Title

Indigenous Language Immersion and Native American Student Outcomes: Quantitative Findings from Three Case Studies

Author

Jacobson, Thomas Abram

Publication Date

2024

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Indigenous Language Immersion and Native American Student Outcomes: Quantitative Findings from Three Case Studies

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy
in Education

by

Thomas Abram Jacobson

2024

ABSTRACT OF THE DISSERTATION

Indigenous Language Immersion and Native American Student Outcomes:
Quantitative Findings from Three Case Studies

by

Thomas Abram Jacobson
Doctor of Philosophy in Education
University of California, Los Angeles, 2024
Professor Michael H. Seltzer, Chair

Indigenous-language immersion (ILI) is a form of schooling where all, or nearly all, classroom instruction in every subject area is conducted in an Indigenous language. This dissertation comprises three case study comparisons of neighboring pairs of ILI and English-medium school programs. The first case study examines two elementary schools in the same community. The second case study consists of two independent co-located schools serving elementary and intermediate grades. The third case study compares the ILI and English-medium programs at an intermediate school serving 6th-8th grades. Various academic achievement measures, including English language arts and math standardized assessment scores, are examined to quantify the contrasting associations between ILI versus English-medium instruction and student outcomes, after accounting for observed student background characteristics. On mainstream English-language measures of academic achievement, we find that with few exceptions, ILI students at the case study sites generally scored as high as, or higher than, their Indigenous peers who experienced English-medium instruction. At the same time, when assessed on their Indigenous language proficiency, the ILI students demonstrated consistent maintenance and growth across various Indigenous-language proficiency domains.

The dissertation of Thomas Abram Jacobson is approved.

Mark P. Hansen

Teresa L. McCarty

Noreen M. Webb

Michael H. Seltzer, Committee Chair

University of California, Los Angeles

2024

Dedicated to the students, families, and educators reclaiming and maintaining Indigenous languages.

Contents

List of Tables

List of Figures

Acknowledgments

None of this work would have been possible without the generous, ongoing support of the educators, students, families, and staff from the various Partner Schools who have participated in the Indigenous-Language Immersion and Native American Student Achievement Study (McCarty et al., 2016) over the past seven years. Special thanks to Joy Au, Michael Fillerup, Keʻalapualoke Fukuda, Robert Hagstrom, Leslie Harper, Jessica Honbo, Kauanoe Kamanā, Lisa LaRonge, Keller Paap, Nāmaka Rawlins, Ann Saucier, Jody Wergeland, and William H. (Pila) Wilson for their assistance with accessing and contextualizing the data that form the basis of this dissertation.

Thanks also to the Spencer Foundation (Chicago) for its generosity in funding the ILI Study. I extend my sincerest appreciation to the ILI Study's principal investigators—Teresa L. McCarty at UCLA, Tiffany S. Lee at the University of New Mexico, Sheilah E. Nicholas at the University of Arizona, and Michael Seltzer at UCLA—for all of the hard work and valuable feedback they have contributed in countless discussions since I joined the team. I am also honored to have collaborated with such a remarkably talented group of graduate research assistants on the study, including James McKenzie, Kyle Halle-Erby, Jayashri Srinivasan, Joaquin Noguera, Kari Chew, Winoka Yepa, Nicole Swentzell, and Ruben Leyva.

Thanks to Michael Seltzer and Teresa McCarty (again), as well as Mark Hansen and Noreen Webb, for serving on my dissertation committee and providing indispensable guidance in formulating and refining this work. All four of them have shared an abundance of wisdom and wide-ranging methodological knowledge with countless students at the UCLA School of Education and Information Studies over the years. May we all live up to their example as teachers and mentors as we move into the next phases of our careers.

Finally, heartfelt, affectionate thanks to Ash Marshall for her steadfast support and encouragement, always.

Vita

Education

2023 University of California Los Angeles
MS, Statistics

2015 University of California Los Angeles
MPP, Public Policy

1998 University of Southern California
BA, English
BA, Spanish

Fellowships, Awards, and Academic Honors

2023-2024 University of California Los Angeles
Dissertation Year Fellowship

2020 University of California Los Angeles
Graduate Summer Research Mentorship

2013-2015 University of California Los Angeles
Luskin Fellowship
David & Mariana Fisher Fellowship
Hal E. Martin Scholarship
Departmental Honors for Capstone Applied Policy Project

1994-1998 University of Southern California
BA in English awarded *cum laude* with departmental honors
Resident Honors Program
Thematic Option Honors Program
Dean's Scholar
National Merit Scholar

Publications

Lee, T., McCarty, T., Nicholas, S., Seltzer, M., Jacobson, T., McKenzie, J., & Halle-Erby, K. (2024). Indigenous-language immersion schools—The link to sustainable and healthy Indigenous community futures. *Wicazo Sa Review.* In press.

Phillips, M., Yamashiro, K., & Jacobson, T. (2017). College going in LAUSD: An analysis of college enrollment, persistence, and completion patterns. Los Angeles, CA: Los Angeles Education Research Institute.

Kleiman, M., Caulkins, J., Jacobson, T., & Rowe, B. (2014). Violence and drug control policy. In Donnelly, P. and Ward, C. (Eds.). *Oxford Textbook of Violence Prevention: Epidemiology, Evidence, and Policy.* (Ch. 42, pp. 297-302). Oxford, UK: Oxford University Press.

Conference Papers and Presentations

Lee, T., McKenzie, J., Jacobson, T., & McCarty, T. (2024, June 6). "I really want to create speakers, but I also really want to create people who know who they are:" Relationality and Educational Sovereignty in Indigenous Language Immersion [Sharing circle] Stabilizing Indigenous Languages Symposium, Victoria, BC, Canada

Jacobson, T. (2024, April 14). Accounting for omitted variable bias in hierarchical linear models with group-varying treatment assignment processes. In A. Davidson (Chair), *Experimental design and hierarchal linear models* [Paper session] American Educational Research Association Annual Meeting, Philadelphia, PA, United States.

Seltzer, M. & Jacobson, T. (2024, April 14). How does Indigenous-language immersion compare with English-medium approaches? Findings from a matched-pair analysis. In A. Marin (Chair), *Academic wellbeing and decolonial racial* *justice: Findings from a national study of Indigenous-language schooling* [Symposium] American Educational Research Association Annual Meeting, Philadelphia, PA, United States.

McCarty, T., Lee, T., Nicholas, S., Seltzer, M., Halle-Erby, K., Jacobson, T., & McKenzie, J. (2024, April 14). "We exist:" Dismantling colonial racial injustice—The possibilities and promise of Indigenous-language immersion. In A. Marin (Chair), *Academic wellbeing and decolonial racial justice: Findings from a national study of Indigenous-language schooling* [Symposium] American Educational Research Association Annual Meeting, Philadelphia, PA, United States.

Seltzer, M. & Jacobson, T. (2024, March 17). How does Indigenous-language immersion compare with English-medium approaches? Findings from a matched-pair analysis. In T. McCarty (Chair), *Indigenous-language immersion, language reclamation, and academic equity: Findings from a US-wide study* [Colloquium] American Association of Applied Linguistics Conference, Houston, TX, United States.

Lee, T., McCarty, T., Nicholas, S., Seltzer, M., Halle-Erby, K., Jacobson, T., McKenzie, J. (2024, February 1). *Indigenous-language immersion and Native American student achievement: Initial findings from a national study* [Panel presentation] American Indian Studies Association Annual Conference, Albuquerque, NM, United States.

Lee, T., McCarty, T., Nicholas, S., Seltzer, M., Halle-Erby, K., Jacobson, T., McKenzie, J. (2023, October 19). *Indigenous-language immersion and Native American student achievement: Initial findings from a national study* [Panel presentation] National Indian Education Association Research Forum, Albuquerque, NM, United States.

Lee, T., McCarty, T., Nicholas, S., Seltzer, M., Jacobson, T., McKenzie, J., LaRonge, L., Paap, K., Ryan, I., Sargent, A., & Yasak, K. (2023, June 23). *Indigenous-language immersion and Native American student achievement: Initial findings and promising practices from a national study* [Panel presentation] Stabilizing Indigenous Languages/American Indian Indigenous Teacher Education Conference, Flagstaff, AZ, United States.

McCarty, T., Lee, T., Nicholas, S., Seltzer, M., Halle-Erby, K., Jacobson, T., & McKenzie, J. (2023, February 23). *Indigenous-language immersion and Native American student achievement* [Panel presentation] National Coalition of Native American Language Schools and Programs Summit & Convening, Hilo, HI, United States.

Seltzer, M. & Jacobson, T. (2022, July 21). Preliminary quantitative findings from the matched-pair component. In McCarty, T., Lee, T., Nicholas, S., Seltzer, M. (Chairs), *Indigenous-language immersion and Native American student achievement: Findings from the national study* [Presentation] Indigenous-language immersion and Native American student achievement: A symposium to advance new research and innovative education practice, Santa Monica, CA, United States.

McCarty, T., Lee, T., Nicholas, S., Seltzer, M., Begay, W., Chew, K., Jacobson, T., Noguera, J., & Srinivasan, J. (2020, April 20). The ethnography of a relational methodology in researching Indigenous-language immersion schooling. In A. Marin (Chair), *With language at the heart: Researcher, school, community, and tribal college collaborations in Indigenous education* [Symposium] American Educational Research Association Annual Meeting, San Francisco, CA, United States.

Phillips, M., Yamashiro, K., Miller, C., Jacobson, T., Lim, C., Hayes, K., Kane, J., Orlick, J., Chau, D., & Alexander, C. (2017, May 1). Using research on college outcomes and college supports to improve practice: Lessons learned from the LAERI research-practice partnership. In L. Wentworth (Chair), *Using research-practice partnerships to improve student success in college and career: Three district and state partnerships share challenges and successes* [Symposium] American Educational Research Association Annual Meeting, San Antonio, TX, United States.

Phillips, M., Yamashiro, K., Lim, C., Miller, C., & Jacobson, T. (2017, April 28). Developing the conditions for research use in Los Angeles: The Los Angeles Education Research Institute-University of California Los Angeles-Los Angeles Unified School District partnership. In M. Holsapple (Chair), *Crossing boundaries and increasing impact: Lessons from successful research-practice partnerships* [Structured poster session] American Educational Research Association Annual Meeting, San Antonio, TX, United States.

Phillips, M., Yamashiro, K., Alexander, C., Jacobson, T., & Miller, C. (2017, April 6). Identifying key educational benchmarks and practices that help students meet them: A project of a university-district partnership. In C. C. Raver (Chair), *Strengthening practice and policy through university-agency partnerships* [Paper symposium] Society for Research in Child Development Biennial Meeting, Austin TX, United States.

1 Introduction

The revitalization and maintenance of Indigenous languages is now widely seen as an urgent priority by both Indigenous and non-Indigenous people worldwide (United Nations Department of Economic and Social Affairs, n.d.). Indigenous language immersion (ILI) schooling is one way that Indigenous communities have undertaken this work for themselves, and over the past fifty years, ILI programs have emerged and flourished in many different places across the colonized Western world. The contemporary ILI landscape spans disparate geographic locations, cultures, and linguistic traditions, and embodies a broad range of pedagogies and epistemologies. What ILI programs typically have in common, however, is that they are multi-year, multifaceted educational programs that start in early childhood and continue into students' adolescence, often encompassing opportunities for adult learners as well. Although ILI curricula and pedagogies take various forms in different schools and places, they typically involve conducting all (or very nearly all) instruction and classroom dialogue in the Indigenous language, with a particular focus on language immersion in early elementary grades.

Language recovery and maintenance is a fundamental animating goal of ILI. But there are other common threads among the range of related outcomes that ILI programs seek to foster, including positive self-identity, relational accountability, and holistic academic wellbeing (Lee et al., 2024).

2 Prior Comparative Research on Language Immersion

A robust and well-developed body of comparative research into the effects of language immersion pedagogy has taken shape over the past 50 years. So far, however, most of it has been limited to comparisons between English-only schooling and immersion programs based around Spanish, French, or other dominant colonial languages.

Some of the first studies on the effects of language immersion centered on French immersion programs in Canada. Barik and Swain (1976) examined the longitudinal effects of French immersion on elementary school students in grades K-4 on IQ test scores. They found that French immersion students obtained higher IQ scores on average than their peers in English-medium classrooms, but the difference appeared to reflect initial differences (*i.e.*, selection effects) between the two groups rather than a statistically significant overall effect of French immersion. However, they also noted some evidence of heterogeneous treatment effects within the subgroup of students in the immersion programs. Specifically, when students in the sample were stratified by their French achievement test scores between grades 1-3, the top third of French achievers exhibited statistically significant positive cognitive growth relative to their peers who scored on the bottom third of the French achievement test.

Genesee (1994) highlighted several lessons that various language immersion studies have consistently observed. First, integrating the second language into instruction on other academic subjects is more effective than approaches that isolate the second language from other instructional content. Second, language immersion programs have tended to be most effective in the development of students' comprehension (listening, reading) in the second language but are often less successful at fostering students' productive language abilities (speaking, writing). Instructional modes that cultivate active discourse between students, peers, and teachers are likely to exhibit stronger effects on students' holistic language development than lecture-based or teacher-centered pedagogies. Third, students' language development depends on how the immersion language infuses the overall curriculum both explicitly (during time devoted to language arts instruction) as well as implicitly during instruction in math, history, science, and other academic subjects. The implicit language curriculum presents a crucial opportunity for students to use the language in an authentic and concrete context rather than isolating it as an abstraction.

While a substantial body of academic literature has examined the effects of language immersion broadly defined, fewer studies have centered on comparative analyses of Indigenous-language immersion specifically. The foundational work of Paul Rosier, Merilyn Farella, and Agnes and Wayne Holm at Navajo Nation schools stands out as a notable exception worth emulating in other ILI contexts.

Rosier and Farella (1976) and Rosier and Holm (1980), whose analyses focused on the Rock Point Community School in the center of the Navajo Nation, were among the first to provide comparative studies of academic achievement between Indigenous students in bilingual (Navajo/English) and monolingual (English as a foreign language) instructional settings. In stark contrast to most contemporary Indigenous language immersion programs in the United States, where students typically come from families where English is the primary language, the students in the analyses of Rosier and Farella (1976) and Rosier and Holm (1980) were predominantly monolingual Navajo speakers when they started kindergarten. However, the bilingual program at Rock Point shared some commonalities with present-day ILI programs in that the students were initially taught to read in Navajo during first grade and then taught to read in English in second grade. (Contemporary ILI programs often defer the introduction of English-medium instruction until third grade or later.) Rosier and Farella (1976) found that Navajo students in the Rock Point bilingual program, after initially learning to read in Navajo, exhibited substantially higher growth rates in English reading between 2nd and 5th grade than comparison students in eight other Navajo Area schools where students were taught to read in English only. Rosier and Holm (1980) similarly found that students in the bilingual Navajo-English program at Rock Point Community School scored higher on English-language standardized tests in reading and mathematics than comparison students from nearby schools who received English-only instruction and higher than prior cohorts of students who had received English-only instruction at Rock Point before it implemented the bilingual program.

Agnes Holm and Wayne Holm (Holm & Holm, 1990) recounted the origins and evolution of the two-language program at Rock Point and described the "fourfold empowerment" they observed among the school board, staff, parents, and students over 25 years at the school (pp. 182-184). As evidence of parent empowerment, they noted that participation in biannual parent conferences rose from around 55% when the two-language program first started in the early 1970s to around 85% by the mid-1980s. The students at Rock Point scored higher on standardized tests than peers at other (English-medium) schools on the Navajo Reservation operated by the Bureau of Indian Affairs, and "did so by a greater margin at each successive grade" (p. 184). While parent conference participation rates (unlike students' standardized test scores) are not part of the quantitative data available for the present study, they would be a measure that educators could potentially observe (and likely already are documenting in many places) relatively easily and non-intrusively.

Holm and Holm (1995) reflected further on the history of the Rock Point program in contrast with the development and implementation of the Navajo Immersion Program at Fort Defiance Elementary School from 1986 onward. The early Navajo Immersion program at Fort Defiance shared much in common with contemporary ILI programs elsewhere in terms of its pedagogical approach and goals. Holm and Holm (1995) identified four key lessons from the experiences at Rock Point and Fort Defiance. First, the programs benefited from selecting a small set of priorities deemed to be particularly important (such as using phonics-based Navajo reading instruction as a foundation for later reading instruction in English) and focusing their energy on doing those things as thoroughly and intensely as possible. Second, the Rock Point program benefited from being a "whole-school program" while the Fort Defiance program faced obstacles as a "supplemental program" co-located with a larger English-medium school. Third, both programs were "total," *i.e.*, full-day, all-year programs. A part-day, limited-year approach would greatly attenuate, if not eliminate any beneficial effects (and may actually be detrimental for students), they argued. Fourth, maintaining high expectations of students and staff was crucial: "[I]n many ways it was communicated to students, through actions not just words, that they were expected to succeed and that they would be helped until they did succeed" (p. 158).

Much of what Agnes and Wayne Holm observed at Navajo Nation schools in the 1980s and 1990s has been corroborated elsewhere in more recent studies. May et al. (2004) comprehensively reviewed research on bilingualism and language immersion from an international perspective with a concurrent focus on immersion schooling in the Māori context. May et al. (2004) noted that existing research overwhelmingly finds cognitive, social, and educational advantages of bilingualism when education programs take an *additive*—as opposed to a *subtractive*—approach to bilingualism. They characterized the subtractive approach as one in which students are required to learn one language "at the specific expense" of another, whereas the additive approach treats bilingualism as a resource and a benefit worth cultivating for individuals and the broader society (p. 1). Prime examples of the subtractive approach are the sorts of English-only and English-as-a-second-language programs that are frequently still imposed on English learners across much of the United States and other parts of the English-speaking world today. May et al. (2004) pointedly characterized these as "English-submersion" programs (p. 48) and noted that existing research shows them to be far less effective with regard to literacy and other important academic outcomes than alternative educational programs that adopt an additive approach to bilingualism. Specifically, they note that "Level 1" Indigenous-language immersion programs, which offer 81-100% of instruction in the Indigenous language, are most beneficial and examples of "good practice"—a quality that aligns with most of the Partner School programs in the ILI Study.

Usborne et al. (2011) examined Mi'kmaq students' language proficiency development in Mi'kmaq and English at a community school in Cape Breton, Nova Scotia. One group of students in this study participated in a Mi'kmaq immersion program; the comparison students were enrolled in a "regular stream" Mi'kmaq-as-a-second-language program where English was the primary language of classroom instruction. Usborne et al. (2011) found that the immersion students were far more advanced in their Mi'kmaq proficiency by first grade than the students who were enrolled in the regular program. Meanwhile, both groups of students demonstrated equivalent levels of English proficiency, on average, by first grade. Usborne et al. (2011) also found a strong, statistically significant positive correlation between Mi'kmaq and English test scores among the immersion students. The correlation they observed between Mi'kmaq and English test scores among students in the regular program, on the other hand, was weak and not statistically significant. This finding lends further support to the theory that language learning in the immersion context can be additive and transferable. Specifically, as young students develop proficiency in an Indigenous language, they acquire skills that support and reinforce their development in the culturally predominant mainstream language.

Despite mounting evidence that shows the benefits of bilingualism and a growing worldwide awareness around the urgency of Indigenous-language revitalization, ILI programs must contend with a barrage of obstacles no matter where they are situated. Hermes and Kawaiʻaeʻa (2014) described the development of ILI programs across three distinct contexts—Māori, Hawaiian, and Ojibwe—and noted several key challenges that ILI programs often face. These include tensions between Indigenous and settler colonial epistemologies, as well as government-mandated student assessment and teacher certification requirements that are frequently mis-aligned, if not totally incompatible, with ILI programs' curricula and goals. In spite of such persistent adversity, Hermes and Kawaiʻaeʻa (2014) concluded that "Much remains to be done, but the tenacity and passion of Indigenous educators are strengthening with each new generation of speakers, and there is much hope for the future of our languages" (p. 317).

As they have flourished in the face of challenging circumstances across diverse lands, languages, and societies in recent decades, ILI programs have exemplified myriad ways that homes, schools, and communities can mutually support the work of language revitalization and reclamation. As examined in McCarty (2021), these benefits include "the transformation of persistent academic disparities...; the cultivation of significant numbers of Indigenous-language users of all ages; enhancing cultural knowledge and Indigenous ways of knowing and being; strengthening family and community ties; and support of Indigenous nation building and self-determination" in addition to fostering youth leadership, documenting languages and literatures, innovations in media and technology, as well as vital policy, linguistic, and human rights reforms (p. 11).

