Mapping the Mechanisms of Interdisciplinary Learning Transfer from Reading to Math Achievement: Evidence from a Large-Scale Randomized Controlled Trial

Joshua B. Gilbert
Harvard University

James S. Kim
Harvard University

Far transfer---the application of learning across distant domains---remains elusive in intervention research, and even when it is found, its mechanisms remain unclear or unexplored. This study analyzes data from the Model of Reading Engagement (MORE), a sustained content literacy intervention implemented in Grades 1-3 that demonstrated positive treatment effects on both near transfer reading and far transfer math outcomes in a prior study. Here, we extend the original analysis to examine the potential mechanisms of the far transfer effects previously observed on math. Latent mediation analysis shows that approximately 50% of the treatment effect on Grade 4 math is explained by Grade 3 reading, leaving the remainder attributable to other factors. The indirect effects on math are driven by broad standardized reading measures rather than narrower content-specific reading comprehension or background knowledge, suggesting that interventions targeting broad, cross-disciplinary skills may be most effective for supporting far transfer. Results are robust to high levels of unobserved confounding, alternative mediators representing reading engagement and social-emotional learning, and alternative model specifications. We conclude with a discussion of how the appropriate methodological choices for assessing transfer depend on intervention characteristics and substantive research questions.

VERSION: January 2026

Mapping the Mechanisms of Interdisciplinary Learning Transfer from Reading to Math Achievement: Evidence from a Large-Scale Randomized Controlled Trial

Joshua B. Gilbert[1] and James S. Kim[1]

[1]Harvard University Graduate School of Education

December 18, 2025

Abstract

Far transfer—the application of learning across distant domains—remains elusive in intervention research, and even when it is found, its mechanisms remain unclear or unexplored. This study analyzes data from the Model of Reading Engagement (MORE), a sustained content literacy intervention implemented in Grades 1-3 that demonstrated positive treatment effects on both near transfer reading and far transfer math outcomes in a prior study. Here, we extend the original analysis to examine the potential mechanisms of the far transfer effects previously observed on math. Latent mediation analysis shows that approximately 50% of the treatment effect on Grade 4 math is explained by Grade 3 reading, leaving the remainder attributable to other factors. The indirect effects on math are driven by broad standardized reading measures rather than narrower content-specific reading comprehension or background knowledge, suggesting that interventions targeting broad, cross-disciplinary skills may be most effective for supporting far transfer. Results are robust to high levels of unobserved confounding, alternative mediators representing reading engagement and social-emotional learning, and alternative model specifications. We conclude with a discussion of how the appropriate methodological choices for assessing transfer depend on intervention characteristics and substantive research questions.

Keywords: learning transfer, mediation, randomized controlled trial, reading, math

1 Introduction

Nearly a century ago, John Dewey argued that "perhaps the greatest of all pedagogical fallacies is the notion that a person learns only the particular thing he is studying at the time" (Dewey, 1938). Today, educational psychologists would call this concept "learning transfer", or the application of skills and concepts learned in one domain or context to another (Barnett & Ceci, 2002; Gick & Holyoak, 1980, 1983, 1987; Hung, 2013; Perkins & Salomon, 1989; Perkins, Salomon, et al., 1992; Perkins & Salomon, 2012, 2018; Sala & Gobet, 2017; Sala et al., 2019; Salomon & Perkins, 1989).

Examples of learning transfer include concrete, physical skills as well as abstract ideas and concepts. For example, learning how to drive a car is sufficiently similar to driving a truck that little additional instruction is needed to perform well in the new context, an example of near transfer (Salomon & Perkins, 1989). At a more abstract level, the concepts of ratio, proportion, and subdivision can be expressed both in terms of mathematical symbols and musical rhythm (Scripp, 2002; Scripp & Gilbert, 2016). Similarly, the concepts, processes, analogical structures, and schemas shared between disciplines can enable learning in one domain to support learning in a more distantly related domain, that is, far transfer (Gick & Holyoak, 1983, 1987; Holyoak, 1985). While transfer has often been neatly dichotomized in terms of "near" and "far", some scholars have argued that the degree of transfer instead forms a continuum depending on the content and context of what is being transferred (Barnett & Ceci, 2002; J. S. Kim et al., 2023, 2024), offering a richer framework for understanding transfer effects in educational and psychological research.

Despite its theoretical and empirical importance, far transfer—skills and concepts that cross disciplinary boundaries such as from reading to math or math to music—remains an elusive phenomenon in contemporary educational research. As Barnett and Ceci (2002, p. 612) observe, "despite a century's worth of research, arguments surrounding the question of whether far transfer occurs have made little progress toward resolution." While several intervention studies and meta-analyses show evidence of positive far transfer effects, in areas such as computer programming, music, executive function, and education generally (Bigand & Tillmann, 2022; Goldin et al., 2014; Neves et al., 2022; Ritchie & Tucker-Drob, 2018; Scherer et al., 2019; Scripp, 2002; Swaminathan & Schellenberg, 2021), other primary studies and meta-analyses of transfer effects across various domains yield opposing results. For example, Banerjee et al. (2025) show that children's arithmetic skills may fail to transfer *within* disciplines, in their case, between applied and academic mathematics. A meta-analysis of executive function skills shows large and positive effects on near transfer outcomes, but no significant effects on far transfer outcomes (Kassai et al., 2019). Two meta-analyses of music interventions find that when methodological quality is controlled for, there are no significant effects on far transfer cognitive outcomes (Cooper, 2020; Sala & Gobet, 2020). Another meta-analysis similarly observes that "far-transfer effects are small or null" and argues that "the lack of generalization of skills acquired by training is thus an invariant of human cognition" (Sala et al., 2019, p. 1), a finding supported by other meta-analyses indicating limited transferability of general cognitive skills (Melby-Lervåg et al., 2016; Sala & Gobet, 2017; Schwaighofer et al., 2015) and theoretical work on the structure of intelligence (Protzko, 2017).

The mixed results of far transfer meta-analyses suggest that interventions successfully promoting interdisciplinary far transfer are rare and, in some cases, may represent false positives. Therefore, when researchers find statistically robust evidence of far transfer, uncovering the mechanisms that explain how and why it occurs is of clear theoretical and practical importance. Interventions purposefully designed to support far transfer may be more likely to show effects that persist across multiple outcomes over time rather than demonstrating "fade out" after the conclusion of intervention activities (D. Bailey, 2019; D. Bailey et al., 2017; D. H. Bailey et al., 2018; Durkin et al., 2022; Hart et al., 2024) and less likely to result in test score inflation where improvement on test items capturing content taught through direct instruction does not generalize to a broader pool of items measuring the latent trait under investigation (Koretz, 2005, 2008).

Interdisciplinary far transfer from reading to math is of particular interest to researchers because these are the primary subjects most often assessed at large scale and across age groups, associated with school accountability measures, and predictive of long-term success in adolescence and adulthood (Duncan et al., 2007; McCoy & Sabol, 2025; Reyna et al., 2009; Ritchie & Bates, 2013; Watts et al., 2014). Accordingly, a substantial body of descriptive and correlational research has explored the nature of reading-math relationships in educational contexts, with meta-analyses showing strong associations between performance in these two domains even when controlling for other factors (Peng et al., 2020). Similarly, longitudinal studies have demonstrated that shared cognitive processes predict longitudinal growth in both reading and math (Cirino et al., 2018), and that most of the covariation between reading and math stems from time-invariant factors, suggesting a strong and stable relationship between performance in both disciplines and that improvement in one may translate into improvement in the other (D. H. Bailey et al., 2014, 2020; Korpipää et al., 2017; Psyridou et al., 2025; Watts et al., 2014, 2017). At a more fundamental cognitive process level, the ability to decode and comprehend language is of critical importance to solving math problems expressed in words (Abedi & Lord, 2001; Koedinger & McLaughlin, 2010), though domain-specific skills and concepts are most predictive of achievement (Mononen et al., 2025).

Similarly, quasi-experimental studies show how long-term policies and interventions designed to improve reading achievement can yield positive spillover effects on math achievement. For example, Novicoff and Dee (2025) find that teacher professional development based on science of reading principles combined with other supports led to a .14 SD increase in ELA test scores and a .11 increase in math test scores for G3 students in California. Similarly, Nichols-Barrer and Haimson (2013) examine the implementation of Expeditionary Learning curricula in elementary schools and find positive effects on reading test scores (.06 SDs) and null effects on math (-.02 SDs) after the first year of implementation, but positive effects in both subjects after two years of implementation (.11 SDs in reading and .09 SDs in math).

However, the mechanisms underlying these transfer effects remain unclear.

Despite this rich body of descriptive and quasi-experimental work, the extent to which improvements in reading achievement mediate improvements in math achievement remains an open question, as existing mediation studies connecting reading and math have mostly been conducted in observational rather than experimental contexts (Austin et al., 2011; D. H. Bailey et al., 2020; Chow & Ekholm, 2019; King & Purpura, 2021; Slusser et al., 2019; Zhang et al., 2017). However, some causal studies provide evidence that vocabulary is a mediator for treatment effects of literacy interventions on more general reading skills such as reading comprehension, suggesting a potentially important role of vocabulary as a mediator of effects on math achievement as well (Language and Reading Research Consortium et al., 2019; Mosher & Kim, 2025; Mosher et al., 2024).

1.1 The Present Study

The purpose of the present study is to evaluate reading achievement as a potential mediator for far transfer effects on math achievement in a causal inference context. We analyze data testing the efficacy of a sustained content literacy intervention called the Model of Reading Engagement (MORE) through a longitudinal randomized controlled trial (RCT) (J. S. Kim et al., 2024). The original results showed both near transfer effects on reading outcomes and far transfer effects on math outcomes. However, the authors did not examine potential mechanisms explaining the far transfer effects on math. Here, we extend the original analysis to determine the extent to which Grade 3 (G3) reading skills mediate the observed treatment effects on Grade 4 (G4) math skills to probe the potential mechanisms of far transfer.

The design and theory of change of the MORE intervention has been previously described in various publications (Gilbert et al., 2023; J. S. Kim et al., 2023, 2024; Mosher & Kim, 2025). In short, MORE emphasizes the development of schemas to build domain knowledge in science and social studies through the implementation of teacher professional development, lessons, read alouds, provision of books to the home, and a digital app for students to practice reading skills. The implementation of MORE evaluated here occurred in a large urban district in the southeastern United States. Thirty schools were randomly assigned to treatment and control conditions. 2,870 students consented to participate in Grade 1 (G1), and baseline data was collected in G1 winter. The MORE intervention was then implemented in treatment schools from G1 spring to Grade 2 (G2) spring while control schools received business as usual instruction. As a response to the COVID-19 pandemic in 2020, MORE lessons and materials were provided to all Grade 3 (G3) students in both conditions during online schooling (Relyea et al., 2025). The treatment-control contrast was therefore a 3-year "full spiral" (G1, G2, G3) of MORE for treatment students compared to a 1-year "partial spiral" (G3 only) for control students. Immediate outcomes were assessed in the spring of G3 and long-term outcomes were assessed one year following program implementation in spring G4.

The results showed moderate and positive intention-to-treat (ITT) and treatment-on-the-treated (TOT) impacts of MORE on a wide range of academic outcomes, including researcher-developed assessments in vocabulary and reading comprehension as well as state standardized assessments in reading. Perhaps most surprisingly, the results also showed positive far transfer effects on state standardized math tests persisting in G4, 14 months following the end of the intervention. The authors explained the unexpected positive treatment effects on math by explicitly invoking theories of learning transfer, arguing that interdisciplinary far transfer requires sustained support over time, a condition met by the three-year MORE intervention. Conceptually, the authors argued that the key ingredient in the MORE intervention's success in promoting far transfer was the focus on developing multiple schemas (Gilbert et al., 2023; J. S. Kim et al., 2023; Mosher & Kim, 2025), which they defined as "intellectual structures that help novice learners build expertise within a given domain ... by making it easier to acquire, organize, connect, and transfer knowledge" (J. S. Kim et al., 2024, p. 1281). The schema development targeted by MORE is a potential "trifecta skill", or one that is "malleable, fundamental, and would not have developed in the absence of the intervention" and thus a necessary prerequisite for long-term effects (D. Bailey et al., 2017, p. 7) and potential interdisciplinary learning transfer. That is, because MORE explicitly focused on building generalized schema for developing reading skills in the context of science and social studies lessons, these skills were more likely to transfer to other content and contexts, such as word problems on math assessments.

In terms of potential mechanisms underlying the far transfer effects on G4 math, the authors suggested that "literacy-focused activities are likely to be key active ingredients driving any observed cross-domain transfer effects." Furthermore, they noted that the state math tests "consisted mostly of word problems requiring strong reading and language skills such as knowledge of quantitative and spatial language" and, "given the emphasis on conceptual knowledge in the full MORE spiral curriculum, it is plausible that the intervention activities enhanced children's ability to comprehend and develop conceptual understanding rather than factual recall" (p. 1293). Table 1 shows publicly available items from G3 reading and G4 math tests and support the plausibility of the authors' arguments, as the necessity of reading skills for the math test items is apparent, with similar levels of text complexity for both tests as measured by the Flesch-Kincaid Grade Level index. However, while the possibility that G3 reading skills provide the mechanism for far transfer effects on G4 math is both plausible and theoretically grounded, the authors did not explicitly test their hypotheses addressing potential mechanisms, leaving them largely as an open question for future exploration. Therefore, the purpose of this study is to extend the original analysis to explore the potential mechanisms promoting far transfer from G3 reading to G4 math achievement.

