robust standard errors were adjusted for classroom-level clustering and are presented in parentheses. In school fixed effect models, standard errors were clustered at the school level. All test scores were standardized to national norms, so coefficients can be interpreted as effect sizes. The Math Composite is the average of the Applied Problems and Math Fluency subscores. For gender, males are the comparison group, and for ethnicity, Whites are the comparison group. Missing dummies for missing baseline test scores were used in all models with control variables.

+ p<0.10 * p< 0.05 ** p < 0.01 *** p < 0.001

Table 3.5

Treatment Impact Heterogeneity and 2-Year Treatment Effects

	Grade 2	Grade 3	Grade 3	Grade 3	Grade 3
	(1)	(2)	(3)	(4)	(5)
Treatment	0.082	-0.172*			
	(0.075)	(0.074)			
Baseline Math Achievement					
Bottom Third	-0.626***	-0.478***			
	(0.061)	(0.080)			
Top Third	0.562***	0.400***			
	(0.081)	(0.089)			
Baseline Math X Treatment					
Bottom Third X Treatment	-0.121	0.071			
	(0.098)	(0.100)			
Top Third X Treatment	0.077	0.259*			
	(0.097)	(0.114)			
Combinations of Treatments over Grades 2 and 3					
Treated Both Years			0.026	-0.064	-0.042
			(0.105)	(0.055)	(0.049)
Treated in G2 Only			0.032	0.045	0.042
			(0.108)	(0.059)	(0.074)
Treated in G3 Only			0.056	0.008	0.017
			(0.165)	(0.070)	(0.063)
Baseline Controls	Inc.	Inc.		Inc.	Inc.
School Fixed Effects					Inc.
F-Test (p-value)	0.229	0.090+			0.504
N	508	462	379	379	379
R-squared	0.760	0.640	0.001	0.710	0.717

Note. Robust standard errors are shown in parentheses. Standard errors were adjusted for classroom level clustering. In school fixed-effect models, standard errors were adjusted for school level clustering. For the full list of baseline controls, see Tables 3.3 and 3.4. In Models 2 and 3, all three non-missing baseline math tests (i.e., WJ-III Applied Problems, WJ-III Math Fluency, *KeyMath*) were averaged, and then split into dummy variables indicating what part of the distribution a given child fell. The distribution was split into thirds, and the middle group was excluded as the control group. In models 1 and 2, F-test p-values indicate whether the two interaction terms were jointly-significantly different from zero. In model 5, the F-test p-value indicates whether the series of longitudinal treatment combination dummies were jointly-significantly different from 0. In all models, missing baseline test score dummies were included.

+ p<0.10 * p< 0.05 ** p < 0.01 *** p < 0.001

Table 3.6

Descriptive Statistics from Classroom Observations

	Second Grade			Third Grade		
	Fall	Winter	Spring	Fall	Winter	Spring
Small Group Time (0 to 3)	0.89	1.87	2.18	1.42	1.31	1.22
	(0.43)	(0.56)	(0.90)	(0.78)	(0.62)	(0.65)
Individualized Time (0 to 3)	0.43	1.32	1.87	1.02	1.06	1.03
	(0.30)	(0.79)	(1.17)	(0.82)	(0.66)	(0.82)
N (Teachers)	17	17	17	17	17	17

Note. Each observation only included teachers assigned to the math arm of the intervention. The "small group time" category consisted of 9 items that were used to rate teachers on how often they used small groups during math instruction and their level of involvement in the groups. The "individualized time" category included 6 items that rated how often teachers individualized instruction during small group time.

Table 3.7

Associations Between Individualized Instruction and Concurrent Math Achievement

	Math Achievement			Individualized Instruction	
	Bivariate	OLS	FIML	OLS	FIML
	(1)	(2)	(3)	(4)	(5)
Observation Composite	0.010	0.035	0.037		
	(0.040)	(0.047)	(0.023)		
Lagged-Math Score		0.917***	0.873***	-0.059	-0.057
		(0.030)	(0.023)	(0.079)	(0.056)
Background Controls		Inc.	Inc.	Inc.	Inc.
Wave Dummies		Inc.	Inc.	Inc.	Inc.
Observations (pooled)	471	286	471	384	471
R-squared	0.000	0.766	-	0.064	-

Note. Robust standard errors are shown in parentheses. Standard errors were adjusted for classroom level clustering. In all models shown, students with multiple individualized instruction observations and corresponding math achievement measures were treated as independent observations. The “observation composite” was the average of the small group and individualized instruction items from the classroom observational measure. The math achievement measure was a composite measure of the average of all non-missing math achievement subtests at a given wave for each student. The "lagged-math score" represents the average of all non-missing math subtests at the previous measurement wave. In both "OLS" models shown, list-wise deletion was used to account for missing data. Background controls included gender and ethnicity and wave dummies were a series of dummy variables indicating the measurement wave of the individualized instruction measure and the math achievement measure.

+ p<0.10 * p< 0.05 ** p < 0.01 *** p < 0.001

Appendix

Chapter 3: Evaluating the Effects of a Two-Year Individualized Instruction Intervention in Mathematics

First Grade Results

Appendix Table 3.1 presents baseline characteristics for classrooms assigned to the math intervention (labeled as treatment) and reading intervention (labeled as control) during first grade. The first-grade math program was not an individualized mathematics program, but was instead a curriculum intervention called Math Pals (Fuchs et al., 1997). The reading arm of the intervention was an ISI program in reading, similar to the one used in second and third grade. As Appendix Table 3.1 reflects, I saw no indication of baseline imbalance.

Appendix Table 3.2 presents treatment impact estimates for the Math Pals intervention in grade 1 following the same modeling techniques used in the main paper. Much like with the other two intervention years, I saw little indication that the intervention affected student math achievement scores.

FIML Results

Appendix Table 3.3 presents treatment impact estimates for the second and third grade years of the study using Full Information Maximum Likelihood (FIML) to adjust for missing data. These estimates are nearly identical to the main treatment impacts presented in the main body of the text. However, note that standard errors for the FIML estimates are smaller because they are not adjusted for any level of clustering (Stata 13.0 does not allow the estimation of clustered standard errors along with FIML).

Appendix Table 3.1

Average Baseline Characteristics for Grade 1 Intervention

	Grade 1		
	Treatment	Control	p-values
Female	0.55	0.51	0.25
Ethnicity			
Asian	0.01	0.02	0.35
African American	0.04	0.07	0.21
Hispanic	0.04	0.05	0.41
White	0.83	0.81	0.42
Other	0.07	0.05	0.35
Age (years) at G1 fall test	6.67	6.61	0.17
	(0.49)	(0.42)	
Cognitive Baseline Tests			
Applied Problems	0.01	0.05	0.67
	(0.83)	(0.82)	
Math Fluency	-0.92	-0.87	0.62
	(0.90)	(0.91)	
Picture Vocabulary	0.09	0.06	0.53
	(0.67)	(0.65)	
Letter-Word ID	0.25	0.41	0.25
	(0.97)	(0.92)	
Passage Comp.	-0.31	-0.16	0.23
	(1.14)	(1.09)	
F-Test	F(12, 27)= 1.49		0.19
Observations	269	317	

Note. Standard deviations are presented in parentheses. All WJ-III (Woodcock Johnson) tests were standardized to national norms. For every test, a nationally representative sample would have a mean of 0 and SD of 1. P-values were generated from a series of regressions in which each baseline characteristic was regressed on treatment status for that given year. The F-statistic was generated from a regression in which treatment status for a given year was regressed on all corresponding baseline characteristics, and a joint-test of statistical significance was used to test whether the covariates jointly related to treatment status.

Appendix Table 3.2

Grade 1 Treatment Impacts on Spring Math Test Scores

	Math Composite			Applied Problems			Math Fluency		
	Biv.	Controls	School FE	Biv.	Controls	School FE	Biv.	Controls	School FE
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Treatment	-0.009	0.075	0.023	-0.050	0.022	-0.016	0.031	0.129	0.063
	(0.134)	(0.092)	(0.073)	(0.125)	(0.082)	(0.050)	(0.169)	(0.136)	(0.103)
Female		-0.020	-0.037		-0.012	-0.017		-0.028	-0.056
		(0.057)	(0.054)		(0.059)	(0.053)		(0.073)	(0.056)
Asian		-0.009	-0.068		-0.102	-0.178		0.085	0.043
		(0.190)	(0.166)		(0.235)	(0.299)		(0.258)	(0.186)
Black		-0.130	-0.252*		-0.269***	-0.352***		0.009	-0.151
		(0.100)	(0.063)		(0.065)	(0.049)		(0.173)	(0.150)
Hispanic		0.114	0.059		0.038	0.008		0.189	0.109
		(0.090)	(0.082)		(0.118)	(0.131)		(0.204)	(0.239)
Other		-0.178	-0.244***		-0.147	-0.175		-0.209	-0.313***
		(0.130)	(0.033)		(0.142)	(0.089)		(0.153)	(0.027)
Age (years)		-0.430***	-0.441**		-0.233*	-0.229*		-0.626***	-0.653***
		(0.080)	(0.066)		(0.090)	(0.067)		(0.095)	(0.077)
Applied Problems		0.384***	0.338***		0.487***	0.437***		0.280***	0.239***
		(0.041)	(0.017)		(0.045)	(0.022)		(0.053)	(0.032)
Math Fluency		0.206***	0.277**		0.196***	0.275**		0.217***	0.279*
		(0.034)	(0.053)		(0.045)	(0.059)		(0.050)	(0.070)
Picture Vocab.		-0.011	-0.010		0.059	0.077		-0.080+	-0.097*
		(0.029)	(0.026)		(0.055)	(0.049)		(0.039)	(0.034)
Letter Word ID		0.140*	0.088		0.129	0.062		0.152*	0.114
		(0.062)	(0.052)		(0.076)	(0.069)		(0.074)	(0.084)
Passage Comp.		0.060	0.078		0.052	0.071		0.068	0.086
		(0.049)	(0.046)		(0.059)	(0.065)		(0.055)	(0.045)
N	472	472	472	472	472	472	472	472	472
R-squared	0.000	0.646	0.672	0.001	0.595	0.613	0.000	0.506	0.533

Note. Robust standard errors were adjusted for classroom-level clustering and are presented in parentheses. For school fixed-effect models, standard errors were clustered at the school level. All test scores were standardized to national norms, so coefficients can be interpreted as effect sizes. The Math Composite is the average of the *Applied Problems* and *Math Fluency* subscores. For gender, males are the comparison group, and for ethnicity, Whites are the comparison group. Missing dummies for missing baseline test scores were included in all models that used control variables. "Biv." stands for bivariate.

Appendix Table 3.3

Grades 2 and 3 Treatment Impact Models with FIML For Missing Data

	Composite	Applied Problems	Math Fluency	Key Math
	A	B	C	D
G2 Treatment	0.063+	0.059	0.162**	0.011
	(0.035)	(0.055)	(0.061)	(0.037)
Controls	Inc.	Inc.	Inc.	Inc.
N	519	519	519	519
G3 Treatment	-0.065+	-0.106*	-0.025	-
	(0.036)	(0.045)	(0.047)	-
Controls	Inc.	Inc.	Inc.	-
N	472	472	472	

Note. Each coefficient and standard error was estimated from a separate model and all baseline controls for that year of treatment were included. Each model was run using the "SEM" with "FIML" commands in Stata 13.0. Standard errors are in parentheses, but they have not been adjusted for clustering due to modeling restrictions with the "SEM" commands in Stata 13.0. For second grade, the "composite" test score included the WJ-III *Applied Problems,* WJ-III *Math Fluency,* and KeyMath tests. In grade 3, the composite score only includes the two WJ-III math tests, as the KeyMath test was not administered. For each model, auxiliary variables included the reading and language WJ-III tests administered at the same time as the dependent variable.

+ $p<0.10$ * $p< 0.05$ ** $p < 0.01$ *** $p < 0.001$

Chapter 4

Will boosting test scores improve labor market outcomes?

Abstract

In educational evaluation, test scores are often used as the primary metric of a program's success or failure because test score gains are expected to translate into better outcomes in adulthood. Relying on nationally-representative data from the U.K., the current paper considers whether high school test scores predict multiple long-run measures of adult earnings assessed when participants were in their 30's, 40's and early 50's. While controlling for an extensive set of characteristics, including childhood socio-emotional skills and personality, IQ, and family background, the current paper found positive associations between high school math and reading scores and adult earnings. However, the size of these associations were highly sensitive to the inclusion of control variables, and results suggest that researchers should be cautious when using test score and earnings correlations to project program impacts.

Will boosting test scores improve labor market outcomes?

Test scores in mathematics and reading are often used as primary measures of academic achievement in educational program evaluation (e.g., Bloom, Canning, & Shenoy, 2012; Chetty et al., 2011; Clements, Sarama, Spitler, & Wolfe, 2013; Connor et al., 2013; Cortes & Goodman, 2014; Curto & Fryer, 2011; Deming, 2009; Krueger, 2003; Loeb, Bridges, Bassok, Fuller, & Rumberger, 2007; Taylor, 2014). This reliance on test score measures for purposes of education evaluation is due in part to the fact that test scores are readily available measures of student knowledge. Every state tests public school students through at least primary school in mathematics and reading, and recent federal guidelines have encouraged states to make these test scores publically available for education research (Every Student Succeeds Act, 2015). Further, test scores have face validity (i.e., they apparently measure knowledge in key subjects taught in school) and strong measurement properties (e.g., they are usually normally distributed and can be standardized to compare students across diverse settings). Thus, for compelling methodological reasons, test scores serve as the go-to measures of educational program success or failure.

However, few would argue that a test score impact *per se* is valuable to any given student. Although standardized tests are used in college admission decisions, adults do not typically report their primary school standardized test scores on their resumes or job applications. Instead, researchers rely on test scores because they assume that test score gains represent actual advances in student knowledge and skills, and these competencies should in turn improve later life outcomes. Correlational studies have shown that adolescent math and reading test scores predict adult health (Reyna, Nelson, Han, & Dieckmann, 2009), educational attainment (Dougherty, 2003; Murnane, Willett, & Levy, 2000), and labor market earnings (Jencks & Phillips, 1999; Ritchie & Bates, 2013; Rose, 2005). This literature implies that if a program

boosts math and reading test scores, then this program might also improve a host of important adult outcomes.

The current study is focused on the hypothesized relation between test score gains in math and reading measured during adolescence and earnings measured across adulthood. A large body of correlational evidence has found positive associations between adolescent math and reading scores and measures of early career earnings (e.g., Chetty et al., 2011; Currie & Thomas, 2001; Dougherty, 2003; Deke & Haimson, 2006; Grogger & Eide, 1995; Heckman, Stixrud, & Urza, 2006; Heckman & Vytacil, 2001; Jencks & Phillips, 1999; Murnane, Willett, Duhaldeborde, & Tyler, 2000; Murnane, Willett, & Levy, 1995; Neal & Johnson, 1996; Rose, 2006). Evaluation studies of educational programs that observe test score gains often rely on this literature to suggest the programs in question should also improve adult earnings; indeed many studies have used the test score to earnings correlation to make explicit back-of-the-envelope cost-benefit calculations based on study subjects' projected future income (e.g., Bartik, Gormley & Adelstein, 2013; Cho, Glewwe, & Whitler, 2012; Curto & Fryer, 2011; Deming, 2009; Krueger, 2003).

However, it is unclear if the test score and earnings correlation should be used this way, as the correlation might not reliably predict a given program's impact on earnings, even if the program had a large effect on test scores. In other words, this approach assumes that the variation observed between students on a given test score following a random assignment study (i.e., the effect size) is in some way the same variation that produces the correlation between test scores and earnings in non-experimental studies.[4] I argue that this is a strong assumption given

[4] This assumption is similar to the *local homogeneity assumption* (see Borsboom, Mellenbergh, & van Herdeen, 2003).

the modeling techniques used by most non-experimental studies that have reported correlations between test scores and earnings.

Drawing on nationally-sampled data from the U.K., I examine the hypothesized link between adolescent test scores and adult labor market success, with a view toward projecting what a given program impact might be if a program were only evaluated with achievement test scores. In other words, I attempt to answer the following question: if an educational program affected high-school math and reading test scores, what impact on earnings might we expect the program to have?