Wilson et al. (2022) charted the history of language shift and Indigenous language reclamation across various Indigenous communities in the United States and Canada. Among the current challenges facing Indigenous-language education, they noted various political hurdles at local, state/province, national, and international levels. Building and maintaining a critical mass of families committed to supporting an Indigenous-language immersion program over many years and grade levels is also often difficult but imperative. Other challenges include barriers imposed by government mandates related to assessment and teacher qualifications; building consensus around a unified academic Indigenous orthography and spoken dialect; teacher recruitment and training; and Indigenous-language curriculum development. Wilson et al. (2022) noted that these challenges, while daunting, are not insurmountable. Indeed, a growing body of evidence shows that the ambitious long-term goals of language revitalization are attainable when sufficiently strong ILI education models are implemented and sustained.

This dissertation contributes to this growing body of scholarship by examining academic outcomes among students from a diverse set of well-established ILI programs and carefully-matched peers enrolled at neighboring English-medium schools.

3 Theoretical Framework

3.1 Causality and Effects

This study is primarily informed by two related theoretical approaches to understanding causes, effects, and inference.

The distinctions between (and threats to) statistical conclusion validity, internal validity, construct validity, and external validity (Campbell, 1957; Shadish et al., 2002) are essential considerations as we go about the design, measurement, data collection, and analysis processes of any educational study. These concepts inform our thinking about exactly what information resides within the available data and the range of inferences we can reasonably draw. For example, is the extent of the data adequate to shed light on what we hope to understand? How congruent are our assessments and other measurement instruments with the key constructs or variables of interest? Where quantifiable effects might be apparent in one direction or another, are there alternative causal explanations that might challenge the hypothesized relationship? If so, how plausible are they? Is it reasonable to extend inferences drawn from one site or study into other similar contexts?

The potential outcomes model of causal inference (Holland, 1986; Rubin, 1974, 1986), with its emphasis on identification and quantification of the effects of causes, is especially relevant to the quantitative data analytic components of this work. Its focus attends to estimating the contrast between the effects of alternative causal conditions, including the bias or uncertainty that can be induced when key assumptions are not tenable. In the present context, we are primarily interested in untangling the relative effects of ILI schooling and mainstream English-medium schooling among peer groups of students sharing a common Indigenous heritage. But this framework could also be relevant for comparing alternative approaches to ILI schooling in relation to each other, as well as other contrasts of substantive interest in the realm of Indigenous-language immersion.

3.2 Critical Indigenous Research Methodologies

My intention is to align this work within a Critical Indigenous Research Methodologies (CIRM) framework (Brayboy et al., 2012; Smith, 2021), recognizing that the principles of Indigenous sovereignty and self-determination are essential to its purpose and value. This work is ultimately accountable to the Indigenous students, teachers, and parents who informed it. Consequently, its worth depends on how useful and relevant it is to those communities.

As stated earlier, a primary intention that all ILI programs seem to share is to guide students toward a sense of holistic well-being in their individual identities and in their connection to language, culture, community, and geography. This is not to say that commonplace notions of academic achievement and skills mastery that dominate mainstream thinking around student progress and school accountability do not matter to ILI educators. Rather, what we hear from many ILI educators is that they regard academic achievement and holistic development as overlapping, interconnected goals. Again and again, we have observed ILI educators demonstrate a profound commitment to helping students develop the knowledge, confidence, and resourcefulness that will empower them to thrive in a future where their relationship to community and cultural heritage is simultaneously enriched.

3.3 Culturally Relevant Assessment and Evaluation

This dissertation also draws inspiration from the scholarly tradition around culturally relevant evaluation (Frazier-Anderson et al., 2012; Hood et al., 2015), especially its practice in Indigenous educational contexts (Cram et al., 2015; Paipa et al., 2015). The relevance of this work depends on how well it attends to the interests of a broad range of stakeholders who engage in determining which outcomes matter, what kinds of evidence are credible, how to design and adapt measurement instruments, and the goals and values that should guide evaluators and decision makers. The culturally relevant assessment/evaluation framework augments our thinking around causation and effects with a fifth validity concern, *i.e.*, the concept of *multicultural validity* (Frazier-Anderson et al., 2012), that stresses the relevance of study findings to the participant population itself.

3.4 My Positionality

I approach this work as an outsider and recognize that my relationship with the participant communities is potentially fraught. My cultural and ethnic heritage is Anglo-Scandinavian and indeed, my own ancestors colonized and settled the Northeast and Midwest regions of what is now known as the United States. My methodological training and most of the analytic approaches that I am able to bring to this work have roots in a Western approach to knowledge-seeking that is sometimes framed in juxtaposition, if not direct conflict, with Indigenous epistemologies in the critical methodological discourse. Undoubtedly, much prior academic research on Indigenous people has been useless if not downright harmful to those communities, eliding criticism by asserting the paramount importance of science for science's sake (Brayboy et al., 2012).

With all that in mind, I try to engage in this work with humility about my own epistemological blind spots and a frank recognition that I could easily be susceptible to delusions of benevolence. One key presumption that is implicit in my research questions and may be worth stating clearly in order to invite critique is this: A quantitative methodology that considers various forms of empirical validity and interrogates alternative explanations to observed phenomena is appropriate and useful in this context as long as the work adheres to the principles of respect, relationality, responsibility, reciprocity, and relevance at every stage. My hope is that this work will demonstrate some ways that quantitative tools can be adapted and used to record and transmit knowledge that is mutually beneficial to Indigenous and non-Indigenous people inside and outside of the academy.

A message that I have heard and continue to hear from various ILI educators in different settings is that the type of research agenda outlined here is useful, welcome, and indeed necessary. I have chosen to engage in this work largely in response to strong and persistent reiterations of that sentiment.

4 Motivating Questions

Contemporary discussions around Indigenous languages often paint them as "endangered" or "on the verge of extinction" (see, for example, United Nations Department of Economic and Social Affairs, n.d.). This is a problematic framing in spite of its ostensibly benevolent intentions. As Leonard, 2023 writes, it is a paradigm that often serves, in effect, to obscure and elide the active, ongoing role that settler colonialism plays in language eradication. Leonard adds, crucially, that what he terms "dominant endangered languages narratives" typically fail to hold themselves accountable to actual Indigenous communities.

ILI education not only represents a promising and hopeful response to what is broadly recognized as an urgent need to recover and sustain Indigenous languages after centuries of colonial injustices and many generations of language shift. Perhaps more importantly, however, Indigenous language immersion represents a reassertion of Indigenous sovereignty, an opportunity for Indigenous communities to re-center their own interests in how they conceive what education is, how they enact it, and what outcomes and results should matter (Lee et al., 2024; McCarty et al., 2021).

ILI represents an opportunity to prepare students to thrive multilingually in the modern world and maintain connections to their Indigenous cultural heritage and community while nurturing their holistic wellbeing. These are not goals that exist in inherent tension with each other like variables that must be balanced in some zero-sum equation. Instead, as many of the ILI Study's Partner Sites have demonstrated, they complement and mutually reinforce each other.

Nevertheless, many stakeholders and policymakers, Indigenous and non-Indigenous alike, express persistent worries that ILI's potential benefits in terms of language reclamation and maintenance must inevitably come at some cost. One such concern, frequently voiced by parents and local leaders, is that ILI takes away from classroom time that would otherwise be devoted to English-language instruction in reading, math, science, and other subjects. Those who share this concern worry that students' achievement in these core areas, at least as measured by mainstream standardized assessments, might be inhibited. As a consequence, the story goes, ILI students run a risk of reaching adolescence and adulthood less well-positioned to thrive in the modern world than if they had adhered to the English-medium status quo in their schooling.

In response to these oft-stated suppositions about the potential downsides of ILI, this dissertation will consider the following research questions:

- Among a select group of well-established ILI programs in the United States, what can available administrative and assessment data tell us about the associations between ILI schooling—relative to local English-medium alternatives—and academic outcomes including Indigenous-language proficiency, English language arts development, and math achievement among elementary and intermediate Indigenous students?

- Are there discernible differences in these associations between different student subgroups?
- What can we reliably infer from the available data and with what degree of certainty?

This dissertation draws upon data from the Indigenous Language Immersion and Native American Student Achievement Study (McCarty et al., 2016).[1] Funded by a major grant from the Spencer Foundation in 2016 (Grant Award 201700054), the ILI study's work has been ongoing for over seven years as of this writing, and involves ILI schools spanning a range of cultures, languages, and geographies—from Hawai'i to the Mohawk lands northeast of the Great Lakes, and places in between. A total of eight ILI Partner Schools participated in the study as case study sites. These schools were selected because they had been in operation for at least 10 years as of 2016, used the local Indigenous language for 50-100% of instruction, and volunteered to participate. The quantitative case studies in this dissertation are derived from work with three of the eight Partner Schools in this larger study.

The ILI study was conceived as a three-part, multiple methods research design with the following goals: first, documenting the current state of Indigenous-language education in the United States; second, developing a thorough qualitative understanding of how ILI is implemented in diverse regional, linguistic, and cultural contexts; and third, to compare carefully matched groups of ILI and non-ILI students on outcomes and opportunities to learn. The study's overarching goals are to understand the similarities and differences in how ILI is practiced in diverse contexts, and how, when, for whom, and why ILI is beneficial.

The data analyzed in the following chapters come from the corpus of student-level administrative data shared with the ILI study by three of the Partner Schools as part of the third component of the study. These data cover a range of student demographic characteristics and academic achievement measures related to Indigenous language development, English language arts, and math achievement.

5 Analytic Framework for Quantifying the Association Between Indigenous-Language Immersion Programs and Standardized Measures of Student Achievement

Students typically take standardized assessments at various time points as they progress through primary and secondary grades. In the United States, these usually consist of, at minimum, federally-mandated end-of-year summative assessments in grades 3-8 and at some point in high school (Every Student Succeeds Act, 2015). In addition, schools often administer shorter formative assessments over the course a school year to gauge students' progress from one semester to the next.

[1] Principal Investigators on the Spencer Award were Teresa L. McCarty and Michael Seltzer (University of California, Los Angeles), Tiffany S. Lee (Diné/Lakota, University of New Mexico), and Sheilah E. Nicholas (Hopi, University of Arizona).

5.1 Longitudinal Ordinary Least Squares Regression Approach

We might consider a few different data analytic techniques for modeling the association between continuous repeatedly-measured outcomes of interest (such as standardized test scores across some time span) and persistence in a multi-year program such as Indigenous-language immersion. One approach would be to regress the result of each assessment instance on students' program status and some additional set of p observed covariates. In other words, we would estimate separate ordinary least squares (OLS) regressions at each time point:

$$y_{it} = \beta_{0_t} + \beta_{1_t} x_{1_{it}} + \beta_{2_t} x_{2_{it}} + \cdots + \beta_{p_t} x_{p_{it}} + e_{it}, \tag{1a}$$

where we assume

$$e_{it} \overset{\text{iid}}{\sim} \mathcal{N}\left(0, \sigma^2\right). \tag{1b}$$

In this way, where $x_{1_{it}}$ denotes the program status of student i at time t, we would obtain unique estimates of the coefficient for program status ($\hat{\beta}_{1_t}$) at each time point, holding constant the additional covariates included in the model. An advantage of this approach is that it allows us to parse out potential time-varying effects of the program. We can use the point estimates and standard errors to compute confidence intervals for $\hat{\beta}_{1_t}$ at each of the t time points to observe whether and how the estimates may change over time.

5.2 Hierarchical Linear Mixed Effects Model Approach

It may be the case that all of the student-level independent variables in the data are *time-invariant*, however. In this situation we might be interested in a model that accounts for the data's nested two-level structure (where repeated measurements are nested within students).

In that situation we could use the following mixed-effects model that allows for a random student-level intercept and accounts for p student-level covariates. The covariates might include an initial test score, which would stand as a measure of each student's baseline achievement level, as well as other observed time-invariant student characteristics (such as gender, socioeconomic status, etc.) that might plausibly be associated with the outcome and with students' program enrollment choice. Such a model might take the following form:

Level 1:

$$y_{it} = \beta_{0_i} + \beta_{1_i} + \beta_{2_i} + \cdots + \beta_{p_i} + e_{it}. \tag{2a}$$

Level 2:

$$\begin{aligned}
\beta_{0_i} &= \gamma_0 + r_i \\
\beta_{1_i} &= \gamma_1 x_{1_i} \\
\beta_{2_i} &= \gamma_2 x_{2_i} \\
&\vdots \\
\beta_{p_i} &= \gamma_p x_{p_i}.
\end{aligned} \tag{2b}$$

Combined level-1 and level-2 models:

$$y_{it} = \gamma_0 + \gamma_1 x_{1_i} + \gamma_2 x_{2_i} + \cdots + \gamma_p x_{p_i} + r_i + e_{it}. \tag{2c}$$

Variance components:

$$r_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \tau); \tag{2d}$$

$$e_{it} \overset{\text{iid}}{\sim} \mathcal{N}\left(0, \sigma^2\right). \tag{2e}$$

Estimating Equation 2c enables us to parse out how much of the total outcome variance lies between students, and how much of it is within-student variance. The parameter estimate of primary substantive interest is $\hat{\gamma_1}$, corresponding to the estimated fixed association between persistence in ILI and the outcome, where we define persistence as student i's continued enrollment in the ILI program throughout each time period t. We can also include interactions between observed variables, such as an ILI $\times$ gender term to investigate subgroup-level variation in ILI's estimated association with the outcome, as well a time period indicator and an ILI $\times$ time interaction to see if we observe meaningful differences in how the expected outcome changes over time between the comparison groups.

5.3 Characteristics and Limitations of the School Administrative Data Used in this Study

The subsequent sections apply variations of these models to the student-level administrative data that three case study sites have shared with the Indigenous Language Immersion and Native American Student Achievement Study (McCarty et al., 2016). These particular data share many features in common with the information that U.S. school districts routinely collect and store in their student information databases, typically including students' gender and ethnicity as well as their free/reduced-price lunch status, special education status, and English learner status.

Although these distinctions can help provide a rough sense of how outcomes vary between different student subgroups, they are inherently limited. For example, gender, as it is construed in the data used here, constrains students to identify as either "male" or "female" with no room for trans or non-binary gender identities. As a consequence, some of the students in the analytic samples from the case study sites are probably misgendered. Unfortunately, it is impossible to know exactly how many misgendered students there are in the data, or who they might be. Nevertheless, male-female differences in academic achievement have been documented extensively (Nowell & Hedges, 1998; Voyer & Voyer, 2014). If gender, narrowly construed, is predictive of the outcomes we are examining, it is important to include it in our analyses while acknowledging the data's shortcomings. In the future, as schools begin to accommodate a greater range of gender identities in their student databases, more nuanced analyses will be possible.

The other student demographic characteristics recorded in the data each have their own limitations as well. Free/reduced-price lunch status is recorded as a simple dichotomous variable that obscures important context. We can probably assume that students who are classified as free/reduced-price lunch program participants are less economically advantaged, on average, than their peers who do not participate in the program. However, we do not have detailed information on students' household income or family structure that might allow for more comprehensive comparisons. Furthermore, some families may have chosen not to share their household income or family size with the school district for privacy reasons. As a consequence, some students who are actually eligible for the free/reduced-price lunch program may be classified as not participating in it.

The variable for special education status is similarly problematic, as it may encompass a range of physical or sensory disabilities and/or mild-to-severe learning differences. The administrative data, however, classify students as either "special education" or "not special education." With larger samples of students and more detailed information about why the special education students qualify as such, we would be able to parse these distinctions more thoroughly. That said, even an imprecise measure may be better than none if it serves some explanatory or predictive function in our quantitative models.

As is the case with the data used in the subsequent analyses, most U.S. school districts categorize students' race and ethnicity according to one of seven categories outlined in federal data standards: American Indian or Alaska Native; Asian, Black or African American; Hispanic or Latino; Native Hawaiian or other Pacific Islander; White; or Other. Students who identify with more than one of these categories are forced to choose one of them as their primary identity, or else classify themselves as "Other." As a consequence, important subtleties inherent to individual racial and ethnic identity are concealed in the data. For the sake of comparing similar groups of ILI and non-ILI students in the analyses that follow, we restrict our case study samples to include only ILI students and Indigenous-heritage non-ILI students. Generally, all (or very nearly all) of the ILI students in each cohort at each case study site are classified as Indigenous. We retain the small number of ILI students who are classified as some other (non-Indigenous) ethnicity on the assumption—supported by conversations and interviews with educators at the Partner Schools—that their participation in the ILI program is linked to Indigenous heritage on some side of their family, or at least some other important personal connection with the local Indigenous community. That said, it is important to note that by only considering non-ILI students who are classified as Indigenous as the comparison group in our analytic sample, we may inadvertently be excluding any non-ILI students who have some Indigenous heritage (but who are not classified as Indigenous by their school) along with all the non-ILI students who have no Indigenous heritage.

6 Site 1 Case Study: Two Elementary Schools

Site 1 comprises two urban elementary schools serving kindergarten through 5th grades. One of the schools is a language-immersion elementary school enrolling approximately 375 students in two separate tracks: an Indigenous-language immersion/revitalization program and a dual-language immersion track where instruction is conducted in a non-Indigenous language and English. The English-medium comparison school for Site 1 is a separate elementary school that enrolls approximately 400 students, and is located on a different campus in the same community.

Since the ILI program at Site 1 shares facilities and resources with a non-Indigenous dual-language immersion program, it does not meet the strict definition of a "whole school" program outlined in Holm and Holm (1995). Given that it is situated at a dedicated language-immersion school, however, the ILI program at Site 1 benefits to some degree by not having to contend with certain challenges that ILI programs often face when they are organized as "streams" or tracks within English-medium schools.

6.1 Data Structure and Descriptive Statistics for the Analytic Sample

The analytic sample for Site 1 consists of four cohorts totaling 94 students—49 in the ILI program and 45 students of Indigenous heritage in the English-medium track—who started kindergarten between the 2012-13 and 2015-16 school years. Although the language-immersion school serves a sizable number of students in its Spanish-immersion program, only one of the Spanish-immersion students is identified as Indigenous among these four cohorts. Therefore, we restrict the following analyses to comparisons between the ILI students from the language-immersion school and the English-medium students of Indigenous heritage from the nearby comparison school.

The data for Site 1 include students' program status (ILI or English-medium), gender, special education status, English learner status, and free/reduced-price lunch status, as well as English-language formative and summative test scores in English language arts and math. The formative assessments were administered three times per year—each fall, winter, and spring—from kindergarten through 5th grade for both groups of students in English reading and from kindergarten through fall of 4th grade for both groups of students in math. The summative assessments were administered in the spring of 3rd, 4th, and 5th grades. The English-language formative and summative assessment data cover the following grade levels and school years:

- 1st cohort: through 5th grade (2017-2018);
- 2nd cohort: through 5th grade (2018-2019);
- 3rd cohort: through 4th grade (2018-2019);

- 4th cohort: through 3rd grade (2018-2019).

Overall, a majority of students were female in both programs (60% of the English-medium students and 53% of the ILI students). Notably, however, the youngest cohort—cohort 4—was unbalanced on gender between the two groups, with 8 girls and 3 boys in the English-medium program, and 2 girls and 8 boys in the ILI program. Table 1 shows the number of male and female students by cohort in each program.

Table 1: Cross-tabulation of gender by cohort and program at Site 1

cohort	female	male	**total**
English			
1)	7	4	**11**
2)	5	4	**9**
3)	7	7	**14**
4)	8	3	**11**
total	**27**	**18**	**45**
ILI			
1)	9	3	**12**
2)	7	2	**9**
3)	8	10	**18**
4)	2	8	**10**
total	**26**	**23**	**49**

Table 2 shows the number of non-special education and special education students by cohort in each program. Overall, the two groups were approximately balanced on this dimension, with about 20% of students having some form of special education designation. Again, the youngest cohort stands out as somewhat anomalous, with half of the ILI group classified as special education students.

Table 2: Cross-tabulation of special education status by cohort and program at Site 1

cohort	not SPED	SPED	**total**
English			
1)	10	1	**11**
2)	7	2	**9**
3)	9	5	**14**
4)	9	2	**11**
total	**35**	**10**	**45**
ILI			
1)	10	2	**12**
2)	9	0	**9**
3)	15	3	**18**
4)	5	5	**10**
total	**39**	**10**	**49**

Very few of the students—under 10%—in either group were classified as English language learners, as shown in Table 3.

Table 3: Cross-tabulation of ELL status by cohort and program at Site 1

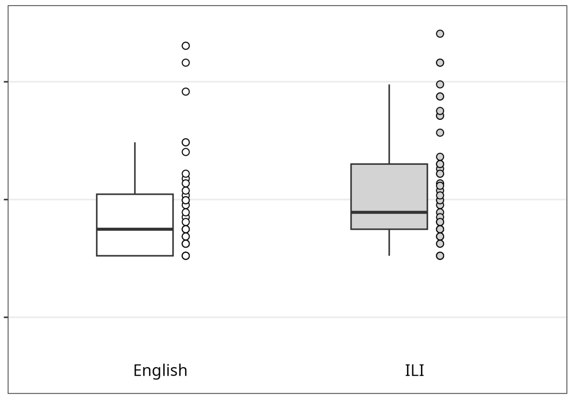
cohort	not ELL	ELL	**total**
English			
1)	8	3	**11**
2)	9	0	**9**
3)	14	0	**14**
4)	10	1	**11**
total	**41**	**4**	**45**
ILI			
1)	12	0	**12**
2)	8	1	**9**
3)	17	1	**18**
4)	10	0	**10**
total	**47**	**2**	**49**

Finally, a large majority of the students in both groups—84% of the English-medium students and 78% of the ILI students—were classified as eligible for free or reduced-price lunch, as shown in Table 4. In fact, however, this may represent an undercount of students actually eligible for free/reduced lunch, as students whose families decline to share their income information with the school for privacy reasons are classified as not participating in the free/reduced-price lunch program.