The design of the MORE study provides a uniquely rigorous opportunity to address potential causal mechanisms by leveraging an RCT within a longitudinal data collection process, in contrast to observational studies where both mediator and outcome measures are collected concurrently (Fairchild & McDaniel, 2017). That is, with baseline demographic and academic achievement measures collected in G1, an intervention implemented from G1 to G3, immediate post-intervention measures of near transfer reading skills collected in

Test	Item Text	Grade Level
G3 EOG Reading	What is the meaning of *pecked* in paragraph 2?	3.6
G3 EOG Reading	According to the text, what is often difficult for the author?	6.9
G3 MAP Reading	Under which heading does the author include information about white-crowned sparrow babies?	12.6
G3 MAP Reading	How are dad's ideas different from Milan's ideas?	6.7
G4 EOG Math	Each day of the work week, Mr. Harbin uses $\frac{3}{4}$ of a gallon of gas. Which estimate *best* describes the amount of gas Mr. Harbin would use in a five-day work week?	6.2
G4 EOG Math	A cafeteria manager ordered 1,251 cartons of milk on Monday. He also ordered cartons of milk on Thursday. He ordered 879 more cartons on Monday than on Thursday. How many cartons did the manager order on Thursday?	9
G4 MAP Math	Manuel paid 34 cents for gum. He gave the clerk $1. Which picture shows the correct change?	1.8
G4 MAP Math	What is the total number of objects in the array? Enter the answer in the box.	3.7

Table 1: Sample Item Text from Measure of Academic Progress (MAP) and End of Grade (EOG) Reading and Math Tests

G3, and long-term measures of far transfer math collected in G4, the design of this study provides a strong basis for exploring mediation and mechanisms in a causal inference context. Furthermore, the availability of multiple potential mediator variables, including assessments of vocabulary, reading comprehension, background knowledge, and broad reading ability allow us to probe which specific reading skills, if any, may be driving the far transfer effects on math, thus building on prior literature that has examined only single mediator models connecting reading to math (King & Purpura, 2021; Slusser et al., 2019). Similarly, the data also include surveys measuring reading engagement and social-emotional learning (SEL) collected at the end of the intervention, which allow us to test alternative, non-cognitive mechanisms of transfer (Barnett, 2004; Sinha, 2013; VanLehn et al., 2017). While causal interpretation of mediation analyses can be challenging even in RCTs because the values of the mediator are potentially at least partially endogenous (Albert et al., 2016; Imai et al., 2010; Keele et al., 2015), they nevertheless can shed light on to what extent the data are more consistent with reading skills explaining the transfer effects on math, compared to other measured and unmeasured factors.

In short, our results show evidence of partial mediation, as G3 reading explains about 50% of the total effect of MORE on G4 math. Decomposing G3 reading into its constituent subscales shows that the indirect effects from reading to math are mostly captured by broad standardized tests rather than narrower content-specific reading comprehension or background knowledge. Results are robust to high levels of unobserved confounding, G3 reading engagement or social-emotional learning (SEL) as alternative non-cognitive mechanisms of transfer, and alternative model specifications. These results suggest a shared role of reading skills and other, unmeasured mediators, potentially including working memory, non-verbal reasoning, problem solving, or parent-child interactions (Ismail et al., 2020; McCoy & Sabol, 2025; Miller et al., 2024; Van der Linden et al., 1999), laying the groundwork for future mediation studies to collect outcomes measuring student abilities and attitudes beyond reading skills, engagement, and SEL.

2 Methods

2.1 Data and Measures

Our analytic sample is comprised of 2073 students with G4 math outcome data in either the End of Grade (EOG) or Measure of Academic Progress (MAP) assessments. EOG and MAP are standardized tests administered by the state in G3 and G4 in both math and reading and report high internal consistencies ($\alpha \approx .90$). The broad state standardized tests are complemented by researcher-developed G3 reading measures including a 30-item assessment of reading content comprehension ($\alpha = .86$), a 36-item assessment of vocabulary knowledge that includes both words explicitly taught through MORE in G1-G3 as well as conceptually related but untaught words ($\alpha = .90$), and a 9-item assessment of background knowledge specific to the reading comprehension test ($\alpha = .55$). All items are scored dichotomously, 1

= correct, 0 = incorrect.[1] The state test scores are derived from item response theory (IRT) models. For comparability, we use expected a posteriori (EAP) scores derived from 2PL IRT models to construct scores for the researcher-developed measures (Bock & Mislevy, 1982).

The data also includes a rich set of demographic control variables, which we include in all models to improve the precision of the point estimates and to align with the analyses in the original study. Demographic controls include race (black, Hispanic, other race), neighborhood socio-economic status (SES; low, middle, high), and dichotomous indicators for gender, limited English proficiency (LEP) status, and individual education plan (IEP) status. We also include fixed effects for school randomization blocks to account for the stratified randomization. To account for baseline academic ability, we include G1 Measure of Academic Progress (MAP) reading and math scores collected before treatment implementation as control variables in all models.

Students also completed affective surveys in G3 to assess levels of reading engagement and social-emotional learning (SEL). To measure reading engagement, after each of three reading comprehension passages on the researcher-developed assessment, students used Likert scales to rate how much they enjoyed the passage (1 = I didn't like it, 2 = it was OK, 3 = I liked it, 4 = I loved it), how difficult they felt the passage was (1 = too easy, 2 = just right, 3 = too hard), and how they felt as a reader (1 = OK reader, 2 = good reader, 3 = great reader). These three sets of three Likert scale items for enjoyment, difficulty, and self-concept were converted to scores using a Graded Response IRT Model. We use these three scores to operationalize a latent variable representing G3 student reading engagement (multiplying the scores for passage difficulty by -1 so that positive scores on each indicator reflect greater reading engagement). To measure SEL, students completed a proprietary measure called the Panorama Student Survey (PSS) (Hao Ma & Elizabeth Cashiola, 2022; Lattke et al., 2022; Panorama Education, 2015). Students rated their agreement with 9 items representing SEL skills and attitudes such as emotional regulation, growth mindset, self-efficacy, and positive

[1] The replication materials from J. S. Kim et al. (2024) contain the full assessments.

teacher-student relationships. The state does not provide item-level data for the PSS, but instead provides a percentage value on how many items within each subdomain each student endorses, which we use as continuous indicators in our models. We standardize each test and survey score in our sample to mean 0 and standard deviation 1 for ease of interpretation.

While the original study did not explore reading engagement or SEL as outcomes or potential mediators, some research has argued for the role of engagement, SEL, and related non-cognitive skills as either predictors of math achievement or potential mechanisms of transfer (Barnett, 2004; Eddy et al., 2021; Liu et al., 2022; Lyashevsky, 2018; McCoy & Sabol, 2025; Sinha, 2013; VanLehn et al., 2017). While MORE did not target these outcomes directly, it is nonetheless plausible that a sustained content literacy intervention could improve student reading engagement by equipping students with more reading skills that makes reading more enjoyable. It is also possible that SEL outcomes would improve for similar reasons. We therefore explore G3 reading engagement and SEL as potential alternative mediators to determine to what extent non-cognitive factors may play a complementary role to reading skills in learning transfer to math.

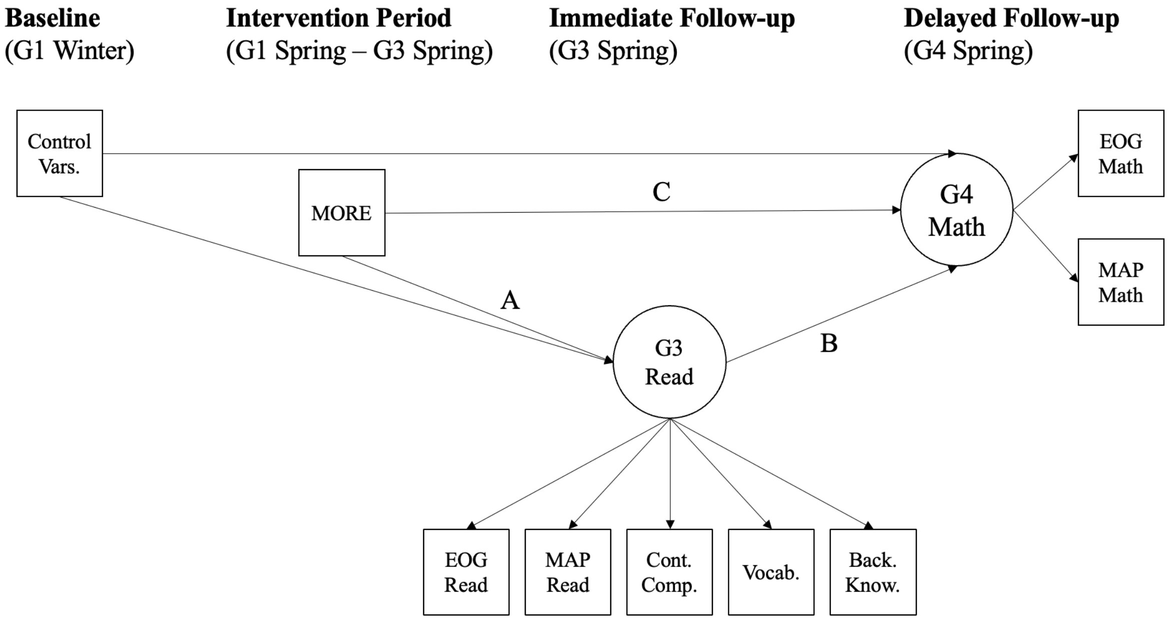
2.2 Statistical Models

To explore the mechanisms of far transfer from reading to math achievement, we use structural equation modeling (SEM), implemented with `lavaan` software in R (Rosseel, 2012). Figure 1 shows a directed acyclic graph (DAG) representing the hypothesized relationships among variables in the primary model. The essential paths are labeled A, B, and C in the diagram. Path A represents the total effect of MORE on G3 reading, operationalized as a latent variable with five observed indicators. Path B represents the potentially causal relationship between G3 reading and G4 math achievement, operationalized as a two-indicator latent variable using G4 EOG and MAP math scores. Path C represents the direct effect of MORE on G4 math, controlling for G3 reading. The total effect of MORE on G4 math is therefore $A \times B + C$ in this linear SEM, where $A \times B$ is the indirect effect of MORE on G4 math mediated by G3 reading. In essence, the original study estimated path A (total effect on G3 reading) and $A \times B + C$ (total effect on G4 math), but not the potential indirect effect $A \times B$ (effect on G4 math mediated by G3 reading).

Because assignment to the MORE intervention was randomized, paths A and C have clear causal interpretations, reflecting the average increase in the outcome caused by assignment to the MORE condition (controlling for any improvements to G3 reading for path C). The interpretation of path B is causal under the assumption that G3 reading achievement and the residuals of G4 math achievement are independent conditional on MORE and the control variables (i.e., G3 reading and G4 math are "sequentially ignorable") (Keele et al., 2015). This assumption would be violated by an unobserved common cause (confounder) of both G3 reading and G4 math, such as general intelligence. However, by including a rich set of demographic control variables including G1 reading and math standardized test scores in the model, we can at least partially control for any confounding, including time-invariant student ability captured by the G1 test scores. We use sensitivity analysis in Section 3.3.2 to determine how robust the results are to unobserved confounders and explore other potential threats to causal identification in Section 4.2.

All latent variables are standardized to mean 0 and SD 1 to facilitate interpretation. We use full information maximum likelihood (FIML) to account for missing data (Enders & Bandalos, 2001) (see Appendix A for additional detail on the patterns of missing data in our sample). Because students were nested in 30 schools and the randomization was carried out at the school level, we apply cluster-robust standard errors at the school level, following the original study.

We follow a sequential model-building strategy to probe the mechanisms of far transfer from G3 reading to G4 math, fitting a taxonomy of five total models. In Models 1 and 2, we replicate the original results on G3 reading and G4 math, respectively, to provide baseline estimates of the total effects of MORE on these two outcomes in the present sample. Model 3 replicates Figure 1 to test the extent to which G3 reading mediates any observed treatment



Notes: Circles indicate latent variables, squares indicate observed variables. EOG = End of Grade. MAP = Measure of Academic Progress.

Figure 1: Directed Acyclic Graph of Primary Mediation Model

effect on G4 math. Model 4 uses each of the five G3 reading indicators as multiple potential mediators for the effect on G4 math to determine which of the constituent elements of reading skills are the more important mediators, providing more insight into the potential cognitive mechanisms underlying far transfer. Finally, Model 5 tests for full mediation by fixing the direct effect of MORE on G4 math (i.e., path C) to 0 and comparing the results to Model 3 to determine whether full or partial mediation is a better fit to the data. We supplement our primary models with a series of robustness tests for alternative mechanisms, sensitivity analyses, and alternative specifications, described in Section 3.3.

3 Results

3.1 Descriptive and Correlational Analyses

Descriptive statistics show that the analytic sample is about 55% treated, 50% male, 75% black or Hispanic, 80% low or moderate SES background, 25% English learners, and 8% students with individual education plans. Table 2 provides a balance test of the baseline test score and demographic variables by treatment condition for the full sample. Following the original study, we conduct the balance tests using school-level averages ($N = 30$ schools) regressed on the treatment indicator and fixed effects for randomization block. There are no significant differences on the demographic variables. There are small but statistically significant differences on the G1 math and reading test score variables, at about 4 points on the scaled test score, in favor of the control group. The student-level SD of the scaled scores is about 17 points, so these differences are about .23 SDs. Given that these baseline imbalances are in favor of the control group, any results favoring the treatment group are likely to be conservative.