In the current data, achievement scores in mathematics and reading were assessed when study subjects were 16 years old, and earnings were measured across the heart of subjects' careers (i.e., the earliest earnings record was taken at age 33, the latest at age 50). This dataset allows me to control for an unusually broad set of covariates, all measured before high school. By controlling for earlier measures of IQ, socio-emotional skills and personality, and family background characteristics, I can assess the association between adolescent achievement scores and earnings holding constant characteristics that many academically-focused educational programs are unlikely to alter. Further, by considering mathematics and reading scores as separate achievement indicators, I am also able to evaluate the independent role of mathematics and reading skills in shaping adult earnings trajectories.

Results from fully-controlled models indicated that test scores positively predicted later earnings throughout subjects' careers, but these associations were much smaller than has been reported in previous studies (e.g., estimates were 50%-70% smaller than Murnane et al., 2000; Lin, Lutter & Ruhm 2016). Between the ages of 33 and 50, I found that a 1-SD gain in adolescent mathematics achievement led to an average earnings boost of 7.6%, whereas a 1-SD

gain in reading achievement led to approximately 5.3% more earnings over this same period, and these associations appeared to be consistent across time. If math and reading scores were averaged together, I found that a 1-SD gain in academic achievement predicted approximately 12.4% more earnings through age 50. Models showed that test scores and earnings correlations are subject to substantial bias, as introducing controls reduced initial correlations by nearly 50%. Even after controlling for a host of important personal and environmental characteristics, it is unclear whether these controls fully accounted for all sources of possible bias. These estimates indicate that although the association between test scores and earnings may be non-zero, studies that imply earnings effects based solely on test score impacts should exercise caution when making this assumption.

Background Literature

Interest in the correlation between test scores and earnings can be found in articles published in journals in education (e.g., Reardon, 2013; Watts, Duncan, Siegler, & Davis-Kean, 2014), psychology (e.g., Bailey, Watts, Littlefield, & Geary, 2014; Marle, Chu, & Geary, 2014; Siegler & Lortie-Forgues, 2015), and economics (Krueger, 2003; Hanushek & Rivkin, 2010; Heckman et al., 2006), with the correlation often cited as empirical evidence of the importance of promoting mathematics and reading skills in American K-12 schools (e.g., Hanushek, 2009). Such studies implicitly reason that through promoting mathematics or reading achievement, interventions should also positively impact affected students' labor market success. This rationale has been explicitly used to make cost-benefit calculations in educational program evaluations. For example, Krueger (2003) used the correlation between test scores and earnings to predict the impact that the Tennessee STAR experiment might have on adult earnings. Krueger relied on three studies to make this calculation: Murnane et al., (1995), Currie and

Thomas (2001) and Neal and Johnson (1996). Based on the varying correlations between adolescent test scores and earnings reported across these three studies, Krueger reasoned that a 1-SD test score impact was likely to have an 8% effect on adult earnings. This led Krueger to project that the Tennessee STAR experiment had an effect of approximately 2% on future earnings. However, years later, Chetty and colleagues (2011) obtained actual measures of early labor market earnings for the children that participated in Tennessee STAR, and they found no impact of the program on earnings (various estimates hovered around a non-statistically significant effect of 1%).

Although Chetty and colleagues' (2011) study calls into question whether the correlation between test scores and earnings should be used to predict program impacts, many researchers continue to make this "back of the envelope" calculation (e.g., Bartik, Gormley & Adelstein, 2013; Cho, Glewwe, & Whitler, 2012; Curto & Fryer, 2011; Deming, 2009; Duncan, Ludwig, & Magnuson, 2010), with many assuming that the test score and earnings correlation is much larger than 8% [e.g., Deming (2009) assumed a 25% increase in earnings per 1-SD boost in test scores]. The desire to link test score impacts to adult earnings is unlikely to fade given that test scores are readily available measures of student achievement, and it may take many years after the end of an educational program to reliably measure earnings impacts. Further, there remain strong theoretical reasons to believe that boosts in academic skills would have impacts on adult economic attainment. Murnane and Levy (2005) argue that the basic skills measured by math and reading tests are likely to become even more important as the economy shifts toward more technical sectors, creating jobs that increasingly require strong analytic and communicative skills (also see Deming, 2015). Indeed, it is difficult to imagine a job in any sector of the economy that does not require some degree of competency in mathematics and literacy.

However, if we hope to understand how impacting academic test scores might affect adult economic attainment, then careful consideration of possible sources of omitted variables bias (OVB) is in order. Unfortunately, the most widely-cited studies that have reported links between adolescent test scores and adult earnings have failed to account for probable sources of confounding variation. For example, using the High School and Beyond data (HS&B), Murnane and colleagues (2000)[5] reported that, conditional only on gender and race, a 1-SD gain in high school math scores led to approximately 20% more earnings for men and women at age 27. Currie and Thomas (2001) reported similar findings, as they found that while controlling for measures of family background characteristics and child demographic variables, a 1-SD increase in high school math scores predicted 14% more earnings at age 33. These findings are similar to the findings of multiple papers with comparable designs (e.g., Blackburn & Neumark, 1995; Chetty et al., 2011; Lin et al., 2016; Neal & Jonson, 1996).

Yet, it is unlikely that these limited sets of control variables account for the possible sources of OVB. In particular, these studies did not control for children's non-academic skills and behaviors, such as sociability- a skill that has been shown to correlate with both adolescent test scores and adult earnings (Deming, 2015). Further, no study has adequately controlled for general cognitive ability (i.e., IQ). A fairly large literature has documented the link between measures of general intelligence and later earnings (see review by Strenze, 2007), and many studies have reported on the link between intelligence and measures of mathematics and reading achievement (Fin et al., 2014; Gathercole et al., 2004; St. Clair-Thompson & Gathercole, 2006). In fact, math and reading test scores have been shown to be excellent indicators of IQ. For example, Kaufman and colleagues (2012) observed nearly a perfect correlation between a "g"

[5] This study was an updated version of the working paper (i.e., Murnane et al., 1995) cited by Krueger (2003).

(i.e., general intelligence) score generated from a factor analysis of math and reading scores and a "g" score generated from cognitive tests more typically used to measure IQ. If general cognitive ability accounts for the positive correlation between achievement scores and earnings, then any educational program that raised reading or math test scores without affecting general cognitive ability would likely fail to impact earnings. Similar omitted variables bias concerns linger for other potential confounding factors like personality and parental inputs.

Further, the current literature reporting links between test scores and later earnings has also been limited by study designers' inability to observe earnings past early adulthood. Indeed, most of the studies that examined relations between test scores and earnings were conducted in the 1990's (e.g., Jencks & Phillips, 1999; Neal & Johnson, 1996) or early 2000's (e.g., Murnane at al., 2000; Currie & Thomas, 2001), and the national datasets upon which these studies relied (e.g., High School and Beyond; National Longitudinal Survey of Youth- 1979; National Child Development Study) lacked earnings measures past early adulthood.

Altonji and Pierret (2001) argue that the returns to cognitive skills should rise as work experience grows because employers are better able to judge the abilities of their employees after they have worked for greater lengths of time. Indeed, Lin, Lutter and Ruhm (2016) recently examined the association between cognitive skills and labor market earnings measured through age 48 using newly available data from the National Longitudinal Survey of Youth- 1979. They found that the association between test scores and earnings increased with age, as a 1-SD boost in test scores predicted approximately 25% more earnings at age 48. This association was detected after controlling for adolescent measures of socio-emotional skills and parent background characteristics. Unfortunately, their study does not provide a clear estimate of the impact of math or reading achievement, because their measure of cognitive skills was the Armed

Forces Qualifying Test (AFQT). The AFQT is a composite measure of cognitive skills that includes tests of general cognitive ability as well as tests of math and reading achievement.

Unfortunately, it remains unclear if an educational program evaluated through math and reading achievement test scores would produce the same impact on earnings observed for the AFQT measure in the NLSY study. Although scores on math and reading tests are correlated with measures of general cognitive ability (e.g., Gathercole et al., 2004) a gain on a mathematics or reading achievement test does not necessarily imply a gain in general cognitive skills. For example, in an evaluation of Massachusetts charter schools, Fin and colleagues (2014) found that although charter schools were successful at raising mathematics and reading achievement test scores, they had no impact on "domain-general" measures of cognitive ability such as working memory, fluid reasoning, or processing speed. Similarly, recent research on early mathematics learning has shown that the correlations observed between early math scores and later cognitive measures in non-experimental datasets are not replicated when variation in early math scores is produced by exogenous processes (i.e., interventions; Watts et al., 2017). Thus, although it remains possible that a given educational program could affect both achievement skills (i.e., math and reading ability) and general cognitive ability (i.e., IQ), understanding the long-run association between academic achievement test scores and earnings while holding constant general cognitive ability can help us better evaluate whether certain programs may have earnings effects.

Of course, educational programs may also have earnings effects through channels that are not captured by test scores. Lindqvist and Vestman (2011) relied on data from the Swedish military enlistment to investigate the link between non-cognitive skills (i.e., personality and socio-emotional skills) and earnings while holding constant measures of cognitive ability. Non-

cognitive skills were measured from an interview conducted by a psychologist, and included measures of persistence, social skills, and emotional stability. Lindqvist and Vestman averaged these skills together, and found that their measure of non-cognitive ability was just as predictive of early career earnings as a composite measure of cognitive ability (effects for both cognitive and non-cognitive skills ranged from 5-10%). Moreover, some recent program evaluation evidence has shown that using the test score and earnings correlation can sometimes lead one to under-predict the effect of the program on earnings (see Fredriksson, Öckert, & Oosterbeek, 2013), further suggesting that non-test score channels may also be responsible for program impacts on economic attainment.

Although the current paper is concerned with the association between test scores and earnings, the evidence on non-cognitive skills and earnings further illustrates the need for careful attention to the myriad of factors that could influence the test score and earnings correlation if one hopes to accurately project program earnings impacts. The success of this approach is likely to depend upon the mechanisms through which a given program affects test scores and the mechanisms through which a program might affect earnings. I return to this issue in more detail below.

Current Study

In the current study, I estimate the associations between high school academic test scores and later earnings with an eye to informing the possible long-run implications of studies of educational interventions that estimate impacts using achievement tests. In the analyses that follow, I draw on nationally representative data from the U.K., as this dataset allows me to control for relatively rich measures of family background characteristics, personality, socio-emotional behaviors, and general cognitive ability (i.e., IQ). Further, this dataset includes

earnings measures collected throughout participants' careers, allowing me to observe the association between high school test scores and earnings measured from age 33 through age 50. I expect to detect a positive association between adolescent achievement test scores and earnings, because higher levels of academic skills should lead to higher worker productivity (see Levy & Murnane, 2005). However, I expect that these associations will be smaller than effects reported by studies that have not controlled for measures of general cognitive ability or socio-emotional skills (e.g., Chetty et al., 2011; Murnane et al., 2000).

Analytic Approach

Because I am interested in examining the earnings impact that could be expected of an educational intervention that boosted math and reading test scores, I must control for factors that might influence test scores and earnings that are also unlikely to be affected by a high school educational program. Although the unadjusted correlation between achievement tests and later earnings may be of interest, it is unlikely to represent the effect that a program might have on earnings if the program narrowly affected test scores but failed to affect the many factors positively correlated with test scores (e.g., motivation, IQ, family income, sociability, etc.). Of course, every educational program is different, and the hypothesized mechanisms through which a program might affect adult attainment are likely to vary widely. For example, intensive interventions, like no-excuse charter schools, may have a much broader set of impacts on a host of personal characteristics (e.g., Dobbie & Fryer, 2013; but see West et al., 2016), when compared with narrowly-focused math or reading curriculum interventions. If a program positively affected socio-emotional behaviors, like sociability, as well as mathematics achievement, then controlling for age-16 sociability in my models may lead to an under-estimation of the potential program impact on earnings.

Consequently, for most of the models considered here, I only control for measures of personal and environmental characteristics measured *prior* to high school (i.e., ages 7 and 11) to allow for the possibility that changes in a host of adolescent characteristics could influence variation in age-16 achievement tests. Thus, I model earnings as a function of math and reading test scores at age 16 and personal and environmental characteristics measured earlier in childhood:

$$Y_i = \alpha_0 + \beta_1 Math_i + \beta_2 Read_i + \emptyset Fam\&Health_i + \delta Beh\&Personality_i + \lambda Cog_i + \gamma Pretests_i + \varepsilon_i$$

where Y_i represents earnings measured at either age 33, 41, 46, or 50, and $Math_i$ and $Read_i$ represent high school math and reading test scores, respectively. $Fam\&Health_i$ represents a vector of control variables that measure aspects of the child's home environment, socio-economic status, and personal health. $Beh\&Personality_i$ represents a set of control variables that include age 7 and 11 measures of socio-emotional skills and personality. Cog_i includes age 7 and age 11 controls for general cognitive ability (i.e., IQ) and motor skills. Finally, $Pretests_i$ represents a vector of math and reading scores measured at ages 7 and 11. Controlling for math and reading tests measured prior to high school allows me to adjust for pre-existing ability in both subjects, as the guiding question is whether we expect to observe an earnings effect for a high school program that *changed* academic achievement scores. Finally, ε_i represents an individual error term, which will only be uncorrelated with the key measures, $Math_i$ and $Read_i$, if the control variables adjust for all possible sources of OVB (a strong assumption).

I also test models that add controls for other child and environmental characteristics measured in high school. This allows me to test the effects that other, non-academic factors, have on adult earnings, and it also allows me to test the effect that a program might have on earnings if it did not affect dimensions of socio-emotional behavior and personality assessed in

the current dataset at age-16 (i.e., this model asks, what is the effect of changes in high school academic skills on earnings holding constant socio-emotional skills and personality). Finally, by adding school characteristics to the model, I also assess the extent to which school differences account for the effects of math and reading scores on later earnings.

High school test scores were transformed to z-scores so that results could be likened to effect sizes, and earnings at each wave were log-transformed. In all models presented that included control variables, missing values were set to the mean for each variable. For each control variable, I then included a dummy variable that was equal to "1" if a person was missing on the corresponding variable to adjust for the missing data imputation. In the appendix, I detail the number of missing cases on each control variable, and I also describe results from analyses that used full information maximum likelihood (FIML) for missing data adjustments. Results from FIML models did not differ substantially from results produced by OLS models with dummy variable adjustments.

Method

Data

Data were drawn from the National Child Development Study (NCDS), a longitudinal study that drew a population sample of approximately 17,000 newborns living in England, Scotland, and Wales during 1 week in 1958. Since recruitment, the study has followed participants throughout the course of their lives, collecting information on physical health and development, education, cognitive ability, economic circumstances, employment, family life, and personality. The current study relies on data collected at ages 7, 11, 16, 33, 42, 46 and 50. Unsurprisingly, many participants have lost touch with the study over time. By the age 16 survey, study administrators collected data on approximately 12,000 participants. By age 33,

most of these participants remained in the study (approximate n= 11,100), and by age 50, approximately 9,550 of the cohort members participated in data collection. The NCDS is a publically available dataset, and it has been widely used to study questions regarding long-run cognitive and socio-emotional development (e.g., Currie & Thomas, 2001; Miles, Savage & Buhlmann, 2011; Takizawa, Maughan & Arseneault, 2014).

In the analyses presented in this paper, I rely only on data collected from male participants in the NCDS. Although theories regarding connections between academic achievement scores and labor market success certainly extend to women, limitations in the current data prevented me from making legitimate comparisons between the adult earnings of men and women. Specifically, women in the NCDS were far less likely to ever work full-time. At each of the four earnings measurement points considered here, approximately 40% of women indicated working full-time, compared with nearly 85% of men. This discrepancy is likely due to historical differences in workforce participation between British men and women, and this has changed over time.[6] However, in the appendix, I present key results shown in the current paper for women that did indicate working full-time at any given earnings wave. In the results section, I briefly discuss some of the differences observed between men and women.

The current analysis sample (n=4,822) is then comprised of men who had valid age-16 mathematics and reading test score data and who had non-missing employment data from at least one of the adult earnings surveys (taken at ages 33, 41, 46, and 50).