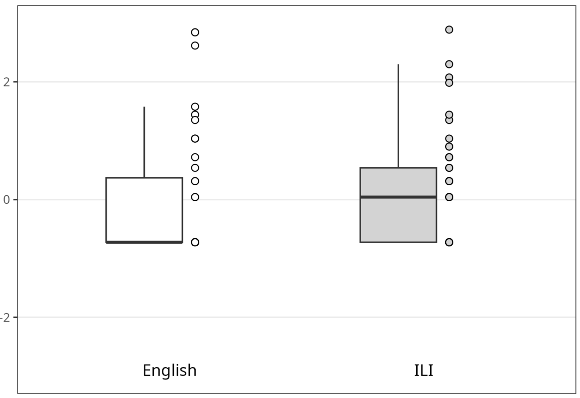
Table 4: Cross-tabulation of free/reduced lunch status by cohort and program at Site 1

cohort	not FRPL	FRPL	**total**
English			
1)	2	9	**11**
2)	2	7	**9**
3)	2	12	**14**
4)	1	10	**11**
total	**7**	**38**	**45**
ILI			
1)	2	10	**12**
2)	2	7	**9**
3)	3	15	**18**
4)	4	6	**10**
total	**11**	**38**	**49**

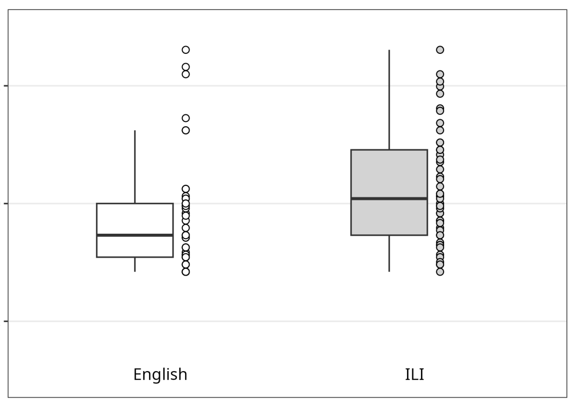
Since the data for Site 1 include students' scores on emergent literacy and numeracy assessments as early as fall of kindergarten, we can compare the relative distributions of the ILI and English-medium students on these measures around the time of their entry into elementary school. As shown in Figure 1, the ILI students scored slightly higher, on average, on all three measures—letter naming fluency, letter sounds fluency, and number identification. The differences in means between the two groups were not statistically significant for either of the emergent literacy measures (for letter naming fluency, $\bar{z}_{\text{EM}} - \bar{z}_{\text{ILI}} = -0.336; t = -1.632; \text{df} = 90.297; p\text{-value} = 0.106$ and for letter sounds fluency, $\bar{z}_{\text{EM}} - \bar{z}_{\text{ILI}} = -0.142; t = -0.677; \text{df} = 86.645; p\text{-value} = 0.501$), but the difference was statistically significant for the number identification measure ($\bar{z}_{\text{EM}} - \bar{z}_{\text{ILI}} = -0.509; t = -2.536; \text{df} = 91.487; p\text{-value} = 0.013$).



(a) letter naming fluency



(b) letter sounds fluency



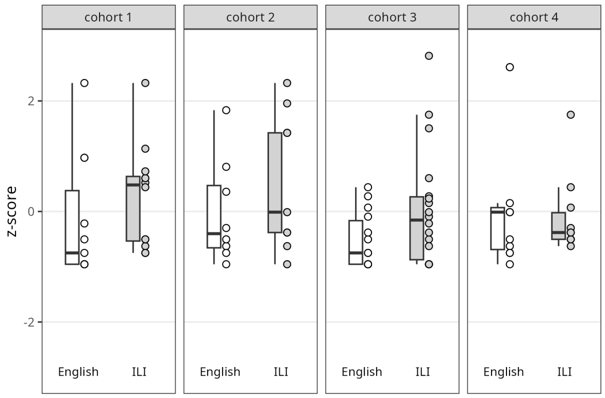
(c) number identification

Figure 1: Fall kindergarten formative assessment z-scores in (a) letter naming fluency, (b) letter sounds fluency, and (c) number identification among ILI and Indigenous-heritage English-medium students at Site 1.

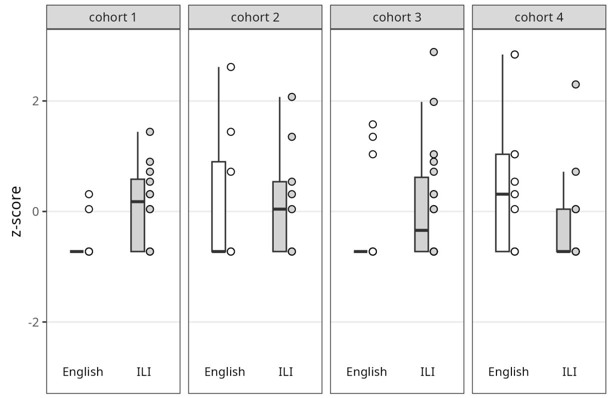
Note: The English-medium group is shaded white and the Indigenous-language immersion group is shaded gray. Dots represent individual students' z-scores. The adjacent box-and-whisker plots summarize each group's distribution of scores, where the lower horizontal border corresponds to the first quartile (25th percentile); the middle horizontal line corresponds to the median (second quartile or 50th percentile); and the upper horizontal border corresponds to the third quartile (75th percentile).

Figure 2 shows the emergent literacy and numeracy score distributions disaggregated by cohort. The relative distributions between the two programs are approximately similar for all four cohorts on the number identification measure and for the first three cohorts for the letter naming and letter sounds fluency measures.

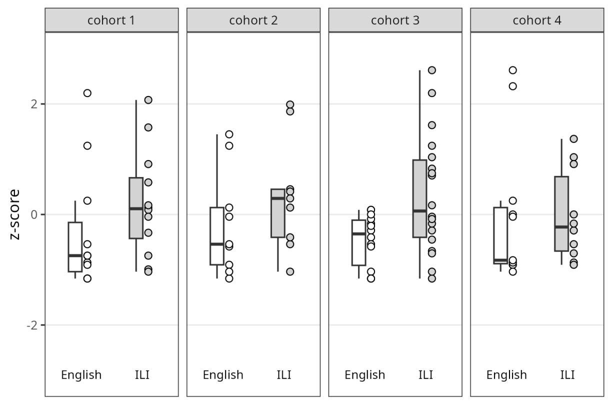
In the fourth cohort, however, the ILI students' median scores in emergent literacy were slightly lower than those of the English-medium students.



(a) letter naming fluency



(b) letter sounds fluency



(c) number identification

Figure 2: Fall kindergarten formative assessment z-scores in (a) letter naming fluency, (b) letter sounds fluency, and (c) number identification among ILI and Indigenous-heritage English-medium students at Site 1, disaggregated by cohort.

Note: The English-medium group is shaded white and the Indigenous-language immersion group is shaded gray. Dots represent individual students' z-scores. The adjacent box-and-whisker plots summarize each group's distribution of scores, where the lower horizontal border corresponds to the first quartile (25th percentile); the middle horizontal line corresponds to the median (second quartile or 50th percentile); and the upper horizontal border corresponds to the third quartile (75th percentile).

Taken together, we might infer from these results that the ILI students at Site 2 started kindergarten slightly more well-prepared, on average, with regard to English language arts and math than their English-medium peers—at least as far as letter and number recognition is concerned. There was little difference between the two groups on emergent phonics as measured by the letter sounds fluency component of the fall kindergarten assessment.

6.2 Formative Assessment Results: Longitudinal Ordinary Least Squares Regression Analysis

Since the data for Site 1 includes a relatively extensive set of student demographic information, we can further examine the differences between the ILI and English-medium students' emergent literacy and numeracy in fall of kindergarten by regressing the formative assessment z-scores on program status and gender, along with special education, free/reduced lunch, and English learner status. Table 5 shows the results of this regression model for each of the three fall kindergarten outcome measures.

Table 5: OLS regressions of standardized letter naming fluency (LNF), letter sounds fluency (LSF), and number identification (NIM) test scores at Site 1, fall kindergarten

	LNF	LSF	NIM
	(1)	(2)	(3)
intercept	0.491*	0.370	0.279
	(0.281)	(0.295)	(0.260)
ILI	0.304	0.116	0.442**
	(0.197)	(0.208)	(0.189)
male	−0.488**	−0.430**	−0.017
	(0.202)	(0.213)	(0.194)
SPED	−0.416*	0.065	−0.741***
	(0.247)	(0.260)	(0.239)
FRPL	−0.400	−0.265	−0.380
	(0.255)	(0.268)	(0.239)
ELL	−0.318	−0.602	−0.579
	(0.405)	(0.426)	(0.391)
Observations	93	93	94
R^2	0.168	0.079	0.223
Adjusted R^2	0.120	0.026	0.179
Residual Std. Error	0.938	0.987	0.906
F Statistic	3.516***	1.489	5.045***

Note: *p<0.1; **p<0.05; ***p<0.01

The differences between the expected z-scores for the two groups shrink slightly closer to zero, relative to the simple t-tests described above, when we account for demographic characteristics in the model, but not by much. All three coefficient estimates for ILI are positive, but only the coefficient for the number identification score is statistically significant after accounting for gender, special education, free/reduced lunch, and English learner status.

After accounting for program and the other demographic variables, boys' expected z-scores are nearly half a standard deviation lower than girls' scores on the emergent literacy measures, but there does not appear to be a meaningful difference between boys and girls on the number identification test.

Special education status has a strong negative association with students' z-scores for number identification after accounting for program, gender, free/reduced lunch, and English learner status. The corresponding coefficient for special education is also negative for the letter naming fluency measure, but not quite at the conventional $\alpha = 0.05$ threshold for statistical significance. For letter sounds fluency, the coefficient estimate for special education status is not significant, but this may be due in part to the relatively small variance in students' scores for that measure. (Many of the raw scale scores in both groups were clustered at or near zero.)

The emergent literacy and numeracy measures provide a rough sense of the ILI and English-medium students' preparedness levels, relative to each other, as they started to learn reading and arithmetic at the beginning of kindergarten. These scores are also moderately predictive of students' later achievement levels in English reading and math, as measured by the same formative assessment module, in later elementary grades. Accordingly, we can include students' fall kindergarten z-scores on the letter naming fluency and number identification assessments as predictors, along with program status and the various demographic characteristics, in longitudinal OLS models where the dependent variable is the English reading or math z-score at each instance that students subsequently took the these tests.

Tables 6, 7, and 8 show the results of these OLS regression models for English reading from fall of 1st grade through spring of 5th grade; Tables 9 and 10 show the corresponding results of regression models for the math formative assessment through fall of 4th grade. (Later math outcomes are omitted from this analysis because the ILI students from these cohorts at Site 1 did not take the math formative assessment from winter of 4th grade onward.)

Table 6: OLS regressions of standardized English reading formative test scores at Site 1, fall 1st to spring 2nd grade

	fall 1st	winter 1st	spring 1st	fall 2nd	winter 2nd	spring 2nd
	(1)	(2)	(3)	(4)	(5)	(6)
intercept	0.482**	0.465*	0.396	0.310	0.603**	0.651***
	(0.240)	(0.245)	(0.239)	(0.244)	(0.243)	(0.243)
ILI	−0.574***	−0.152	−0.218	−0.138	−0.274	−0.350**
	(0.168)	(0.172)	(0.168)	(0.174)	(0.173)	(0.172)
fall K LNF z-score	0.488***	0.403***	0.407***	0.389***	0.327***	0.354***
	(0.090)	(0.092)	(0.089)	(0.092)	(0.091)	(0.091)
male	−0.250	−0.190	−0.013	−0.064	−0.189	−0.175
	(0.175)	(0.179)	(0.176)	(0.182)	(0.180)	(0.179)
SPED	−0.397*	−0.730***	−0.916***	−0.870***	−0.962***	−0.916***
	(0.209)	(0.215)	(0.210)	(0.220)	(0.217)	(0.216)
FRPL	0.062	−0.140	−0.058	0.004	−0.187	−0.221
	(0.214)	(0.221)	(0.215)	(0.219)	(0.218)	(0.217)
ELL	−0.460	−0.430	−0.434	−0.455	−0.329	−0.245
	(0.338)	(0.348)	(0.338)	(0.345)	(0.344)	(0.342)
Observations	91	92	91	88	89	88
R^2	0.434	0.398	0.431	0.415	0.416	0.426
Adjusted R^2	0.393	0.356	0.390	0.371	0.373	0.383
Residual Std. Error	0.779	0.803	0.781	0.793	0.792	0.785
F Statistic	10.717***	9.375***	10.606***	9.562***	9.733***	10.004***

Note: *p<0.1; **p<0.05; ***p<0.01

Table 7: OLS regressions of standardized English reading formative test scores at Site 1, fall 3rd to spring 4th grade

	fall 3rd	winter 3rd	spring 3rd	fall 4th	winter 4th	spring 4th
	(1)	(2)	(3)	(4)	(5)	(6)
intercept	0.445*	0.576**	0.748***	0.506	0.621*	0.430
	(0.256)	(0.235)	(0.247)	(0.314)	(0.322)	(0.327)
ILI	−0.161	−0.166	−0.339*	−0.099	−0.234	−0.109
	(0.179)	(0.164)	(0.173)	(0.219)	(0.221)	(0.226)
fall K LNF z-score	0.319***	0.374***	0.323***	0.359***	0.356***	0.330***
	(0.096)	(0.088)	(0.093)	(0.115)	(0.114)	(0.119)
male	−0.137	−0.149	−0.131	0.001	0.104	−0.051
	(0.187)	(0.172)	(0.180)	(0.218)	(0.220)	(0.228)
SPED	−0.875***	−0.937***	−0.913***	−0.706**	−0.687**	−0.772**
	(0.223)	(0.205)	(0.215)	(0.280)	(0.288)	(0.292)
FRPL	−0.103	−0.218	−0.337	−0.326	−0.428	−0.180
	(0.229)	(0.210)	(0.221)	(0.278)	(0.290)	(0.291)
ELL	−0.377	−0.432	−0.256	−0.470	−0.520	−0.523
	(0.360)	(0.331)	(0.348)	(0.411)	(0.412)	(0.418)
Observations	91	91	90	69	67	65
R^2	0.355	0.454	0.398	0.320	0.326	0.309
Adjusted R^2	0.309	0.415	0.354	0.255	0.258	0.237
Residual Std. Error	0.831	0.764	0.802	0.864	0.860	0.871
F Statistic	7.715***	11.637***	9.146***	4.871***	4.829***	4.317***

Note: *p<0.1; **p<0.05; ***p<0.01

Table 8: OLS regressions of standardized English reading formative test scores at Site 1, fall to spring 5th grade

	fall 5th	winter 5th	spring 5th
	(1)	(2)	(3)
intercept	0.329	0.526	0.399
	(0.449)	(0.446)	(0.410)
ILI	−0.121	−0.089	−0.266
	(0.299)	(0.294)	(0.271)
fall K LNF z-score	0.348**	0.349**	0.409***
	(0.144)	(0.144)	(0.133)
male	−0.307	0.047	−0.169
	(0.331)	(0.329)	(0.303)
SPED	−0.919*	−0.970**	−1.130**
	(0.455)	(0.456)	(0.420)
FRPL	−0.074	−0.498	−0.092
	(0.391)	(0.392)	(0.361)
ELL	−0.055	0.200	−0.163
	(0.490)	(0.489)	(0.450)
Observations	39	40	40
R^2	0.351	0.338	0.443
Adjusted R^2	0.229	0.217	0.342
Residual Std. Error	0.871	0.875	0.805
F Statistic	2.879**	2.803**	4.380***

Note: *p<0.1; **p<0.05; ***p<0.01

Table 9: OLS regressions of standardized math formative test scores at Site 1, fall 1st to spring 2nd grade

	fall 1st	winter 1st	spring 1st	fall 2nd	winter 2nd	spring 2nd
	(1)	(2)	(3)	(4)	(5)	(6)
intercept	0.268	0.547*	0.155	0.432	0.420	0.722*
	(0.298)	(0.287)	(0.296)	(0.357)	(0.359)	(0.361)
ILI	0.179	−0.630***	−0.176	−0.151	−0.716***	−0.845***
	(0.208)	(0.200)	(0.208)	(0.265)	(0.261)	(0.253)
fall K NIM z-score	0.233**	0.364***	0.221*	0.389***	0.308**	0.284**
	(0.113)	(0.108)	(0.111)	(0.136)	(0.134)	(0.132)
male	−0.259	−0.185	0.119	−0.134	0.160	−0.111
	(0.207)	(0.198)	(0.207)	(0.258)	(0.260)	(0.257)
SPED	−0.686**	−0.435	−0.595**	−0.282	−0.258	−0.158
	(0.289)	(0.271)	(0.281)	(0.334)	(0.346)	(0.349)
FRPL	−0.145	−0.077	0.075	−0.317	−0.175	−0.354
	(0.280)	(0.270)	(0.280)	(0.331)	(0.332)	(0.339)
ELL	−0.252	−0.515	−0.998**	−0.104	−0.356	−0.607
	(0.400)	(0.386)	(0.399)	(0.451)	(0.452)	(0.441)
Observations	80	82	80	60	60	58
R^2	0.254	0.297	0.255	0.231	0.229	0.281
Adjusted R^2	0.192	0.241	0.193	0.143	0.141	0.196
Residual Std. Error	0.899	0.871	0.898	0.926	0.927	0.896
F Statistic	4.132***	5.293***	4.154***	2.646**	2.617**	3.323***

Note: *p<0.1; **p<0.05; ***p<0.01

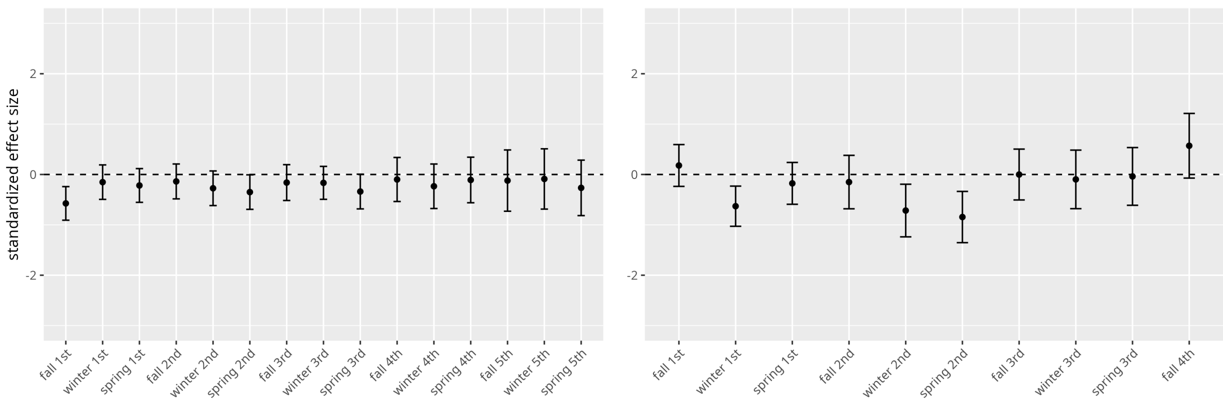
Table 10: OLS regressions of standardized math formative test scores at Site 1, fall 3rd to fall 4th grade

	fall 3rd	winter 3rd	spring 3rd	fall 4th
	(1)	(2)	(3)	(4)
intercept	0.587*	0.706*	0.749**	0.462
	(0.340)	(0.359)	(0.322)	(0.370)
ILI	−0.001	−0.098	−0.039	0.570*
	(0.252)	(0.289)	(0.285)	(0.316)
fall K NIM z-score	0.335**	0.412***	0.350***	0.345**
	(0.130)	(0.136)	(0.130)	(0.166)
male	−0.232	−0.019	0.035	0.007
	(0.240)	(0.246)	(0.245)	(0.278)
SPED	−0.497	−0.358	−0.779**	−0.409
	(0.307)	(0.312)	(0.308)	(0.336)
FRPL	−0.433	−0.610*	−0.684**	−0.559
	(0.315)	(0.345)	(0.317)	(0.347)
ELL	−0.351	−0.555	0.223	−0.213
	(0.428)	(0.465)	(0.458)	(0.541)
Observations	62	53	53	43
R^2	0.296	0.359	0.375	0.354
Adjusted R^2	0.220	0.275	0.294	0.246
Residual Std. Error	0.883	0.851	0.840	0.868
F Statistic	3.863***	4.293***	4.607***	3.282**
Note:				*p<0.1; **p<0.05; ***p<0.01

After accounting for the full array of demographic variables—gender, special education, free/reduced lunch, and English learner status—as well as fall kindergarten z-scores on the letter naming fluency test, the point estimates for the association between ILI participation and 1st-5th grade English reading z-scores were consistently negative at each time point. In most cases, however, the estimates were relatively small in magnitude (*i.e.*, around -0.2 standard deviations or less) and were only statistically significant and larger in fall of 1st grade and spring of 2nd grade. The ILI-versus-English-medium contrasts tended to be smaller in 4th and 5th grade, suggesting that the two groups of students were approaching parity on English language reading as they reached the end of elementary school.