Variable	Control	Treatment	Adjusted Mean Difference
G1 Math	170.72 (6.72)	166.64 (7.78)	-4.08 (1.92)*
G1 Reading	169.62 (5.97)	165.58 (7.62)	-4.03 (1.81)*
Prop. Black	0.4 (0.19)	0.41 (0.22)	0.01 (0.08)
Prop. Hispanic	0.3 (0.18)	0.35 (0.2)	0.05 (0.06)
Prop. IEP	0.11 (0.04)	0.08 (0.04)	-0.03 (0.01)
Prop. LEP	0.21 (0.14)	0.25 (0.17)	0.04 (0.05)
Prop. Male	0.52 (0.07)	0.49 (0.04)	-0.02 (0.02)
Prop. Other Race	0.04 (0.03)	0.03 (0.02)	0 (0.01)
Prop. High SES	0.18 (0.32)	0.16 (0.29)	-0.02 (0.1)
Prop. Low SES	0.39 (0.39)	0.52 (0.37)	0.13 (0.1)
Prop. Middle SES	0.42 (0.36)	0.31 (0.27)	-0.11 (0.11)

Notes: Cell entries are school-level means and SDs or regression coefficients and SEs ($N = 30$ schools). Adjusted mean differences are derived from regression models of the school mean outcome on the treatment indicator and fixed effects for randomization block. $^*p < .05$

Table 2: Balance Test on Baseline Variables

Density plots, bivariate scatter plots, and correlations between all test score variables are displayed in Appendix B. We observe large, positive, and statistically significant associations between all reading and math test scores at all time points, though cross-domain (reading to math) and temporally distant correlations are weaker than same-domain, concurrent correlations. The relationships among the variables appear linear, and the univariate distributions reveal no extreme deviations from normality, though the G3 content comprehension test reveals some right skew. These results suggest that the standard residual normality assumptions of SEM are likely to be reasonable in our analyses. We include analogous plots for the three reading engagement and nine SEL variables in Appendix B and find that the indicators are generally positively correlated with one another, though some SEL variables show some left skew.

3.2 Structural Equation Models

Table 3 shows the SEM results, following the model building strategy described earlier. Models 1 and 2 essentially replicate the original findings on G3 reading and G4 math, respectively, though in a latent variable framework disattenuating standardized effect size estimates for measurement error (Gilbert, 2025; Hedges, 1981; Shear & Briggs, 2024). We see positive, statistically significant, and moderate main effects of MORE on both G3 reading ($\beta = .13, p < .01$) and G4 math ($\beta = .14, p < .01$), in line with the original analyses in J. S. Kim et al. (2024). The subsequent models extend the original analysis to probe potential mechanisms of the total MORE effect on G4 math.

Model 3 includes overall G3 reading as a potential mediator of the MORE treatment effect on G4 math, and we see evidence consistent with at least partial mediation. That is, we see a reduction in the direct effect of MORE on G4 math to non-significance and a strong association between G3 reading and G4 math. Holding constant G3 reading (and the covariates), MORE had no significant impact on G4 math ($\beta = .08, p = .07$), and a positive one SD difference in G3 reading predicts a .56 SD difference in G4 math ($p < .001$). The indirect effect from treatment to G4 math through G3 reading is positive and statistically significant ($\beta = .07, p < .01$, cluster bootstrap 95% CI [.02, .12]). Expressed as a proportion, G3 reading mediates about 50% of the total effect of MORE on G4 math. While the same pattern of results would obtain if G3 reading were a confounder (i.e., common cause) rather than a mediator of the relationship between the MORE treatment and G4 math, G3 reading cannot be a confounder because it was measured after treatment implementation.

Model 4 disaggregates the latent G3 reading mediator into its five constituent indicators (i.e., we allow for separate treatment effects and indirect effects through each of the G3 reading indicators, with corrections for measurement error in the observed scores). We see that the overall direct effect is the same, direct effects of MORE are positive and statistically significant for all indicators except G3 MAP, and only the indirect effect for G3 EOG scores is statistically significant, with the largest indirect effect coefficient on the vocabulary scores, though the effect is not statistically significant ($\beta = .05, p > .05$). Interestingly, the strongest path from the mediators to the G4 math outcome is the researcher designed vocabulary measure ($\beta = .30, p < .05$), with a stronger relationship than either of the state standardized MAP or EOG reading scores in G3. These results underscore the potential importance of broad, unconstrained skills in promoting transfer, compared to, for example, the specific background knowledge relevant for particular reading passages (Fitzgerald et al., 2020, 2022; Perfetti, 2007; Wright & Cervetti, 2017). Model 5 allows for full mediation by fixing the path from treatment to G4 math to 0. While the model fit is only slightly affected, a χ^2 test shows the partial mediation of Model 3 is preferred ($\chi^2_1 = 8.79, p < .01$).

Our preferred specification is Model 3, allowing for partial mediation of the MORE treatment effect through a unidimensional G3 reading latent variable. The goodness of fit statistics are adequate (RMSEA = .041, SRMR = .015, TLI = .964, CFI = .975) (Hu & Bentler, 1999) and we see a positive and statistically significant indirect effect from the treatment to G4 math through the G3 reading mediator. Figure 2 shows a graphical depiction of Model 3.

path	Model 1	Model 2	Model 3	Model 4	Model 5
MORE Effect on G3 Lang.	0.13 (0.042)**		0.126 (0.04)**		0.141 (0.042)***
MORE Effect on G4 Math		0.143 (0.05)**	0.078 (0.042)	0.082 (0.037)*	
G4 Math on G3 Lang.			0.56 (0.033)***		0.571 (0.035)***
MORE Indirect Effect through G3 Lang.			0.071 (0.023)**		0.081 (0.025)**
G4 Math on G3 EOG				0.175 (0.069)*	
G4 Math on G3 MAP				0.224 (0.067)***	
G4 Math on G3 Content Comp.				-0.208 (0.151)	
G4 Math on G3 Vocab.				0.304 (0.128)*	
G4 Math on G3 Back. Know.				0.075 (0.072)	
MORE Effect on G3 EOG				0.133 (0.042)**	
MORE Effect on G3 MAP				0.069 (0.036)	
MORE Effect on G3 Content Comp.				0.147 (0.061)*	
MORE Effect on G3 Vocab.				0.16 (0.054)**	
MORE Effect on G3 Back. Know.				0.107 (0.051)*	
MORE Indirect Effect through G3 EOG				0.023 (0.011)*	
MORE Indirect Effect through G3 MAP				0.015 (0.01)	
MORE Indirect Effect through G3 Content Comp.				-0.03 (0.027)	
MORE Indirect Effect through G3 Vocab.				0.049 (0.031)	
MORE Indirect Effect through G3 Back. Know.				0.008 (0.009)	
N	2073	2073	2073	2073	2073
RMSEA	0.035	0.05	0.041	0.056	0.041
SRMR	0.014	0.007	0.015	0.006	0.015
TLI	0.968	0.966	0.964	0.931	0.964
CFI	0.975	0.984	0.975	0.99	0.974

Notes: Cell entries are standardized coefficients and standard errors; p-values are derived from unstandardized models (Gonzalez & Griffin, 2001; Kline, 2023). RMSEA = root mean square error of approximation. SRMR = standardized root mean residual. TLI = Tucker-Lewis index. CFI = confirmatory fit index. $^{*}p < .05,^{**}p < .01,^{***}p < .001$

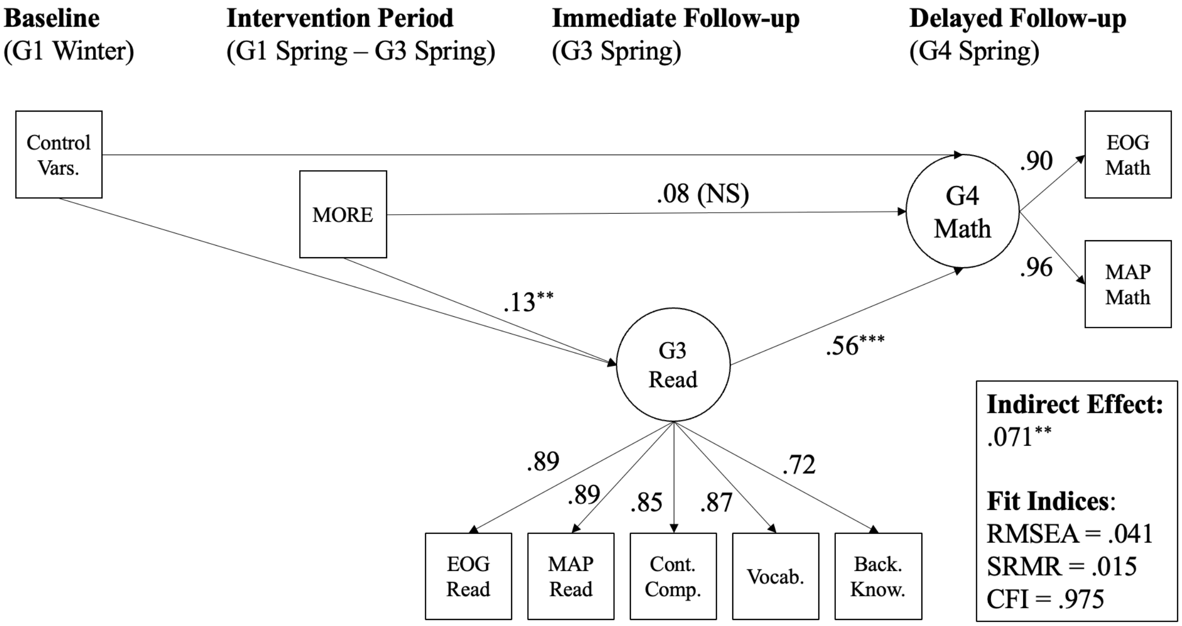
Table 3: Results of Structural Equation Models fit to the MORE Data

3.3 Robustness Checks

We next test the robustness of our preferred specification (Model 3) through a series of tests, including alternative non-cognitive mediators, sensitivity analysis, and alternative model specifications.

3.3.1 Alternative Mechanisms: Reading Engagement and SEL

To probe alternative mechanisms for far transfer to math that do not include G3 reading skill, we extend Model 3 with two additional latent mediators (separately) in a parallel mediation framework (Gervais et al., 2025; Hayes, 2022): G3 reading engagement and SEL. We allow for correlated residuals among the two mediators. The latent reading engagement mediator is represented by the survey scores of reading self-concept, enjoyment, and perceived easiness described in Section 2.1. We find a positive, moderate, and statistically significant treatment effect on latent G3 reading engagement ($\beta = .13, p < .05$). However, conditional on G3



Note: $^{**}p < .01$, $^{***}p < .001$. Factor loadings are labeled without p-values for clarity; all are $p < .001$. RMSEA = root mean square error of approximation. SRMR = standardized root mean residual. CFI = confirmatory fit index.

Figure 2: Results of fitted Structural Equation Model 3

reading and the baseline covariates, there is no significant association between G3 reading engagement and G4 math ($\beta = .001, p > .05$), perhaps an unsurprising result given that some research shows that *math* interest and engagement is not conditionally associated with math achievement (Liu et al., 2022; Zhu & Chiu, 2019). Accordingly, the indirect effect of MORE on G4 math through G3 reading engagement is a precise null ($\beta = .001$, SE = .002). The latent SEL mediator is represented by the nine survey indicators described in Section 2.1. We find no statistically significant treatment effect on SEL ($\beta = .06, p > .05$), a small but significant association between SEL and G4 math ($\beta = .04, p < .05$), and precise null indirect effect from MORE to G4 math ($\beta = .002$, SE = .002). Accordingly, the indirect effect of MORE on G4 math through G3 reading is essentially unchanged in these models from Model 3 ($\beta = .07, p < .01$). We include path diagrams for these alternative mechanism models in Appendix C.

These results suggest that while MORE had a positive impact on G3 reading engagement and that G3 SEL is positively conditionally associated with G4 math—arguably important ends in themselves that were not explored in the original study—improvements in G3 reading engagement or SEL do not appear to function as mechanisms of far transfer to G4 math achievement or to affect the strong positive indirect effect from G3 reading to G4 math observed in our main specification.

3.3.2 Sensitivity Analysis

To what extent can we interpret Model 3 as a fully causal mediation model? The randomized treatment assignment provides unconfounded causal paths from the MORE intervention to G3 reading and the direct effect on G4 math, in expectation. Causal interpretation of the indirect effect requires an unconfounded relationship between G3 reading and G4 math. While the assumption of unconfoundedness is not testable because an unobserved common cause could explain the positive correlation between these two measures (i.e., omitted variable bias), the issue is substantially mitigated by including both G1 reading and G1 math baseline scores as control variables in the model, because any time-invariant student ability in both reading and math would be partialled out of the association between G3 reading and G4 math. That is, the strong association between G3 reading and G4 math is already conditional on baseline achievement in both domains and demographic characteristics. Prior longitudinal research showing the relationship between reading and math ability is stable over time further protects our results from omitted confounders (Keele et al., 2015; Korpipää et al., 2017).

Nonetheless, we probe the robustness of our results by conducting a sensitivity analysis to determine how strong an unobserved confounder must be to attenuate the path from G3 reading to G4 math to 0, or at least non-significance (Blackwell, 2014; Imai et al., 2010; Keele et al., 2015; Tchetgen & Shpitser, 2012). Sensitivity analysis is easily conducted in the SEM context because we can simply add a latent variable as a common cause of both G3 reading and G4 math, fix the paths to each latent variable a range of values, and determine what level of correlation would be needed to fully remove the effect of interest.

Testing a range of correlations between unobserved confounder U and G3 reading and G4 math shows that the path from G3 reading to G4 math becomes non-significant when U is correlated .34 with each variable, implying partial $R^2 = 11.6\%$. How should we interpret this result? First, we can benchmark the critical value of .34 against the coefficients on the standardized G1 pretest scores. For example, in Model 3, both G1 reading and G1 math are positively correlated with G3 reading, with standardized regression coefficients of .51 and .25, respectively. Therefore, U would be equivalent to another broad standardized test similarly correlated with G3 reading. Prior research on the alternative potential confounders provide additional benchmarks for interpreting the results. For example, meta-analytic evidence suggests that executive function may have similar predictive power for academic performance to the levels explored here (Cortés Pascual et al., 2019, $\bar{r} = .37$), though these moderate correlations are themselves potentially inflated by unobserved confounders (Spiegel et al., 2021). In contrast, a variable such as teachers' self-efficacy is unlikely to be strong enough to explain the observed relationship (K. R. Kim & Seo, 2018, $\bar{r} = .07$).