Measures

Mathematics achievement. My primary measure of mathematics achievement was assessed at age 16. The NCDS *Mathematics Test* contained 31 questions that covered both

[6] Recent data from the U.S. Bureau of Labor Statistics suggests that in 2014, 57% of adult women were in the labor force, compared with 70% of men (U.S. Bureau of Labor Statistics, 2015).

numerical and geometric concepts and procedures, and 27 of the questions were multiple choice items while 4 were true-or-false questions. The measure was constructed specifically for use in this study, and it was found to have strong internal reliability (alpha = 0.85).

I also rely on mathematics achievement measures taken at ages 11 and 7 as control variables. At age 7, students were given an arithmetic test that was orally administered by participating teachers. At age 11, students were given a pencil and paper arithmetic assessment that contained 40 items involving numerical and geometric procedures (alpha= 0.94).

All three mathematics tests have been used by previous studies that investigated associations between academic skills and adult outcomes (e.g., Currie & Thomas, 2001), and more information regarding the development of the measures can be found in Shepherd, 2012.

Reading Achievement. At age 16, students were administered the NCDS *Reading Comprehension Test,* a 35-item measure that asked students to complete sentences by filling in a missing word. For each item, students were provided with a selection of 5 words, from which one word would correctly complete the sentence presented. The measure was found to have strong internal reliability (alpha = 0.86).

As with mathematics achievement, I also rely on reading measures assessed at ages 7 and 11 as control variables. The age 7 measure of reading, called the *Southgate Group Reading Test,* assessed students on word recognition and comprehension. At age 11, students were given an exam similar to the age 16 *Reading Comprehension Test,* as they were presented with sentences missing a key word and asked to fill in the blank (alpha = 0.82). For more information regarding the reading achievement measures, see Shepherd, 2012.

Earnings. Adult labor-market earnings were measured via a telephone survey administered at ages 33, 41, 46 and 50. At each survey, study examiners asked participants

about their current employment. If a participant indicated that they were currently employed either part- or full-time (self-employment was included), then study examiners asked them to report the amount of their last take-home pay after deductions for tax. Respondents then indicated the period over which this pay was assessed.

For each wave, I used this information to generate several different measures indicating labor market participation and income. I converted all reported earnings to monthly earnings, and then transformed this monthly income amount to 2016 U.S. dollars. I also created a measure that included respondents who indicated unemployment at a given wave as having "0" earnings. In the models that follow, I present results that include unemployed workers with "0" earnings, as well as models that were conditional on full-time employment and reporting an actual earnings amount [these models can be most closely compared with widely cited estimates from Murnane and colleagues (2000) and Currie and Thomas (2001)].

In Table 4.1, I present descriptive statistics for the various measures of earnings and employment at each wave. As mentioned above, the sample is restricted to participants who had at least one non-missing adult earnings survey and valid math and reading scores at age 16 (n= 4,822). Because earnings measures were positively skewed (i.e., there are small numbers of earners that earn far more than the median income amount), I dropped the top 1% of earners from each wave (n= 140 excluded participants). As Table 4.1 reflects, the average participant responded to 3.14 follow-up earnings surveys, and they verbally reported earnings on 2.21 follow-up surveys.

[Insert Table 4.1]

Not surprisingly, average monthly earnings (presented in 2016 U.S. dollars) grew over time. At age 33, 91% of respondents indicated employment, and this corresponded with an

average of $2,577 in monthly earnings. By age 50, 89% of respondents attested to working full-time, and average monthly earnings had grown to $3,400.

Control Variables

In order to control for as many sources of confounding variation as possible, I include a host of control variables assessed at ages 7, 11, and 16. These variables were primarily measured via parent and teacher surveys, and test scores were collected from the students themselves. In the following section, I give brief descriptions of the measures used. In the appendix, I present descriptive statistics for every control variable included in the models (see Appendix Table 4.1).

Family Background and Child Health. Family background characteristics were measured via parent surveys administered at ages 7, 11 and 16. At all three waves, family socioeconomic status (SES) was assessed as a measure of father's occupational prestige, income, and educational level. Children were then placed into one of 7 discrete SES levels, and I control for the set of SES dummies at ages 7 and age 11. I also control for indicators of free lunch, whether the family owned their home, whether the father was employed, and whether the father was present in the child's life; all of these indicators were assessed at both age 7 and age 11. I also included age 7 measures of the child's birth order, whether they were adopted, whether they attended formal daycare, the number of rooms in the house, and the region of birth.

Characteristics related to the child's health and physical development were assessed at ages 7 and 11. At both ages, I created an index of the number of diseases (e.g., whooping cough, chicken pox) that the child was reported as ever having contracted. I also included an index of how many times the child had visited a hospital by age 11, and I included a dummy variable indicating whether they had been diagnosed with epilepsy. Teachers also rated children on their

health at ages 7 and 11 via a "yes" or "no" question, where "yes" indicated that the child was generally "in good health."

Child Characteristics. Measures of general cognitive ability and motor skills were assessed at age 11. General cognitive ability (i.e., IQ) was measured via a non-verbal test that asked students to interpret and predict patterns and a verbal test that involved identifying how sets of 4 words related to one another (alpha = 0.94). Motor ability was assessed by a copying test in which the child was asked to copy various designs using a pencil and paper.

Various socio-emotional skills and behaviors were measured at age 11 via the *British Social Adjustment Guide*. Teachers rated children on their attention, social skills, anti-social behavior and affect, among other characteristics, and these items were primarily assessed via Likert-scales that ranged from 1 to 3 or 1 to 5. At age 16, teachers rated children on their personality characteristics, and these characteristics were assessed via items that asked teachers to indicate where a child fell on various personality continua. For example, teachers rated children on a 1 to 5 scale for work ethic, where a score of "1" indicated extreme "laziness" and a score of "5" indicated that the child was extremely "hardworking." Teachers rated children on 6 different personality characteristics that included aggression, flexibility, sociability, impulsivity, work ethic, and affability.

Teachers also rated children's academic abilities in math, reading, language, science and social studies at ages 7 and 11. These ratings were scaled from 1 to 5, with a score of "1" indicating that the child was a "very good learner" and a score of "5" indicating that they had "very little ability" in a given subject.

School Characteristics. In high school, teachers were surveyed about school characteristics. In the following models, I include measures of school enrollment, whether the

school was coed, the proportion of boys enrolled, whether the school had an active PTA, the proportion of students taking O-level exams, the proportion of students expected to continue education past secondary school, student to teacher ratio, whether the school had adequate facilities, and an index measuring the number of punitive disciplinary methods regularly used by the school.

Results

Descriptive Results

In Table 4.2, I present descriptive statistics for various high school characteristics for all men in the sample (n=4,822), and I also present descriptive characteristics disaggregated by whether a participant was, on average, in the top or bottom 50% of earners across the four measurement waves. As Table 4.2 reflects, subjects in the sample correctly answered an average of 13.74 (SD= 7.17) of the 31 items on the age-16 mathematics test compared with an average of 25.75 (SD= 7.06) of the 35 items on the reading test. Teacher-rated personality and behavior measures, which were scaled 1 to 5, tended to hover around 2.5 across the six personality dimensions assessed. Average monthly family income at age 16 was $2,518 (in 2016 U.S. dollars), and only 9% of the sample qualified for free school meals. Among the school characteristics presented, schools reported that approximately 27% of enrolled boys were registered to take O-levels, which were the exams required in the U.K. for further pursuing academic study past the age of 16.

[Insert Table 4.2]

Men who would go on to be above-median earners scored higher on both the math and reading tests at age 16, though differences in age-16 personality measures do not appear to be as pronounced. Family demographic characteristics reflect relative income stability over-time, as

above-median earners tended to come from higher-earning families and from schools that had more students registered to take O-level tests.

Effects of Math and Reading Skills on Earnings

In the following analyses, I add sets of control variables to each model in order to gauge how much bias is attenuated when covariates are adjusted. In general, I begin with models with no controls, then I add measures of family background and child health, followed by socio-emotional skills and behaviors. I then add earlier measures of math and reading skills (assessed at ages 7 and 11, respectively) and the IQ measures assessed at age 11. Finally, I add age-16 measures of socio-emotional skills and family background characteristics.

Table 4.3 presents associations between adult measures of log-earnings and age-16 measures of math and reading achievement, both standardized across the entire sample. These models are conditional on working full-time at a given earnings survey, which has been the typical modeling specification of studies that have examined the link between test scores and earnings (e.g., Currie & Thomas, 2001). In Column A of Table 4.3, I present results from models in which math and reading test scores were entered separately, thus each coefficient was derived from a bivariate model in which a measure of log-earnings was regressed on a single test score. Looking across the estimates shown in Column A, the bivariate association between achievement scores and earnings is large and positive, and it appears to grow slightly with age. A 1-SD gain in age-16 math achievement predicts 12% more earnings at age 33, and 15% more earnings at age 50. The trajectory for reading achievement is similar.

[Insert Table 4.3]

However, Table 4.3 suggests that there is bias in the unadjusted association between test scores and earnings. In Columns B, I entered the math and reading tests together, and both

coefficients fell considerably (e.g., for earnings measured at age 50, the standardized effect for math fell from 15% to 11% when the reading score was added to the model). This indicates that the effect of a 1-SD gain in mathematics achievement on earnings is smaller than the effect of a gain in both mathematics and reading scores. It also suggests that simply assessing the association between earnings and achievement in one academic domain without controlling for achievement in the other domain may overstate the effect of gains in a single academic subject.

In models C through F, I progressively added more control variables, beginning with measures of family background and child health. In the final set of models, which controlled for a host of variables measured prior to high school as well as high school measures of socio-emotional skills and family characteristics, a 1-SD boost in the mathematics test predicted almost 6% more earnings at age 33 and 8% more earnings at age 50. In comparison, the effect of reading was nearly unchanged (statistically significant 5.2% at age 33 (*SE*= 0.015); non-statistically significant 5.0% at age 50 (*SE*= 0.027)). Although the math test score statistically significantly predicted earnings at age 50 and the reading test did not, the coefficients produced by the two test scores were not statistically significantly different from one another (p=0.44).

Although these models certainly indicate a positive and significant association between high school test scores and later earnings, the models also demonstrate that unadjusted estimates between test scores and earnings contain substantial bias. For math achievement, the coefficients fell by approximately 50% between the bivariate models and the fully-controlled models shown in column F. For reading achievement, coefficients fell even more when the full set of controls were added, with the coefficients falling between 58 and 68% depending on the age at which earnings were assessed.

[Insert Figures 4.1 and 4.2]

Figures 1and 2 illustrate the reduction in bias achieved when adding the control variables, as they display the coefficients generated from columns A and F of Table 4.3. Figure 1 presents estimates for the age 16 math test, and Figure 2 shows estimates for the age 16 reading test. In general, the graphs also show that for the fully-controlled estimates, there is no clear growth in the returns to math and reading skills with advancement in age. For both math and reading achievement, the labor market returns are fairly stable over time when all the control variables are included in the model, though the unadjusted associations appear to grow linearly with age. This may indicate that although there is growth with age in the return to general cognitive skills, the returns for the narrower set of skills measured by math and reading scores, when generally cognitive ability is controlled, may be stable over time.

Table 4.4 presents results from models with the same specifications as the models presented in Table 4.3, but with the requirement for full-time work relaxed. Here, I included men that indicated either part-time work or unemployment at any given earnings wave, and I imputed a value of "0" for the earnings of men who claimed to be unemployed. Earnings were again converted to 2016 USD and log-transformed, and a value of $500 was added to the logarithmic function (values of $100 and $1,000 were also tested, and results did not differ). With part-time and unemployed workers added to the model, standard errors were slightly smaller as sample sizes increased. Further, results differed slightly, but in perhaps interesting ways. In these models, the fully-controlled association between math and earnings increased from 7% at age 33 to 10% at age 50. However, the association between reading and earnings clearly fell with age, from 8% at age 33 to non-statistically significant 2% (SE= 0.23) at age 50. The difference between the 10% math effect and the 2% reading effect at age 50 was also statistically significant ($p = 0.02$). Thus, when compared with the models presented in Table 4.3,

these results indicate that some of the association between test scores and earnings may be mediated by level of employment.

[Insert Table 4.4]

Composite Measure of Academic Achievement

Because many of the papers that have investigated links between achievement tests and earnings relied on achievement measures composed of both math and reading scores (e.g., Chetty et al., 2011; Lin et al., 2016), I also tested the association between earnings and a standardized average of the age-16 math and reading tests. Table 4.5 presents results from models that included only men working full-time at any given earnings wave, and as expected, the returns for a composite measure of math and reading achievement are larger than for a single measure of math or reading considered independently. Bivariate associations between the achievement composite and later earnings grow from 13% at age 33 to 17% at age 50, and these relations are again attenuated by the addition of control variables. When all controls were added to the model, the achievement composite effect was 12% at age 50.

[Insert Table 4.5]

Pooled Models

Finally, I tested models in which each earnings measurement was considered as an independent observation, allowing me to test the achievement score effect on earnings averaged across the four earnings measures. For these models, subjects were included in the sample if they had at least one non-missing earnings measurement, and a series of dummy variables indicating the age at which each earnings measurement occurred were included. Standard errors were adjusted to take the non-independence of the repeated earnings measures into account (i.e., person-level clustering).

In columns 1, 3, and 5 of Table 4.6, I present models that coincide to estimates shown in column F of Tables 4.3, 4.4, and 4.5, respectively (i.e., fully-controlled models). Here, I also present coefficients produced by some other age-16 measures of characteristics of interest, namely standardized measures of family income and personality. I present these coefficients not to estimate the causal effect of personality on earnings, but to give a sense of how the coefficients produced by the achievement tests compare with other factors thought to influence adult economic attainment.

[Insert Table 4.6]

The results shown in column 1 of Table 4.6 indicate that, conditional on working full-time, a 1-SD gain in mathematics achievement has an average return of 7.6% more earnings between the ages of 33 and 50, and a 1-SD gain in reading achievement had an average return of 5.3% more earnings measured over the same period (these effects were not statistically significant different from one another, p= 0.22). When part-time and unemployed workers were included in the model, the effects were very similar, as a 1-SD gain in math achievement predicted 8% more adult earnings and the reading effect was unchanged (column 3 of Table 4.6). The average effect of a 1-SD gain in the achievement composite variable between age 33 and age 50 was 12% (column 5 of Table 4.6).

When compared with other characteristics, the achievement test effects were slightly larger than most of the other age-16 measures assessed, but other non-academic characteristics also appear to have consistent effects on earnings. A 1-SD gain in family income (SD= $1,208 in 2016 U.S.D.) at age 16 had a 3% effect on adult earnings between ages 33 and 50, and the measure of sociability (negatively scaled as “withdrawn”) had a consistent effect of approximately -3.5 to -4% on average adult earnings. Similarly, the variable that assessed

conscientiousness (negatively scaled as "lazy") had an effect that ranged from -2.1 to -3.1% depending on the model. In the models that were conditional on full-time employment, the personality dimension that assessed timidity and aggression had a 2.4% effect on earnings, with men rated by their teachers as being more aggressive (i.e., less timid) earning more.

Finally, in columns 2, 4, and 6 of Table 4.6, I add controls for high school characteristics measured during the age-16 survey. By adding these controls to the model, we can gauge how much school differences account for the association between achievement test scores and earnings. Across each of the 3 sets of models, the inclusion of the school characteristics accounted for very little of the achievement to earnings effect, as coefficients remained nearly unchanged. This indicates that school differences, at least in the dimensions measured by the NCDS, are unlikely to account for the association between test scores and earnings. However, it is important to note that although the NCDS measured an interesting set of school characteristics that were likely to measure some dimensions of school quality (e.g., percent of students taking O-level exams, student to teacher ratio, whether teachers rated facilities as adequate), these variables do not measure many of the school and classroom policies investigated by educational researchers today (i.e., specific curricula, charter schools, etc.).

Additional Results

In the appendix, I present results from robustness checks that relaxed different assumptions present in the models displayed here. Specifically, I tested analogous models with women who indicated working full-time (Appendix Tables 4.2 and 4.3). Interestingly, results for women are nearly identical to men when the achievement composite variable (i.e., the average of the age-16 math and reading test) was used, as a 1-SD gain in math and reading scores predicted 11% more earnings at age 33, and 12.7% more earnings at age 50. However, it appears that for

women, much of this effect was driven by the reading test rather than the math test. When the reading and math tests were considered independently, the reading test had a large association with age 50 earnings in the fully controlled model (11%) but the math test had no effect on age 50 earnings (4.0%, SE= 0.22). This heterogeneity was unexpected and is certainly interesting, but it is difficult to draw strong claims based on this dataset when only 40% of women were ever working full-time.