The corresponding point estimates for the association between ILI participation and 1st-4th grade math z-scores followed a somewhat similar pattern, in that the expected scores approached parity from 3rd grade onward. The coefficients were negative from the middle of 1st grade through the end of 2nd grade, and statistically significant at the winter 1st, winter 2nd, and spring 2nd grade assessment instances. However, throughout 3rd grade, there appeared to be no meaningful difference between the two groups of students after accounting for fall kindergarten number identification z-score and the other demographic variables. In fall of 4th grade, the last occasion when both groups of students took the math formative test, the coefficient for ILI was positive with nearly all of its 95% confidence interval above zero.

The point estimates and confidence intervals of the coefficients on ILI from the longitudinal regressions in Tables 6, 7, 8, 9, and 10 are displayed graphically in Figure 3. The point estimates for separate regressions for female and male students are plotted in Figure 4.

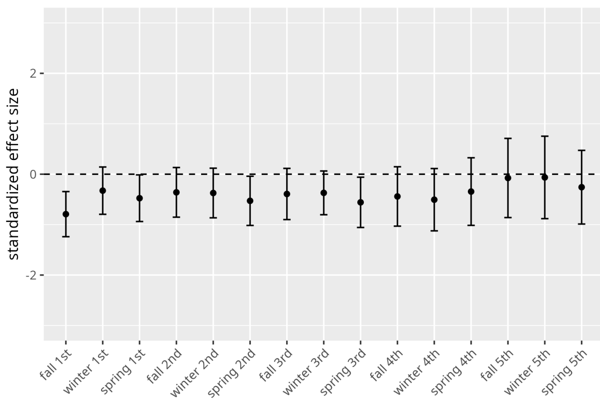


(a) English reading (all students)

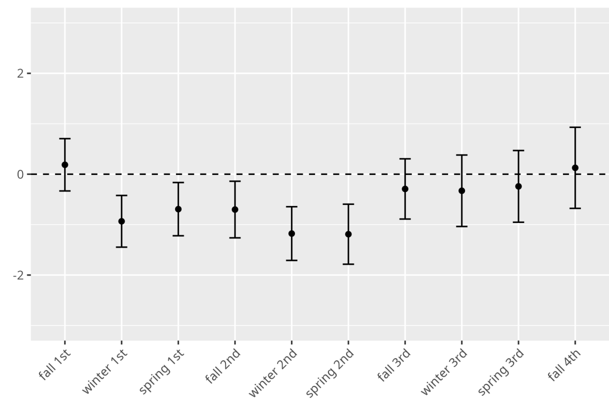
(b) math (all students)

Figure 3: Coefficient estimates and 95% confidence intervals of persistence in ILI on English reading and math formative assessments in elementary school at Site 1, after accounting for fall kindergarten emergent literacy/numeracy test scores, gender, special education status, free/reduced lunch eligibility, and English learner status.

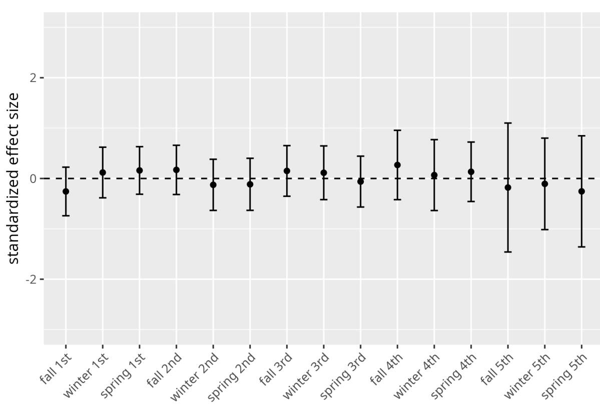
Note: Coefficient estimates at each time point are represented as dots. 95% confidence intervals appear as lines, with the vertical line segments representing the spans of the confidence intervals and the horizontal line segments representing the upper and lower bounds of the confidence intervals. The horizontal dashed line corresponds to zero, *i.e.*, an estimate of no difference between ILI students' and English-medium students' achievement levels after accounting for fall kindergarten test scores, gender, special education status, free/reduced lunch eligibility, and English learner status.



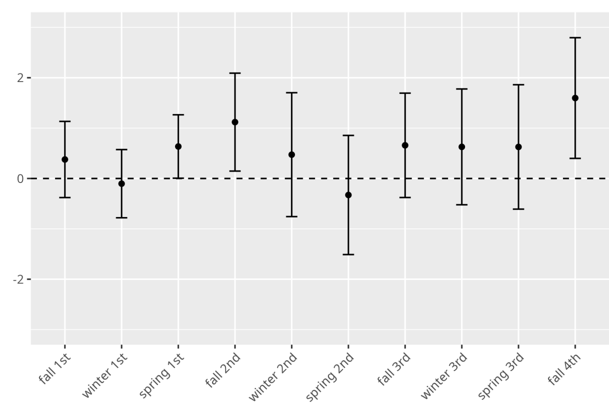
(a) English reading (female students only)



(b) math (female students only)



(c) English reading (male students only)



(d) math (male students only)

Figure 4: Coefficient estimates and 95% confidence intervals of persistence in ILI among female and male students on English reading and math formative assessments in elementary school at Site 1, after accounting for fall kindergarten emergent literacy/numeracy test scores, special education status, free/reduced lunch eligibility, and English learner status.

Note: Coefficient estimates at each time point are represented as dots. 95% confidence intervals appear as lines, with the vertical line segments representing the spans of the confidence intervals and the horizontal line segments representing the upper and lower bounds of the confidence intervals. The horizontal dashed line corresponds to zero, *i.e.*, an estimate of no difference between female or male ILI and English-medium students' achievement levels after accounting for fall kindergarten test scores, special education status, free/reduced lunch eligibility, and English learner status.

Some notable patterns are apparent when we estimate separate regressions for male and female students. Specifically, most of the negative association between ILI persistence and English reading and math z-scores that remains after accounting for fall kindergarten z-scores, special education, free/reduced lunch, and English learner status is concentrated in the contrast between ILI and English-medium girls, particularly in the early elementary grades. For boys, on the other hand, the estimated differences between ILI students and English-medium students in English reading are negligible—all clustered around zero. And for math, the ILI coefficient estimates are positive for boys at all assessment instances except for winter 1st grade and spring 2nd grade. That said, the estimated contrast between the ILI and English-medium groups are approximately similar for boys and girls in English reading in 5th grade.

6.3 Formative Assessment Results: Hierarchical Linear Mixed Effects Analysis with Random Intercepts

The longitudinal OLS approach in the previous section is useful for providing discrete estimates at each formative testing instance and observing how they change over time for the sample writ large. But we can also take advantage of the data's hierarchical structure—individual observations at each time point nested within students—to estimate variations on the following mixed-mixed effects model, where coefficients on the student-level variables and the $\text{trimester}_{.t}$ time index are treated as fixed-effects parameters and the intercept is allowed to vary randomly at the student level:

$$\begin{aligned} y_{it} = {} & \hat{\gamma}_0 + \hat{\gamma}_1(\text{ILI}_{i.}) + \hat{\gamma}_2(\text{fall K z-score}_{i.}) + \hat{\gamma}_3(\text{male}_{i.}) + \hat{\gamma}_4(\text{SPED}_{i.}) + \hat{\gamma}_5(\text{FRPL}_{i.}) \\ & + \hat{\gamma}_6(\text{ELL}_{i.}) + \hat{\gamma}_7(\text{trimester}_{.t}) + \hat{\gamma}_8(\text{ILI}_{i.} \times \text{male}_{i.}) + \hat{\gamma}_9(\text{ILI}_{i.} \times \text{trimester}_{.t}) + r_{i.} + e_{it}. \end{aligned} \tag{3a}$$

$$\widehat{\text{Var}}[r_{i.}] = \hat{\tau}; \tag{3b}$$

$$\widehat{\text{Var}}[e_{it}] = \hat{\sigma}^2. \tag{3c}$$

Tables 11 and 12 show various permutations on the model expressed in Equations 3a, 3b, and 3c, starting initially with a simple one-way ANOVA with random effects (model 1), and iteratively adding student-level variables for ILI participation, fall kindergarten z-score, gender (male), special education status (SPED), and an ILI × male interaction term.

Tables 13 and 14 expand on these models by adding in fixed coefficients for trimester (representing the change in scores from one assessment instance t to the next), an ILI × trimester parameter that isolates an overall contrast in change-over-time slopes between ILI and English-medium programs, and finally student-level indicators for free/reduced-price lunch (FRPL) and English-learner status (ELL). The outcome y_{it} in each case is the standardized English reading (Tables 11 and 13) or math (12 and 14) formative test score for student i at time t.

Table 11: Random intercept models: standardized English reading formative test scores at Site 1

	(1)	(2)	(3)	(4)	(5)	(6)
Fixed effects:						
intercept	−0.008	0.034	0.124	0.228	0.366***	0.477***
	(0.098)	(0.143)	(0.123)	(0.139)	(0.129)	(0.146)
ILI		−0.079	−0.251	−0.222	−0.225	−0.438**
		(0.197)	(0.170)	(0.170)	(0.153)	(0.202)
fall K LNF z-score			0.512***	0.474***	0.398***	0.398***
			(0.085)	(0.088)	(0.081)	(0.080)
male				−0.269	−0.148	−0.400*
				(0.174)	(0.159)	(0.223)
SPED					−0.884***	−0.918***
					(0.191)	(0.191)
ILI × male						0.484
						(0.302)
Variance components:						
Random Intercept ($\hat{\tau}$)	0.881	0.889	0.639	0.630	0.510	0.501
Level-1 Residual ($\hat{\sigma}^2$)	0.126	0.126	0.126	0.126	0.126	0.126
Data structure:						
Number of Students	93	93	93	93	93	93
Number of Observations	1131	1131	1131	1131	1131	1131
Selection criteria:						
AIC	1285.463	1288.716	1263.178	1264.453	1248.528	1248.538
BIC	1300.555	1308.839	1288.332	1294.638	1283.744	1288.785
Note:						*p<0.1; **p<0.05; ***p<0.01

Table 12: Random intercept models: standardized math formative test scores at Site 1

	(1)	(2)	(3)	(4)	(5)	(6)
Fixed effects:						
intercept	−0.025	−0.043	0.064	0.080	0.156	0.295**
	(0.083)	(0.112)	(0.097)	(0.115)	(0.115)	(0.119)
ILI		0.042	−0.188	−0.189	−0.203	−0.505***
		(0.169)	(0.151)	(0.152)	(0.147)	(0.175)
fall K NIM z-score			0.405***	0.404***	0.345***	0.334***
			(0.074)	(0.074)	(0.076)	(0.072)
male				−0.040	0.008	−0.328*
				(0.148)	(0.145)	(0.181)
SPED					−0.493**	−0.512***
					(0.194)	(0.185)
ILI × male						0.795***
						(0.278)
Variance components:						
Random Intercept ($\hat{\tau}$)	0.512	0.519	0.365	0.369	0.342	0.305
Level-1 Residual ($\hat{\sigma}^2$)	0.469	0.469	0.469	0.470	0.469	0.470
Data structure:						
Number of Students	84	84	84	84	84	84
Number of Observations	747	747	747	747	747	747
Selection criteria:						
AIC	1752.688	1756.354	1736.024	1739.937	1737.101	1731.982
BIC	1766.536	1774.818	1759.104	1767.634	1769.414	1768.911

Note: *p<0.1; **p<0.05; ***p<0.01

Table 13: Additional random intercept models: standardized English reading formative test scores at Site 1

	(7)	(8)	(9)	(10)
Fixed effects:				
intercept	0.497***	0.527***	0.652***	0.675***
	(0.147)	(0.148)	(0.228)	(0.229)
ILI	−0.437**	−0.493**	−0.499**	−0.506**
	(0.202)	(0.205)	(0.205)	(0.205)
fall K LNF z-score	0.398***	0.397***	0.388***	0.380***
	(0.080)	(0.080)	(0.082)	(0.082)
male	−0.400*	−0.400*	−0.413*	−0.409*
	(0.223)	(0.223)	(0.225)	(0.225)
SPED	−0.920***	−0.920***	−0.916***	−0.884***
	(0.191)	(0.191)	(0.191)	(0.194)
FRPL			−0.142	−0.147
			(0.197)	(0.197)
ELL				−0.321
				(0.311)
trimester	−0.003	−0.008**	−0.008**	−0.008**
	(0.003)	(0.004)	(0.004)	(0.004)
ILI × male	0.481	0.484	0.480	0.465
	(0.302)	(0.302)	(0.303)	(0.303)
ILI × trimester		0.009*	0.009*	0.009*
		(0.006)	(0.006)	(0.006)
Variance components:				
Random Intercept ($\hat{\tau}$)	0.501	0.501	0.504	0.504
Level-1 Residual ($\hat{\sigma}^2$)	0.126	0.126	0.126	0.126
Data structure:				
Number of Students	93	93	93	93
Number of Observations	1131	1131	1131	1131
Selection criteria:				
AIC	1259.052	1266.802	1269.694	1271.124
BIC	1304.329	1317.111	1325.033	1331.495

Note: *p<0.1; **p<0.05; ***p<0.01

Table 14: Additional random intercept models: standardized math formative test scores at Site 1

	(7)	(8)	(9)	(10)
Fixed effects:				
intercept	0.264**	0.282**	0.506**	0.524***
	(0.127)	(0.128)	(0.198)	(0.200)
ILI	−0.490***	−0.556***	−0.553***	−0.552***
	(0.176)	(0.190)	(0.189)	(0.189)
fall K NIM z-score	0.335***	0.335***	0.317***	0.305***
	(0.072)	(0.072)	(0.073)	(0.074)
male	−0.328*	−0.328*	−0.334*	−0.330*
	(0.181)	(0.180)	(0.179)	(0.179)
SPED	−0.510***	−0.511***	−0.516***	−0.492***
	(0.185)	(0.185)	(0.184)	(0.186)
FRPL			−0.265	−0.273
			(0.180)	(0.180)
ELL				−0.236
				(0.264)
trimester	0.005	0.002	0.002	0.002
	(0.007)	(0.008)	(0.008)	(0.008)
ILI × male	0.800***	0.815***	0.809***	0.806***
	(0.278)	(0.278)	(0.276)	(0.276)
ILI × trimester		0.019	0.019	0.019
		(0.020)	(0.020)	(0.020)
Variance components:				
Random Intercept ($\hat{\tau}$)	0.306	0.304	0.298	0.299
Level-1 Residual ($\hat{\sigma}^2$)	0.470	0.471	0.470	0.470
Data structure:				
Number of Students	84	84	84	84
Number of Observations	747	747	747	747
Selection criteria:				
AIC	1741.485	1748.574	1750.003	1752.031
BIC	1783.029	1794.734	1800.779	1807.424
Note:				*p<0.1; **p<0.05; ***p<0.01

We can obtain an estimate of the proportion of the total outcome variance that is concentrated between students from the one-way ANOVA models in column 1 of Tables 11 and 12. The total variance is 1.01 in English reading and 0.98 in math. Before accounting for any student background characteristics, the between-student variance accounts for 87% of the total variance in English reading but only about half (52%) of the total variance in math.

When we account for ILI status in the second iteration of the models (column 2), the fixed effects estimates for ILI are both very close to zero (negative for reading and positive for math) but not statistically significant. ILI status also does not appear to explain any additional between-student variance relative to the one-way ANOVA models in either subject area.

When we add fall kindergarten z-scores into the model in column 3—letter naming fluency (LNF) in the reading model and the number identification measure (NIM) for math—the fixed effect estimate of ILI becomes larger in a negative direction (but not statistically significantly so) in both cases. Fall kindergarten scores explain about 30% of the between-student variance in both subjects.

Gender, by itself, does not explain any additional between-student variance after accounting for program status and fall kindergarten test scores. The coefficient for special education status, however, is negative and statistically significant for both subjects. And when we add the ILI $\times$ male interaction term, its fixed effect estimate is positive—statistically significantly so for math. The free/reduced lunch (FRPL) and English language learner (ELL) indicators do not seem to provide any additional explanatory power once the other variables are included (columns 9 and 10 of Tables 13 and 14). Nor does there appear to be a meaningful difference in the change-over-time slopes (ILI $\times$ trimester) for the ILI students relative to the English-medium students.

In terms of model fit, the Akaike and Bayesian information criteria (AIC and BIC) suggest that model 6, which accounts for ILI status, fall kindergarten scores, gender, special education status and the ILI-gender interaction strikes the best balance between model parsimony and fit. The inferences that we might draw from this model mirror the findings from the preceding longitudinal OLS analysis: namely that the negative association that we observe between ILI persistence and English reading and math formative test scores at Site 1 seems to be located mostly among female students.

6.4 Summative Assessment Results: Ordinary Least Squares and Logistic Regression Analysis

The data for Site 1 also include students' spring summative assessment scores in English language arts and math in 3rd, 4th, and 5th grades. Although they are administered only once per year, these summative assessments are considerably more comprehensive—and consequential—for students and their schools than the relatively low-stakes quarterly formative assessments detailed in the preceding section.

The summative assessment results are represented in the data from Site 1 in two different ways: as quasi-continuous scale scores (which can be standardized for easier interpretation) and as categorical grade-level proficiency levels ranging from "minimally proficient" to "partially proficient" to "proficient" to "highly proficient." (Tables 69 through 74 in Chapter 10: Appendix 1 show the counts of students in each proficiency level by program, subject and grade for Site 1.)

Tables 15 and 16 show the estimates from selected OLS regression models with the following parametrization:

$$
\begin{aligned}
y_{it} = \beta_{0t} &+ \beta_{1t}(\text{ILI}_{i.}) + \beta_{2t}(\text{fall K z-score}_{i.}) + \beta_{3t}(\text{male}_{i.}) \\
&+ \beta_{4t}(\text{SPED}_{i.}) + \beta_{5t}(\text{FRPL}_{i.}) + \beta_{6t}(\text{ELL}_{i.}) + \beta_{7t}(\text{ILI} \times \text{male}_{i.}) + e_{it},
\end{aligned} \tag{4}
$$

where y_{it} represents student i's z-score on the summative assessment in grade t in English language arts or math and "fall K z-score" corresponds to student i's standardized fall kindergarten letter naming fluency or number identification measure score.

Table 15: Standardized 3rd-5th grade English language arts summative test scores at Site 1

	3rd grade		4th grade		5th grade	
	(1)	(2)	(3)	(4)	(5)	(6)
intercept	−0.034	0.143	0.423	0.705**	−0.031	−0.114
	(0.210)	(0.215)	(0.288)	(0.292)	(0.325)	(0.327)
ILI	0.219	−0.110	−0.095	−0.549**	0.051	0.238
	(0.148)	(0.202)	(0.199)	(0.258)	(0.214)	(0.254)
fall K LNF z-score	0.453***	0.431***	0.167	0.194**	0.115	0.105
	(0.086)	(0.083)	(0.102)	(0.096)	(0.105)	(0.104)
male	−0.126	−0.466**	−0.077	−0.578**	0.004	0.257
	(0.156)	(0.211)	(0.189)	(0.266)	(0.240)	(0.304)
SPED	−0.792***	−0.863***	−0.574**	−0.640***	−0.810**	−0.814**
	(0.169)	(0.165)	(0.250)	(0.235)	(0.332)	(0.329)
FRPL	−0.126	−0.161	−0.703***	−0.709***	−0.329	−0.355
	(0.193)	(0.185)	(0.254)	(0.237)	(0.286)	(0.283)
ELL	−0.361	−0.256	−0.506	−0.476	−0.179	−0.191
	(0.394)	(0.380)	(0.455)	(0.425)	(0.356)	(0.352)
ILI × male		0.642**		0.879**		−0.579
		(0.282)		(0.348)		(0.438)
Observations	53	53	44	44	40	40
R^2	0.678	0.711	0.415	0.503	0.296	0.332
Adjusted R^2	0.636	0.666	0.320	0.407	0.168	0.186
Residual Std. Error	0.507	0.485	0.597	0.558	0.637	0.630
F Statistic	16.146***	15.838***	4.380***	5.212***	2.309*	2.273*
AIC	86.854	83.076	87.805	82.624	85.735	85.609
BIC	102.616	100.809	102.079	98.682	99.246	100.809
Note:				*p<0.1; **p<0.05; ***p<0.01		

When male and female students are pooled together (see columns 1, 3, and 5 of Table 15), we can see that there is no statistically significant association between ILI persistence and summative assessment scores in English language arts in grades 3-5 after accounting for fall kindergarten scores, gender, and the other observed covariates. There is a positive statistically significant positive coefficient for the ILI × male interaction in 3rd and 4th grade in English language arts (columns 2 and 4). Also in 4th grade, there appears to be a negative association between ILI persistence and English language arts scores for girls, as is apparent from the non-interacted coefficient for ILI in column 4. However, by 5th grade, this pattern is not apparent: neither the non-interacted ILI nor the ILI × male coefficient is statistically significant (see column 6).