We can also use the residual variances of the G3 reading and G4 math latent variables to determine how much variance there is left for a potential confounder to explain. In Model 3, 29% of the residual variance is unexplained for G3 reading, and 22% is unexplained for G4 math. Therefore, U must explain about 40-50% of the *remaining* variance in these variables ($\frac{11.9}{29} = 40\%$ for reading, $\frac{11.9}{22} = 53\%$ for math), which is quite a high bar in social science research where over 70% of the variation has already been captured by observed variables. In sum, U must be quite strong as a confounding variable to change the results of the analysis, both substantively and in terms of statistical significance. Therefore, the results of the sensitivity analysis provide substantial protection against omitted variable bias.

Assuming for the moment that the path from G3 reading to G4 math is a causal one, how should we interpret the indirect effect of MORE on G4 math? The indirect effect suggests that improving G3 reading ability by .13 SDs would cause G4 math ability to improve by .07 SDs, holding the treatment constant (Pearl & Mackenzie, 2018), providing a measure of the mechanism of far transfer linking the MORE content literacy intervention with long-term math outcomes. That is, treated students were able to successfully apply the reading skills they developed through MORE to a new context of math problem solving.

3.3.3 Alternative Model Specifications

To further probe the stability of our results across alternative model specifications, we fit concurrent mediator models of the G3 and G4 scores separately, each of which provides evidence consistent with our underlying hypothesis that MORE promotes far transfer across disciplines. For example, models of G3 reading mediating treatment effects on G3 math ($\beta = .10, p < .01$), and G4 reading mediating treatment effects on G4 math ($\beta = .06, p < .05$) yield positive indirect effects similar in magnitude to our main specification. Thus, however specified, the data are consistent with reading skills as an underlying mechanism for the far transfer effects observed on math performance.

Last, given the evidence that researcher-developed (RD) measures and independently-developed (ID) measures function somewhat differently in educational RCTs (Gilbert & Soland, 2024; Halpin & Gilbert, 2024; Wolf & Harbatkin, 2023), we fit a parallel mediator model that estimates two latent variables for G3 reading, one loading on the RD measures of vocabulary, content comprehension, and background knowledge, and the other loading on the two state reading tests. We find positive direct effects of MORE on both the G3 reading outcomes, at .15 SDs for the RD latent variable and .11 for the ID latent variable (both $p < .01$), consistent with evidence that RD outcome measures are often more sensitive to treatments than ID outcome measures. The path from G3 ID reading to G4 math is stronger than in our main specification, at .63 SDs ($p < .001$) while the path from G3 RD reading to G4 math is not significant, suggesting that the RD assessments do not provide substantial predictive power for G4 math performance when broad G3 reading achievement measured by state tests is controlled for. Thus, only the indirect effect through the G3 ID reading latent variable is positive and statistically significant ($\beta = .06, p < .05$). These results again underscore the importance of broad G3 reading skills captured by state tests as an important mechanism on the path to G4 math achievement. We include a path diagram for this model in Appendix C.

4 Discussion

Far transfer is an elusive yet intriguing educational and psychological phenomenon. The present study leverages secondary data from a large-scale longitudinal randomized controlled trial of the MORE sustained content literacy intervention to explore the potential mechanisms of interdisciplinary far transfer from G3 reading to G4 math. While the original study showed positive and statistically significant treatment effects of MORE on both near transfer G3 reading and far transfer G4 math, the potential mechanisms of the causal effects on the G4 math outcome remained an open question. This study tested the hypothesis that improvements in G3 reading caused by the MORE treatment that fully or partially mediate the effect on G4 math.

Results of structural equation models provide strong evidence of partial mediation, with about 50% of the treatment effect on G4 math explained by G3 reading, with a positive and statistically significant indirect effect of .07 SDs, suggesting that at least one of the ingredients in interdisciplinary far transfer from reading to math is the reading skills required to understand math problems. Disaggregating the G3 reading mediator into its constituent indicators showed that EOG reading scores and vocabulary were the strongest mediators, in contrast to content-specific background knowledge and reading comprehension, suggesting that broad foundations of unconstrained, generalized reading skills are necessary to support far transfer, rather than constrained or content-specific skills such as background knowledge (D. Bailey et al., 2017; J. Kim et al., 2021). Furthermore, the indirect effect size of .07 SDs is substantial, considering that it represents about 25% of the effect of doubling math class time (Cortes et al., 2015; Taylor, 2014), suggesting interdisciplinary instruction can be a potential lever for improvements in student learning across the curriculum.

What mechanisms could the remaining 50% of the treatment effect on G4 math not explained by G3 reading reflect? While the results of Section 3.3.1 suggest that we can rule out improved reading engagement and SEL as alternative mechanisms, by definition, the remaining direct effect of MORE on G4 math represents all other, unmeasured mechanisms from the treatment to the outcome, so we can only speculate and suggest that future studies collect data on multiple potential mediators beyond standardized test scores, student engagement, and SEL. For example, prior research has suggested that psychological measures of executive function, working memory, motivation, attitudes toward school, general problem solving, procedural flexibility, pattern recognition, parental involvement, social skills, curiosity, creativity, or other skills may explain shared variation in reading and math skills and thus may provide more insight into the mechanisms of far transfer in future research (Burgoyne et al., 2017; Deming, 2017; Farhi et al., 2024; Liu et al., 2022; Mak et al., 2024; McClelland et al., 2007; McCormick & Shira, 2022; McCoy & Sabol, 2025; Mielicki et al., 2024; Nakijoba et al., 2024; Shvartsman & Shaul, 2024; Silla et al., 2024; Whitehead et al., 2024).

4.1 Methodological Considerations when Defining and Measuring Transfer

Our analysis provides new insight into the mechanisms of learning transfer from reading to math in an experimental context, but nevertheless raises some important conceptual questions on the nature of quantitatively assessing far transfer. That is, we used the indirect effect of MORE on G4 math through the mediating variable of G3 reading to empirically capture far transfer, but this is not the only defensible approach to assess far transfer. Even within an SEM context applied to a single data set, careful conceptual work is needed to determine which paths in the model substantively reflect far transfer.

The first question to return to is, what is learning transfer? The research cited in Section 1 provides a range of perspectives, of which we highlight three here. Proceeding with the reading to math transfer example used in the present study, one perspective would suggest that *any* effect of a reading intervention on math outcomes would reflect transfer, because students are demonstrating gains in subject areas not directly trained. Such an approach is consistent with the analysis of the original study that examined outcomes on G4 math as evidence of far transfer. Statistically, this view of transfer is captured by the *total* effect of the intervention in one domain on an outcome in another domain.

A second perspective would suggest that transfer is about the application of specific skills to new contexts. For example, when students develop reading skills through reading instruction, it is only when these skills are applied to solve math problems that transfer has occurred. This conceptualization is more in line with the analytic approach of the present study, which examined *indirect* effects as evidence of far transfer, because, if we can defend a causal interpretation of the mediation model, the indirect effect tells us how an improvement in the mediating or near transfer outcome translates to an improvement in the ultimate or far transfer outcome. In other words, the indirect effect plausibly captures how students were able to apply their reading skills in the new context of math problem-solving.

A third perspective suggests that transfer is defined not simply by new contexts or applications, but much more generalized cognitive skills, for example, the “mindful abstraction” of principles (Perkins & Salomon, 1989). In this case, reading skills employed on the math test, e.g., decoding and comprehension of word problems with complex text, as shown in the sample items Table 1, might be better categorized as relatively near transfer where only the context, but not the content, is changed. That is, the foundation of far transfer would be more general problem solving or cognitive skills because such skills could be applied regardless of content area in multiple domains. Under this view, it is the *direct* effect of the treatment on G4 math, holding constant G3 reading, that could truly represent far transfer, because the direct effect represents all *other* pathways from the treatment to the outcome, such as general problem solving skills or motivation, which could be better candidates for true generalization underlying far transfer in multiple domains.

Which of these divergent but plausible perspectives should we accept? While SEM provides a statistically rigorous framework for evaluating all of these possibilities in a latent variable context, the determination of when a statistical relationship represents far transfer is a normative or substantive question and depends on the research question and research context for meaningful resolution. In the context of this study, we argue that the indirect effect of MORE on G4 math through G3 reading is the most meaningful measure of far transfer, because it reveals the extent to which students were able to leverage their improved reading skills as a result of a sustained content literacy intervention and apply them to a new context of mathematical problem solving, which is ultimately what the designers of the MORE intervention hoped for in supporting long term academic success in multiple disciplines. The indirect effect framework also demonstrates the importance of multiple mediators, for example, we observed that the unconstrained reading skills measured by the standardized G3 EOG reading test and vocabulary knowledge, rather than more narrow, domain-specific background knowledge emerged as the strongest components of the mediator in this analysis. In other contexts, direct or total effects might better represent far transfer, depending on the hypothesized mediators or moderators and the nature of the intervention itself. For example, if in this study our only mediator included measures of working memory, engagement, or motivation in G3, then the direct effect would better represent far transfer because it would by definition reflect unmeasured G3 reading skills. In contrast, in an intervention that focused on a general cognitive ability such as working memory, the total effect on any outcome other than working memory would likely be of most interest.

Beyond the interpretational issues in the interpretation of far transfer in SEM, other families of statistical methods entirely have been used to evaluate far transfer effects in both causal and descriptive contexts. In addition to total effects on far transfer outcomes or mediation analysis that is the methodological basis of this study (Abenavoli, 2019; J. S. Kim et al., 2021; Melby-Lervåg et al., 2016; Mohohlwane et al., 2024; Pages et al., 2023; Watts et al., 2017), prior literature suggests two additional approaches to statistically evaluating far transfer. First, applications of explanatory item response modeling (Wilson & De Boeck, 2004; Wilson et al., 2008) have operationalized transfer as a type of latent treatment heterogeneity driven by individual item characteristics within a single assessment, such as whether a vocabulary word was taught or untaught, or by the number of shared words and concepts contained in a reading passage (Ahmed et al., 2025; Gilbert, 2024; Gilbert et al., 2023, 2024; J. S. Kim et al., 2023; Sales et al., 2021). Second, moderation (i.e., statistical interactions) can also represent transfer, with higher correlations between domains such as music, reading, and math representing mutually beneficial, bidirectional relationships between disciplines (Scripp, 2002; Scripp, 2007; Scripp & Gilbert, 2016). Such approaches are conceptually similar to network psychometric models that estimate direct interactions between indicators rather than postulate latent variables as common causes explaining correlations between indicators (Borsboom & Cramer, 2013; Gilbert, Domingue, & Kim, 2025; Isvoranu et al., 2022). In sum, evaluating transfer demands careful consideration of which method may be most appropriate for the research question and context at hand. Ultimately, what empirical quantity and which statistical model best represents far transfer is a substantive rather than statistical decision, and interpreting transfer as a two-dimensional continuum of content and context, rather than a dichotomy, provides a more nuanced and realistic view of the cognitive issues at play (Barnett & Ceci, 2002).

4.2 Limitations

While the present study is bolstered by the strengths of its design as an RCT in a longitudinal setting with multiple mediators and rich covariates, the results are nonetheless tempered by the following limitations. First, item-level data for the standardized test outcomes is not available, preventing analyses that could reveal whether the treatment had stronger effects on, e.g., word problems or computation items. Such analyses would provide a more fine-grained perspective on how transfer can manifest at the assessment item level (Gilbert, Himmelsbach, et al., 2025; Gilbert et al., 2023). We view item-level analysis as a promising potential extension to the analysis presented here. For example, if our hypotheses about the importance of the G3 reading mediator are correct, we would theoretically observe stronger indirect effects on math word problems compared to computation only problems. Second, we could only examine the two competing mediator variables of student reading engagement and SEL, each of which is based on noisy self-report data, so we can only speculate about what additional effects of the MORE intervention might explain the 50% of the total effect on G4 math not mediated by G3 reading, though the literature cited previously provides some candidate hypotheses. Third, while sensitivity analysis provides some protection against confounding, the estimated path from G3 reading to G4 math is not likely to provide an unbiased causal relationship, complicating the interpretation of the magnitude of the indirect effect of MORE on G4 math. Finally, mediation analyses carried out at the aggregate level do not necessarily capture individual cognitive processes (Bogdan et al., 2023; Borsboom, 2005), and as such, qualitative work such as think-aloud protocols may be necessary to provide insight into the realization of far transfer at the individual level.

It is also worth considering whether alternative mechanisms altogether, other than learning transfer, could explain the pattern of effects observed in our data. For example, the teacher professional development provided through MORE could in principle improve overall teaching ability, which would improve both reading and math outcomes for students, without any reading to math transfer occurring at the student level. While theoretically possible, we view this alternative explanation of our findings as implausible for two reasons. First, the content of the MORE professional development provided to treatment teachers was tightly focused on the logistics and implementation of MORE lesson slides. MORE professional development did not emphasize generalized teaching strategies that would be likely to impact student learning across multiple subjects. Second, qualitative interviews with MORE teachers revealed that only a small subset of teachers adapted MORE strategies to other settings, and none of the examples provided suggested MORE impacted math instruction (Mosher et al., 2024). As such, we believe that learning transfer is the most parsimonious explanation for our results.

However, without data on teacher practices in the classroom, we cannot directly test the hypothesis that general teaching ability was impacted by MORE, suggesting yet another area for future exploration.