I also test the sensitivity of the results to measurement error, as measurement error in predictor variables will drive coefficients toward 0 (see Appendix Table 4.4). Using the "errors in variables" adjustment in Stata, I tested the pooled models with adjustments for the reliability of the age-16 math (alpha= 0.85) and reading (alpha=0.86) measures. As expected, coefficients were larger, with the math coefficient rising to 13% and the reading coefficient rising to 7%. This indicates that measurement error may attenuate the coefficient sizes reported in the main results. However, it is important to remember that control variables were also measured with error, and error in control variables could bias coefficients upward because controls are unable to perfectly capture sources of confounding variation (e.g., the "personality" variables from age 16 are one item indicators with no reported reliability).

Finally, I tested whether using the SEM with full-information maximum likelihood (FIML) commands in Stata changed the results (recall that current models used dummy-variables to adjust for missing data on covariates). I show results from models that tested associations between math and reading scores at ages 33 and 50, respectively, and results were very similar (coefficients were within 1-1.5 percentage points of the coefficients from the analogous models shown in Table 4.3).

Discussion

Across most of the models tested, I found that standardized gains in age-16 math and reading scores had respective effects of about 5-8% on measures of adult earnings, and a composite score of math and reading achievement had an effect of approximately 12% on earnings. These effects were fairly consistent across earnings measured between ages 33 and 50, though there was some indication, especially when part-time and unemployed workers were included, that the return to math scores rose slightly over time while the return to reading scores diminished with age. In the following section, I consider how my findings compare with previous literature, and I discuss the possible implications of these findings for educational program evaluation.

Comparisons with Previous Literature

When compared with most previous studies that have tested relations between adolescent cognitive test scores and later earnings, my estimates are substantially smaller. However, drawing direct comparisons with previous literature is difficult, because the studies that have investigated the test score and earnings relation all vary in the measures used, the ages at which earnings and test scores were measured, and the modeling specifications employed (e.g., Currie & Thomas, 2001; Deke & Haimson, 2006; Dougherty, 2003; Jencks & Phillips, 1999; Lin et al., 2016; Murnane et al., 2000; Neal & Johnson, 1996; Rose, 2006; Tyler 2004). Further, it should be noted that many of the papers most commonly cited as evidence that a test score impact should translate into adult earnings did not apparently set out to investigate the "causal" effect of test score gains on economic attainment. For example, one of the most widely cited papers in this regard (Murnane et al., 2000) cautions against causal interpretations, as the authors observe that "in none of the (models) did test scores explain as much as 25 percent of the variation in

annual earnings… There is enormous variation in the earnings of workers with the same high school skills."

Still, there is no doubt that the findings from studies like Murnane et al., 2000 have been used to imply that observed program impacts on test scores should translate to effects on later adult earnings, and my results suggest that the correlations reported in previous studies should probably not be interpreted this way (I would also caution against such interpretations of the correlations reported here). In considering my results alongside previous literature, a few comparisons can be made. Currie and Thomas (2001) also used the NCDS data, and they tested models that regressed earnings at age 33 on math test scores and family background controls. They reported that, conditional on working full-time, a 1-SD gain in math achievement at age 16 predicted 14% more earnings at age 33. In comparison, in my fully-controlled age-33 model that only included full-time workers, I found that a 1-SD gain in math achievement led to only 5.8% more earnings at age 33. However, Currie and Thomas did not control for reading scores at age 16, so it is possible their reported effect for math could be larger because of the unadjusted correlation between math and reading scores. If I remove the age-16 reading test from my fully-controlled model, the math effect at age 33 rises to only 7%, an estimate which is still much smaller than what was reported by Currie and Thomas.

My estimates are also much smaller than the findings reported by studies that have relied on U.S. data. For example, using the National Longitudinal Survey of Youth 1979 (NLSY79) data, Lin and colleagues examined links between an aggregated measure of cognitive skills (the AFQT) taken in late adolescence and early adulthood and earnings measured through the late-40's. Controlling for family background measures and some non-cognitive characteristics, they reported earnings effects as large as 25% at age 48. A similarly sized effect was recently

reported by Chetty et al. (2011), as they found that, conditional on family background demographics, a 1-SD gain in elementary school math and reading test scores predicted 18% more earnings in adulthood. In comparison, I found that the unadjusted association between the achievement composite variable and age-50 earnings was 17%. Once controls were added to the model, this association dropped to 12%, which was over 50% smaller than the association reported by Lin and colleagues, and approximately 33% smaller than the effect reported by Chetty and colleagues.

Interestingly, some papers have reported associations between test scores and earnings that were smaller than the associations reported here. In particular, Rose (2006) and Deke and Haimson (2006) both reported effects for standardized gains in math achievement on earnings that ranged between -1% and 3% depending on the model. However, their studies only considered earnings measures taken when participants were in their early and mid-20's, leading to the strong possibility that many subjects included in their samples were not entirely participating in the labor market due to postsecondary schooling. Indeed, the NCDS also surveyed subjects regarding their earnings at age 23, and regressing the age-23 log-earnings measure on the age-16 mathematics and reading tests, respectively, leads to bivariate correlations of -1.3% for math and 0.6% for reading (both non-statistically significant).

Implications for Program Evaluation

Although I found that results were smaller than has been previously reported, earnings effects across most models were still positive and statistically significant. Assessed at the mean of age-50 earnings, a 1-SD boost in age-16 mathematics achievement would be predicted to lead to approximately $2,800 in additional earnings across the year. As this approximate association was found for every earnings wave considered between ages 33 and 50, this would be a

substantial earnings boost (approximately $50,000 over 17 years). Further, if a program boosted average math *and* reading scores by a full SD, then my estimates would project an earnings increase of approximately $84,000 over the time earnings were assessed here (i.e., this estimate comes from the 12.4% effect reported for the composite achievement score in Table 4.6). These results suggest that boosts in high school academic skills could have important effects on long-run economic attainment.

However, a few caveats should be kept in mind. First, these effects were not derived from random variation, and it is unclear how much OVB could remain unaccounted for. It is unlikely that the NCDS control measures, although numerous, perfectly captured all of the potential sources of confounding variation. For example, NCDS assessed IQ at age 11, and it is unlikely that the unmeasured variation in IQ at age 16 is completely captured by the age-11 measure. Further, other socio-emotional and personality dimensions of interest now (e.g., motivation, grit) were also unassessed. When assessing how OVB might affect estimates, any source of OVB probably had a positive correlation with both earnings and test scores. Thus, if estimates were biased, they were likely to be biased upwards.

Second, the mechanisms through which test score impacts affect later earnings remain murky. Murane and Levy (2005) argue that the basic skills measured by achievement tests hold value on the labor market, as a growing share of jobs require higher-level communication and analytic skills. Yet, even if educational programs produce real gains in achievement skills at the end of high school, research on fadeout suggests that these skill impacts are likely to diminish over time (for a review see Bailey, Duncan, Odgers, & Yu, 2017). Thus, it is difficult to know exactly what mechanism produces the stable association between achievement skills and labor market earnings observed here, but signaling could also be a plausible mechanism. If

postsecondary institutions use test scores for determining enrollment, then a test score gain may help a marginal student obtain a better placement in higher education. This could then lead to real changes in their economic attainment, but the implications for academic interventions in this scenario are less clear. If educational programs were to positively shift the distribution of test scores, it seems likely that colleges would follow suit and shift the test score requirements for admission. Nevertheless, future research should carefully examine the specific mechanisms that link measured gains in achievement tests to economic attainment, as identifying these mechanisms will be key for developing successful academic programs.

Further, when interpreting the magnitude of the coefficients presented here, it important to remember that test score effects as large as one full standard deviation are hardly ever observed. For example, Krueger (2003) estimated that spending a year in a small class had an effect of 0.20 SD's on test scores, and Angrist and colleagues (2010) found that charter schools had an effect of approximately 0.35 SD's on mathematics scores. If the results reported here contained no bias, then these programs may still be expected to have worthwhile positive impacts on long-run earnings. However, given that the correlations reported here probably do not fully represent causal impacts, then it remains possible that achievement score impacts on the order of ¼ of a standard deviation may only produce negligible to small effects on earnings.

Finally, even if observed program impacts on achievement tests do not reliably imply large adult earnings impacts, this does not mean that academic programs cannot affect adult earnings. In a recent analysis of small-class size, Fredriksson, Öckert, and Oosterbeek (2013) used the correlation between earnings and test scores to estimate the effect of small class size or earnings using experimental data collected in Norway. They showed that if they had relied only on the earnings and test score correlation, they would have concluded that small classes had no

effect on earnings. However, analyses of program effects on actual earnings measures were statistically significant and substantial, indicating that the small class program affected other skills unmeasured by achievement test scores. Thus, although achievement tests remain important indicators of educational inputs, program evaluators should also focus on the impact of educational interventions on non-academic measures, as adult attainment may be affected through other channels.

Limitations

Unfortunately, the current analysis was unable to adequately examine the association between test scores and earnings in women, due to the low labor force participation patterns of women in the NCDS. This issue leads to a broader limitation of the study – participants in the data first entered the labor force nearly 40 years ago. Study subjects were in high school in the late 1970s, and it is possible that certain skills are valued today that were not valued when NCDS subjects first began working. However, we cannot accurately measure the long-run economic trajectories for today's young adults. Thus, it is imperative that future research studies continue to track the development of individuals over time, as assessing subjects' trajectories from childhood well into adulthood will remain critically important to researchers and policy-makers.

Further, these results were likely influenced by measurement error, as adjustments for measurement error did increase the size of the coefficients for the reading and mathematics test scores. It is difficult to gauge how measurement error and OVB both affect the results, as OVB probably positively inflates coefficient estimates, while measurement error is likely to push estimates toward 0. Also, because the sample was restricted to non-attriting men, the confidence intervals around most estimates were not unsubstantial. The size of the standard errors was also partly due to the low R-squared values in most of the models. As the pooled estimates in Table

4.6 show, even with all the control variables included, R-squared values hovered around 0.14 for the models conditional on working full-time, and they were 0.30 for the models that included unemployed workers. The low R-squared values underscore the aforementioned point made by Murnane and colleagues (2000): most of the variation in earnings is left unexplained in these models. Thus, research still has a long way to go to fully explain variation in adult attainment.

Conclusion

There remain strong theoretical reasons to believe that the skills measured by mathematics and reading tests should translate into adult economic success. However, assuming that program impacts observed on mathematics and reading tests will necessarily lead to large economic gains is a strong assumption, and results from the current study suggest that the association between test scores and earnings contains substantial bias. It is unlikely that all bias in the test score and earning correlation was accounted for here, and researchers should be cautious when using the correlation to project program impacts on adult attainment.

References

Altonji, J. G., & Pierret, C. (2001). Employer learning and statistical discrimination. *Quarterly Journal of Economics, 116,* 313-350.

Bailey, D., Duncan, G. J., Odgers, C. L., & Yu, W. (2017). Persistence and fadeout in the impacts of child and adolescent interventions. *Journal of Research on Educational Effectiveness, 10*(1), 7-39.

Bailey, D. H., Watts, T. W., Littlefield, A. K., & Geary, D. C. (2014b). State and trait effects on individual differences in children's mathematical development. *Psychological Science, 25*(11), 2017-2026. doi: 10.1177/0956797614547539

Bartik, T. J., Gormley, W., & Adelstein, S. (2012). Earnings benefits of Tulsa's pre-K program for different income groups. *Economics of Education Review, 31*(6), 1143-1161.

Blackburn, M. L., & Neumark, D. (1995). Are OLS estimates of the return to schooling biased downward? Another look. *Review of Economics and Statistics, 77*(2), 217-230.

Bloom, D. E., Canning, D., & Shenoy, E. S. (2012). The effect of vaccination on children's physical and cognitive development in the Philippines. *Applied Economics*, *44*(21), 2777-2783.

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

Cho, H., Glewwe, P., & Whitler, M. (2012). Do reductions in class size raise students' test scores? Evidence from population variation in Minnesota's elementary schools. *Economics of Education Review*, *31*(3), 77-95.

Clements, D. H., Sarama, J., Wolfe, C. B., & Spitler, M. E. (2013). Longitudinal evaluation of a scale-up model for teaching mathematics with trajectories and technologies: Persistence of effects in the third year. *American Educational Research Journal, 50*(4), 812-850.

Connor, C. M., Morrison, F. J., Fishman, B. J., Crowe, E. C., Al Otaiba, S., & Schatschneider, C. (2013). A longitudinal cluster-randomized controlled study on the accumulating effects of individualized literacy instruction on students' reading from first through third grade. *Psychological Science, 24*(8), 1408-1419. doi:10.1177/0956797612472204

Cortes, K. E., & Goodman, J. S. (2014). Ability-tracking, instructional time, and better pedagogy: The effect of Double-Dose Algebra on student achievement. *The American Economic Review, 104*, 400-405. doi:http://dx.doi.org/10.1257/aer.104.5.400

Currie, J. & Thomas, D. (2001). Early test scores, school quality and SES: Longrun effects on wage and employment outcomes. In S. W. Polachek (Ed.), *Worker wellbeing in a changing labor market*. Amsterdam: Elsevier Science.

Curto, V. E., & Fryer Jr, R. G. (2014). The potential of urban boarding schools for the poor: Evidence from SEED. *Journal of Labor Economics, 32*(1), 65-93.

Deke, J. & Haimson, J. (2006). *The relationship between competencies observed in high school and postsecondary education and income: Is there more to student achievement than test scores?* Mathematica Working Paper 4872, Mathematica Policy Research. Retrieved from: http://www.mathematica-mpr.com/~/media/publications/PDFs/studentachieve.pdf

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

Deming, D. J. (2015). *The growing importance of social skills in the labor market* (No. w21473). National Bureau of Economic Research.

Dobbie, W. & Fryer, R. (2015). The medium-term impact of high-achieving charter schools. *Journal of Political Economy, 123*(5), 985-1037.

Dougherty, C. (2003). Numeracy, literacy, and earnings: Evidence from the National Longitudinal Survey of Youth. *Economics of Education Review, 22*, 511-521.

Duncan, G., Ludwig, J., & Magnuson, K. (2010). Child development. In P.B. Levine, & D. J. Zimmerman (Eds.), *Targeting investments in children fighting poverty when resources are limited* (pp. 27–58). Chicago: University of Chicago Press.

Finn, A. S., Kraft, M. A., West, M. R., Leonard, J. A., Bish, C. E., Martin, R. E., ... & Gabrieli, J. D. (2014). Cognitive skills, student achievement tests, and schools. *Psychological Science, 25*(3), 736-744.

Fredriksson, P., Öckert, B., & Oosterbeek, H. (2013). Long-term effects of class size. *The Quarterly Journal of Economics*, *128*(1), 249-285.

Gathercole, S. E., Pickering, S. J., Knight, C., & Stegmann, Z. (2004). Working memory skills and educational attainment: Evidence from national curriculum assessments at 7 and 14 years of age. *Applied Cognitive Psychology, 18*(1), 1-16.

Grogger, J., & Eide, E. (1995). Changes in college skills and the rise in the college wage premium. *Journal of Human Resources*, *30*(2), 280-310.

Hanushek, E. A. (2009). The economic value of education and cognitive skills. *In: Handbook of Education Policy Research.* New York, NY: Routledge; 2009:39-56.

Hanushek, E. A., & Rivkin, S. G. (2010). Generalizations about using value-added measures of teacher quality. *The American Economic Review, 100*(2), 267-271.

Heckman, 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.

Heckman, J. & Vytlacil (2001). Identifying the role of cognitive ability in explaining the level of and change in the return of schooling. *Review of Economics and Statistics, 83*(1), 1-12.

Jencks, C., & Phillips, M. (1999). Aptitude or achievement: Why do test scores predict educational attainment and earnings? In S. Mayer & P. E. Peterson (Eds.), *Earning and learning: How schools matter* (pp. 15–47). Washington, DC: Brookings Institution Press.

Kaufman, S. B., Reynolds, M. R., Liu, X., Kaufman, A. S., & McGrew, K. S. (2012). Are cognitive g and academic achievement g one and the same g? An exploration on the Woodcock–Johnson and Kaufman tests. *Intelligence, 40*(2), 123-138.