We can also see that the fall kindergarten letter naming score becomes less predictive of summative test scores as students advance toward the later elementary grades.

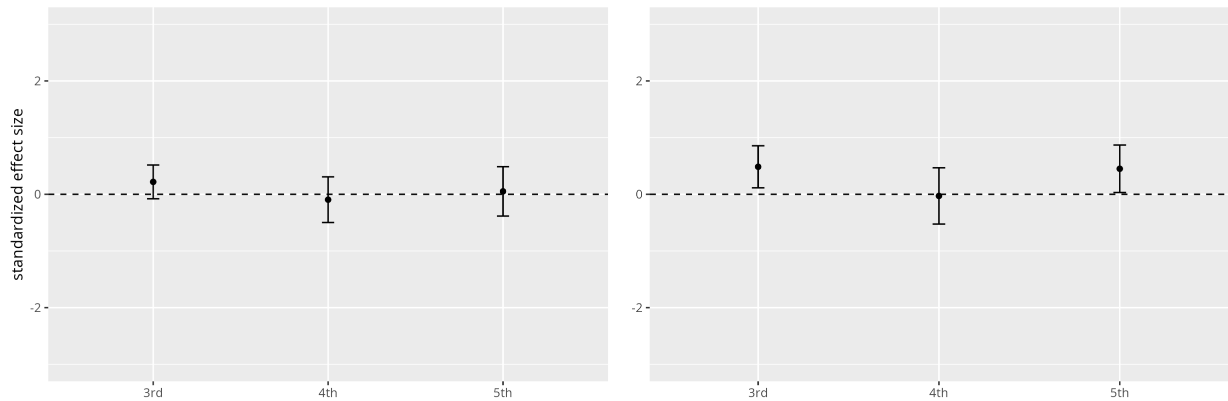
Table 16: Standardized 3rd-5th grade math summative test scores at Site 1

	3rd grade		4th grade		5th grade	
	(1)	(2)	(3)	(4)	(5)	(6)
intercept	−0.101	−0.001	0.158	0.109	−0.471	−0.445
	(0.250)	(0.267)	(0.312)	(0.339)	(0.283)	(0.291)
ILI	0.485**	0.293	−0.029	0.060	0.450**	0.378
	(0.184)	(0.259)	(0.245)	(0.334)	(0.206)	(0.250)
fall K NIM z-score	0.315***	0.305***	0.326**	0.321**	0.196*	0.201*
	(0.103)	(0.104)	(0.130)	(0.132)	(0.106)	(0.108)
male	−0.069	−0.263	−0.052	0.052	0.373	0.274
	(0.181)	(0.257)	(0.220)	(0.344)	(0.226)	(0.296)
SPED	−0.715***	−0.761***	−0.636**	−0.625*	−0.523	−0.522
	(0.223)	(0.227)	(0.307)	(0.311)	(0.324)	(0.327)
FRPL	−0.334	−0.350	−0.507*	−0.512*	−0.357	−0.341
	(0.231)	(0.231)	(0.288)	(0.291)	(0.260)	(0.264)
ELL	−0.310	−0.254	0.066	0.059	−0.292	−0.283
	(0.477)	(0.479)	(0.569)	(0.576)	(0.350)	(0.354)
ILI × male		0.378		−0.178		0.229
		(0.357)		(0.450)		(0.438)
Observations	53	53	45	45	41	41
R^2	0.580	0.591	0.371	0.374	0.420	0.424
Adjusted R^2	0.526	0.527	0.272	0.255	0.317	0.302
Residual Std. Error	0.618	0.617	0.727	0.735	0.624	0.630
F Statistic	10.606***	9.274***	3.734***	3.152**	4.097***	3.475***
AIC	107.853	108.552	107.364	109.173	85.970	87.631
BIC	123.615	126.284	121.817	125.433	99.678	103.053

Note: *p<0.1; **p<0.05; ***p<0.01

For math, we observe a positive and statistically significant association between ILI persistence and summative test scores in 3rd and 5th grades (but not 4th) in columns 1, 3, and 5 of Table 16, after accounting for the fall kindergarten number identification measure score, gender and the other observed covariates. That said, there does not appear to be a statistically significant interaction effect between gender and ILI status on the math summative test (see columns 2, 4, and 6). As with English language arts (though to a lesser degree), the predictiveness of the fall kindergarten assessment in relation to math summative assessment scores attenuates over time.

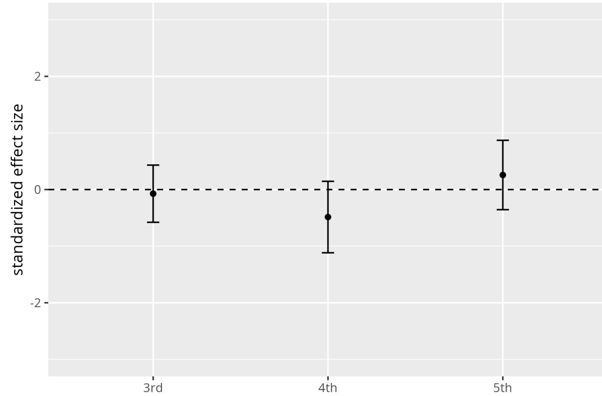
The coefficient estimates and corresponding 95% confidence intervals of the coefficient for ILI are represented graphically in Figure 5 for male and female students pooled together, and in Figure 6 for male and female students separately.



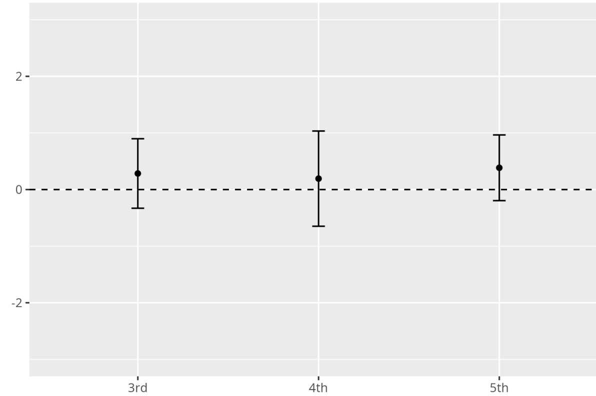
(a) English language arts (b) math

Figure 5: Coefficient estimates and 95% confidence intervals of persistence in ILI 3rd, 4th, and 5th grade English language arts and math summative assessments in elementary school at Site 1, after accounting for fall kindergarten emergent literacy/numeracy test scores, gender, special education status, free/reduced lunch eligibility, and English learner status.

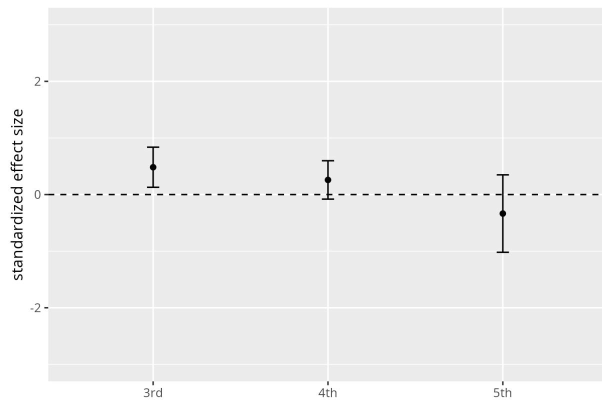
Note: Coefficient estimates at each time point are represented as dots. 95% confidence intervals appear as lines, with the vertical line segments representing the spans of the confidence intervals and the horizontal line segments representing the upper and lower bounds of the confidence intervals. The horizontal dashed line corresponds to zero, *i.e.*, an estimate of no difference between ILI and English-medium students' achievement levels after accounting for fall kindergarten test scores, gender, special education status, free/reduced lunch eligibility, and English learner status.



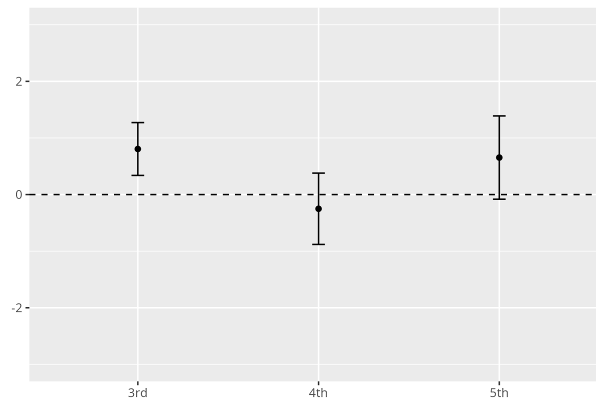
(a) English language arts (female students only)



(b) math (female students only)



(c) English language arts (male students only)



(d) math (male students only)

Figure 6: Coefficient estimates and 95% confidence intervals of persistence in ILI among female and male students on 3rd, 4th, and 5th grade English language arts and math summative assessments in elementary school at Site 1, after accounting for fall kindergarten emergent literacy/numeracy test scores, special education status, free/reduced lunch eligibility, and English learner status.

Note: Coefficient estimates at each time point are represented as dots. 95% confidence intervals appear as lines, with the vertical line segments representing the spans of the confidence intervals and the horizontal line segments representing the upper and lower bounds of the confidence intervals. The horizontal dashed line corresponds to zero, *i.e.*, an estimate of no difference between female or male ILI and English-medium students' achievement levels after accounting for fall kindergarten test scores, special education status, free/reduced lunch eligibility, and English learner status.

We can also use logistic regression to compare ILI and English-medium students' relative odds of scoring "proficient" or higher on the 3rd, 4th and 5th grade English language arts and math summative assessments by running the following model:

$$\begin{aligned}\log\frac{p(y_{it})}{1-p(y_{it})} &= \beta_{0t} + \beta_{1t}(\text{ILI}_{i.}) + \beta_{2t}(\text{fall K z-score}_{i.}) + \beta_{3t}(\text{male}_{i.}) \\ &\quad + \beta_{4t}(\text{SPED}_{i.}) + \beta_{5t}(\text{FRPL}_{i.}) + \beta_{6t}(\text{ELL}_{i.}) + \beta_{7t}(\text{ILI} \times \text{male}_{i.}) + e_{it},\end{aligned} \tag{5}$$

where the outcome y_{it} in this case is a binary dependent variable coded 0 if student i scored in the "minimally proficient" or "partially proficient" range in grade t and 1 if they scored in the "proficient" or "highly proficient" range, and $p(y_{it})$ corresponds to the predicted probability that $y_{it} = 1$.

Table 17 shows the results of the logistic regression models where the outcome is scoring "proficient" or higher on the English language arts summative test in 3rd, 4th or 5th grade; Table 18 shows the corresponding logistic regression estimates for math. Based on these results, we do not see evidence that there is a statistically significant difference between ILI and English-medium students in the probability of scoring proficient or higher in English language arts or math in grades 3, 4 or 5 after accounting for gender, fall kindergarten emergent literacy/numeracy, special education status, free/reduced-price lunch status, or English learner status, whether or not an ILI × male interaction term is included in the model.

Table 17: Log odds of scoring "proficient" or higher on English language arts summative test at Site 1 in 3rd-5th grade

	3rd grade		4th grade		5th grade	
	(1)	(2)	(3)	(4)	(5)	(6)
intercept	−0.383	0.024	0.377	0.932	1.386	1.243
	(1.400)	(1.416)	(1.107)	(1.195)	(1.298)	(1.441)
ILI	−0.840	−2.080	−0.016	−0.864	−0.268	0.237
	(0.980)	(1.294)	(0.855)	(1.125)	(0.786)	(0.934)
fall K LNF z-score	1.131**	1.157**	0.034	0.077	0.180	0.167
	(0.507)	(0.581)	(0.401)	(0.413)	(0.347)	(0.358)
male	−0.609	−19.379	−0.347	−1.492	−0.777	0.006
	(0.922)	(5,160.041)	(0.808)	(1.363)	(0.955)	(1.214)
SPED	−18.084	−19.336	0.402	0.305	−17.399	−17.371
	(2,783.626)	(4,487.653)	(1.072)	(1.126)	(2,618.111)	(2,593.455)
FRPL	0.005	0.090	−1.861**	−1.950**	−1.615	−1.795
	(1.276)	(1.288)	(0.922)	(0.954)	(1.110)	(1.240)
ELL	0.114	1.506	−16.083	−16.063	−18.438	−18.381
	(7,757.900)	(12,016.520)	(2,797.291)	(2,764.446)	(2,756.969)	(2,832.649)
ILI × male		20.262		1.925		−1.876
		(5,160.041)		(1.732)		(1.821)
Observations	53	53	44	44	40	40
Log Likelihood	−17.530	−15.385	−20.621	−19.956	−20.612	−20.041
AIC	49.060	46.769	55.242	55.912	55.224	56.082
BIC	62.852	62.532	67.732	70.185	67.046	69.593
Note:						*p<0.1; **p<0.05; ***p<0.01

Table 18: Log odds of scoring "proficient" or higher on math summative test at Site 1 in 3rd-5th grade

	3rd grade		4th grade		5th grade	
	(1)	(2)	(3)	(4)	(5)	(6)
intercept	18.382	18.913	1.340	0.936	−2.202	−1.785
	(2,550.157)	(2,514.209)	(1.082)	(1.200)	(1.365)	(1.437)
ILI	0.754	−0.286	0.405	1.357	1.857*	1.336
	(0.795)	(1.109)	(0.902)	(1.219)	(1.011)	(1.217)
fall K NIM z-score	0.311	0.281	0.475	0.449	0.338	0.350
	(0.433)	(0.439)	(0.438)	(0.459)	(0.435)	(0.436)
male	−0.135	−1.324	−1.179	0.302	1.110	0.339
	(0.764)	(1.246)	(0.855)	(1.359)	(1.038)	(1.599)
SPED	−19.055	−19.353	−0.839	−0.757	−17.343	−18.487
	(2,550.157)	(2,514.209)	(1.429)	(1.369)	(2,573.826)	(4,150.513)
FRPL	−19.315	−19.441	−2.522**	−2.782**	−0.351	−0.351
	(2,550.157)	(2,514.209)	(0.984)	(1.082)	(1.034)	(1.065)
ELL	0.384	0.596	−15.626	−15.741	−16.269	−17.349
	(7,867.086)	(7,403.362)	(2,630.636)	(2,482.122)	(2,936.325)	(4,876.863)
ILI × male		2.271		−2.448		1.259
		(1.667)		(1.854)		(2.019)
Observations	53	53	45	45	41	41
Log Likelihood	−21.027	−20.015	−19.586	−18.650	−17.023	−16.826
AIC	56.055	56.031	53.172	53.300	48.045	49.652
BIC	69.847	71.793	65.818	67.753	60.040	63.360

Note: *p<0.1; **p<0.05; ***p<0.01

Additional tables detailing findings from Site 1 are compiled in Chapter 10: Appendix 1, including results of different permutations of the longitudinal formative assessment OLS regressions, contingency tables of the summative assessment proficiency levels, and different permutations of the summative assessment OLS and logistic regression models.

7 Site 2 Case Study: Two Co-Located Schools Serving Elementary and Intermediate Grades

Site 2 comprises two Native nation schools serving elementary through intermediate grades, both located in the same school facility on a single campus in a rural area. The combined elementary and intermediate enrollment of the two schools is approximately 300. One of the schools is an autonomous Indigenous-language immersion school with its own administration, faculty, and staff; the other is an administratively separate English-medium school.

7.1 Data Structure and Descriptive Statistics for the Analytic Sample

The following comparative analyses of the ILI and English-medium programs at Site 2 are based on a sample of 73 students—26 in the ILI program and 47 students of Indigenous heritage in the English-medium track—who belong to three cohorts that started kindergarten between the 2013-14 and 2015-16 school years.

The data for Site 2 include students' program status (ILI or English-medium) and gender, as well as English-language formative test scores in reading and math. The formative assessments were administered three times per year—each fall, winter, and spring—from 4th grade onward for both groups of students, and cover the following grade levels, ending with the 2021-2022 school year:

- 1st cohort: through 8th grade;
- 2nd cohort: through 7th grade;
- 3rd cohort: through 6th grade.

The sample for the comparative analyses is restricted to include only students from either program with non-missing fall 4th grade formative assessment scores, in order to include those early scores as covariates in models predicting later test scores.

The data for Site 2 also include Indigenous-language oral proficiency assessment results for the ILI students only.

A slight majority of the English-medium students were female (57%), but the ILI students were predominantly male (62%). Most of the gender discrepancy in both programs was concentrated in the first cohort, *i.e.*, the oldest group of students. The subsequent cohorts were much more closely balanced on gender. Table 19 shows the number of male and female students by cohort in each program.

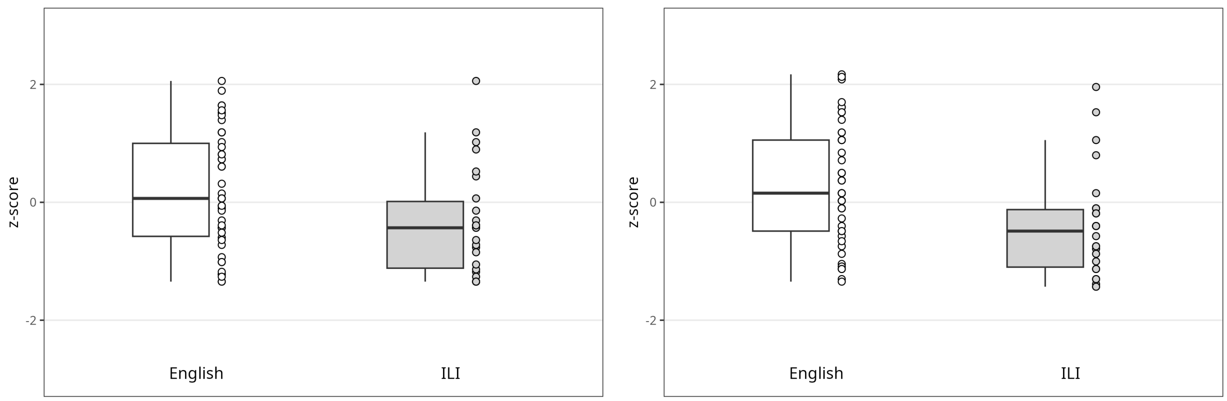
The earliest time point at which English-language formative assessment scores in reading and math are available for both groups of students is in 4th grade. The distributions of the English-medium and ILI students' fall 4th grade formative assessment z-scores are shown in Figure 7. On average, the English-medium students scored higher on both reading and math than the ILI students, and the differences in mean

Table 19: Cross-tabulation of gender by cohort and program at Site 2

cohort	female	male	**total**
English			
1)	11	5	**16**
2)	9	9	**18**
3)	7	6	**13**
total	**27**	**20**	**47**
ILI			
1)	1	9	**10**
2)	4	4	**8**
3)	5	3	**8**
total	**10**	**16**	**26**

z-scores are statistically significant. (For reading, $\bar{z}_{\text{EM}} - \bar{z}_{\text{ILI}} = 0.549; t = 2.383; \text{df} = 56.253; p\text{-value} = 0.021$; and for math, $\bar{z}_{\text{EM}} - \bar{z}_{\text{ILI}} = 0.700; t = 3.080; \text{df} = 55.654; p\text{-value} = 0.003$.) Notably, however, 4th grade is typically the first time that the ILI students at Site 2 encounter these English-language formative tests, whereas the English-medium students start taking them in 3rd grade or earlier. So called "practice effects" (Greene, 1941; Wing, 1980), which refer to a boost in students' scores resulting from familiarity with the test format rather than increased proficiency in the reading or math content that the test ostensibly measures, could be driving some of the difference between the two groups. Formative assessments similar to the ones used at Site 2 have been known to exhibit some degree of measurement bias due to practice effects (Shepard, 2017). Another important factor is that these assessments are also administered in the English-medium students' language of instruction and are more closely aligned with the English-medium curriculum. Even if the English-medium teachers at Site 2 did not deliberately "teach to the test," these assessments would certainly have been much more relevant and familiar to the English-medium students' school experience than they would have been for the ILI students.