4.3 Conclusion

Interdisciplinary far transfer is theoretically important yet empirically rare. Therefore, when researchers find strong evidence of causal far transfer effects in intervention research, we should seek to understand the potential mechanisms by which far transfer operates. This study showed that about half of the treatment effect of a sustained content literacy intervention on G4 math was mediated by improvements in G3 reading and that this effect was driven primarily by broad, unconstrained reading ability rather than content-specific reading comprehension or background knowledge. Results are robust to high levels of unobserved confounding, alternative mechanisms such as reading engagement and SEL, and alternative model specifications. Our results underscore the importance of long-term and sustained interventions targeting broad skills to lay the foundation of transfer, measuring multiple outcomes and mediators, and thoughtful consideration of the appropriate statistical framework to evaluate learning transfer in educational and psychological research.

5 Data and Code Availability

The data, code, and online supplemental materials (OSM) for this study are available at the following URL: [will be added at publication].

Publicly available sample test items are available at the following URLs. G3 EOG Reading: https://www.dpi.nc.gov/documents/accountability/testing/eog/bog3-eog-reading-grade-3-released-form. G4 EOG Math: https://www.dpi.nc.gov/documents/accountability/testing/eog/eog-mathematics-grade-4-re MAP: https://teach.mapnwea.org/impl/maphelp/Content/Testing/PracticeTest.htm

6 Acknowledgments

The authors wish to thank Kenji Kitamura, Ben Domingue, Eric Taylor, Ashutosh Bhuradia, Doug Mosher, Bill Young, and Jin Li Lim for their helpful comments on drafts of this paper.

References

Abedi, J., & Lord, C. (2001). The language factor in mathematics tests. *Applied Measurement in Education*, *14*(3), 219–234. https://doi.org/10.1207/S15324818AME1403_2

Abenavoli, R. M. (2019). The mechanisms and moderators of "fade-out": Towards understanding why the skills of early childhood program participants converge over time with the skills of other children. *Psychological Bulletin*, *145*(12), 1103–1127.

Ahmed, I., Bertling, M., Zhang, L., Ho, A. D., Loyalka, P., Xue, H., Rozelle, S., & Benjamin, W. (2025). Heterogeneity of item-treatment interactions masks complexity and generalizability in randomized controlled trials. *Journal of Research on Educational Effectiveness*, *18*(4), 854–877. https://doi.org/10.1080/19345747.2024.2361337

Albert, J. M., Geng, C., & Nelson, S. (2016). Causal mediation analysis with a latent mediator. *Biometrical Journal*, *58*(3), 535–548.

Austin, A. M. B., Blevins-Knabe, B., Ota, C., Rowe, T., & Lindauer, S. L. K. (2011). Mediators of preschoolers' early mathematics concepts. *Early Child Development and Care*, *181*(9), 1181–1198.

Bailey, D. (2019). Explanations and implications of diminishing intervention impacts across time. In *Cognitive foundations for improving mathematical learning* (pp. 321–346). Elsevier.

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. J., Watts, T., Clements, D. H., & Sarama, J. (2018). Risky business: Correlation and causation in longitudinal studies of skill development. *American Psychologist*, *73*(1), 81–94.

Bailey, D. H., Oh, Y., Farkas, G., Morgan, P., & Hillemeier, M. (2020). Reciprocal effects of reading and mathematics? Beyond the cross-lagged panel model. *Developmental Psychology*, *56*(5), 912–921.

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.

Banerjee, A. V., Bhattacharjee, S., Chattopadhyay, R., Duflo, E., Ganimian, A. J., Rajah, K., & Spelke, E. S. (2025). Children's arithmetic skills do not transfer between applied and academic mathematics. *Nature*, *639*(8055), 673–681. https://doi.org/10.1038/s41586-024-08502-w

Barnett, S. M. (2004). *Teaching for far transfer: How interactive teaching promotes engagement and learning* [Doctoral dissertation, Cornell University].

Barnett, S. M., & Ceci, S. J. (2002). When and where do we apply what we learn?: A taxonomy for far transfer. *Psychological Bulletin*, *128*(4), 612–637. https://doi.org/10.1037/0033-2909.128.4.612

Bigand, E., & Tillmann, B. (2022). Near and far transfer: Is music special? *Memory & Cognition*, *50*(2), 339–347.

Blackwell, M. (2014). A selection bias approach to sensitivity analysis for causal effects. *Political Analysis*, *22*(2), 169–182.

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement*, *6*(4), 431–444. https://doi.org/10.1177/014662168200600405

Bogdan, P., Cervantes, V., & Regenwetter, M. (2023). What does a population-level mediation reveal about individual people? *Behavior Research Methods*. https://doi.org/10.3758/s13428-023-02298-9

Borsboom, D. (2005). *Measuring the mind: Conceptual issues in contemporary psychometrics.* Cambridge University Press. https://doi.org/10.1017/cbo9780511490026

Borsboom, D., & Cramer, A. O. (2013). Network analysis: An integrative approach to the structure of psychopathology. *Annual Review of Clinical Psychology*, *9*(1), 91–121. https://doi.org/10.1146/annurev-clinpsy-050212-185608

Burgoyne, K., Witteveen, K., Tolan, A., Malone, S., & Hulme, C. (2017). Pattern understanding: Relationships with arithmetic and reading development. *Child Development Perspectives*, *11*(4), 239–244.

Chow, J. C., & Ekholm, E. (2019). Language domains differentially predict mathematics performance in young children. *Early Childhood Research Quarterly*, *46*, 179–186.

Cirino, P. T., Child, A. E., & Macdonald, K. T. (2018). Longitudinal predictors of the overlap between reading and math skills. *Contemporary Educational Psychology*, *54*, 99–111.

Cooper, P. K. (2020). It's all in your head: A meta-analysis on the effects of music training on cognitive measures in schoolchildren. *International Journal of Music Education*, *38*(3), 321–336.

Cortes, K. E., Goodman, J. S., & Nomi, T. (2015). Intensive math instruction and educational attainment: Long-run impacts of double-dose algebra. *Journal of Human Resources*, *50*(1), 108–158.

Cortés Pascual, A., Moyano Muñoz, N., & Quílez Robres, A. (2019). The relationship between executive functions and academic performance in primary education: Review and meta-analysis. *Frontiers in Psychology*, *10*, 449759.

Deming, D. J. (2017). The growing importance of social skills in the labor market. *The Quarterly Journal of Economics*, *132*(4), 1593–1640. https://doi.org/10.1093/qje/qjx022

Dewey, J. (1938). *Experience and education*. Macmillan.

Duncan, G. J., Dowsett, C. J., Claessens, A., Magnuson, K., Huston, A. C., Klebanov, P., Pagani, L. S., Feinstein, L., Engel, M., Brooks-Gunn, J., Sexton, H., Duckworth, K., & Japel, C. (2007). School readiness and later achievement. *Developmental Psychology*, *43*(6), 1428–1446. https://doi.org/10.1037/0012-1649.43.6.1428

Durkin, K., Lipsey, M. W., Farran, D. C., & Wiesen, S. E. (2022). Effects of a statewide pre-kindergarten program on children's achievement and behavior through sixth grade. *Developmental Psychology*, *58*(3), 470.

Eddy, M., Blatt-Gross, C., Edgar, S. N., Gohr, A., Halverson, E., Humphreys, K., & Smolin, L. (2021). Local-level implementation of Social Emotional Learning in arts education: Moving the heart through the arts. *Arts Education Policy Review*, *122*(3), 193–204. https://doi.org/10.1080/10632913.2020.1788681

Enders, C., & Bandalos, D. (2001). The relative performance of full information maximum likelihood estimation for missing data in structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, *8*(3), 430–457. https://doi.org/10.1207/S15328007SEM0803_5

Fairchild, A. J., & McDaniel, H. L. (2017). Best (but oft-forgotten) practices: Mediation analysis , *The American Journal of Clinical Nutrition*, *105*(6), 1259–1271. https://doi.org/10.3945/ajcn.117.152546

Farhi, M., Gliksman, Y., & Shalev, L. (2024). Cognitive control among primary-and middle-school students and their associations with math achievement. *Education Sciences*, *14*(2), 159.

Fitzgerald, J., Elmore, J., Relyea, J. E., & Stenner, A. J. (2020). Domain-specific academic vocabulary network development in elementary grades core disciplinary textbooks. *Journal of Educational Psychology*, *112*(5), 855–879. https://doi.org/10.1037/edu0000386

Fitzgerald, J., Relyea, J. E., & Elmore, J. (2022). Academic vocabulary volume in elementary grades disciplinary textbooks. *Journal of Educational Psychology*, *114*(6), 1257–1276. https://doi.org/10.1037/edu0000735

Gervais, J., Lefebvre, G., & Moodie, E. E. M. (2025). Causal mediation analysis with two mediators: A comprehensive guide to estimating total and natural effects across various multiple mediators setups. *Psychological Methods*. https://doi.org/10.1037/met0000781

Gick, M. L., & Holyoak, K. J. (1980). Analogical problem solving. *Cognitive psychology*, *12*(3), 306–355.

Gick, M. L., & Holyoak, K. J. (1983). Schema induction and analogical transfer. *Cognitive psychology*, *15*(1), 1–38.

Gick, M. L., & Holyoak, K. J. (1987). The cognitive basis of knowledge transfer. In *Transfer of learning* (pp. 9–46). Elsevier.

Gilbert, J. B. (2024). Modeling item-level heterogeneous treatment effects: A tutorial with the glmer function from the lme4 package in R. *Behavior Research Methods*, *56*(5), 5055–5067. https://doi.org/10.3758/s13428-023-02245-8

Gilbert, J. B. (2025). How measurement affects causal inference: Attenuation bias is (usually) more important than outcome scoring weights. *Methodology*, *21*(2), 91–122. https://doi.org/10.5964/meth.15773

Gilbert, J. B., Domingue, B. W., & Kim, J. S. (2025). Estimating causal effects on psychological networks using item response theory. *Psychological Methods*. https://doi.org/10.1037/met0000764

Gilbert, J. B., Himmelsbach, Z., Soland, J., Joshi, M., & Domingue, B. W. (2025). Estimating heterogeneous treatment effects with item-level outcome data: Insights from item response theory. *Journal of Policy Analysis and Management*, *44*(4), 1417–1449. https://doi.org/10.1002/pam.70025

Gilbert, J. B., Kim, J. S., & Miratrix, L. W. (2023). Modeling item-level heterogeneous treatment effects with the explanatory item response model: Leveraging large-scale online assessments to pinpoint the impact of educational interventions. *Journal of Educational and Behavioral Statistics*, *48*(6), 889–913. https://doi.org/10.3102/10769986231171710

Gilbert, J. B., Kim, J. S., & Miratrix, L. W. (2024). Leveraging item parameter drift to assess transfer effects in vocabulary learning. *Applied Measurement in Education*, *37*(3), 240–257. https://doi.org/10.1080/08957347.2024.2386934

Gilbert, J. B., & Soland, J. (2024). Mechanisms of effect size differences between researcher developed and independently developed outcomes: A meta-analysis of item-level data. https://doi.org/10.26300/8AXS-Y713

Goldin, A. P., Hermida, M. J., Shalom, D. E., Elias Costa, M., Lopez-Rosenfeld, M., Segretin, M. S., Fernández-Slezak, D., Lipina, S. J., & Sigman, M. (2014). Far transfer to language and math of a short software-based gaming intervention. *Proceedings of the National Academy of Sciences*, *111*(17), 6443–6448.

Gonzalez, R., & Griffin, D. (2001). Testing parameters in structural equation modeling: Every "one" matters. *Psychological Methods*, *6*(3), 258.

Halpin, P., & Gilbert, J. (2024). Testing whether reported treatment effects are unduly dependent on the specific outcome measure used. https://doi.org/10.48550/ARXIV.2409.03502

Hao Ma & Elizabeth Cashiola. (2022). Longitudinal measurement invariance analysis of panorama student survey in a large, urban school district in Texas. *International Journal of Intelligent Technologies and Applied Statistics*, *15*(2&3). https://doi.org/10.6148/IJITAS.202209_15(2_3).0004

Hart, E. R., Bailey, D. H., Luo, S., Sengupta, P., & Watts, T. W. (2024). Fadeout and persistence of intervention impacts on social–emotional and cognitive skills in children and adolescents: A meta-analytic review of randomized controlled trials. *Psychological Bulletin*, *150*(10), 1207–1236. https://doi.org/10.1037/bul0000450

Hayes, A. F. (2022). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach* (3rd ed.). The Guilford Press.

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, *6*(2), 107–128.

Holyoak, K. J. (1985). The pragmatics of analogical transfer. In *Psychology of learning and motivation* (pp. 59–87, Vol. 19). Elsevier.

Hu, L.-t., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Hung, W. (2013). Problem-based learning: A learning environment for enhancing learning transfer. *New Directions for Adult & Continuing Education*, *2013*(137).

Imai, K., Keele, L., & Yamamoto, T. (2010). Identification, inference and sensitivity analysis for causal mediation effects. *Statistical Science*, *25*(1), 51–71.

Ismail, Z., Halias, N., Saad, R. M., & Mohamed, M. F. (2020). Motivation as the mediator in relationship between non-verbal communication of Arabic language teachers and student learning outcomes. *Universal Journal of Educational Research*, *8*(2), 700–708.

Isvoranu, A.-M., Epskamp, S., Waldorp, L. J., & Borsboom, D. (2022). *Network psychometrics with R: A guide for behavioral and social scientists*. Routledge. https://doi.org/10.4324/9781003111238

Kassai, R., Futo, J., Demetrovics, Z., & Takacs, Z. K. (2019). A meta-analysis of the experimental evidence on the near-and far-transfer effects among children's executive function skills. *Psychological Bulletin*, *145*(2), 165.

Keele, L., Tingley, D., & Yamamoto, T. (2015). Identifying mechanisms behind policy interventions via causal mediation analysis. *Journal of Policy Analysis and Management*, *34*(4), 937–963.