Krueger, A. B. (2003). Economic considerations and class size. *The Economic Journal, 113*(485), F34-F63.

Levy, F., & Murnane, R. J. (2005). *The new division of labor: How computers are creating the next job market.* New York, NY: Russell Sage Foundation.

Lindqvist, E., & Vestman, R. (2011). The labor market returns to cognitive and noncognitive ability: Evidence from the Swedish enlistment. *American Economic Journal: Applied Economics, 3*(1), 101-128.

Loeb, S., Bridges, M., Bassok, D., Fuller, B., & Rumberger, R. W. (2007). How much is too much? The influence of preschool centers on children's social and cognitive development. *Economics of Education Review, 26*(1), 52-66.

Marle, K., Chu, F. W., Li, Y., & Geary, D. C. (2014). Acuity of the approximate number system and preschoolers' quantitative development. *Developmental Science, 17*(4), 492-505.

Miles, A., Savage, M., & Bühlmann, F. (2011). Telling a modest story: accounts of men's upward mobility from the National Child Development Study. *The British Journal of Sociology, 62*(3), 418-441.

Murnane, R. J., Willett, J. B., Duhaldeborde, Y., & Tyler, J. H. (2000). How important are the cognitive skills of teenagers in predicting subsequent earnings? *Journal of Policy Analysis and Management, 19*(4), 547-568.

Murnane, R. J., Willett, J. B., & Levy, F. (1995). The growing importance of cognitive skills in wage determination. *Review of Economics and Statistics, 78,* 251-266. Retrieved from: http://www.jstor.org/stable/2109863

Neal, D. & Johnson, W. R. (1996). The role of premarket factors in black-white wage differences. *Journal of Political Economy, 104*(5), 869-895.

Reyna, V. F., Nelson, W. L., Han, P. K., & Dieckmann, N. F. (2009). How numeracy influences risk comprehension and medical decision making. *Psychological Bulletin, 135,* 943-973. doi:10.1037/a0017327

Ritchie, S. J., & Bates, T. C. (2013). Enduring links from childhood mathematics and reading achievement to adult socioeconomic status. *Psychological Science*, *24*(7), 1301-1308. doi: 10.1177/0956797612466268

Reardon, S. The widening income achievement gap. *Educational Leadership, 70*(8), 10-16.

Rose, H. (2006). Do gains in test scores explain labor market outcomes? *Economics of Education Review, 25,* 430-446.

Shepherd, P. (2012). *1958 National Child Development Study user guide: Measures of ability at ages 7 to 16.* London: Center for Longitudinal Studies.

Siegler, R. S., & Lortie-Forgues, H. (2014). An integrative theory of numerical development. *Child Development Perspectives, 8*(3), 144-150.

St. Clair-Thompson, H. L., & Gathercole, S. E. (2006). Executive functions and achievements in school: Shifting, updating, inhibition, and working memory. *The Quarterly Journal of Experimental Psychology, 59*(4), 745-759.

Strenze, T. (2007). Intelligence and socioeconomic success: A meta-analytic review of longitudinal research. *Intelligence, 35*(5), 401-426.

Takizawa, R., Maughan, B., & Arseneault, L. (2014). Adult health outcomes of childhood bullying victimization: evidence from a five-decade longitudinal British birth cohort. *American Journal of Psychiatry, 171*(7), 777-784.

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

Watts, T. W., Duncan, G. J., Clements, D. H., Sarama, J. (2017). What is the long-run impact of learning mathematics during preschool? *Child Development.* Advance online publication. doi:10.1111/cdev.12713

Watts, T. W., Duncan, G. J., Siegler, R. S., & Davis-Kean, P. E. (2014). What's past is prologue: Relations between early mathematics knowledge and high school achievement. *Educational Researcher, 43,* 352-360. doi:10.3102/0013189X14553660

West, M. R., Kraft, M. A., Finn, A. S., Martin, R. E., Duckworth, A. L., Gabrieli, C. F., & Gabrieli, J. D. (2016). Promise and paradox: Measuring students' non-cognitive skills and the impact of schooling. *Educational Evaluation and Policy Analysis, 38*(1), 148-170.

Table 4.1

Descriptive Statistics for Employment and Earnings Measures in Men

	M	SD	N
No. Valid Employment Waves	3.14	1.08	4822
No. Valid Earnings Waves	2.21	1.47	4822
Employment at Age 33 (1991)			
Employed	0.91		4082
Full-Time	0.90		4082
Part-Time	0.01		4082
Monthly Earnings (2016 USD)	2576.55	1131.13	3114
Employment at Age 41 (1999)			
Employed	0.91		4094
Full-Time	0.88		4094
Part-Time	0.02		4094
Monthly Earnings (2016 USD)	3131.67	1756.49	2890
Employment at Age 46 (2004)			
Employed	0.92		3434
Full-Time	0.90		3434
Part-Time	0.02		3434
Monthly Earnings (2016 USD)	3583.79	1957.96	2366
Employment at Age 50 (2008)			
Employed	0.89		3514
Full-Time	0.86		3514
Part-Time	0.03		3514
Monthly Earnings (2016 USD)	3400.72	1925.96	2269

Note. For all values presented, the sample was limited to men with non-missing high school math and reading scores, and the top 1% of earners at any measurement point were also excluded.

Table 4.2

High School Descriptive Characteristics by Average Adult Earnings for Men

	Full Sample		Top 50% Earners		Bottom 50% Earners	
	M	SD	M	SD	M	SD
Math (0-31)	13.74	7.17	16.47	7.13	11.47	6.42
Reading (0-35)	25.75	7.06	28.37	5.45	23.59	7.55
Personality and Behaviors- Scaled 1 through 5						
Timid to Aggressive	2.97	0.75	2.99	0.67	2.94	0.83
Flexible to Rigid	2.79	0.78	2.69	0.77	2.86	0.79
Sociable to Withdrawn	2.43	1.05	2.28	1.01	2.57	1.06
Cautious to Impulsive	2.76	0.92	2.75	0.85	2.76	0.96
Hardworking to Lazy	2.86	1.21	2.61	1.19	3.06	1.18
Moody to Affable	2.43	1.17	2.27	1.11	2.55	1.19
Demographics						
Free School Meals (prop.)	0.07		0.04		0.10	
Financial Problems in Past Year (prop.)	0.07		0.05	Co	0.10	
Family Monthly Income (2016 USD)	2517.78	1207.70	2667.80	1244.78	2382.78	1121.19
Family Owns Home (prop.)	0.42		0.49		0.35	
School Characteristics						
School Enrollment	916.65	418.09	922.27	410.85	918.84	428.74
Student/Teacher Ratio	16.93	8.78	16.62	2.44	17.22	12.52
Percent of Boys in School Taking O-Levels	26.87	34.56	33.39	37.65	21.55	30.73
Percent of Girls in School Taking O-Levels	19.82	29.23	22.87	31.61	18.22	27.87
Observations	4822		1975		2266	

Note. The sample is limited to men with valid math and reading scores in high school, and at least one earnings measure. The top 50% of earners represent men that, on average, were at or above the 50th percentile in earnings across each of the earnings waves. The behavioral and personality scales were measured on a continuum from "1" to "5" (e.g., for "Timid to Aggressive" a value of "1" would indicated maximum timidness and a value of "5" would indicate maximum aggressiveness).

Table 4.3

Associations Between High School Test Scores and Log-Monthly Earnings Conditional on Working Full Time

	Bivariate	**Math & Reading Only**	**Family Back & Health**	**Behaviors and Personality**	**Cognitive Ability and Pre-Tests**	**Concurrent Characteristics**
	A	**B**	**C**	**D**	**E**	**F**
				Age 33		
	(1)	(2)	(3)	(4)	(5)	(6)
Math	0.115***	0.070***	0.063***	0.065***	0.054***	0.058***
	(0.009)	(0.011)	(0.011)	(0.012)	(0.014)	(0.015)
Reading	0.124***	0.075***	0.072***	0.069***	0.056***	0.052***
	(0.008)	(0.011)	(0.011)	(0.011)	(0.014)	(0.015)
p-value of difference						0.8
N				2840		
				Age 41		
	(7)	(8)	(9)	(10)	(11)	(12)
Math	0.144***	0.107***	0.098***	0.101***	0.093***	0.091***
	(0.013)	(0.018)	(0.018)	(0.018)	(0.023)	(0.024)
Reading	0.139***	0.063***	0.059**	0.050**	0.047*	0.043
	(0.013)	(0.017)	(0.019)	(0.019)	(0.022)	(0.023)
p-value of difference						0.203
N				2836		
				Age 46		
	(13)	(14)	(15)	(16)	(17)	(18)
Math	0.154***	0.110***	0.094***	0.097***	0.080***	0.070***
	(0.010)	(0.013)	(0.014)	(0.014)	(0.017)	(0.017)
Reading	0.163***	0.081***	0.079***	0.072***	0.073***	0.061***
	(0.011)	(0.014)	(0.015)	(0.015)	(0.018)	(0.018)
p-value of difference						0.76
N				2324		
				Age 50		
	(19)	(20)	(21)	(22)	(23)	(24)
Math	0.153***	0.112***	0.097***	0.106***	0.094***	0.083***
	(0.013)	(0.017)	(0.017)	(0.018)	(0.022)	(0.024)
Reading	0.153***	0.072**	0.076**	0.069**	0.047	0.050
	(0.017)	(0.022)	(0.024)	(0.026)	(0.026)	(0.027)
p-value of difference						0.435
N				2202		

Note. Robust standard errors are presented in parentheses. Test scores were transformed to z-scores, so coefficients can be interpreted as the effect of a 1-SD gain in a high school test score on monthly earnings for a given age. The "p-value of difference" rows list p-values from post-hoc tests that tested whether the math and reading coefficients were equal to one another. The first row of the table lists the additional control variables added in the models presented in each column (e.g., for the models listed in Column C, behavioral and personality measures were added to the already included set of family background and health controls). Family background and health characteristics were measured at age 7 and 11. Measures of socio-emotional behaviors and personality were taken at ages 7 and 11 from teacher reports. Measures of cognitive and motor ability were taken at age 11, and this set of variables also includes teacher ratings of academic skills at ages 7 and 11. The pretests include math and reading scores assessed at ages 7 and 11. Finally, "concurrent characteristics" includes age-16 measures of socio-emotional skills and family characteristics.

* p<0.05 ** p<0.01 *** p<0.001

Table 4.4

Associations Between High School Test Scores and Log-Monthly Earnings With Unemployed Earnings Imputed as "0"

	Bivariate	**Math & Reading Only**	**Family Background & Health**	**Behaviors and Personality**	**Cognitive Ability and Pre-Tests**	**Concurrent Characteristics**
	A	**B**	**C**	**D**	**E**	**F**
				Age 33		
	(1)	(2)	(3)	(4)	(5)	(6)
Math	0.148***	0.078***	0.072***	0.070***	0.071***	0.069***
	(0.008)	(0.010)	(0.011)	(0.011)	(0.014)	(0.014)
Reading	0.164***	0.110***	0.105***	0.095***	0.085***	0.081***
	(0.009)	(0.012)	(0.012)	(0.013)	(0.015)	(0.016)
p-value of difference						0.599
N				3230		
				Age 41		
	(7)	(8)	(9)	(10)	(11)	(12)
Math	0.176***	0.113***	0.096***	0.090***	0.084***	0.076***
	(0.009)	(0.012)	(0.013)	(0.013)	(0.016)	(0.016)
Reading	0.179***	0.101***	0.091***	0.076***	0.066***	0.055***
	(0.010)	(0.013)	(0.013)	(0.013)	(0.016)	(0.016)
p-value of difference						0.417
N				3271		
				Age 46		
	(13)	(14)	(15)	(16)	(17)	(18)
Math	0.220***	0.150***	0.119***	0.114***	0.107***	0.097***
	(0.014)	(0.018)	(0.019)	(0.019)	(0.023)	(0.024)
Reading	0.230***	0.122***	0.119***	0.101***	0.073**	0.058*
	(0.016)	(0.021)	(0.022)	(0.022)	(0.027)	(0.027)
p-value of difference						0.329
N				2644		
				Age 50		
	(19)	(20)	(21)	(22)	(23)	(24)
Math	0.202***	0.149***	0.129***	0.130***	0.114***	0.101***
	(0.012)	(0.016)	(0.017)	(0.017)	(0.021)	(0.021)
Reading	0.190***	0.087***	0.081***	0.065***	0.035	0.022
	(0.013)	(0.017)	(0.018)	(0.018)	(0.022)	(0.023)
p-value of difference						0.022*
N				2644		

Note. Robust standard errors are presented in parentheses. Test scores were transformed to z-scores, so coefficients can be interpreted as the effect of a 1-SD gain in a high school test score on monthly earnings for a given age. The "p-value of difference" rows list p-values from post-hoc tests that tested whether the math and reading coefficients were equal to one another. The first row of the table lists the additional control variables added in the models presented in each column (e.g., for the models listed in Column C, behavioral and personality measures were added to the already included set of family background and health controls). For description of sets of control measures, see Table 4.3.

* p<0.05 ** p<0.01 *** p<0.001

Table 4.5

Associations Between Composite Math and Reading Scores and Log-Monthly Earnings Conditional on Working Full-Time

	Bivariate	**Family Back & Health**	**Behaviors and Personality**	**Cognitive Ability and Pre-Tests**	**Concurrent Characteristics**
	A	**B**	**C**	**D**	**E**
			Age 33		
	(1)	(2)	(3)	(4)	(5)
Achievement Composite	0.132***	0.123***	0.122***	0.100***	0.100***
	(0.009)	(0.010)	(0.010)	(0.015)	(0.016)
N			2840		
			Age 41		
	(6)	(7)	(8)	(9)	(10)
Achievement Composite	0.157***	0.145***	0.140***	0.130***	0.124***
	(0.013)	(0.014)	(0.015)	(0.024)	(0.025)
N			2836		
			Age 46		
	(11)	(12)	(13)	(14)	(15)
Achievement Composite	0.176***	0.159***	0.157***	0.140***	0.120***
	(0.010)	(0.011)	(0.012)	(0.019)	(0.019)
N			2324		
			Age 50		
	(16)	(17)	(18)	(19)	(20)
Achievement Composite	0.170***	0.158***	0.161***	0.132***	0.124***
	(0.014)	(0.019)	(0.022)	(0.024)	(0.026)
N			2202		

Note. Robust standard errors are presented in parentheses. The "achievement composite" variable is the standardized average of the age-16 math and reading tests. The first row of the table lists the additional control variables added in the models presented in each column (e.g., for the models listed in Column C, behavioral and personality measures were added to the already included set of family background and health controls). For description of sets of control measures, see Table 4.3.