Figure 8 shows the fall 4th grade English reading and math formative assessment scores for ILI and English-medium students at Site 2, disaggregated by cohort. The distributions of scores are similar for the English-medium students in all three cohorts and for the ILI students in cohorts 1 and 2. The ILI students in the youngest group, cohort 3, scored slightly higher on average than the preceding two cohorts of ILI students. But the median scores were lower for ILI students than English-medium students for both subjects in all three cohorts.



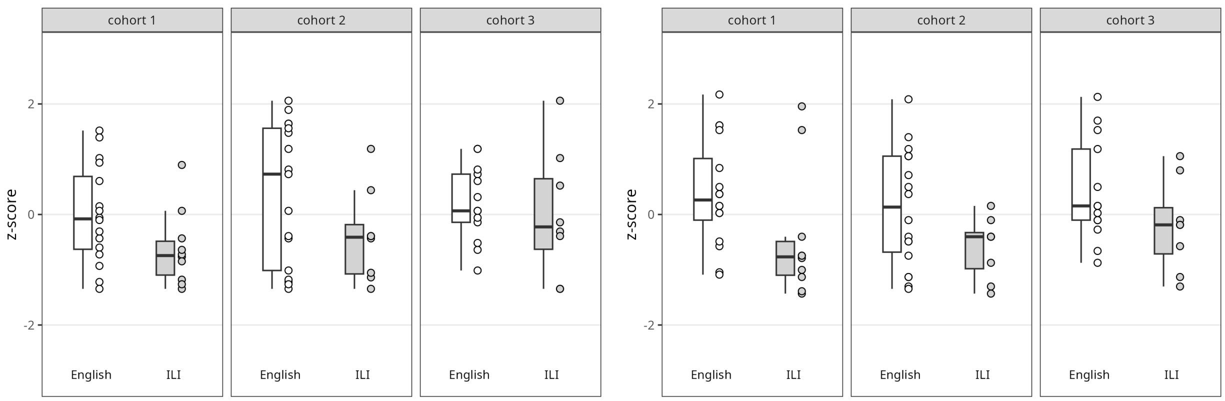
(a) reading (b) math

Figure 7: Fall 4th grade formative assessment z-scores in (a) reading and (b) math among ILI and Indigenous-heritage English-medium students at Site 2.

Note: The English-medium group is shaded white and the Indigenous-language immersion group is shaded gray. Dots represent individual students' z-scores. The adjacent box-and-whisker plots summarize each group's distribution of scores, where the lower horizontal border corresponds to the first quartile (25th percentile); the middle horizontal line corresponds to the median (second quartile or 50th percentile); and the upper horizontal border corresponds to the third quartile (75th percentile).

7.2 Formative Assessment Results: Longitudinal Ordinary Least Squares Regression Analysis

Table 20 shows the results of ordinary least squares (OLS) regressions of fall 4th grade z-scores in English reading (column 1) and math (column 2) on program and gender. After accounting for gender, the coefficient estimates for ILI are both negative. Although program status is only statistically significant when math is the outcome, it is close to the margin of statistical significance for English reading. The 95% confidence intervals of the point estimates for ILI range from -0.940 to 0.007 for reading and -1.121 to -0.179 for math.



(a) reading (b) math

Figure 8: Fall 4th grade formative assessment z-scores in (a) reading and (b) math among ILI and Indigenous-heritage English-medium students at Site 2, disaggregated by cohort.

Note: The English-medium group is shaded white and the Indigenous-language immersion group is shaded gray. Dots represent individual students' z-scores. The adjacent box-and-whisker plots summarize each group's distribution of scores, where the lower horizontal border corresponds to the first quartile (25th percentile); the middle horizontal line corresponds to the median (second quartile or 50th percentile); and the upper horizontal border corresponds to the third quartile (75th percentile).

Table 20: OLS regressions of standardized English reading and math test scores at Site 2, fall 4th grade

	reading	math
	(1)	(2)
intercept	0.359**	0.405**
	(0.169)	(0.169)
ILI	−0.466*	−0.650***
	(0.237)	(0.236)
male	−0.406*	−0.261
	(0.228)	(0.226)
Observations	72	73
R^2	0.112	0.129
Adjusted R^2	0.086	0.104
Residual Std. Error	0.949	0.951
F Statistic	4.361**	5.196***
Note:	*p<0.1; **p<0.05; ***p<0.01	

Kindergarten or pre-kindergarten English-language standardized test scores are not available for either group of students at Site 2, so we are not able to make direct comparisons of emergent literacy or numeracy skills for these students upon entry to elementary school, as we did with the students at Site 1. However, students' scores in fall of 4th grade are strongly predictive of their scores on the same set of assessments in later grades, and we can include these initial scores along with gender as covariates in OLS regressions of the outcomes at each subsequent testing instance for which scores are available between winter of 4th grade and spring of 8th grade. Tables 21, 22, and 23 show the results of these longitudinal models for English reading and Tables 24, 25, and 26 show the corresponding models with math test scores as the outcome. (Additional tables detailing results of different permutations of the longitudinal formative assessment OLS regressions from Site 2 are compiled in Chapter 11: Appendix 2.) We can interpret the coefficient estimates for ILI in these models as the association between persistence in ILI and formative test scores, relative to English-medium instruction, after accounting for fall 4th grade achievement level and gender.

Table 21: OLS regressions of standardized English reading formative test scores at Site 2, winter 4th to spring 5th grade

	winter 4th	spring 4th	fall 5th	winter 5th	spring 5th
	(1)	(2)	(3)	(4)	(5)
intercept	−0.072	0.218*	0.459**	0.194	0.277
	(0.110)	(0.125)	(0.181)	(0.137)	(0.167)
ILI	0.124	−0.144	0.030	−0.269	0.074
	(0.157)	(0.173)	(0.242)	(0.205)	(0.254)
fall 4th ELA z-score	0.785***	0.676***	0.612***	0.625***	0.897***
	(0.079)	(0.086)	(0.141)	(0.116)	(0.144)
male	−0.032	−0.298*	−0.854***	−0.312	−0.345
	(0.145)	(0.170)	(0.246)	(0.226)	(0.277)
Observations	66	48	33	25	24
R^2	0.633	0.680	0.635	0.765	0.761
Adjusted R^2	0.615	0.658	0.598	0.731	0.725
Residual Std. Error	0.573	0.543	0.625	0.420	0.514
F Statistic	35.622***	31.186***	16.852***	22.783***	21.229***
Note:					*p<0.1; **p<0.05; ***p<0.01

Table 22: OLS regressions of standardized English reading formative test scores at Site 2, fall 6th to spring 7th grade

	fall 6th	winter 6th	spring 6th	fall 7th	winter 7th	spring 7th
	(1)	(2)	(3)	(4)	(5)	(6)
intercept	−0.134	−0.311	−0.434	−0.180	−0.268	−0.288
	(0.207)	(0.295)	(0.323)	(0.244)	(0.238)	(0.269)
ILI	0.079	0.678	0.635	0.159	0.228	0.164
	(0.255)	(0.406)	(0.444)	(0.362)	(0.354)	(0.399)
fall 4th ELA z-score	0.837***	0.781***	0.652**	0.681***	0.688***	0.636***
	(0.142)	(0.225)	(0.246)	(0.150)	(0.149)	(0.169)
male	0.039	0.206	0.353	−0.206	0.051	0.007
	(0.245)	(0.422)	(0.461)	(0.326)	(0.335)	(0.375)
Observations	27	18	18	20	19	17
R^2	0.611	0.517	0.409	0.615	0.616	0.548
Adjusted R^2	0.560	0.413	0.282	0.542	0.539	0.444
Residual Std. Error	0.615	0.837	0.915	0.686	0.671	0.726
F Statistic	12.019***	4.986**	3.226*	8.509***	8.007***	5.259**
Note:						*p<0.1; **p<0.05; ***p<0.01

Table 23: OLS regressions of standardized English reading formative test scores at Site 2, fall to spring 8th grade

	fall 8th	winter 8th	spring 8th
	(1)	(2)	(3)
intercept	0.198	0.292	0.628***
	(0.280)	(0.348)	(0.200)
ILI	0.386	0.177	−0.016
	(0.431)	(0.510)	(0.316)
fall 4th ELA z-score	0.910***	0.747*	0.895***
	(0.286)	(0.352)	(0.207)
male	0.046	−0.051	−0.579
	(0.463)	(0.549)	(0.330)
Observations	20	19	19
R^2	0.441	0.282	0.738
Adjusted R^2	0.336	0.138	0.685
Residual Std. Error	0.813	0.963	0.580
F Statistic	4.209**	1.964	14.056***
Note:			*p<0.1; **p<0.05; ***p<0.01

After accounting for fall 4th grade English reading score and gender, we do not see statistically significant differences in the expected English reading z-scores of ILI students relative to English-medium students at Site 2 between winter of 4th grade and spring of 8th grade. Notably, although none of the coefficient estimates for ILI are statistically significant, only three of the point estimates are negative; the other eleven point estimates all trend in a positive direction.

Table 24: OLS regressions of standardized math formative test scores at Site 2, winter 4th to spring 5th grade

	winter 4th	spring 4th	fall 5th	winter 5th	spring 5th
	(1)	(2)	(3)	(4)	(5)
intercept	−0.140	−0.036	−0.262	−0.233	−0.121
	(0.093)	(0.119)	(0.204)	(0.154)	(0.169)
ILI	0.313**	0.205	0.492*	0.077	0.271
	(0.137)	(0.166)	(0.268)	(0.205)	(0.239)
fall 4th math z-score	0.896***	0.856***	0.690***	0.689***	0.709***
	(0.064)	(0.079)	(0.128)	(0.091)	(0.099)
male	−0.059	−0.143	−0.375	−0.070	−0.353
	(0.121)	(0.159)	(0.261)	(0.216)	(0.251)
Observations	67	50	33	26	24
R^2	0.772	0.754	0.571	0.780	0.788
Adjusted R^2	0.761	0.738	0.526	0.750	0.757
Residual Std. Error	0.486	0.522	0.666	0.447	0.479
F Statistic	71.235***	47.011***	12.854***	26.034***	24.829***

Note: *p<0.1; **p<0.05; ***p<0.01

Table 25: OLS regressions of standardized math formative test scores at Site 2, fall 6th to spring 7th grade

	fall 6th	winter 6th	spring 6th	fall 7th	winter 7th	spring 7th
	(1)	(2)	(3)	(4)	(5)	(6)
intercept	−0.674***	−0.646***	−0.814***	−0.206	−0.166	−0.131
	(0.167)	(0.162)	(0.211)	(0.157)	(0.185)	(0.186)
ILI	0.750***	0.703***	1.290***	0.415	0.653**	0.318
	(0.212)	(0.234)	(0.304)	(0.240)	(0.284)	(0.282)
fall 4th math z-score	0.568***	0.751***	0.559***	0.842***	1.048***	1.125***
	(0.105)	(0.137)	(0.178)	(0.116)	(0.137)	(0.138)
male	0.065	−0.044	0.197	−0.157	0.033	0.198
	(0.199)	(0.228)	(0.296)	(0.219)	(0.268)	(0.264)
Observations	27	18	18	20	19	17
R^2	0.605	0.691	0.606	0.786	0.804	0.840
Adjusted R^2	0.553	0.625	0.522	0.745	0.765	0.803
Residual Std. Error	0.496	0.453	0.589	0.466	0.548	0.521
F Statistic	11.732***	10.437***	7.191***	19.536***	20.480***	22.760***

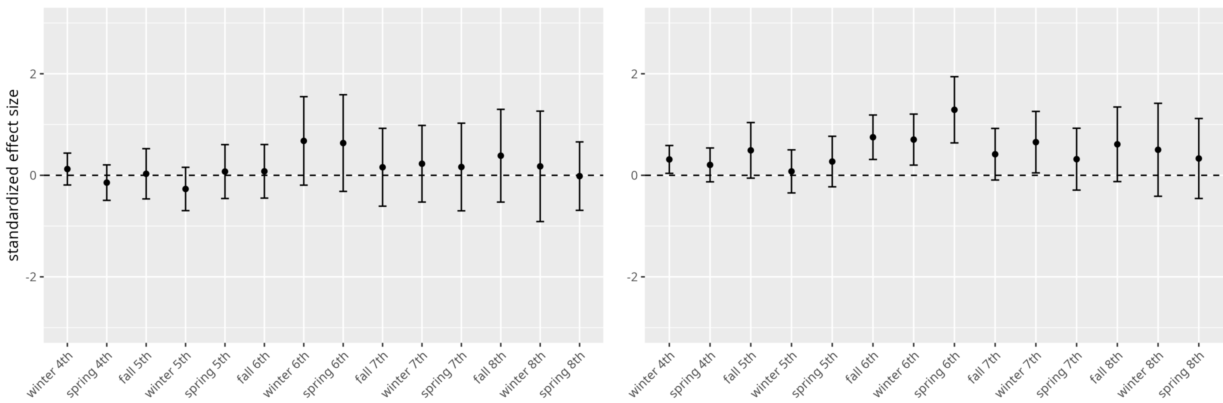
Note: *p<0.1; **p<0.05; ***p<0.01

Table 26: OLS regressions of standardized math formative test scores at Site 2, fall to spring 8th grade

	fall 8th	winter 8th	spring 8th
	(1)	(2)	(3)
intercept	0.590**	0.418	0.385
	(0.254)	(0.310)	(0.242)
ILI	0.611*	0.503	0.331
	(0.345)	(0.430)	(0.370)
fall 4th math z-score	0.481**	0.519**	0.658***
	(0.181)	(0.219)	(0.183)
male	−1.205***	−0.861*	−0.794**
	(0.357)	(0.440)	(0.359)
Observations	19	19	19
R^2	0.672	0.498	0.649
Adjusted R^2	0.606	0.398	0.579
Residual Std. Error	0.639	0.799	0.670
F Statistic	10.246***	4.963**	9.254***
Note:			*p<0.1; **p<0.05; ***p<0.01

After accounting for fall 4th grade math score and gender, we see evidence that ILI persistence has a statistically significant and positive association with math formative assessment z-scores relative to English-medium instruction at Site 2, at least for certain time periods (winter 4th grade, fall/winter/spring 6th grade, and winter 7th grade). And although the other nine point estimates for ILI are not statistically significant at the 95% confidence level, all of them are positive.

The point estimates and confidence intervals of the coefficients on ILI from the longitudinal regressions in Tables 21, 22, 23, 24, 25, and 26 are plotted in Figure 9. The point estimates for separate regressions for female and male students are plotted in Figures 10 and 11, respectively.

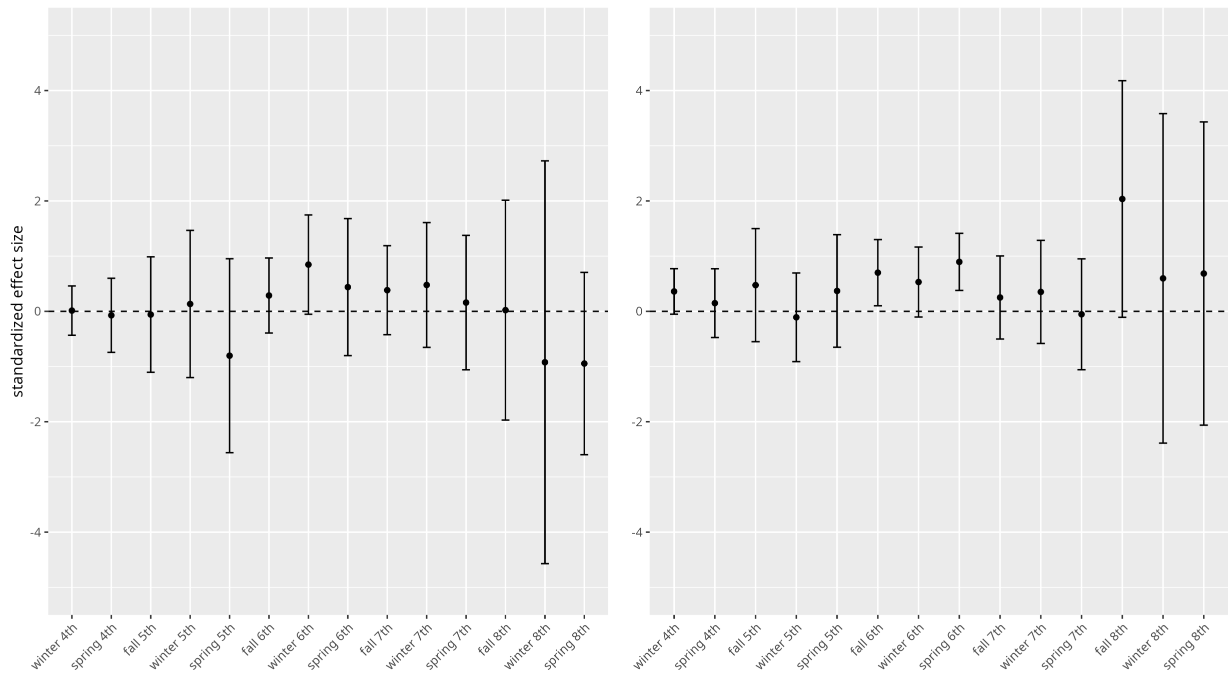


(a) English reading (all students)

(b) math (all students)

Figure 9: Coefficient estimates and 95% confidence intervals of persistence in ILI on English reading and math formative assessments in elementary and middle school at Site 2, after accounting for fall 4th grade English reading/math test scores and gender.

Note: Coefficient estimates at each time point are represented as dots. 95% confidence intervals appear as lines, with the vertical line segments representing the spans of the confidence intervals and the horizontal line segments representing the upper and lower bounds of the confidence intervals. The horizontal dashed line corresponds to zero, *i.e.*, an estimate of no difference between ILI students' and English-medium students' achievement levels after accounting for fall 4th grade test scores and gender.



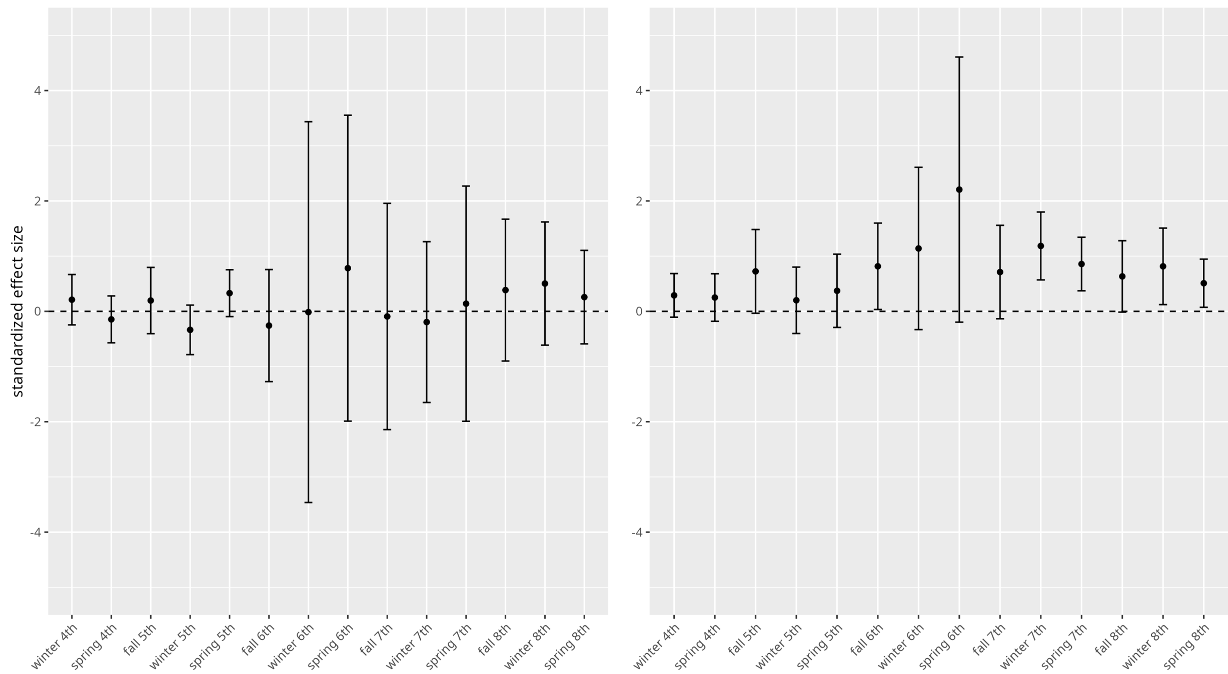
(a) English reading (female students only)

(b) math (female students only)

Figure 10: Coefficient estimates and 95% confidence intervals of persistence in ILI among female students on English reading and math formative assessments in elementary and middle school at Site 2, after accounting for fall 4th grade English reading/math test scores.