Kim, J., Gilbert, J., Yu, Q., & Gale, C. (2021). Measures matter: A meta-analysis of the effects of educational apps on preschool to grade 3 children's literacy and math skills. *AERA Open*, *7*, 23328584211004183.

Kim, J. S., Burkhauser, M. A., Mesite, L. M., Asher, C. A., Relyea, J. E., Fitzgerald, J., & Elmore, J. (2021). Improving reading comprehension, science domain knowledge, and reading engagement through a first-grade content literacy intervention. *Journal of Educational Psychology*, *113*(1), 3–26.

Kim, J. S., Burkhauser, M. A., Relyea, J. E., Gilbert, J. B., Scherer, E., Fitzgerald, J., Mosher, D., & McIntyre, J. (2023). A longitudinal randomized trial of a sustained content literacy intervention from first to second grade: Transfer effects on students' reading comprehension. *Journal of Educational Psychology*, *115*(1), 73–98.

Kim, J. S., Gilbert, J. B., Relyea, J. E., Rich, P., Scherer, E., Burkhauser, M. A., & Tvedt, J. N. (2024). Time to transfer: Long-term effects of a sustained and spiraled content literacy intervention in the elementary grades. *Developmental Psychology*, *60*(7), 1279–1297.

Kim, K. R., & Seo, E. H. (2018). The relationship between teacher efficacy and students' academic achievement: A meta-analysis. *Social Behavior and Personality: An International Journal*, *46*(4), 529–540.

King, Y. A., & Purpura, D. J. (2021). Direct numeracy activities and early math skills: Math language as a mediator. *Early Childhood Research Quarterly*, *54*, 252–259.

Kline, R. B. (2023). *Principles and practice of structural equation modeling* (5th ed.). Guilford Publications.

Koedinger, K., & McLaughlin, E. (2010). Seeing language learning inside the math: Cognitive analysis yields transfer [Issue: 32]. *Proceedings of the annual meeting of the cognitive science society*, *32*.

Koretz, D. (2005). Alignment, high stakes, and the inflation of test scores. *Teachers College Record*, *107*(14), 99–118.

Koretz, D. (2008). *Measuring up*. Harvard University Press.

Korpipää, H., Koponen, T., Aro, M., Tolvanen, A., Aunola, K., Poikkeus, A.-M., Lerkkanen, M.-K., & Nurmi, J.-E. (2017). Covariation between reading and arithmetic skills from Grade 1 to Grade 7. *Contemporary Educational Psychology*, *51*, 131–140.

Language and Reading Research Consortium, Jiang, H., & Logan, J. (2019). Improving reading comprehension in the primary grades: Mediated effects of a language-focused classroom intervention. *Journal of Speech, Language, and Hearing Research*, *62*(8), 2812–2828.

Lattke, L. S., De Lorenzo, A., Settanni, M., & Rabaglietti, E. (2022). PE-Iv (Panorama Education-Italian version): The adaptation/validation of 5 scales, a step towards a SEL approach in Italian schools. *Frontiers in Psychology*, *13*, 1026264. https://doi.org/10.3389/fpsyg.2022.1026264

Liu, A. S., Rutherford, T., & Karamarkovich, S. M. (2022). Numeracy, cognitive, and motivational predictors of elementary mathematics achievement. *Journal of Educational Psychology*, *114*(7), 1589–1607. https://doi.org/10.1037/edu0000772

Lyashevsky, I. (2018). *Teaching to transfer in the social emotional learning context: The case for an instructional model of the human emotion system* [Doctoral Dissertation]. Columbia University.

Mak, C., Tang, J., & Chan, W. W. L. (2024). The importance of home numeracy environment in preschoolers' spontaneous focusing on numerosity: A mediation study. *Early Childhood Research Quarterly*, *68*, 45–53.

McClelland, M. M., Cameron, C. E., Connor, C. M., Farris, C. L., Jewkes, A. M., & Morrison, F. J. (2007). Links between behavioral regulation and preschoolers' literacy, vocabulary, and math skills. *Developmental Psychology*, *43*(4), 947–959.

McCormick, M., & Shira, M. (2022). *Learning more by measuring more: Building better evidence on pre-K programs by assessing the full range of children's skills. Measures for early success* (tech. rep.). MDRC. https://files.eric.ed.gov/fulltext/ED617725.pdf

McCoy, D. C., & Sabol, T. J. (2025). Overcoming the streetlight effect: Shining light on the foundations of learning and development in early childhood. *American Psychologist*, *80*(2), 135–147. https://doi.org/10.1037/amp0001432

Melby-Lervåg, M., Redick, T. S., & Hulme, C. (2016). Working memory training does not improve performance on measures of intelligence or other measures of "far transfer" evidence from a meta-analytic review. *Perspectives on Psychological Science*, *11*(4), 512–534.

Mielicki, M. K., Mbarki, R., & Wang, J. J. (2024). Understanding the social-emotional components of our "number sense": Insights from a novel non-symbolic numerical comparison task. *Frontiers in Psychology*, *15*, 1175591.

Miller, E. B., Canfield, C. F., Roby, E., Wippick, H., Shaw, D. S., Mendelsohn, A. L., & Morris-Perez, P. A. (2024). Enhancing early language and literacy skills for racial/ethnic minority children with low incomes through a randomized clinical trial: The mediating role of cognitively stimulating parent–child interactions. *Child Development*, *95*(4), 1172–1185.

Mohohlwane, N., Taylor, S., Cilliers, J., & Fleisch, B. (2024). Reading skills transfer best from home language to a second language: Policy lessons from two field experiments in South Africa. *Journal of Research on Educational Effectiveness*, *17*(4), 687–710. https://doi.org/10.1080/19345747.2023.2279123

Mononen, R., Korhonen, J., Hægeland, K., Younesi, M., Goebel, S. M., & Niemivirta, M. (2025). Domain-specific and domain-general skills as predictors of arithmetic fluency development: The moderating effect of gender. *Learning and Individual Differences*, *117*.

Mosher, D. M., Burkhauser, M. A., & Kim, J. S. (2024). Improving second-grade reading comprehension through a sustained content literacy intervention: A mixed-methods study examining the mediating role of domain-specific vocabulary. *Journal of Educational Psychology*, *116*(4), 550–568. https://doi.org/10.1037/edu0000868

Mosher, D. M., & Kim, J. S. (2025). Building a science of teaching reading and vocabulary: Experimental effects of structured supplements for a read aloud lesson on third graders' domain-specific reading comprehension. *Scientific Studies of Reading*, *29*(1), 7–31. https://doi.org/10.1080/10888438.2024.2368145

Nakijoba, R., Biirah, J., Akullo, T., & Mugimu, C. B. (2024). Parental involvement and children acquisition of literacy and numeracy skills in Uganda. *Futurity Education*, *4*(1), 53–70.

Neves, L., Correia, A. I., Castro, S. L., Martins, D., & Lima, C. F. (2022). Does music training enhance auditory and linguistic processing? A systematic review and meta-analysis of behavioral and brain evidence. *Neuroscience & Biobehavioral Reviews*, 104777.

Nichols-Barrer, I., & Haimson, J. (2013). *Impacts of five expeditionary learning middle schools on academic achievement* (tech. rep. No. 40207.400). Mathematica. Cambridge MA. https://files.eric.ed.gov/fulltext/ED618299.pdf

Novicoff, S., & Dee, T. S. (2025). The Achievement Effects of Scaling Early Literacy Reforms. *Educational Evaluation and Policy Analysis*, 01623737251349178. https://doi.org/10.3102/01623737251349178

Pages, R., Bailey, D. H., & Duncan, G. J. (2023). The impacts of Abecedarian and Head Start on educational attainment: Reasoning about unobserved mechanisms from temporal patterns of indirect effects. *Early Childhood Research Quarterly*, *65*, 261–274.

Panorama Education. (2015). *Validity brief: Panorama Student Survey* (tech. rep.). https://go.panoramaed.com/hubfs/Panorama_January2019%20/Docs/validity-brief.pdf

Pearl, J., & Mackenzie, D. (2018). *The book of why: The new science of cause and effect*. Basic books.

Peng, P., Lin, X., Ünal, Z. E., Lee, K., Namkung, J., Chow, J., & Sales, A. (2020). Examining the mutual relations between language and mathematics: A meta-analysis. *Psychological Bulletin*, *146*(7), 595–634.

Perfetti, C. (2007). Reading ability: Lexical quality to comprehension. *Scientific Studies of Reading*, *11*(4), 357–383. https://doi.org/10.1080/10888430701530730

Perkins, D. N., & Salomon, G. (1989). Are cognitive skills context-bound? *Educational Researcher*, *18*(1), 16–25.

Perkins, D. N., Salomon, G., et al. (1992). Transfer of learning. *International encyclopedia of education*, *2*, 6452–6457.

Perkins, D. N., & Salomon, G. (2012). Knowledge to Go: A Motivational and Dispositional View of Transfer. *Educational Psychologist*, *47*(3), 248–258. https://doi.org/10.1080/00461520.2012.693354

Perkins, D. N., & Salomon, G. (2018). Transfer and teaching thinking. In *Thinking* (pp. 285–303). Routledge.

Protzko, J. (2017). Effects of cognitive training on the structure of intelligence. *Psychonomic Bulletin & Review*, *24*(4), 1022–1031. https://doi.org/10.3758/s13423-016-1196-1

Psyridou, M., Tolvanen, A., Koponen, T., Aunola, K., Lerkkanen, M.-K., Poikkeus, A.-M., & Torppa, M. (2025). Directional associations in reading and arithmetic fluency development across grades 1 to 9: A random intercept cross-lagged panel model. *Developmental Psychology*, *61*(6), 1196–1209. https://doi.org/10.1037/dev0001944

Relyea, J. E., Gilbert, J. B., Burkhauser, M., Scherer, E., Mosher, D. M., Wei, Z., Tvedt, J., & Kim, J. S. (2025). Asset-based implementation of structured adaptations in an online third-grade content literacy intervention. *Reading Research Quarterly*, *60*(4), e70048. https://doi.org/10.1002/rrq.70048

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*(6), 943–973. https://doi.org/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.

Ritchie, S. J., & Tucker-Drob, E. M. (2018). How much does education improve intelligence? A meta-analysis. *Psychological Science*, *29*(8), 1358–1369.

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*, 1–36.

Sala, G., Aksayli, N. D., Tatlidil, K. S., Tatsumi, T., Gondo, Y., & Gobet, F. (2019). Near and far transfer in cognitive training: A second-order meta-analysis. *Collabra: Psychology*, *5*(1), 18.

Sala, G., & Gobet, F. (2017). Does far transfer exist? Negative evidence from chess, music, and working memory training. *Current Directions in Psychological Science*, *26*(6), 515–520.

Sala, G., & Gobet, F. (2020). Cognitive and academic benefits of music training with children: A multilevel meta-analysis. *Memory & Cognition*, *48*(8), 1429–1441.

Sales, A., Prihar, E., Heffernan, N., & Pane, J. F. (2021). The effect of an intelligent tutor on performance on specific posttest problems. *International Educational Data Mining Society*. https://eric.ed.gov/?id=ED615618

Salomon, G., & Perkins, D. N. (1989). Rocky roads to transfer: Rethinking mechanism of a neglected phenomenon. *Educational Psychologist*, *24*(2), 113–142.

Scherer, R., Siddiq, F., & Sánchez Viveros, B. (2019). The cognitive benefits of learning computer programming: A meta-analysis of transfer effects. *Journal of Educational Psychology*, *111*(5), 764.

Schwaighofer, M., Fischer, F., & Bühner, M. (2015). Does working memory training transfer? A meta-analysis including training conditions as moderators. *Educational Psychologist*, *50*(2), 138–166.

Scripp, L. (2002). An overview of research on music and learning. *Critical links: Learning in the arts and student academic and social development*, 132–136.

Scripp, L. (2007). The Conservatory Lab Charter School-NEC Research Center "learning through music" partnership (1999–2003). *Journal for Music-in-Education*, *2*, 202–223.

Scripp, L., & Gilbert, J. (2016). Music Plus Music Integration: A model for music education policy reform that reflects the evolution and success of arts integration practices in 21st century American public schools. *Arts Education Policy Review*, *117*(4), 186–202. https://doi.org/10.1080/10632913.2016.1211923

Shear, B. R., & Briggs, D. C. (2024). Measurement issues in causal inference. *Asia Pacific Education Review*, *25*, 719–731. https://doi.org/10.1007/s12564-024-09942-9

Shvartsman, M., & Shaul, S. (2024). Working memory profiles and their impact on early literacy and numeracy skills in kindergarten children. *Child & Youth Care Forum*, *53*(5), 1141–1171.

Silla, E. M., Barbieri, C. A., & Newton, K. J. (2024). Procedural flexibility on fraction arithmetic and word problems predicts middle-schoolers' differential algebra skills. *Journal of Educational Psychology*, *116*(2), 195–211.

Sinha, S. (2013). *Exploring student engagement and transfer in technology mediated environments* [Doctoral dissertation, Rutgers University].

Slusser, E., Ribner, A., & Shusterman, A. (2019). Language counts: Early language mediates the relationship between parent education and children's math ability. *Developmental Science*, *22*(3), e12773.

Spiegel, J. A., Goodrich, J. M., Morris, B. M., Osborne, C. M., & Lonigan, C. J. (2021). Relations between executive functions and academic outcomes in elementary school children: A meta-analysis. *Psychological Bulletin*, *147*(4), 329–351.

Swaminathan, S., & Schellenberg, E. G. (2021). Music training. *Cognitive training: An overview of features and applications*, 307–318.

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.

Tchetgen, E. J. T., & Shpitser, I. (2012). Semiparametric theory for causal mediation analysis: Efficiency bounds, multiple robustness, and sensitivity analysis. *Annals of Statistics*, *40*(3), 1816.