* p<0.05 ** p<0.01 *** p<0.001

Table 4.6

Pooled Models- Associations Between High School Test Scores and Log-Average Earnings Between Age 33 and Age 50

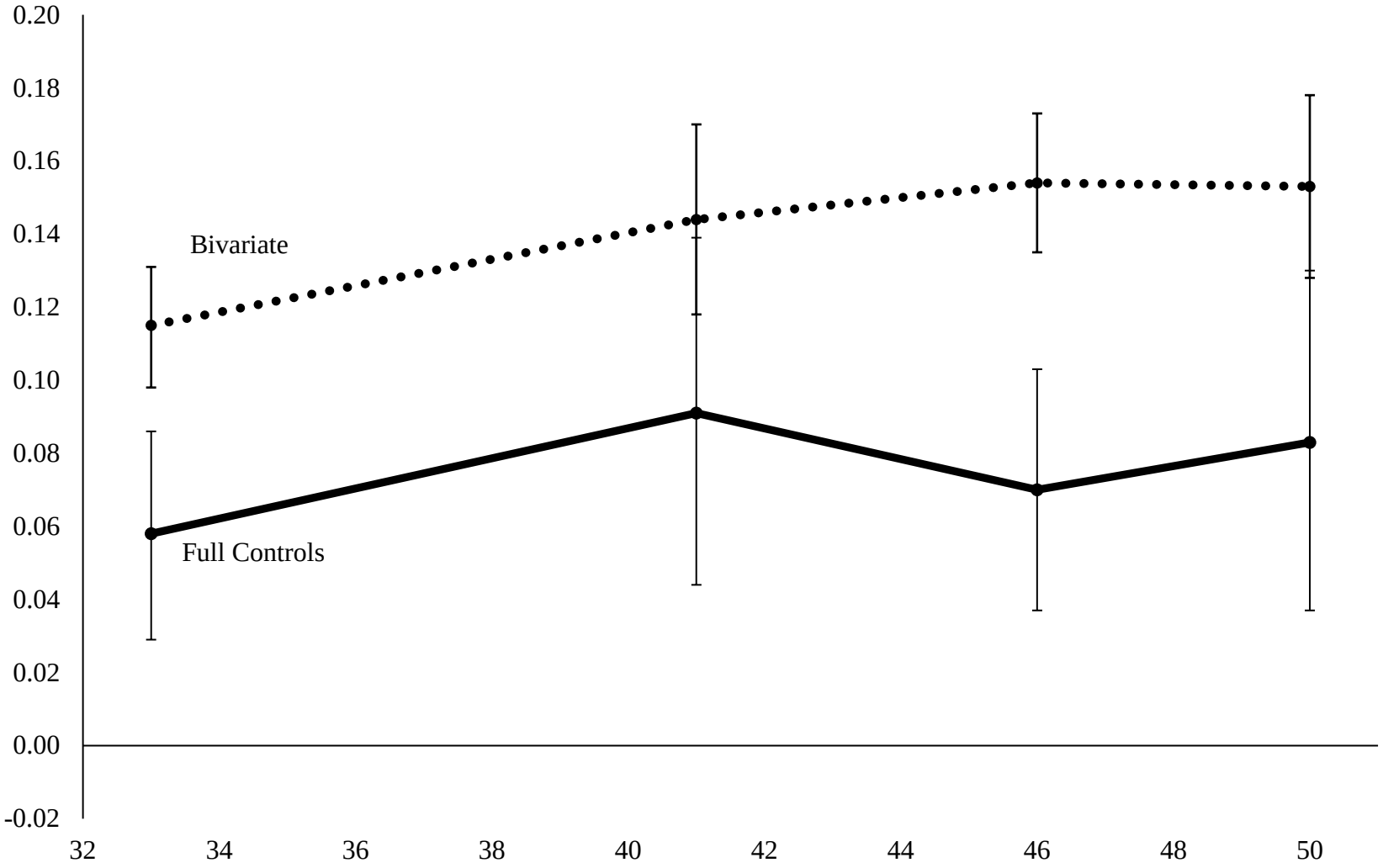
	Full-Time Only		**"0's" For Unemployed Earnings**		**Achievement Composite / Full-Time Only**	
	All Controls	School Characteristics	All Controls	School Characteristics	All Controls	School Characteristics
	(1)	(2)	(3)	(4)	(5)	(6)
Math	0.076***	0.067***	0.080***	0.078***		
	(0.012)	(0.012)	(0.012)	(0.013)		
Reading	0.053***	0.051***	0.053***	0.053***		
	(0.012)	(0.012)	(0.013)	(0.013)		
Achievement Composite					0.124***	0.114***
					(0.014)	(0.015)
Other H.S. Characteristics						
Family Income	0.030**	0.029**	0.026**	0.028**	0.030**	0.029**
	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)
Timid to Aggressive	0.024**	0.024**	0.006	0.007	0.024**	0.024*
	(0.009)	(0.009)	(0.010)	(0.010)	(0.009)	(0.009)
Moody to Affable	0.008	0.008	-0.001	0.000	0.008	0.008
	(0.009)	(0.010)	(0.010)	(0.010)	(0.009)	(0.010)
Flexible to Rigid	-0.000	0.001	0.010	0.010	0.000	0.001
	(0.009)	(0.009)	(0.008)	(0.008)	(0.009)	(0.009)
Sociable to Withdrawn	-0.036***	-0.038***	-0.045***	-0.045***	-0.036***	-0.038***
	(0.008)	(0.008)	(0.009)	(0.009)	(0.008)	(0.008)
Hardworking to Lazy	-0.021*	-0.025**	-0.029**	-0.031**	-0.023*	-0.026**
	(0.009)	(0.009)	(0.010)	(0.010)	(0.009)	(0.009)
Cautious to Impulsive	0.018	0.016	0.002	0.001	0.018*	0.017
	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)
Observations	10202	10202	11789	11789	10202	10202
R-squared	0.141	0.145	0.300	0.303	0.141	0.145

Note. Robust standard errors are presented in parentheses. All independent variables shown were transformed to z-scores. All models include the full set of controls used in the "Column F" models of Table 4.3. Models 2, 4, and 6 add school-level measures of high-school quality to the regressions. Pooled models were generated by treating each respective earnings measure (taken at ages 33, 41, 46, and 50, respectively) as independent observations. Models controlled for a "time of earnings measurement" fixed effect and standard errors were adjusted for person-level clustering. Models 1 and 2 correspond to the models shown in Table 4.3, as only adult men who indicated full-time employment at a given earnings measurement were included. Models 3 and 4 correspond to the estimates shown in Table 4.4, with men who indicated that they were working part-time or that they were unemployed included in the models. Earnings for men who indicated unemployment were imputed to be "0." Models 5 and 6 correspond to the models shown in Table 4.5, with age-16 math and reading scores averaged and standardized. All high school characteristics were measured at age 16, along with the high school test scores.

* p<0.05 ** p<0.01 *** p<0.001

Figure 4.1

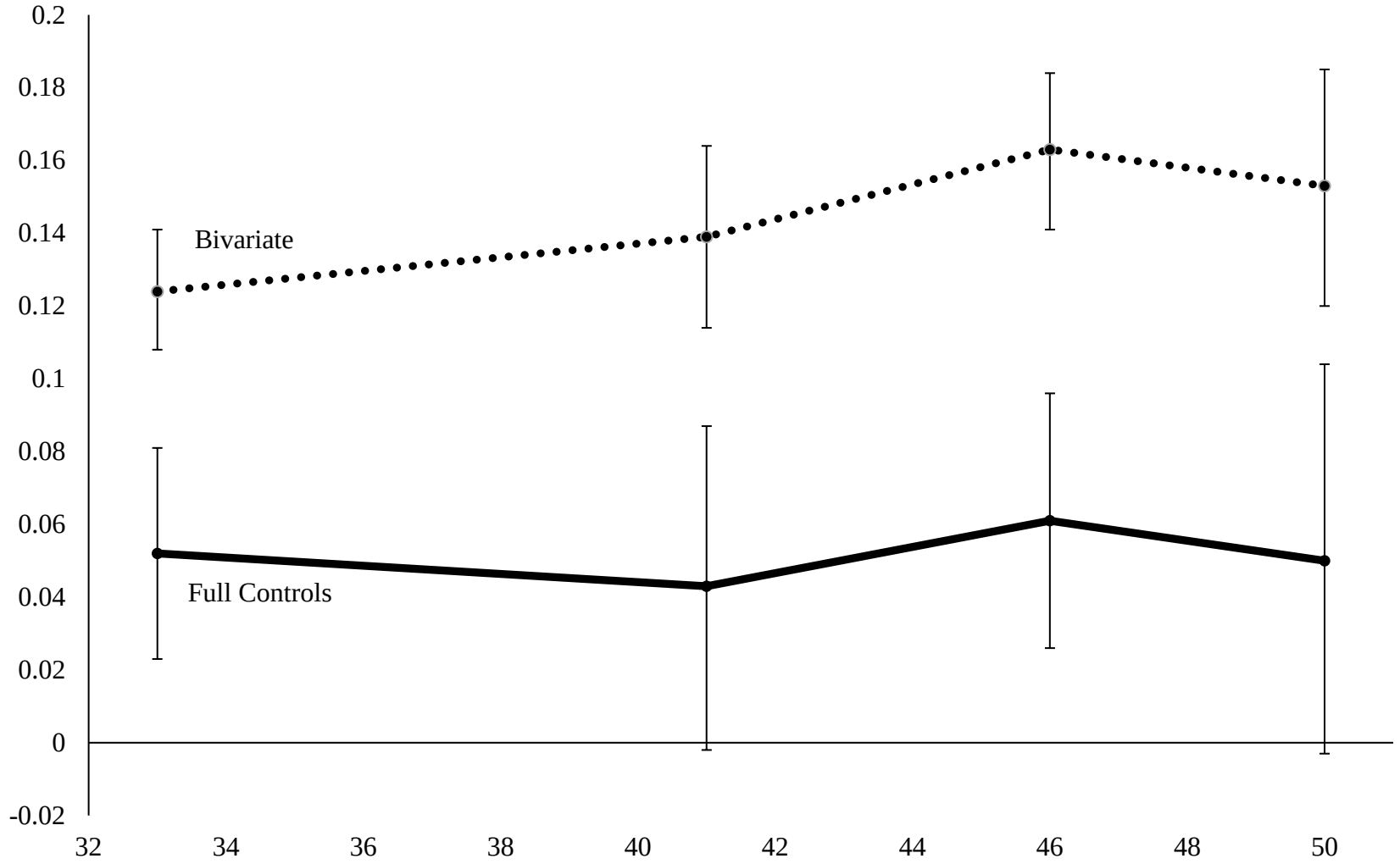
Plotted Associations Between Math Test Scores and Log-Earnings



Note. Error bars represent 95% confidence intervals. Bivariate lines correspond to models shown in in Column A of Table 4.3. Control lines correspond to models shown in Column F of Table 4.3.

Figure 4.2

Plotted Associations Between Reading Test Scores and Log-Earnings



Note. Error bars represent 95% confidence intervals. Bivariate lines correspond to models shown in in Column A of Table 4.3. Control lines correspond to models shown in Column F of Table 4.3.

Appendix

Control Measures

Appendix Table 4.1 presents all control measures used throughout each of the models presented in the main paper. Control variables are organized via the categories shown in the columns of each main results table. The descriptive statistics presented were generated using all of the available cases for each variable in the analysis sample (n = 4,822).

Results for Women

Appendix Table 4.2 presents estimates of the association between math and reading test scores and log-earnings for women who indicated working full-time at any of the earnings waves. These estimates can be compared with the estimates shown in Table 4.3 of the main text. Finally, Appendix Table 4.3 presents estimates of the association composite achievement test score and log-earnings for women working full-time (compare with table 4.6 from the main text).

Additional Results

Appendix Table 4.4 presents results from models that used Full Information Maximum Likelihood (FIML) to adjust for missing data. Results were similar to estimates shown in Table 4.3 of the main text. Column 2 of Appendix Table 4.4 presents results from pooled models that used the "errors in variables" adjustment in Stata 13.0 to account for measurement error in the mathematics and reading test scores.

Appendix Table 4.1 (Panel A)
Descriptive Statistics for All Control Variables Used

	Source/ Wave	% Missing	M	SD	Min	Max
Family Background and Health						
Child 8 Years	P1	8%	0.97	0.17	0	1
Child Second Born	P1	8%	0.32	0.47	0	1
Child Third Born	P1	8%	0.14	0.35	0	1
Child's Acting Mom is Birth Mom	P1	12%	0.97	0.16	0	1
Child's Acting Dad is Birth Dad	P1	12%	0.94	0.23	0	1
Attend private nursery	P1	18%	0.06	0.23	0	1
Attend Local Authority Nursery	P1	17%	0.03	0.16	0	1
Attend Other Organized Preschool	P1	14%	0.04	0.20	0	1
No. Rooms in Home	P1	13%	4.83	1.32	1	15
Father Occupational Category						
Professional	P1	14%	0.15	0.36	0	1
Managerial	P1	14%	0.10	0.30	0	1
Skilled Non-Manual	P1	14%	0.45	0.50	0	1
Skilled Manual	P1	14%	0.02	0.14	0	1
Partly Skilled	P1	14%	0.16	0.37	0	1
Unskilled	P1	14%	0.06	0.23	0	1
No. Children in Household	P2	14%	3.02	1.54	1	9
Father Region of Birth						
Northern	P2	16%	0.08	0.27	0	1
E & W Ridings	P2	16%	0.09	0.28	0	1
North Midlands	P2	16%	0.06	0.24	0	1
Eastern	P2	16%	0.05	0.22	0	1
London & S. Eastern	P2	16%	0.17	0.38	0	1
Southern	P2	16%	0.03	0.18	0	1
South Western	P2	16%	0.05	0.22	0	1
Midlands	P2	16%	0.08	0.27	0	1
Wales	P2	16%	0.07	0.25	0	1
Scotland	P2	16%	0.13	0.34	0	1
Don't Know	P2	16%	0.08	0.27	0	1
Father Occupational Category						
Professional	P2	16%	0.19	0.39	0	1
Managerial	P2	16%	0.10	0.30	0	1
Skilled Non-Manual	P2	16%	0.43	0.50	0	1
Skilled Manual	P2	16%	0.02	0.13	0	1
Partly Skilled	P2	16%	0.15	0.36	0	1
Unskilled	P2	16%	0.05	0.23	0	1
Mother Region of Birth						
Northern	P2	15%	0.08	0.28	0	1
E & W Ridings	P2	15%	0.09	0.28	0	1
North Midlands	P2	15%	0.06	0.24	0	1
Eastern	P2	15%	0.06	0.23	0	1
London & S. Eastern	P2	15%	0.17	0.38	0	1
Southern	P2	15%	0.04	0.19	0	1
South Western	P2	15%	0.05	0.22	0	1

Note. See Panel D for full table note.

Appendix Table 4.1 (Panel B)
Descriptive Statistics for All Control Variables Used- Continued

	Source/ Wave	% Missing	M	SD	Min	Max
Family Background and Health						
Mother Region of Birth						
Midlands	P2	15%	0.09	0.28	0	1
Wales	P2	15%	0.06	0.24	0	1
Scotland	P2	15%	0.12	0.33	0	1
Don't Know	P2	15%	0.07	0.26	0	1
Father Role in Childcare	P2	18%	2.60	0.64	1	3
One Person Per Room in Home	P2	10%	0.59	0.49	0	1
Father Provides Family Income	P2	15%	0.96	0.20	0	1
Child Receive Free School Meals	P2	15%	0.08	0.28	0	1
Does Child Have Health Problems	P2	6%	0.21	0.41	0	1
Index of Diseases from ages 7-11	P2	14%	3.01	1.17	0	8
Child Epileptic	P2	15%	0.05	0.23	0	1
Index of Health Problems	P2	14%	0.08	0.13	0	1
No. Times Child Admitted to Hospital	P2	14%	0.76	1.04	0	9
Behaviors and Personality						
Coordination Problems	T2	14%	2.77	0.42	1	4
Speech Difficulties	T2	15%	2.87	0.38	1	4
Cannot Sit Still	T2	14%	2.66	0.52	1	4
British Social Adjustment Guide						
Inconsequential Behavior	T2	13%	1.72	2.17	0	12
Nervous Symptoms	T2	13%	0.13	0.41	0	4
Anxiety	T2	13%	0.52	1.15	0	10
Acceptance	T2	13%	0.42	0.91	0	7
Hostility Toward Other Children	T2	13%	0.28	0.82	0	9
Writes Off Adults	T2	13%	1.12	1.80	0	12
Hostility Toward Adults	T2	13%	0.91	1.88	0	19
Miscellaneous Symptoms	T2	13%	0.57	0.98	0	7
Restlessness	T2	13%	0.26	0.60	0	4
Unforthcoming	T2	13%	1.54	2.02	0	12
Depression	T2	13%	1.09	1.55	0	10
Withdrawn	T2	13%	0.37	0.85	0	8
Cognitive Ability and Pre-Tests						
Motor Ability- Design Copy Test	C1	11%	7.14	1.97	0	12
Motor Ability- Draw A Man Test	C1	13%	23.81	7.00	0	51
Motor Ability- Design Copy Test	C2	13%	8.47	1.44	0	12
IQ- General Ability Verbal Test	C2	13%	21.94	9.24	0	40
IQ- General Ability Non-Verbal Test	C2	13%	21.46	7.45	0	40
Teacher Ratings						
Number Work (Math)	T1	10%	3.07	0.88	1	5
Oral Ability	T1	10%	2.93	0.95	1	5
Reading Ability	T1	11%	3.00	0.89	1	5
Good Grasp of English	T1	11%	3.94	0.40	1	4
General Knowledge	T2	13%	2.93	0.87	1	5
Use of Books	T2	13%	2.95	0.86	1	5

Note. See Panel D for full table note.