Note: Coefficient estimates at each time point are represented as dots. 95% confidence intervals appear as lines, with the vertical line segments representing the spans of the confidence intervals and the horizontal line segments representing the upper and lower bounds of the confidence intervals. The horizontal dashed line corresponds to zero, *i.e.*, an estimate of no difference between female ILI and English-medium students' achievement levels after accounting for fall 4th grade test scores.

Some of the confidence intervals in Figures 10 and 11 are very wide, and in these cases, it would not be justifiable to draw strong inferences specific to male or female students. That said, the majority of the point estimates are near (or slightly above) zero for boys and girls in both English reading and math. Also, the point estimates for boys in math in 7th-8th grades are consistently positive, and nearly all of them are statistically significant if not on the margin.



(a) English reading (male students only) (b) math (male students only)

Figure 11: Coefficient estimates and 95% confidence intervals of persistence in ILI among male students on English reading and math formative assessments in elementary and middle school at Site 2, after accounting for fall 4th grade English reading/math test scores.

Note: Coefficient estimates at each time point are represented as dots. 95% confidence intervals appear as lines, with the vertical line segments representing the spans of the confidence intervals and the horizontal line segments representing the upper and lower bounds of the confidence intervals. The horizontal dashed line corresponds to zero, *i.e.*, an estimate of no difference between male ILI and English-medium students' achievement levels after accounting for fall 4th grade test scores.

7.3 Formative Assessment Results: Hierarchical Linear Mixed Effects Analysis with Random Intercepts

As with the other two case study sites, the nested structure of the formative assessment data lends itself to a mixed effects modeling approach where we allow the intercept term to vary randomly at the student level and estimate fixed effects for the student-level variables and a time period indicator, as expressed in the following set of equations:

$$y_{it} = \hat{\gamma_0} + \hat{\gamma}_1(\text{ILI}_{i.}) + \hat{\gamma}_2(\text{fall 4th z-score}_{i.}) + \hat{\gamma}_3(\text{male}_{i.}) + \hat{\gamma}_4(\text{ILI}_{i.} \times \text{male}_{i.}) + r_{i.} + e_{it}. \tag{6a}$$

$$\widehat{\text{Var}}[r_{i.}] = \hat{\tau}; \tag{6b}$$

$$\widehat{\text{Var}}[e_{it}] = \hat{\sigma}^2. \tag{6c}$$

Tables 27 and 28 show the results from various permutations of this model where the respective outcomes are student i's score at trimester t in English reading or math, starting with a simple one-way ANOVA with random effects (model 1), and iteratively adding student-level variables for ILI participation, fall 4th grade z-score in reading or math, gender (male), an ILI $\times$ male interaction term, a trimester variable corresponding to the expected change in scores from one assessment instance to the next, and finally an ILI $\times$ trimester interaction term to isolate the difference in change-over-time slopes for ILI students relative to English-medium students.

Table 27: Random intercept models: standardized English reading formative test scores at Site 2

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Fixed effects:							
intercept	−0.059	0.054	−0.049	0.013	0.003	−0.063	−0.037
	(0.104)	(0.129)	(0.078)	(0.095)	(0.104)	(0.109)	(0.112)
ILI		−0.310	0.060	0.080	0.111	0.114	0.041
		(0.214)	(0.131)	(0.132)	(0.195)	(0.195)	(0.206)
fall 4th reading z-score			0.720***	0.704***	0.705***	0.707***	0.706***
			(0.065)	(0.066)	(0.067)	(0.067)	(0.067)
male				−0.144	−0.122	−0.112	−0.116
				(0.127)	(0.161)	(0.161)	(0.161)
trimester						0.013**	0.008
						(0.006)	(0.008)
ILI $\times$ male					−0.058	−0.057	−0.046
					(0.260)	(0.261)	(0.260)
ILI $\times$ trimester							0.014
							(0.013)
Variance components:							
Random Intercept ($\hat{\tau}$)	0.688	0.676	0.203	0.201	0.205	0.207	0.205
Level-1 Residual ($\hat{\sigma}^2$)	0.257	0.257	0.258	0.258	0.258	0.255	0.256
Data structure:							
Number of Students	69	69	69	69	69	69	69
Number of Observations	373	373	373	373	373	373	373
Selection criteria:							
AIC	743.325	744.484	680.549	683.559	686.371	692.668	700.308
BIC	755.089	760.170	700.157	707.089	713.822	724.040	735.602

Note: *p<0.1; **p<0.05; ***p<0.01

Table 28: Random intercept models: standardized math formative test scores at Site 2

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Fixed effects:							
intercept	−0.092	−0.011	−0.230***	−0.191**	−0.157	−0.232**	−0.204*
	(0.110)	(0.137)	(0.076)	(0.094)	(0.102)	(0.106)	(0.107)
ILI		−0.225	0.362***	0.372***	0.255	0.258	0.179
		(0.229)	(0.131)	(0.132)	(0.193)	(0.194)	(0.200)
fall 4th math z-score			0.812***	0.805***	0.801***	0.805***	0.802***
			(0.063)	(0.064)	(0.064)	(0.065)	(0.065)
male				−0.086	−0.163	−0.148	−0.153
				(0.122)	(0.153)	(0.153)	(0.153)
trimester						0.015***	0.009
						(0.005)	(0.006)
ILI × male					0.211	0.208	0.221
					(0.252)	(0.254)	(0.253)
ILI × trimester							0.016
							(0.010)
Variance components:							
Random Intercept ($\hat{\tau}$)	0.808	0.809	0.214	0.215	0.216	0.220	0.218
Level-1 Residual ($\hat{\sigma}^2$)	0.142	0.142	0.142	0.142	0.142	0.138	0.137
Data structure:							
Number of Students	70	70	70	70	70	70	70
Number of Observations	376	376	376	376	376	376	376
Selection criteria:							
AIC	575.913	578.059	499.366	503.241	505.456	506.899	513.844
BIC	587.702	593.777	519.014	526.818	532.963	538.336	549.210
Note:							*p<0.1; **p<0.05; ***p<0.01

The one-way ANOVA models in column 1 of each table provide an estimate of how much of the total outcome variance is located between students: 73% for English reading and 85% for math. The indicator for ILI persistence, added in column 2, explains a negligible amount of the between-student variance for either subject, and neither of the fixed effects coefficients is statistically significant when we don't account for fall 4th grade scores or gender.

Fall 4th grade z-scores (added in column 3) explain 70% of the between-student variance in English reading and 74% in math. The fixed effects estimates for the fall 4th grade scores are large (0.72 for English reading and 0.81 for math) and statistically significant. In other words, fall 4th grade scores are strongly predictive of students' scores on the subsequent assessments.

When we account for fall 4th grade scores, the coefficient for ILI persistence is approximately zero and not statistically significant for English reading, but it is statistically significant and positive (+0.36) for math, indicating that ILI persistence is associated with an increase of approximately one-third of a standard deviation in math formative assessment scores after accounting for 4th grade test results. In both cases—English reading and math—the Akaike and Bayesian information criteria (AIC and BIC) suggest that this parametrization (column 3) strikes the optimal balance between fit and parsimony among the various permutations considered in Tables 27 and 28. Adding gender as a covariate does not meaningfully change the coefficient for ILI or explain any substantial residual variance. And neither the ILI $\times$ male nor the ILI $\times$ trimester coefficient estimates are statistically significant for either subject. In other words, there is no apparent evidence that the association between ILI persistence and English reading or math scores, relative to English-medium instruction, differs by gender or over time between the two programs.

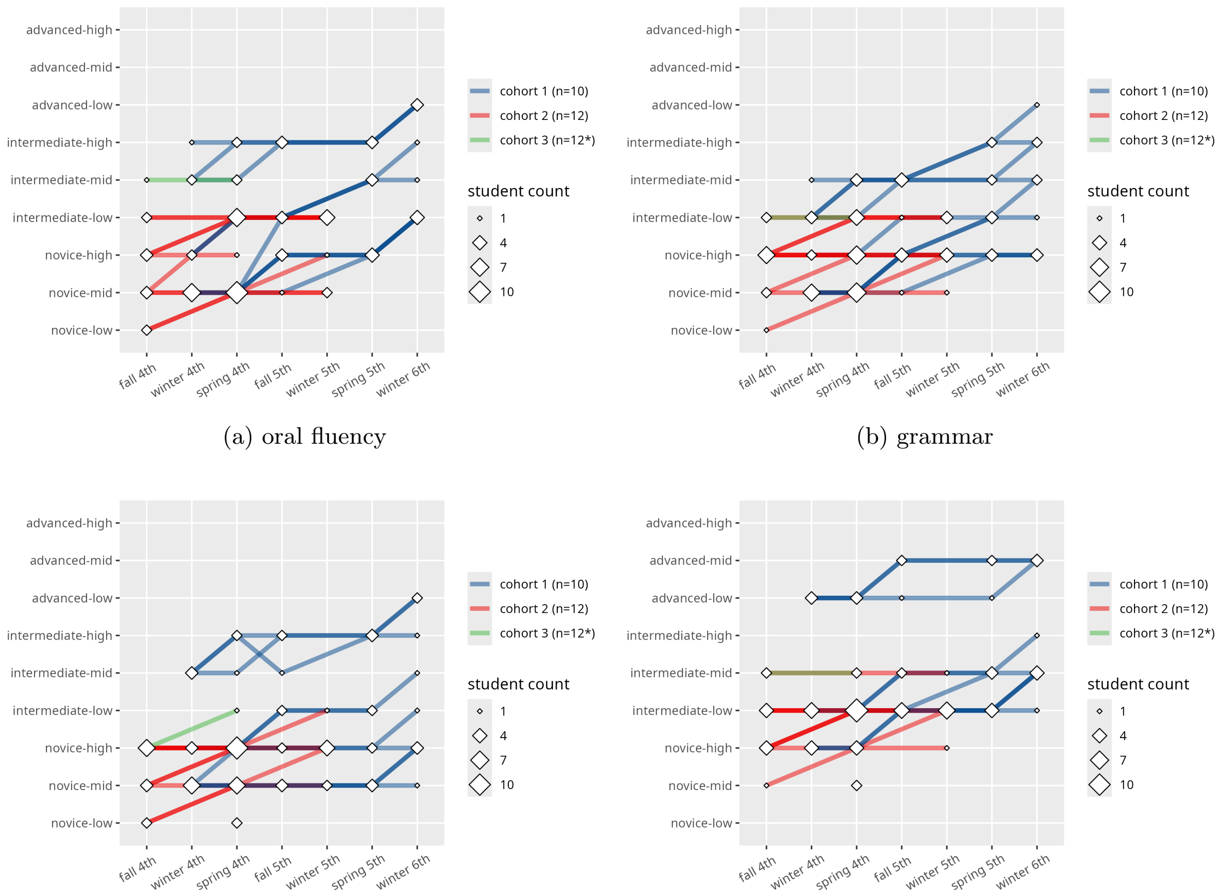
7.4 Indigenous-Language Assessment Results

ILI students at Site 2 are periodically assessed on their developing proficiency in the Indigenous language used at their school. This assessment covers four verbal proficiency domains: oral fluency, grammar, vocabulary, and listening comprehension. The rubric for this assessment is adapted from Swender et al. (2012) and Center for Applied Linguistics (1999) and classifies speakers according to a nine-point scale ranging from "novice-low/medium/high" to "intermediate-low/medium/high" to "advanced-low/medium/high."

According to the rubric, novice speakers can demonstrate consistent production and comprehension of memorized speech patterns, exhibiting less success with complex sentence structures or verb forms. Intermediate speakers are characterized as having an ability to express opinions on familiar topics in complete sentences and to be capable of comprehending speech on unfamiliar topics with some contextual support. However, intermediate speakers may struggle with unfamiliar topics as well as complex syntactic or narrative structures. Advanced speakers, on the other hand, are capable of expressing and justifying opinions on abstract topics using complex sentences organized into paragraph-length discourse with fluency approaching a native speaker's level.

Figure 12 shows the results of the Indigenous oral proficiency assessment between 4th and 6th grades from March 2018 to March 2020 for students who belonged to the three cohorts discussed in the preceding section. (A few students in cohorts 2 and 3 who were omitted from the comparative analyses above due to missing fall 4th grade formative assessment scores are included here; hence the slightly larger headcounts for those subgroups in Figure 12.)

In 4th grade, most ILI students were classified as having "novice-low" to "intermediate-low" proficiency in oral fluency, grammar, vocabulary, and listening comprehension. By the middle of 5th grade, nearly all of the students in the oldest two cohorts had attained at least "novice-high" higher levels of proficiency in all four domains, with many students attaining intermediate-level proficiency in at least one domain. By the middle of 6th grade, all of the students in cohort 1 had attained intermediate-level proficiency on the oral fluency and listening comprehension dimensions, with most students also demonstrating the same degree of proficiency in grammar and vocabulary as well.



(a) oral fluency

(b) grammar

(c) vocabulary

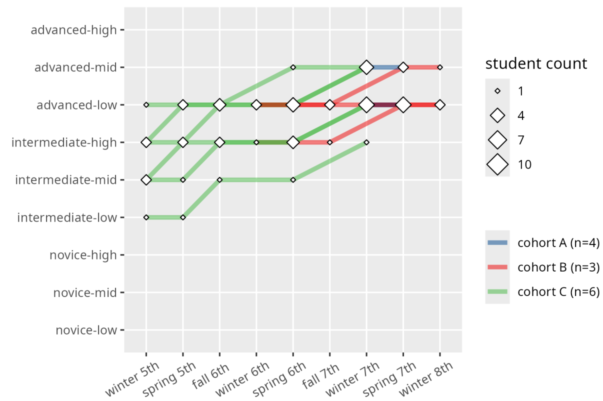
(d) listening comprehension

Figure 12: Indigenous-language proficiency assessment levels among ILI students at Site 2 in oral fluency, grammar, vocabulary, and listening comprehension in 4th-6th grades.

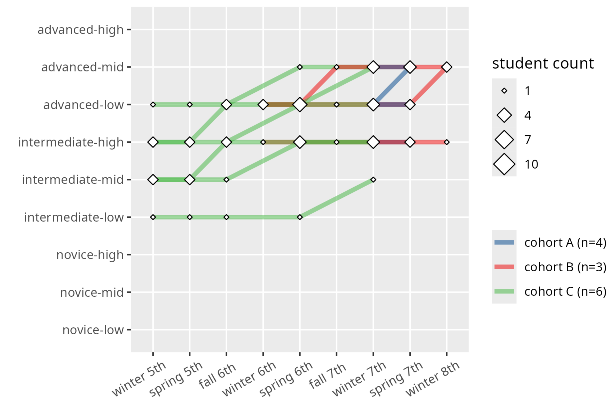
Note: Individual students' progress over time are represented by connecting lines, with different colors corresponding to each cohort. The number of students progressing from one level to the next is represented by the darkness of the line: darker lines correspond to more students; lighter lines correspond to fewer students sharing that path.

*Due to the COVID-19 school closure, only 3 out of 12 students in the youngest cohort were assessed.

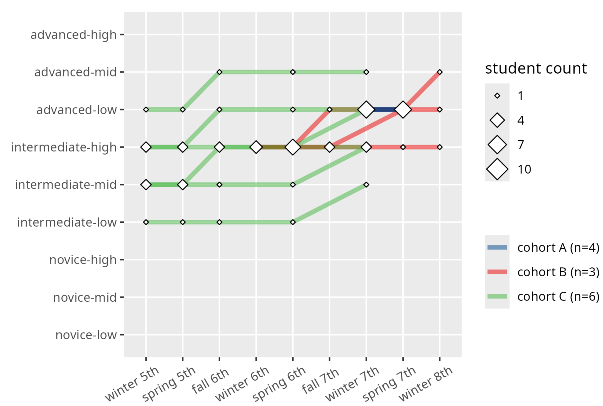
The data for Site 2 also include Indigenous-language proficiency results from three older cohorts of students who were in 5th through 8th grade between March 2018 and March 2020, displayed in Figure 13. As of the middle of 5th grade, all of these students exhibited at least "intermediate-low" proficiency in all four domains, with most students at the "intermediate-mid" level or higher. By the end of 7th grade, all of the students were at the "intermediate-high"—if not "advanced" proficiency in all four domains. And all of the students had reached the advanced level in listening comprehension by the end of 6th grade.



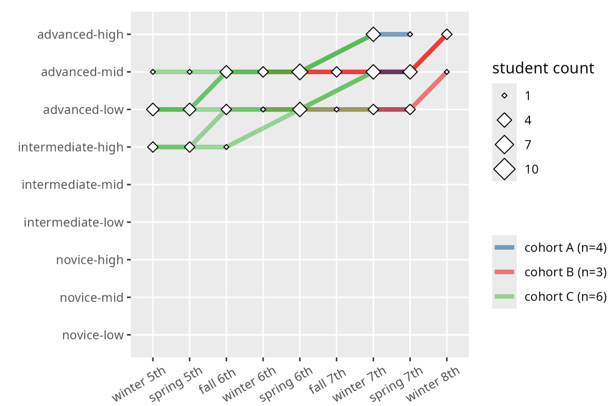
(a) oral fluency



(b) grammar



(c) vocabulary



(d) listening comprehension

Figure 13: Indigenous-language proficiency assessment levels among ILI students at Site 2 in oral fluency, grammar, vocabulary, and listening comprehension in 5th-8th grades.

Note: Individual students' progress over time are represented by connecting lines, with different colors corresponding to each cohort. The number of students progressing from one level to the next is represented by the darkness of the line: darker lines correspond to more students; lighter lines correspond to fewer students sharing that path.

Reviewing the oral proficiency assessment results, a clear pattern emerges: Over time, the students from all six cohorts at Site 2 exhibited consistent maintenance and growth in their Indigenous-language proficiency in all four domains. The singular exception was one student who briefly dipped from "intermediate-high" to "intermediate-mid" on the vocabulary component between the end of 4th grade and the beginning of 5th grade (see Figure 12(c)). By spring of 5th grade, however, this student was back at the "intermediate-high" level and by winter of 6th grade, they had progressed to the "advanced-low" level on the vocabulary assessment.

8 Site 3 Case Study: A Middle School with ILI and English-Medium Tracks

Site 3 is a semi-urban intermediate school serving approximately 700 students in grades 6 through 8, where the Indigenous-language immersion program shares the same campus, facilities, and administrators as the English-medium comparison program. Students at Site 3 are typically alumni of one of several elementary feeder schools in the area, one of which offers an ILI track through 5th grade. The other four elementary schools provide English-medium instruction only.

8.1 Data Structure and Descriptive Statistics for the Analytic Sample

The analytic sample for Site 3 consists of a cohort of 92 students—27 in the ILI program and 65 students of Indigenous heritage in the English-medium track—who started 6th grade in fall 2017.

The data for Site 3 include students' program status (ILI or English-medium), gender, socioeconomic status (where "low SES" corresponds to free/reduced price lunch eligibility), as well as students' scores on English language arts and mathematics formative assessments administered once per academic term from grades 6 through 8. The formative assessments that Site 3 uses are administered in English only, but students at Site 3 also normally take yearly summative assessments which have been administered in the students' language of instruction since the 2018-2019 school year (English for the English-medium students and the Indigenous language for the ILI students). The English-medium students in this cohort took the English-language summative assessment in 6th and 7th grade. The ILI students were only tested in 7th grade because the Indigenous-language summative assessment was still under development while they were in 6th grade. No students took the 8th grade summative assessments due to the COVID-19 school closures in spring 2020.

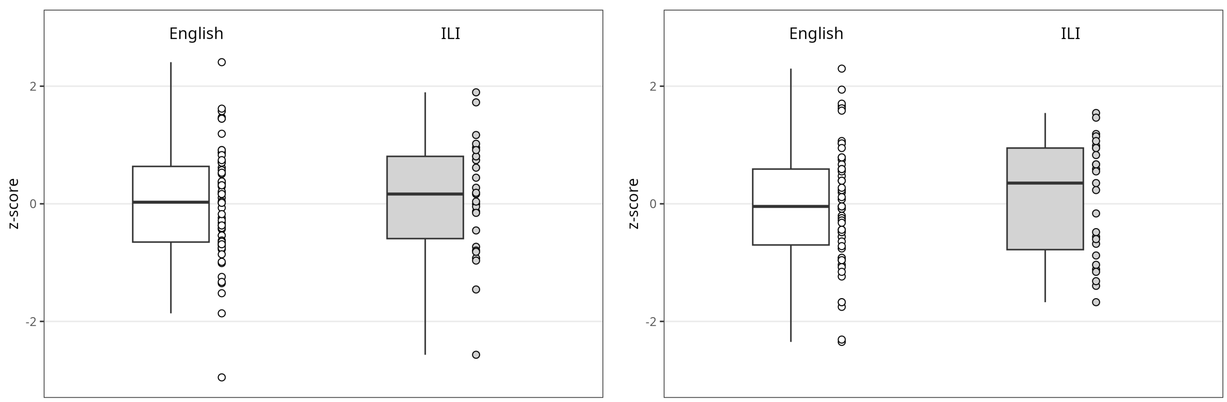
ILI students from this cohort at Site 3 are majority female (63%) and approximately evenly split between low SES (48%) and non-low SES (52%). The Indigenous-heritage English-medium students, in contrast, are majority low SES (57%) and roughly evenly split between female (48%) and male (52%). Table 29 shows the number of English-medium and ILI students in each SES and gender subgroup.