Van der Linden, M., Hupet, M., Feyereisen, P., Schelstraete, M.-A., Bestgen, Y., Bruyer, R., Lories, G., El Ahmadi, A., & Seron, X. (1999). Cognitive mediators of age-related differences in language comprehension and verbal memory performance. *Aging, Neuropsychology, and Cognition*, *6*(1), 32–55.

VanLehn, K., Zhang, L., Burleson, W., Girard, S., & Hidago-Pontet, Y. (2017). Can a non-cognitive learning companion increase the effectiveness of a meta-cognitive learning strategy? *IEEE Transactions on Learning Technologies*, *10*(3), 277–289. https://doi.org/10.1109/TLT.2016.2594775

Watts, T. W., Clements, D. H., Sarama, J., Wolfe, C. B., Spitler, M. E., & Bailey, D. H. (2017). Does early mathematics intervention change the processes underlying children's learning? *Journal of Research on Educational Effectiveness*, *10*(1), 96–115.

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.

Whitehead, H. L., Ball, M.-C., Brice, H., Wolf, S., Kembou, S., Ogan, A., & Jasinska, K. K. (2024). Variability in the age of schooling contributes to the link between literacy and numeracy in Côte d'Ivoire. *Child Development*, *95*(2), e93–e109.

Wilson, M., & De Boeck, P. (2004). Descriptive and explanatory item response models. In P. De Boeck & M. Wilson (Eds.), *Explanatory item response models: A generalized linear and nonlinear approach* (pp. 43–74). Springer.

Wilson, M., De Boeck, P., & Carstensen, C. H. (2008). Explanatory item response models: A brief introduction. In J. Hartig, E. Klieme, & D. Leutner (Eds.), *Assessment of competencies in educational contexts* (pp. 91–120). Hogrefe & Huber Publishers.

Wolf, B., & Harbatkin, E. (2023). Making sense of effect sizes: Systematic differences in intervention effect sizes by outcome measure type. *Journal of Research on Educational Effectiveness*, *16*(1), 134–161.

Wright, T. S., & Cervetti, G. N. (2017). A Systematic Review of the Research on Vocabulary Instruction That Impacts Text Comprehension. *Reading Research Quarterly*, *52*(2), 203–226. https://doi.org/10.1002/rrq.163

Zhang, J., Fan, X., Cheung, S. K., Meng, Y., Cai, Z., & Hu, B. Y. (2017). The role of early language abilities on math skills among Chinese children. *PloS one*, *12*(7), e0181074.

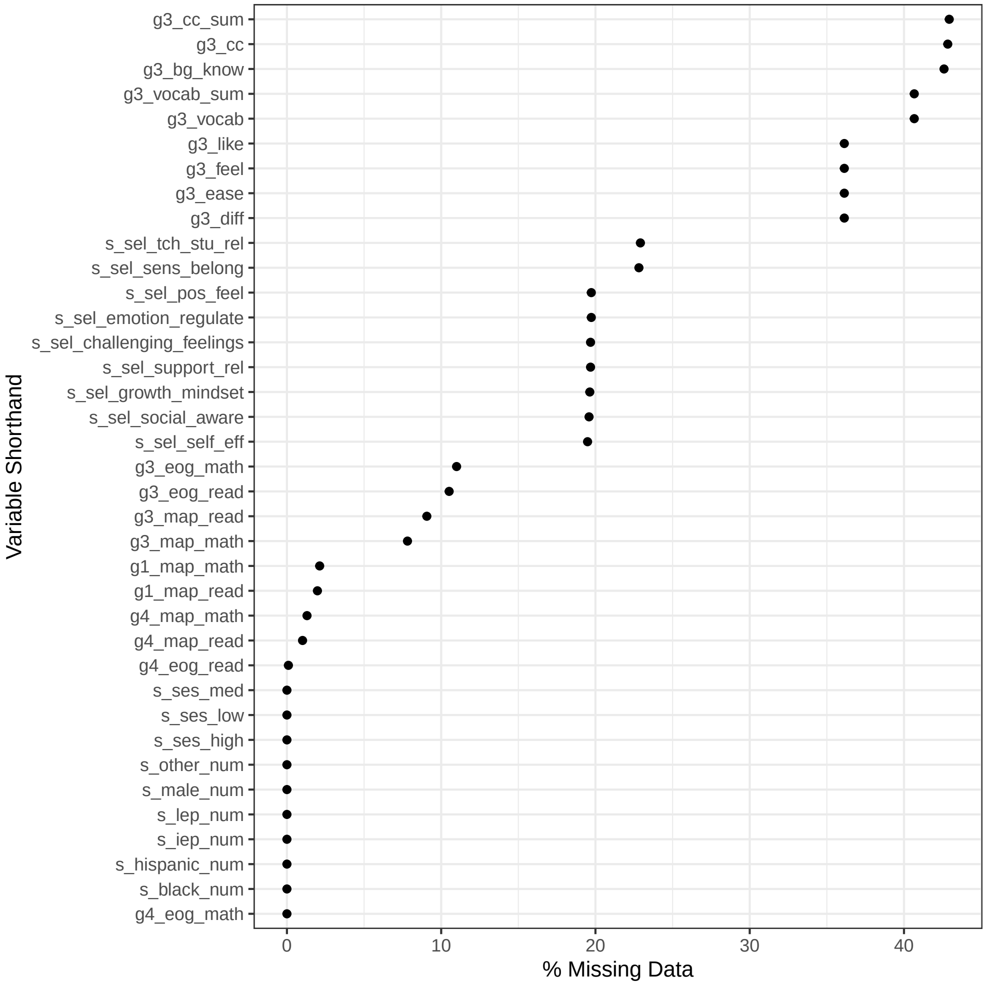
Zhu, J., & Chiu, M. M. (2019). Early home numeracy activities and later mathematics achievement: Early numeracy, interest, and self-efficacy as mediators. *Educational Studies in Mathematics*, *102*(2), 173–191. https://doi.org/10.1007/s10649-019-09906-6

Appendices

A Additional Details on Missing Data in Our Sample

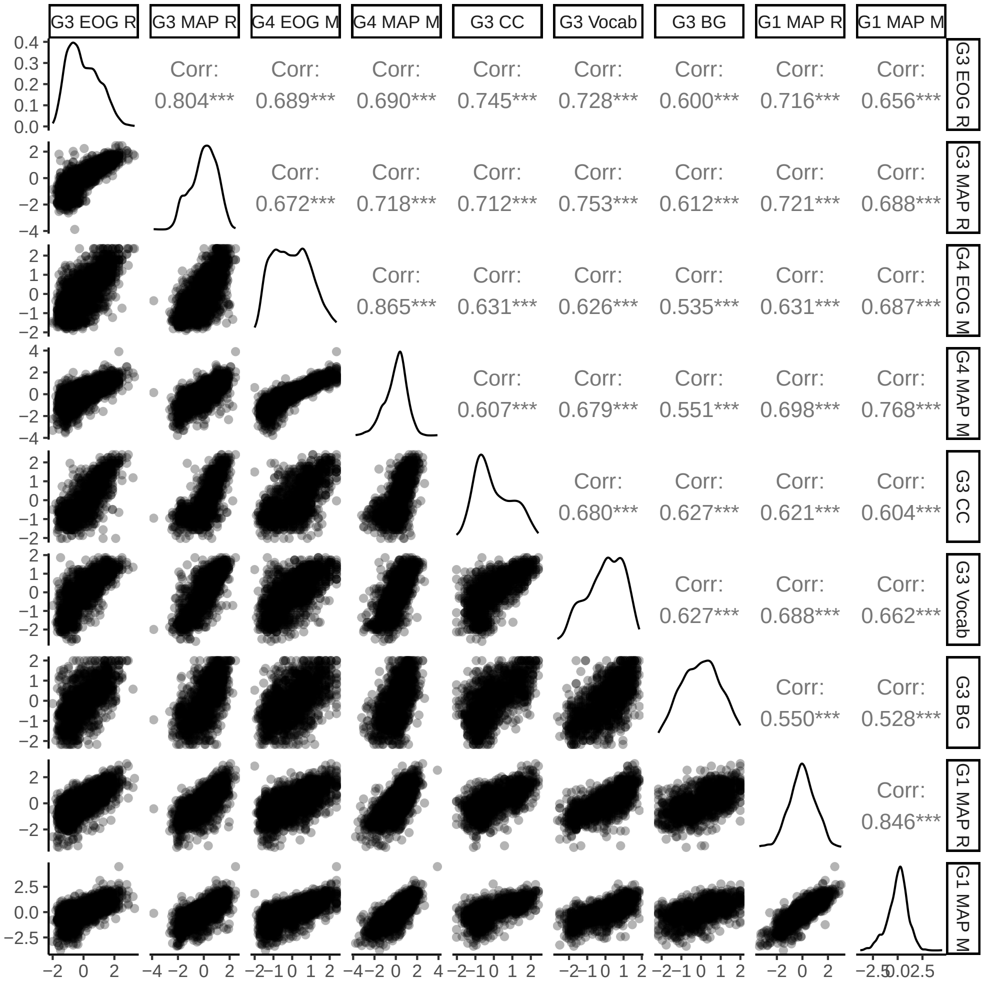
Figure A1 shows the proportion of students that have missing values for each variable. Missing data is highest for the researcher-developed measures administered in G3, due to the COVID-19 pandemic in the 2020-2021 school year (G3 of our sample). Accordingly, in all of our models, we use Full Information Maximum Likelihood (FIML) to account for missing data (Enders & Bandalos, 2001).

B Correlation Matrices for Test Score, Student Engagement, and SEL Variables



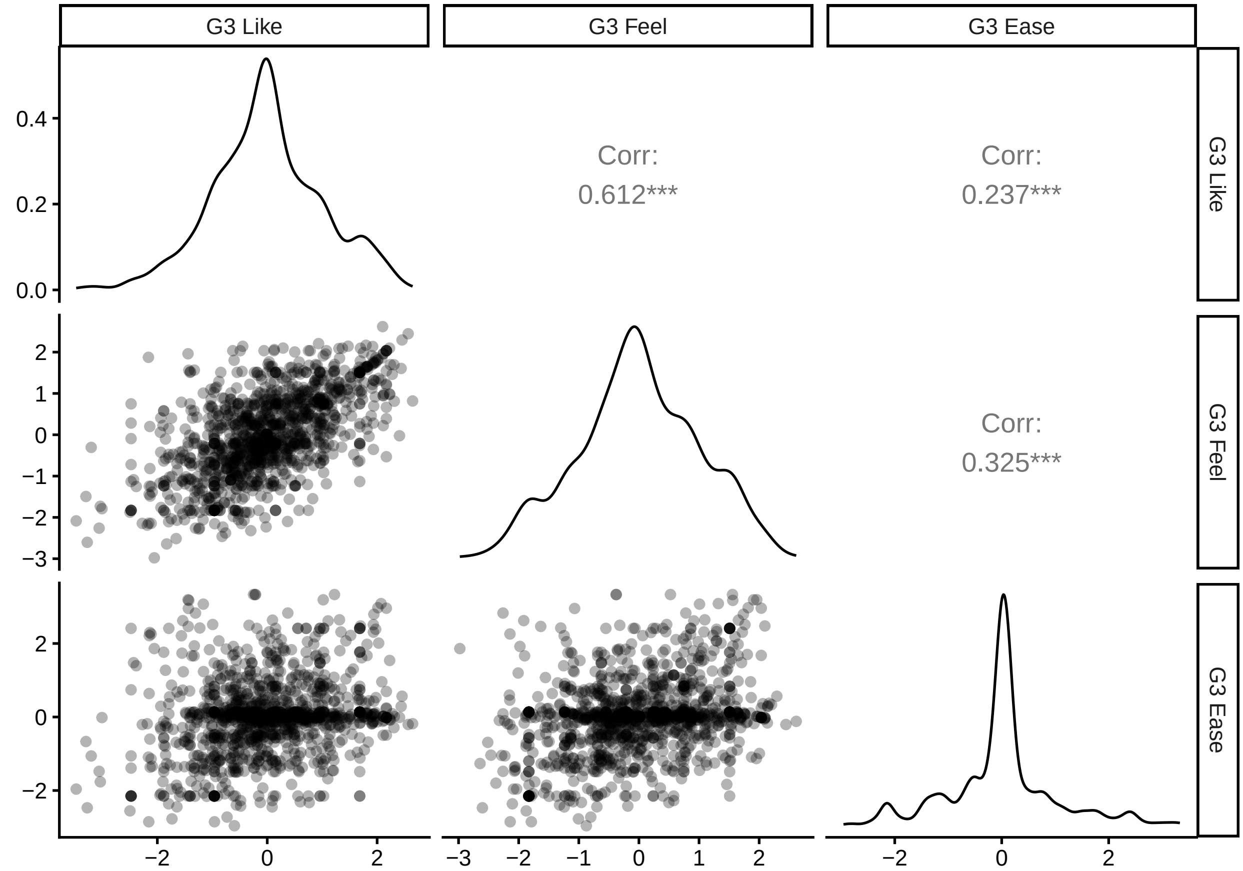
Note: The x-axis shows the percent of students missing data for a particular variable, and the y-axis shows the variable shorthand label. G = grade, s = student.