Appendix Table 4.1 (Panel C)

Descriptive Statistics for All Control Variables Used- Continued

	Source/ Wave	% Missing	M	SD	Min	Max
Cognitive Ability and Pre-Tests						
Teacher Ratings						
Oral Ability	T2	13%	3.00	0.76	1	5
Good Grasp of English	T2	13%	2.95	0.24	1	4
Number Work (Math)	T2	13%	3.08	0.95	1	5
Math and Reading Pre-Tests						
Math	C1	11%	5.32	2.44	0	10
Reading	C1	11%	23.14	6.97	0	30
Math	C2	13%	17.85	10.46	0	40
Reading	C2	13%	16.50	6.37	0	35
Concurrent Characteristics						
Family Monthly Income (2016 USD)	P3	27%	2517.78	1207.70	121	7311
No. of Diseases Contracted	P3	23%	1.15	0.47	1	5
No. of Siblings	P3	19%	2.37	1.80	0	14
Mother Works	P3	20%	0.67	0.47	0	1
Parents Wish Child Left School at Age 15	P3	21%	0.25	0.43	0	1
No. Schools Attended Since Age 11	P3	19%	1.24	0.51	1	5
Does Any Child Get Free Meals	P3	20%	0.09	0.28	0	1
Serious Financial Trouble in Last Year	P3	21%	0.09	0.29	0	1
Owns Home	P3	19%	0.52	0.50	0	1
No. of Rooms in Home	P3	19%	4.97	1.59	1	39
Child Shares Bedroom	P3	19%	0.39	0.49	0	1
No. Family Moves Since Birth	P3	19%	1.83	1.86	0	9
Current Region of Residence						
North West	P3	0%	0.11	0.32	0	1
E & W Riding	P3	0%	0.09	0.28	0	1
North Midlands	P3	0%	0.08	0.27	0	1
Midlands	P3	0%	0.10	0.29	0	1
East	P3	0%	0.09	0.28	0	1
South East	P3	0%	0.15	0.36	0	1
South	P3	0%	0.07	0.25	0	1
South Western	P3	0%	0.07	0.26	0	1
Wales	P3	0%	0.06	0.24	0	1
Scotland	P3	0%	0.11	0.32	0	1
Personality Ratings						
Timid to Aggressive	T3	3%	2.97	0.75	1	5
Moody to Affable	T3	3%	2.43	1.17	1	5
Flexible to Rigid	T3	3%	2.79	0.78	1	5
Sociable to Withdrawn	T3	3%	2.43	1.05	1	5
Hardworking to Lazy	T3	3%	2.86	1.21	1	5
Cautious to Impulsive	T3	3%	2.76	0.92	1	5
School Characteristics						
Grammar School	T3	0%	0.1	0.31	0	1
Secondary Modern School	T3	0%	0.22	0.41	0	1

Note. See Panel D for full table note.

Appendix Table 4.1 (Panel D)

Descriptive Statistics for All Control Variables Used- Continued

	Source/ Wave	% Missing	M	SD	Min	Max
School Characteristics						
Other School Type	T3	0%	0.1	0.29	0	1
School Enrollment	T3	2%	916.65	418.09	1	2674
School Co-Ed	T3	1%	0.24	0.43	0	1
School has PTA	T3	1%	0.66	0.48	0	1
Percent Boys Taking O-Levels	T3	6%	26.87	34.56	0	100
Percent Girls Taking O-Levels	T3	29%	19.82	29.23	0	100
Percent Boys Studying CSE Only	T3	6%	28.53	26.58	0	100
Percent Girls Studying CSE Only	T3	29%	30.38	25.92	0	100
Percent Boys Studying Both O-Levels & CSE	T3	6%	23.43	23.57	0	100
Percent Girls Studying Both O-Levels & CSE	T3	29%	24.69	24.33	0	100
Proportion Taking A-Levels	T3	37%	0.02	0.03	0	0
Proportion Taking Degree Courses	T3	39%	0.01	0.02	0	0
Percentage of Boys Staying in School	T3	7%	60.93	27.61	0	100
Percentage of Girls Staying in School	T3	31%	53.75	26.66	0	100
Full Time Teachers	T3	2%	54.46	24.51	1	190
Student to Teacher Ratio	T3	3%	16.93	8.78	2	590
No. Teachers to Quit Last Year	T3	4%	7.62	6.15	0	46
School Facilities Lacking	T3	0%	1.53	1.62	0	7
Index of School Disciplinary Methods	T3	1%	6.89	1.44	0	9

Note. The "Source/Wave" column marks the wave and survey source for each measure. "P" represents parent survey, "C" represents direct child assessment, and "T" represents teacher assessment. Wave 1 was assessed at age 7; Wave 2 at ages 10 and 11; Wave 3 at age 16. The "% Missing" column percent of cases (rounded to the nearest whole number) missing from the sample of 4,822.

Appendix Table 4.2

Associations Between High School Test Scores and Log-Monthly Earnings Conditional on Working Full Time - Women

	Bivariate	Math & Reading Only	Family Back & Health	Behaviors and Personality	Cognitive Ability and Pre-Tests	Concurrent Characteristics
	A	**B**	**C**	**D**	**E**	**F**
				Age 33		
	(1)	(2)	(3)	(4)	(5)	(6)
Math	0.139***	0.085***	0.076***	0.064***	0.054**	0.037
	(0.012)	(0.015)	(0.016)	(0.017)	(0.020)	(0.022)
Reading	0.169***	0.105***	0.087***	0.083***	0.100***	0.086***
	(0.015)	(0.018)	(0.019)	(0.020)	(0.024)	(0.025)
p-value of difference						0.182
N				1318		
				Age 41		
	(7)	(8)	(9)	(10)	(11)	(12)
Math	0.121***	0.066**	0.071**	0.074**	0.068*	0.062*
	(0.022)	(0.025)	(0.024)	(0.024)	(0.030)	(0.031)
Reading	0.148***	0.104***	0.089***	0.091***	0.087*	0.084*
	(0.022)	(0.025)	(0.026)	(0.026)	(0.035)	(0.037)
p-value of difference						0.671
N				1673		
				Age 46		
	(13)	(14)	(15)	(16)	(17)	(18)
Math	0.147***	0.074***	0.068***	0.064***	0.038	0.031
	(0.013)	(0.016)	(0.017)	(0.017)	(0.025)	(0.026)
Reading	0.189***	0.140***	0.135***	0.133***	0.119***	0.116***
	(0.013)	(0.017)	(0.019)	(0.020)	(0.023)	(0.022)
p-value of difference						0.023*
N				1552		
				Age 50		
	(19)	(20)	(21)	(22)	(23)	(24)
Math	0.155***	0.087***	0.073***	0.068**	0.048*	0.040
	(0.014)	(0.019)	(0.021)	(0.021)	(0.021)	(0.022)
Reading	0.188***	0.127***	0.121***	0.127***	0.117***	0.111**
	(0.017)	(0.022)	(0.024)	(0.025)	(0.032)	(0.034)
p-value of difference						0.129
N				1602		

Note. Robust standard errors are presented in parentheses. Test scores were transformed to z-scores, so coefficients can be interpreted as the effect of a 1-SD gain in a high school test score on monthly earnings for a given age. The "p-value of difference" rows list p-values from post-hoc tests that tested whether the math and reading coefficients were equal to one another. The first row of the table lists the additional control variables added in the models presented in each column (e.g., for the models listed in Column C, behavioral and personality measures were added to the already included set of family background and health controls). Family background and health characteristics were measured at age 7 and 11. Measures of socio-emotional behaviors and personality were taken at ages 7 and 11 from teacher reports. Measures of cognitive and motor ability were taken at age 11, and this set of variables also includes teacher ratings of academic skills at ages 7 and 11. The pretests include math and reading scores assessed at ages 7 and 11. Finally, "concurrent characteristics" includes age-16 measures of socio-emotional skills and family characteristics.

* p<0.05 ** p<0.01 *** p<0.001

Appendix Table 4.3

Associations Between Composite Math and Reading Scores and Log-Monthly Earnings Conditional on Working Full Time- Women

	Bivariate	**Family Back & Health**	**Behaviors and Personality**	**Cognitive Ability and Pre-Tests**	**Concurrent Characteristics**
	A	**B**	**C**	**D**	**E**
			Age 33		
	(1)	(2)	(3)	(4)	(5)
Achievement Composite	0.171***	0.148***	0.131***	0.133***	0.105***
	(0.013)	(0.015)	(0.016)	(0.023)	(0.025)
N			1318		
			Age 41		
	(6)	(7)	(8)	(9)	(10)
Achievement Composite	0.152***	0.144***	0.149***	0.139***	0.130***
	(0.023)	(0.024)	(0.025)	(0.037)	(0.039)
N			1673		
			Age 46		
	(11)	(12)	(13)	(14)	(15)
Achievement Composite	0.190***	0.180***	0.174***	0.132***	0.123***
	(0.012)	(0.015)	(0.015)	(0.028)	(0.029)
N			1552		
			Age 50		
	(16)	(17)	(18)	(19)	(20)
Achievement Composite	0.192***	0.172***	0.172***	0.140***	0.127***
	(0.015)	(0.017)	(0.018)	(0.026)	(0.028)
N			1602		

Note. Robust standard errors are presented in parentheses. The "achievement composite" variable is the standardized average of the age-16 math and reading tests. The first row of the table lists the additional control variables added in the models presented in each column (e.g., for the models listed in Column C, behavioral and personality measures were added to the already included set of family background and health controls). For description of sets of control measures, see Table 4.3.

* p<0.05 ** p<0.01 *** p<0.001

Appendix Table 4.4

Additional Models Using FIML Adjustments for Missing Data and EIV Adjustments for Measurement Error

	FIML		Meas. Error Corrected
	Age 33	Age 50	Pooled
	(1)	(2)	(3)
Math	0.041**	0.066**	0.130***
	(0.015)	(0.026)	(0.017)
Reading	0.056**	0.056	0.066***
	(0.017)	(0.029)	(0.017)
Control Variables	Inc.	Inc.	Inc.
Observations	3675	3017	10202
R-squared	-	-	0.123

Note. Models 1 and 2 can be compared with the models shown in Column F of Table 4.3. These models were estimated using the structural equation modeling (SEM) commands with full information maximum likelihood (FIML) to adjust for missing data. For each model, the sample was restricted to men who indicated full-time employment at the given earnings wave and who had non-missing math and reading test score data at age 16. The estimates shown in Column 3 were drawn from pooled models, and can be compared with estimates shown in Column 1 of Table 4.6 from the main text. The models were estimated with the "errors in variables" (EIV) option in Stata 13.0, which was used to adjust the coefficients on age-16 math and reading for measurement error.

* p<0.05 ** p<0.01 *** p<0.001

Chapter 5

Conclusion

Across all three chapters, I focused on theoretical and practical issues surrounding the promotion of academic skills during K-12 schooling. In study 2, I examined the effects of an intervention designed to boost mathematics achievement in early elementary school. The intervention tested an individualized instruction program in mathematics, and I found little indication that the program positively affected student math achievement. This study revealed some of the inherent difficulties in designing and implementing successful academic interventions, and it also raised questions concerning teachers' willingness to adopt researcher-developed instructional methods.

In studies 1 and 3, I focused on the possible long-run effects of interventions that do successfully boost academic skills. In study 1, I relied on instrumental variables to produce exogenous variation in school-entry math skills. Although I found some indication that instrumented early math scores predicted late-elementary school achievement, the association between early and later math achievement was much smaller, and less consistent, than what had been reported by previous studies (e.g., Duncan et al., 2007; Watts et al., 2014). In study 3, I evaluated the association between adolescent academic test scores and adult earnings. I attempted to attenuate possible sources of bias in the association between test scores and earnings by controlling for a host of child and family characteristics, many of which had not been considered by prominent studies in the previous literature (e.g., Currie & Thomas, 2001; Murnane et al., 2000). Much like in study 1, although I found positive associations between academic test scores and later earnings, these effects were much smaller than had been previously reported.

In the following section, I briefly review the motivation for all three chapters, and I address common themes that arose across the studies. In particular, I focus on what these findings could mean for theories of skill building, the design of academic interventions, and our understanding of the role that academic interventions can play in shaping adult trajectories. Finally, I discuss next steps and directions I plan to take in my future research.

Implications for Skill-Building Theories

In all three chapters, I discussed the strong correlation reported by many studies between children's early academic competencies and later achievement (e.g., Bailey, Siegler, & Geary, 2014; Duncan et al., 2007; Watts et al., 2014). This finding drives the motivation behind each of the studies included here, as all three chapters attempted to address some implication of this apparent statistical relationship. This correlation implies that earlier academic skills lead to the acquisition of later skills through a skill-building process that unfolds over time, as children with higher levels of skills in "time 1" should be better equipped to learn new skills in "time 2" (see Cunha & Heckman, 2008). Further, many argue that this skill acquisition process should be supported by academic interventions during K-12 schooling, because adolescent academic skills correlate with adult earnings (e.g., Currie & Thomas, 2001; Murnane et al., 2000) much in the same way that early academic skills correlate with later skills. Thus, previous literature on academic skill development has painted a compelling picture for educational policy-makers: early academic abilities lead to the development of later academic abilities through a skill building process that can be enhanced by interventions, and such efforts should pay off through adult labor market returns. Indeed, this message has not been lost on policy-makers, as yearly testing in mathematics and reading has become standard procedure in U.S. public schooling (e.g., Every Student Succeeds Act, 2015).

However, the findings reported here, especially in studies 1 and 3, suggest that the skill-building process does not unfold in a purely autoregressive process, and previous correlational studies probably oversold the ability of academic interventions to spur skill attainment in later periods. Perhaps the most-cited theory of skill building is the model put forth by Cunha and Heckman (2008), which predicts that skill acquisition in a given period is dependent upon the skills obtained in a previous period. This model is intuitive, and it can be easily applied to the development of both mathematics and reading skills. For example, children who learn the alphabet at "time 1" should be better-equipped to learn to pronounce simple words at "time 2." In mathematics, mastering basic calculation skills should better prepare someone to learn more advanced skills like algebra. This model provides an appealing explanation to the aforementioned statistical relation observed between early test scores and later measures of achievement (e.g., Duncan et al., 2007).

Yet, these correlational studies are susceptible to omitted variables bias, and early academic interventions are typically followed by a steep pattern of fadeout that suggests that early skill gains might not necessarily lead to later skill gains (see Bailey, Duncan, Odgers, & Yu, 2017). In study 1, I tested whether the association between early and later math achievement was robust to an alternative modeling technique that relied solely on exogenous variation in early math achievement. I found that across many models tested, instrument-produced variation in early math achievement did predict later achievement, but this association was much smaller than had been reported by correlational studies. Interestingly, I found that although instrumented gains in early math skills did not predict achievement at fourth grade, they did predict achievement at fifth grade. Of course, this result should be further tested through replication with other samples and alternative interventions, but even if other studies confirm that gains in

preschool math achievement lead to better outcomes at grade 5, but not grade 4, this would still imply a need for serious revision to current skill-building theories. Following Cunha and Heckman's model (2008), there is no reason to expect that early skill gains would have no impact at one time-point but have a reemerging impact at a later time-point.

It is difficult to imagine a theoretical model of skill development that would predict such a pattern of results. However, Clements and colleagues (under review) proposed an alternative model that they called the "latent foundation hypothesis," which predicts that the returns to comprehensive early investments in skills may only be observed when subsequent environments place uniquely challenging demands on children's capacities to learn. Thus, the remerging fifth grade effect reported in study 1 may be due to the increased difficulty of the math curriculum encountered by students in grade 5. Yet, it is unclear what made the fifth-grade curriculum uniquely more challenging than the fourth-grade curriculum, when no such effect was observed.

Further, when the findings from study 1 are viewed alongside long-run findings from other early intervention studies (see Bailey et al., 2017), the reemerging fifth grade effect appears to be more of an aberration than a finding typical of the larger literature. Most studies of early academic interventions with long-run follow-ups find fadeout and not reemergence. In a series of papers, Bailey and Watts (Bailey, Watts, Littlefield and Geary, 2014; Watts, Clements, Sarama, Spitler, Wolfe, & Bailey, 2016) argue that this is because one-time boosts in achievement fail to affect the stable underlying factors that cause achievement tests to be so highly correlated in non-experimental studies. Thus, it is unlikely that interventions focused on narrow skill development will necessarily lead to long-run skill boosts without further instructional supports during later periods.