Figure A1: Missing Data by Variable in Our Sample



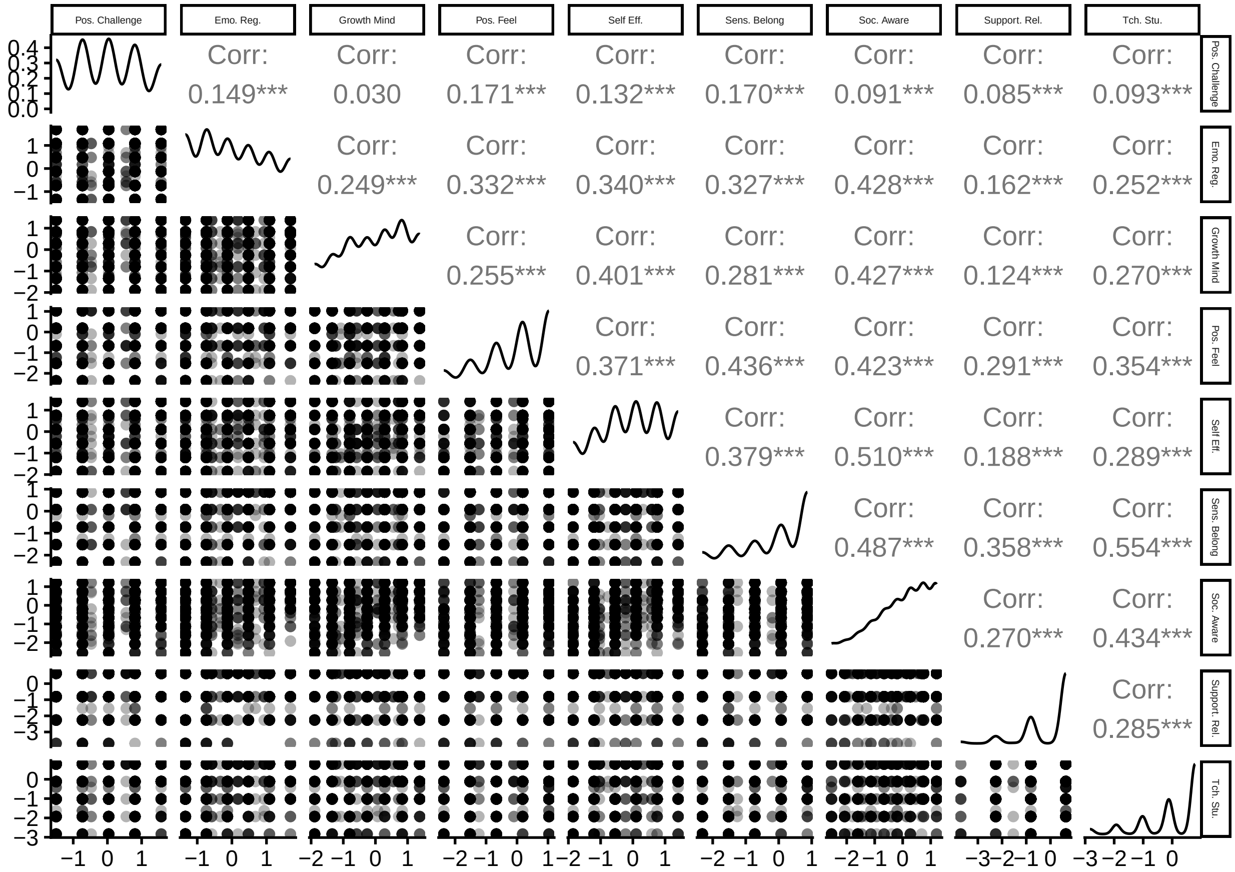
Note: $^{***}p < .001$. MAP = measure of academic progress, EOG = end of grade, CC = content comprehension, BG = background knowledge, M = math, R = reading.

Figure B1: Correlations Among Test Score Variables



Note: ***$p < .001$. G3 = grade 3. All scores are derived from Graded Response Models fit to Likert scale survey data.

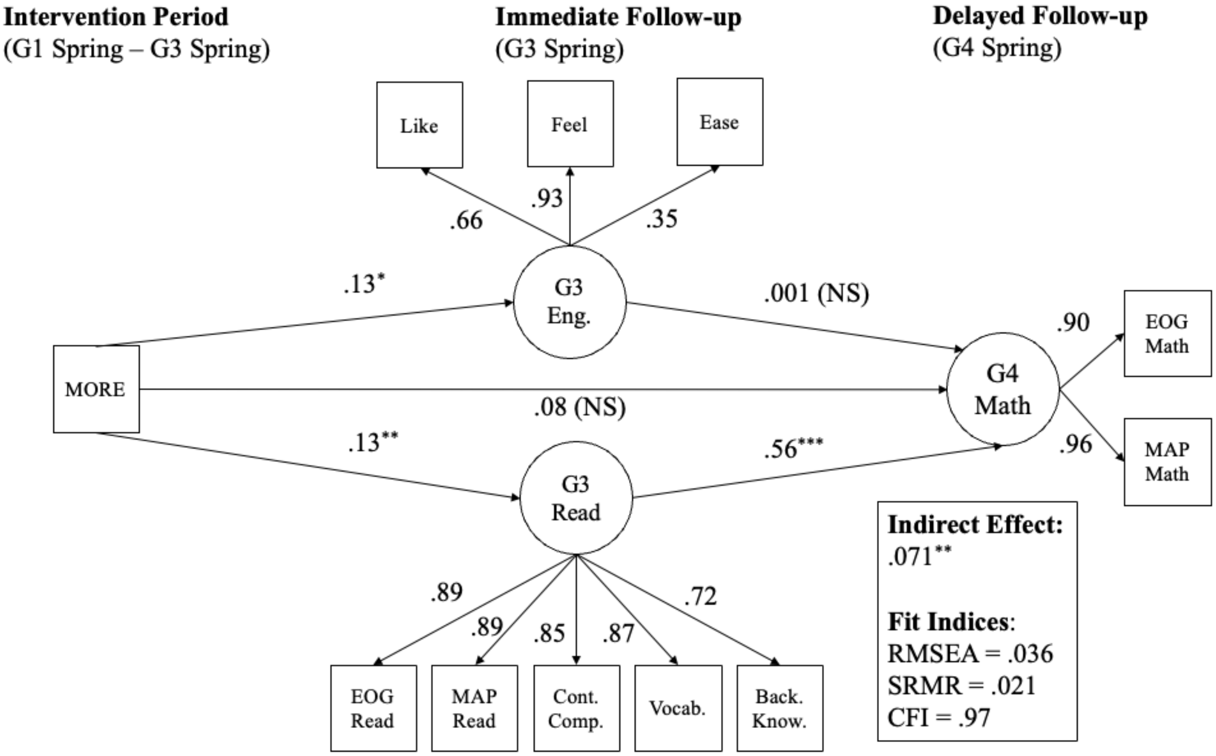
Figure B2: Correlations Among Engagement Variables



Note: ***$p < .001$. Scores represent proportions of positive responses to sets of indicators representing each SEL subdomain (individual item responses not available).

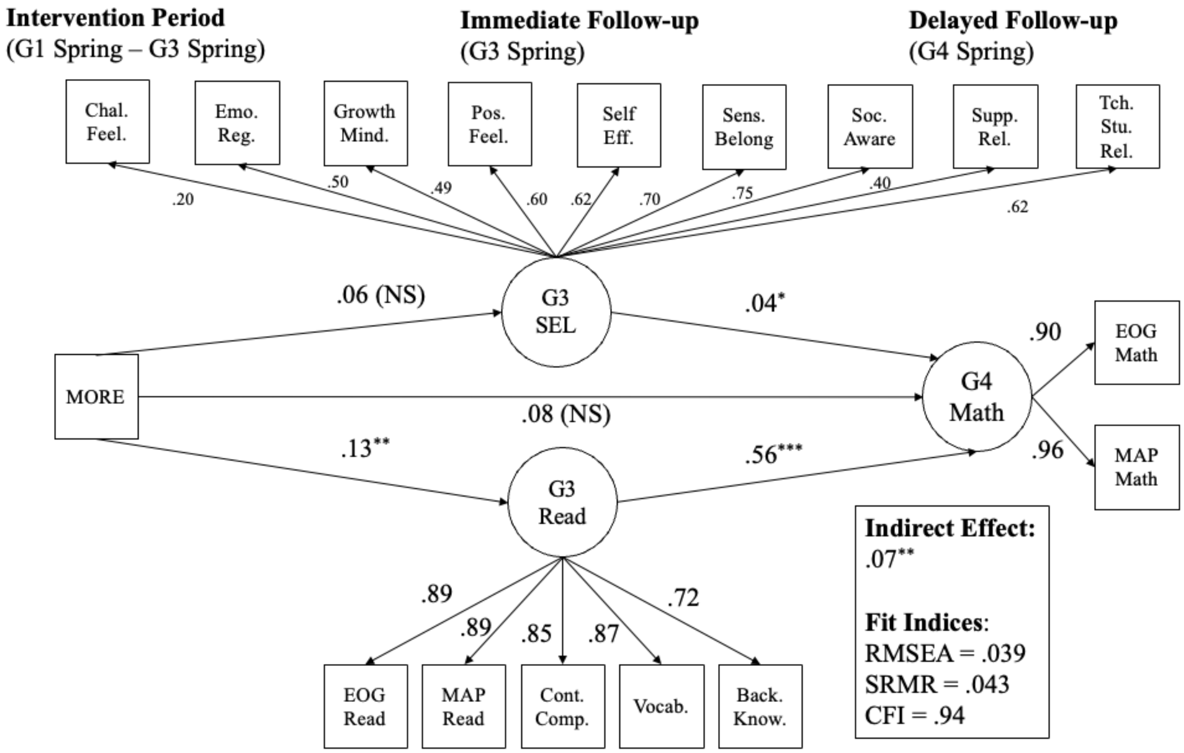
Figure B3: Correlations Among SEL Variables

C Path Diagram for Alternative Mediator Models



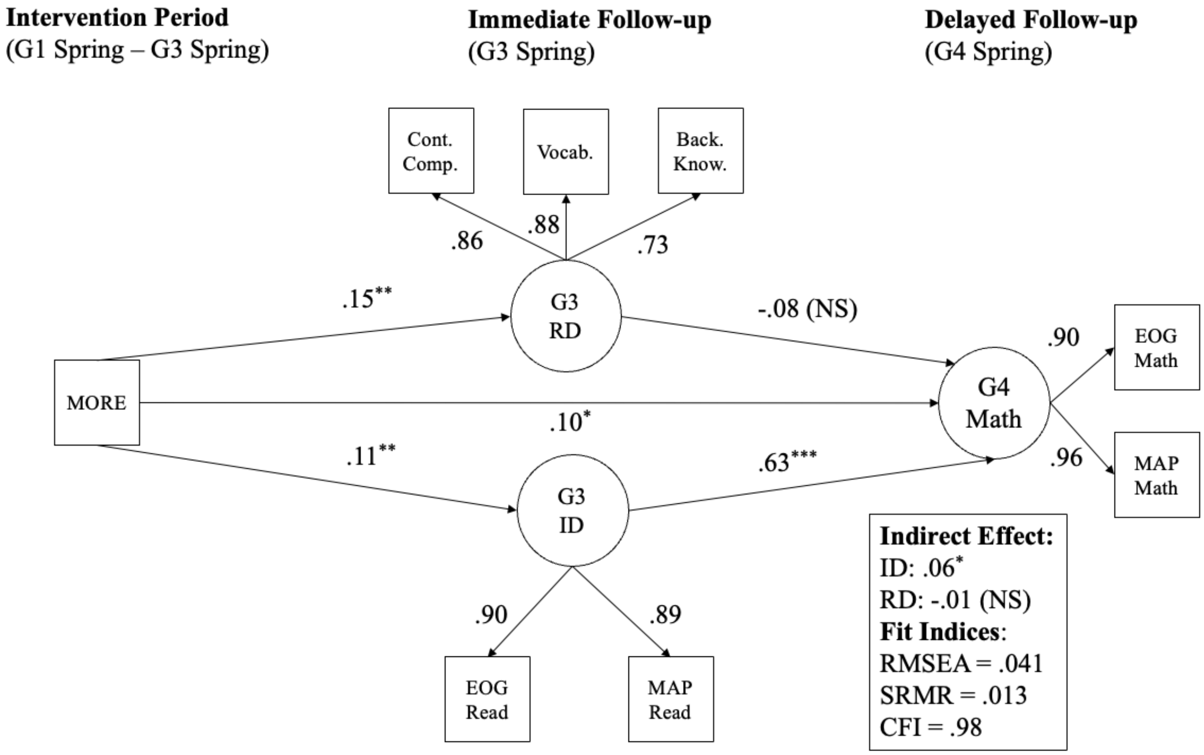
Note: **p* < .05, ***p* < .01, ****p* < .001. Factor loadings are labeled without p-values for clarity; all are $p < .001$. RMSEA = root mean square error of approximation. SRMR = standardized root mean residual. CFI = confirmatory fit index. Like = enjoyment of reading passages, Feel = reading self-concept, Ease = perceived ease of reading passage. G1 control variables are included in the model but omitted from the figure for clarity. The residual correlation between the mediators is $r = .14, p < .01$.

Figure C1: Results of fitted Structural Equation Model Including the Reading Engagement Mediator



Note: $*p < .05$, $**p < .01$, $***p < .001$. Factor loadings are labeled without p-values for clarity; all are $p < .001$. RMSEA = root mean square error of approximation. SRMR = standardized root mean residual. CFI = confirmatory fit index. G1 control variables are included in the model but omitted from the figure for clarity. The residual correlation between the mediators is $r = .20, p < .001$.

Figure C2: Results of fitted Structural Equation Model Including the Social-Emotional Learning Mediator



Indirect Effect:
ID: .06*
RD: -.01 (NS)
Fit Indices:
RMSEA = .041
SRMR = .013
CFI = .98

Note: *p < .05, **p < .01, ***p < .001. Factor loadings are labeled without p-values for clarity; all are $p < .001$. RMSEA = root mean square error of approximation. SRMR = standardized root mean residual. CFI = confirmatory fit index. RD = researcher-developed. ID = independently-developed. G1 control variables are included in the model but omitted from the figure for clarity. The residual correlation between the mediators is $r = .89, p < .001$.

Figure C3: Results of fitted Structural Equation Model Separating the Researcher-Developed and Independently Developed G3 Reading Measures