Implications for Intervention

If early interventions will only lead to later skill development (i.e., no fadeout) in the presence of continued instructional support, what kinds of instructional supports are needed? Many have argued that differentiated instruction can produce consistent skill gains following early intervention because such instructional approaches would allow teachers to tailor instruction to the needs of students who received early skill boosts (e.g., Stipek, 2017). As was discussed in study 2, descriptive studies have found that early-grade teachers appear to teach mathematics through a one-size-fits-all model that targets instruction at only the lowest achieving students in the class (Engel, Claessens, Watts, & Farkas, 2016). If this were the case, then students that received an early math intervention may encounter subsequent instruction targeted well below their skill level, and this could cause the fadeout patterns observed following many early academic and cognitive interventions.

In study 2, I examined the effects of a 2-year individualized instruction program in mathematics that was implemented in low-income schools in northern Florida. Unfortunately, I found little indication that the program positively affected mathematics achievement, and I found limited evidence that the program differentially benefited students who began the school year with higher levels of skills. These results somewhat dim hopes that differentiated instruction could be the answer for sustaining early intervention gains, but this intervention was a newly-developed pilot program. Thus, future attempts at individualizing instruction in mathematics may prove more successful, and a similarly-designed program targeted at reading achievement produced substantial positive effects on student reading achievement (Connor et al., 2013).

Yet, teachers' apparent resistance to implementing the program should not be overlooked when considering whether differentiated instruction might be a scalable alternative to current curricular approaches in early-grade classrooms. Early-grade teachers have been described as

having high levels of anxiety and discomfort teaching mathematics (e.g., Bursal & Paznokas, 2006), and the findings of Engel and colleagues (2013; 2016) suggest that they are more comfortable teaching very basic mathematics than teaching more advanced skills better aligned with levels of student knowledge. Thus, as enrollment rates in academic preschool programs rise (e.g., Lipsey, Farran, & Hofer, 2015; Yoshikawa, H., Weiland, C., & Brooks-Gunn, J., 2016), and as students enter kindergarten with higher levels of academic skills (Reardon & Portilla, 2016), early-grade teachers may need a substantial level of pedagogical support if we hope to alter instruction in order to build upon the gains that academic preschool might produce.

Findings reported from other studies also paint a rather mixed picture of the role that subsequent environments might play in sustaining early intervention gains, as some studies have found evidence supporting the hypothesis that high-quality instructional environments will help sustain early intervention boosts (e.g., Clements et al., 2013; Johnson & Jackson, 2017), whereas others have reported disconfirming evidence (Claessens, Engel, & Curran, 2014; Jenkins et al., under review). In studies that have shown some sustained treatment impacts for children who encountered higher quality educational environments following early intervention (e.g., Johnson & Jackson, 2017), the instructional factors that led to the sustained treatment impact remain unknown. Thus, more work is needed to better understand the role that subsequent environments can play in sustaining growth from early interventions, and I will further address my plans for continued work on this topic in the "future directions" section below.

Long-Run Returns to Academic Interventions

If we are able to identify interventions that produce sustained growth in academic skills from early childhood through adolescence, then how might such programs affect adult outcomes? In study 3, I turned to this question, as I examined the association between

adolescent measures of math and reading skills and adult labor market earnings. In the presence of a large set of child and environmental control variables, I found positive associations between adolescent math and reading scores and measures of adult earnings through age 50. Yet, it was unclear if all the bias in the correlation between test scores and earnings had been attenuated by control variables. Further, even if the reported correlations represent a type of causal effect, it remains unclear what mechanisms account for the association between adolescent skills and labor market productivity.

Understanding these mechanisms remains a crucial task for educational researchers and policy makers, as educational programs designed to boost academic skills might have very different effects on adult outcomes depending on which mechanisms account for the association between skills and earnings. For example, Levy and Murnane (2005) argue that test scores predict earnings because test scores measure skills that are valued on the labor market (also see Deming, 2015). In other words, employers value students who are high math achievers because math skills are actually necessary for performing regular job-related tasks. Thus, the Levy and Murnane perspective argues that adolescent academic skills affect labor market earnings through a human capital effect, and they also argue that this effect will grow larger as the economy shifts toward jobs that require high levels of technical skills.

If this were the case, then we would expect to see higher returns for test scores in more technical job sectors, and previous research has typically reported that the test score and earnings correlation is consistent across various industries (e.g., Grogger & Eide, 1995). Further, on employer surveys, math and literacy skills are regularly rated as some of the least important skills required for any job (Lerman, 2013). Although this evidence certainly does not rule out Levy and Murnane's (2005) hypothesis, empirical support for the idea that academic skills

constitute key competencies for labor market success remains surprisingly limited outside of the literature reviewed in study 3 (i.e., correlational studies reporting associations between test scores and earnings).

As was also discussed in study 3, test scores could also affect earnings because of a signaling process in which employers or postsecondary institutions respond to signs of math and reading skills because these skills signal latent cognitive abilities that will enable a student to be productive. This path is most likely to operate through postsecondary institutions, where test scores are explicitly used to make admissions decisions. To the extent that mathematics or reading achievement boosts later earnings through some sort of signaling effect, then the implications for wide-ranging K-12 educational policy are dubious at best. For example, if one were to raise the average mathematics achievement level of all U.S. high school students, then any characteristic that was once signaled through a certain level of mathematics achievement (e.g. completion of Algebra II) would now require a higher level of achievement (e.g. completion of Calculus) to produce the same competitive advantage. Thus, in this scenario, even highly successful math interventions would fail to produce strong earnings effects.

More research should seek to better understand these mechanisms, as they carry important implications for K-12 educational policy. Although the prevalence of testing in mathematics and reading is not likely to go away, better understanding the link between academic skills and adult outcomes can help us design interventions that have a legitimate chance of making positive long-run changes in students' lives. In my future work, I plan to continue investigating the link between academic skills and adult outcomes by examining the effects of programs that required students to take additional coursework in mathematics and reading. I turn to my plans for future work in the next section.

Future Directions

These three chapters have shed light on new areas of inquiry that I plan to pursue in the next phase of my research career. In particular, I plan to investigate the long-run returns to investments in other areas of early childhood development, like the promotion of socio-emotional skills and executive functioning. I also plan to continue investigating mechanisms of fadeout and persistence in early academic interventions, and I hope to examine the returns to high school coursework in mathematics and reading. All of these future research projects should help address issues raised by the three studies presented here. Below, I briefly describe my plans in each area of inquiry.

First, my postdoctoral work will largely focus on the Chicago School Readiness Project (CSRP), an early childhood socio-emotional intervention that sought to promote self-regulation and executive function skills in a sample of low-income preschoolers living in impoverished neighborhoods in Chicago (see Raver et al., 2011). The program produced large impacts on measures of socio-emotional development and academic achievement measured at the end of preschool, but these effects largely faded by early elementary school. I plan to examine the impacts of the intervention on socio-emotional and academic outcomes measured in early high school. Although it is unlikely that effects could have reemerged, this would allow me to further test Clements and colleagues (under review) "latent foundation hypothesis," as key skills promoted during the intervention could have become more prominent during the challenging transition into high school. Further, students in the sample were re-randomly assigned to a mindset intervention (Yeager & Walton, 2011) that attempted to shift adolescents' perceptions of their own academic ability. As part of my postdoctoral work, I will also examine the effects of

participation in this mindset intervention, which will allow me to investigate another promising approach to promoting academic skills.

I also plan to continue pursuing questions regarding early intervention fadeout and persistence through a collaboration with colleagues Jade Jenkins and Ken Dodge. In this project, I will reexamine the TRIAD data used in study 1 to test whether fadeout effects differ based on the composition of subsequent classroom environments. Because previous research has identified that teachers often teach to the lowest achieving children in the class (e.g., Engel et al., 2016), if academic preschool programs change the composition of classes by raising the skill level of an entire cohort of students, then program impacts might persist if teachers are better positioned to adapt their curriculum accordingly. In other words, if preschool programs raise the achievement level of enough students, will kindergarten and first grade teachers adapt and teach better-aligned content? In this study, I will test whether the percent share of students who received the initial preschool treatment in a kindergarten class moderates preschool impact fadeout. This will give some indication as to how learning trajectories may be affected as preschool programs continue to scale up across the country, and it can also better elucidate the mechanisms behind early impact fadeout.

Finally, I requested data from the state of Florida to test the causal impact of taking an additional mathematics or reading course during high school on postsecondary outcomes. During the early 2000's, Florida tested every eighth-grade student in math and reading. Students who failed either test were required to take an additional class in that same subject during the following year. So, if a student failed the eighth-grade reading test, they were required to take two reading courses in ninth grade. This would allow me to use a quasi-experimental research design to test the causal effect of taking additional academic courses on later student outcomes.

Cortes and Goodman (2014) reported positive effects for a similar program targeted at ninth grade algebra in the Chicago Public School district, though the program was smaller in scope when compared with Florida, and it is unclear if requiring additional reading coursework would provide similar benefits. If given access to the data, I will be able to test whether taking an additional math or reading course affected student skill levels, postsecondary success and adult earnings. This work could provide much needed causal evidence on the long-run returns to programs designed to boost mathematics and reading skills.

Conclusion

These new research directions should help answer questions raised by the three studies presented here, and this future work should also allow me to expand my research program into promising new directions. I plan to continue investigating the role that educational programs can have in shaping children's long-run developmental trajectories. Although the studies presented here elucidate many of the challenges that academic program developers face when trying to design interventions that can meaningfully boost long-run achievement, these studies also further our understanding of how academic skills develop over time. I remain optimistic that research can lead us to developmental models that can inform the design of programs that can meaningfully transform children's lives for the better.

References

Bailey, D., Duncan, G. J., Odgers, C. L., & Yu, W. (2017). Persistence and fadeout in the impacts of child and adolescent interventions. *Journal of Research on Educational Effectiveness, 10*(1), 7-39.

Bailey, D. H., Siegler, R. S., & Geary, D. C. (2014). Early predictors of middle school fraction knowledge. *Developmental Science, 17*(5), 775-785. doi: 10.1111/desc.12155

Bailey, D. H., Watts, T. W., Littlefield, A. K., & Geary, D. C. (2014). State and trait effects on individual differences in children's mathematical development. *Psychological Science, 25*(11), 2017-2026. doi: 10.1177/0956797614547539

Bursal, M., & Paznokas, L. (2006). Mathematics anxiety and preservice elementary teachers' confidence to teach mathematics and science. *School Science and Mathematics, 106*(4), 173-180.

Claessens, A., Engel, M., & Curran, F.C. (2014). Academic content, student learning, and the persistence of preschool effects. *American Educational Research Journal, 51*(2), 403-34.

Clements, D. H., Sarama, J., Layzer, C., Unlu, F., Wolfe, C. B., Fesler, L., . . . Spitler, M. E. (under review). Effects of TRIAD on mathematics achievement: Long-term impacts.

Clements, D. H., Sarama, J., Wolfe, C. B., & Spitler, M. E. (2013). Longitudinal evaluation of a scale-up model for teaching mathematics with trajectories and technologies: Persistence of effects in the third year. *American Educational Research Journal, 50*(4), 812-850.

Connor, C. M., Morrison, F. J., Fishman, B. J., Crowe, E. C., Al Otaiba, S., & Schatschneider, C. (2013). A longitudinal cluster-randomized controlled study on the accumulating effects of individualized literacy instruction on students' reading from first through third grade. *Psychological Science, 24*(8), 1408-1419. doi:10.1177/0956797612472204

Cortes, K. E., & Goodman, J. S. (2014). Ability-tracking, instructional time, and better pedagogy: The effect of Double-Dose Algebra on student achievement. *The American Economic Review, 104,* 400-405. doi:http://dx.doi.org/10.1257/aer.104.5.400

Cunha, F., & Heckman, J. J. (2008). Formulating, identifying and estimating the technology of cognitive and noncognitive skill formation. *Journal of Human Resources, 43*(4), 738-782.

Currie, J. & Thomas, D. (2001). Early test scores, school quality and SES: Longrun effects on wage and employment outcomes. In S. W. Polachek (Ed.), *Worker wellbeing in a changing labor market*. Amsterdam: Elsevier Science.

Duncan, G. J., Dowsett, C. J., Claessens, A., Magnuson, K., Huston, A. C., Klebanov, P., et al. (2007). School readiness and later achievement. *Developmental Psychology, 43,* 1428-1446. doi: 10.1037/0012-1649.43.6.1428

Deming, D. J. (2015). *The growing importance of social skills in the labor market* (No. w21473). National Bureau of Economic Research.

Engel, M., Claessens, A., Watts, T., & Farkas, G. (2016). Mathematics content coverage and student learning in kindergarten. *Educational Researcher, 45*(5), 293-300.

Every Student Succeeds Act, 20 U.S.C. §§ 1111 (2015).

Grogger, J., & Eide, E. (1995). Changes in college skills and the rise in the college wage premium. *Journal of Human Resources, 30*(2), 280-310.

Jenkins, J. M., Watts, T. W., Magnuson, K., Gershoff, E., Clements, D. H., Sarama, J., Wolfe, C. B., & Spitler, M. E. (under review). *Preventing preschool fadeout through instructional intervention in kindergarten and first grade.*

Lerman, R. I. (2013). Are employability skills learned in US youth education and training programs? *IZA Journal of Labor Policy*, *2*(1), 1-20.

Levy, F., & Murnane, R. J. (2005). *The new division of labor: How computers are creating the next job market.* New York, NY: Russell Sage Foundation.

Lipsey, M. W., Farran, D. C., & Hofer, K. G. (2015). A randomized control trial of a statewide voluntary prekindergarten program on children's skills and behaviors through third grade. Nashville, TN: Vanderbilt University, Peabody Research Institute. Retrieved from http://peabody.vanderbilt.edu/research/pri

Murnane, R. J., Willett, J. B., Duhaldeborde, Y., & Tyler, J. H. (2000). How important are the cognitive skills of teenagers in predicting subsequent earnings? *Journal of Policy Analysis and Management*, *19*(4), 547-568.

Raver, C. C., Jones, S. M., Li-Grining, C., Zhai, F., Bub, K., & Pressler, E. (2011). CSRP's impact on low-income preschoolers' preacademic skills: self-regulation as a mediating mechanism. *Child Development*, *82*(1), 362-378.

Reardon, S. F., & Portilla, X. A. (2016). Recent trends in income, racial, and ethnic school readiness gaps at kindergarten entry. *AERA Open*, *2*(3). doi: 10.1177/2332858416657343

Rucker, C. J. & Jackson, K. (2017). Reducing inequality through dynamic complementarity: Evidence from Head Start and public school spending. Berkeley: University of California, Berkeley, Goldman School of Public Policy. Retrieved from http://socrates.berkeley.edu/~ruckerj/RJabstract_LRHeadStartSchoolQuality.pdf

Stipek, D. (2017, March 17). The preschool fade-out effect is not inevitable. *Education Week.* Retrieved from http://www.edweek.org/ew/articles/2017/03/17/the-preschool-fade-out-effect-is-not-inevitable.html?cmp=soc-edit-tw&intc=es

Watts, T. W., Clements, D. H., Sarama, J., Wolfe, C. B., Spitler, M. E., & Bailey, D. H. (2016). Does early mathematics intervention make lower-achieving children learn like higher-

achieving children? *Journal of Research on Educational Effectiveness, 10*(1), 95-115. doi: 10.1080/19345747.2016.1204640

Watts, T. W., Duncan, G. J., Siegler, R. S., & Davis-Kean, P. E. (2014). What's past is prologue: Relations between early mathematics knowledge and high school achievement. *Educational Researcher, 43*, 352-360. doi:10.3102/0013189X14553660

Yeager, D. S., & Walton, G. M. (2011). Social-psychological interventions in education: They're not magic. *Review of Educational Research, 81*(2), 267-301.

Yoshikawa, H., Weiland, C., & Brooks-Gunn, J. (2016). When Does Preschool Matter? *The Future of Children*, *26*(2), 21-35.