The distributions of the English-medium and ILI students' formative assessment z-scores are approximately similar, relative to each other, at the beginning of 6th grade. As depicted in Figure 14, the ranges of English-medium students' z-scores in ELA and math are slightly wider than those of the ILI students, and the median z-scores of the ILI students are slightly higher than those of the English-medium students. However, the differences in mean z-scores between the two groups are not statistically significant. (For ELA, $\bar{z}_{\text{EM}} - \bar{z}_{\text{ILI}} = -0.142; t = -0.623; \text{df} = 49.373; p\text{-value} = 0.536$; and for math, $\bar{z}_{\text{EM}} - \bar{z}_{\text{ILI}} = -0.113; t = -0.492; \text{df} = 50.195; p\text{-value} = 0.625$.) In other words, the students who attended the ILI elementary program began 6th grade with roughly equivalent achievement levels in English language

Table 29: Cross-tabulation of SES by gender and program at Site 3

	female	male	**total**
English			
non-low SES	11	17	**28**
low SES	20	17	**37**
total	**31**	**34**	**65**
ILI			
non-low SES	9	5	**14**
low SES	8	5	**13**
total	**17**	**10**	**27**

arts and math as their Indigenous-heritage peers who received English-only instruction in grades K-5. However, as these assessments are administered in English—not the ILI students' language of instruction—they may understate the ILI students' actual math and verbal achievement levels.



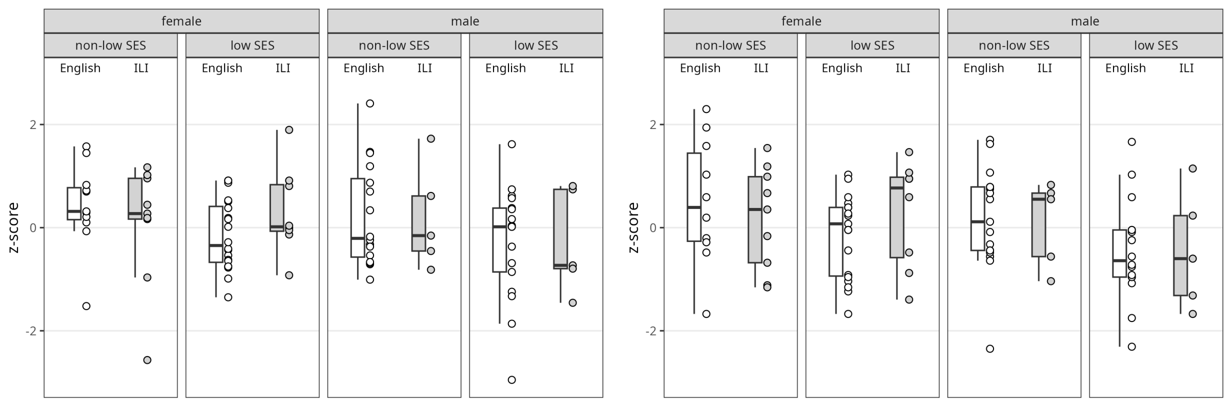
(a) English language arts

(b) math

Figure 14: Fall 6th grade formative assessment z-scores in (a) English language arts and (b) math among ILI and Indigenous-heritage English-medium students at Site 3.

Note: The English-medium group is shaded white and the Indigenous-language immersion group is shaded gray. Dots represent individual students' z-scores. The adjacent box-and-whisker plots summarize each group's distribution of scores, where the lower horizontal border corresponds to the first quartile (25th percentile); the middle horizontal line corresponds to the median (second quartile or 50th percentile); and the upper horizontal border corresponds to the third quartile (75th percentile).

Since the data for Site 3 include information on students' gender and socioeconomic status, we can compare the relative achievement levels of English-medium and ILI students within each demographic subgroup. Figure 15 shows the students' z-score distributions disaggregated by gender and SES. In English language arts, the median z-scores for non-low SES girls and boys are approximately the same for both groups, slightly higher for ILI students than English-medium students among low SES girls, and slightly lower for ILI students than English-medium students among low SES boys. In math, the median z-scores are approximately the same for both groups among non-low SES girls and low SES boys, and slightly higher for ILI students than English-medium students among low SES girls and non-low SES boys. In all cases, there is substantial overlap between both groups in the z-scores corresponding to the 25th-75th percentiles (represented by the rectangular segments of the box-and-whisker plots). Even after disaggregating the data by gender and SES, there do not appear to be substantial subgroup-specific differences between the ILI and Indigenous-heritage English-medium students' achievement levels in English language arts and math at the beginning of 6th grade.



(a) English language arts (b) math

Figure 15: Fall 6th grade formative assessment z-scores in (a) English language arts and (b) math among ILI and Indigenous-heritage English-medium students at Site 3, disaggregated by gender and socioeconomic status.

Note: The English-medium group is shaded white and the Indigenous-language immersion group is shaded gray. Dots represent individual students' z-scores. The adjacent box-and-whisker plots summarize each group's distribution of scores, where the lower horizontal border corresponds to the first quartile (25th percentile); the middle horizontal line corresponds to the median (second quartile or 50th percentile); and the upper horizontal border corresponds to the third quartile (75th percentile).

8.2 Longitudinal Ordinary Least Squares Regression Analysis

Another way to compare the differences between the average achievement levels of the ILI and Indigenous English-medium students at the beginning of 6th grade is to regress students' z-scores in each subject on program, gender, and SES. Table 30 shows the corresponding ordinary least squares (OLS) regression results for English language arts (column 1) and math (column 2). After accounting for gender and SES, the coefficient estimates for ILI are slightly positive but very close to zero and not statistically significant. That said, the 95% confidence intervals of the point estimates for ILI range from -0.379 to 0.542 for English language arts and -0.433 to 0.477 for math—a relatively large degree of uncertainty spanning differences of nearly half a standard deviation in either direction.

Table 30: OLS regressions of standardized ELA and math formative test scores at Site 3, fall 6th grade

	ELA	math
	(1)	(2)
intercept	0.250	0.404*
	(0.211)	(0.212)
ILI	0.081	0.022
	(0.232)	(0.229)
male	−0.223	−0.321
	(0.211)	(0.211)
low SES	−0.307	−0.474**
	(0.211)	(0.210)
Observations	92	89
R^2	0.037	0.076
Adjusted R^2	0.004	0.043
Residual Std. Error	0.998	0.978
F Statistic	1.122	2.326*
Note:	*p<0.1; **p<0.05; ***p<0.01	

As is normally the case with achievement tests, students' early scores on the English language arts and math formative assessments are strongly predictive of later scores in the same subject. Thus we can include fall 6th grade scores along with program, gender, and SES as covariates in OLS regressions where the outcomes are measured at each subsequent assessment instance between winter of 6th grade and winter of 8th grade. The resulting coefficient estimates for ILI correspond to the association between persistence in ILI and formative test scores, relative to English-medium instruction, at each time point in middle school after accounting for fall 6th grade achievement level, gender, and SES. Tables 31 and 32 show the longitudinal OLS regression results for English language arts and math, respectively. (Additional tables detailing results of different permutations of the longitudinal formative assessment OLS regressions from Site 3 are compiled in Chapter 12: Appendix 3.)

Table 31: OLS regressions of standardized ELA formative test scores at Site 3, winter 6th to winter 8th grade

	winter 6th (1)	spring 6th (2)	fall 7th (3)	winter 7th (4)	spring 7th (5)	fall 8th (6)	winter 8th (7)
intercept	0.109	−0.010	0.213**	0.068	0.096	0.130	0.282**
	(0.106)	(0.099)	(0.106)	(0.142)	(0.133)	(0.117)	(0.141)
ILI	0.197*	0.171	−0.078	0.014	−0.105	0.159	0.006
	(0.115)	(0.108)	(0.116)	(0.155)	(0.145)	(0.129)	(0.154)
fall 6th grade z-score	0.834***	0.875***	0.852***	0.768***	0.801***	0.790***	0.718***
	(0.053)	(0.050)	(0.056)	(0.074)	(0.070)	(0.059)	(0.074)
male	−0.144	0.006	−0.066	−0.053	0.036	0.060	−0.148
	(0.106)	(0.099)	(0.107)	(0.142)	(0.133)	(0.117)	(0.141)
low SES	−0.181*	−0.080	−0.302***	−0.127	−0.202	−0.197*	−0.316**
	(0.106)	(0.099)	(0.107)	(0.142)	(0.133)	(0.118)	(0.142)
Observations	92	92	91	91	91	91	90
R^2	0.765	0.793	0.750	0.569	0.622	0.702	0.586
Adjusted R^2	0.754	0.783	0.739	0.549	0.605	0.688	0.566
Residual Std. Error	0.496	0.466	0.500	0.665	0.624	0.548	0.652
F Statistic	70.651***	83.091***	64.633***	28.406***	35.436***	50.573***	30.063***

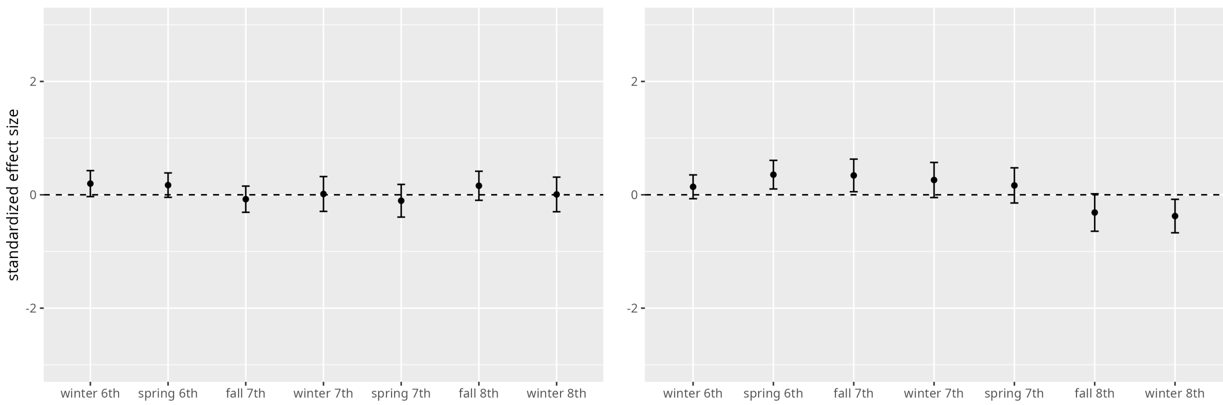
Note: *p<0.1; **p<0.05; ***p<0.01

Table 32: OLS regressions of standardized math formative test scores at Site 3, winter 6th to winter 8th grade

	winter 6th (1)	spring 6th (2)	fall 7th (3)	winter 7th (4)	spring 7th (5)	fall 8th (6)	winter 8th (7)
intercept	0.122	−0.052	−0.205	−0.115	−0.187	0.076	0.160
	(0.100)	(0.119)	(0.137)	(0.147)	(0.147)	(0.141)	(0.141)
ILI	0.141	0.355***	0.342**	0.261*	0.165	−0.313*	−0.375**
	(0.106)	(0.127)	(0.145)	(0.156)	(0.156)	(0.166)	(0.149)
fall 6th grade z-score	0.793***	0.748***	0.784***	0.746***	0.784***	0.717***	0.741***
	(0.050)	(0.060)	(0.068)	(0.074)	(0.073)	(0.071)	(0.071)
male	−0.167*	0.024	−0.112	0.015	0.053	0.054	−0.108
	(0.099)	(0.117)	(0.135)	(0.146)	(0.145)	(0.139)	(0.138)
low SES	−0.078	−0.011	0.259*	0.078	0.196	0.054	0.049
	(0.100)	(0.119)	(0.137)	(0.147)	(0.147)	(0.143)	(0.140)
Observations	89	88	88	89	88	78	88
R^2	0.778	0.686	0.639	0.573	0.591	0.597	0.599
Adjusted R^2	0.768	0.671	0.622	0.552	0.571	0.575	0.580
Residual Std. Error	0.451	0.535	0.617	0.665	0.660	0.591	0.628
F Statistic	73.676***	45.286***	36.777***	28.146***	29.935***	27.057***	31.057***

Note: *p<0.1; **p<0.05; ***p<0.01

The point estimates and confidence intervals of the coefficients on ILI from the longitudinal regressions in Tables 31 and 32 are plotted in Figure 16.



(a) English language arts

(b) math

Figure 16: Coefficient estimates and 95% confidence intervals of persistence in ILI on English language arts and math formative assessments in middle school at Site 3, after accounting for fall 6th grade test score, free/reduced lunch eligibility, and gender.

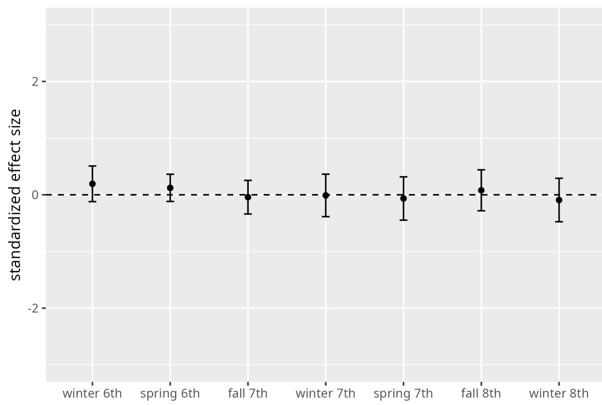
Note: Coefficient estimates at each time point are represented as dots. 95% confidence intervals appear as lines, with the vertical line segments representing the spans of the confidence intervals and the horizontal line segments representing the upper and lower bounds of the confidence intervals. The horizontal dashed line corresponds to zero, *i.e.*, an estimate of no difference between ILI students' and English-medium students' achievement levels after accounting for fall 6th grade test score, free/reduced-price lunch eligibility, and gender.

When we compare the OLS estimates of differences over time between ILI students' achievement levels in English language arts and math relative to their Indigenous-heritage English-medium peers (after accounting for the other covariates), we can see that there is more variation in math than in English language arts.

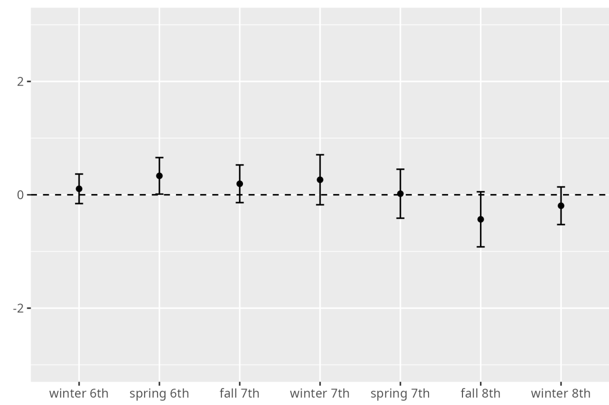
Figure 16(a) shows that in English language arts, there is a slightly positive difference (*i.e.*, higher expected scores for the ILI students relative to their English-medium peers) at three of the seven testing instances after fall of 6th grade: winter 6th grade, spring 6th grade, and fall 8th grade. There is a slightly negative difference (*i.e.*, lower expected scores for the ILI students relative to their English-medium peers) at two of the time points: fall 7th grade and spring 7th grade. In winter 7th grade and winter 8th grade, there is approximately zero difference in expected English language arts scores between the ILI and English-medium students. None of the seven estimates of the difference in English language arts scores between the two groups is statistically significant at the 95% confidence level.

Figure 16(b) shows that in math, of the seven testing instances after fall of 6th grade, there is a positive difference for ILI students at the first five time points and a negative difference for ILI students at the last two time points. Two of the positive estimates in the first five time points are statistically significant (spring 6th grade and fall 7th grade). And while the remaining three estimated differences through the end of 7th grade (winter 6th, winter 7th, and spring 7th) are not statistically significant at the 95% confidence level, nearly all of their respective confidence intervals are in the positive range. In the last two testing instances in math, however (fall and winter of 8th grade), the ILI students have lower expected scores than their English-medium peers. The winter 8th grade estimate is statistically significant at the 95% confidence level, and the fall 8th grade estimate—while not quite statistically significant—is right at the margin.

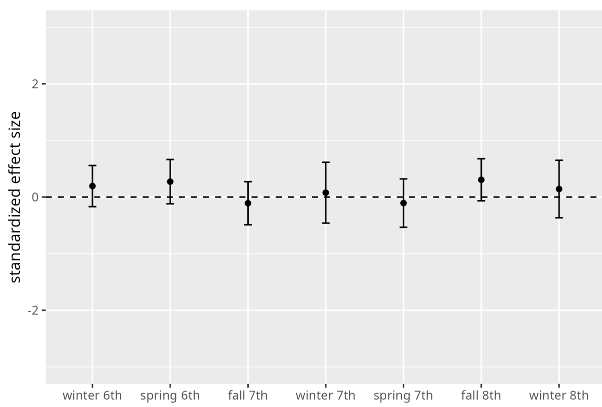
When we disaggregate the longitudinal OLS regression estimates by gender, as in Figure 17, it is evident that the negative association between ILI participation and math achievement in 8th grade is concentrated among female students in the fall term, and among male students in the winter term. There are not strong apparent differences between boys and girls at other time points in math or at any time point in English language arts.



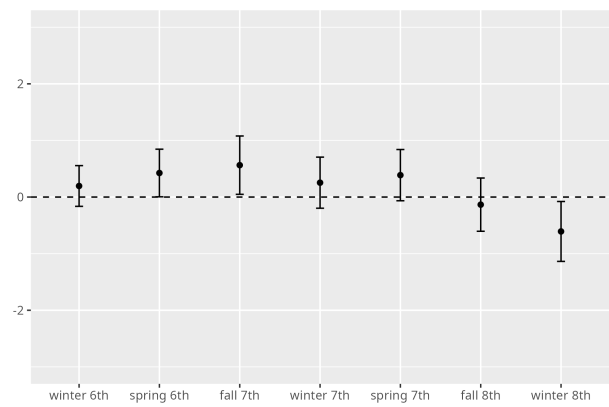
(a) English language arts (female students only)



(b) math (female students only)



(c) English language arts (male students only)



(d) math (male students only)

Figure 17: Coefficient estimates and 95% confidence intervals of persistence in ILI among female and male students on English language arts and math formative assessments in middle school at Site 3, after accounting for fall 6th grade test score and free/reduced lunch eligibility.

Note: Coefficient estimates at each time point are represented as dots. 95% confidence intervals appear as lines, with the vertical line segments representing the spans of the confidence intervals and the horizontal line segments representing the upper and lower bounds of the confidence intervals. The horizontal dashed line corresponds to zero, *i.e.*, an estimate of no difference between female or male ILI students' and English-medium students' achievement levels after accounting for fall 6th grade test score and free/reduced-price lunch eligibility.

The pattern of contrasts in math achievement bears further examination. It would be useful to collect data from additional cohorts of ILI and English-medium students to understand whether the sudden negative difference for ILI students in 8th grade that we observe in this cohort is simply an aberration. If not, there may be some other explanation that would be beneficial for educators at Site 3 to understand and address. Perhaps the 8th grade math formative assessments and the 8th grade ILI math curriculum were misaligned in ways that the 6th and 7th grade assessments were not. Teacher effects may also play some part. Or there could be other factors at play.

8.3 Hierarchical Linear Mixed Effects Analysis with Random Intercepts

Moving beyond the discrete point-in-time snapshots provided by the ordinary least squares framework above, we might also consider analyses that account for the nested structure of the formative assessment data. Given that each student was tested multiple times, we can construct various two-level hierarchical linear models, estimating fixed-effects coefficients for each of the student-level covariates and allowing the intercept to vary randomly at the student level.

Tables 33 and 34 show fixed effect and variance component estimates for various mixed effects models where the outcome is student i's standardized formative test score in English language arts and math, respectively, at trimester t between winter of 6th grade and winter of 8th grade. The first model (column 1) in each table is a simple one-way ANOVA with random effects. In columns 2-7 we iteratively add student-level covariates, starting with ILI program status, then fall 6th grade z-score in the relevant subject area, then gender and SES, an $\text{ILI}_{i.} \times \text{male}_{i.}$ interaction term, then a "trimester" covariate that indexes the change in score over time from trimester t to trimester $t+1$, and finally an $\text{ILI}_{i.} \times \text{trimester}_{.t}$ interaction term. This latter interaction term allows us to isolate the contrast between the time-indexed slope fixed effect estimates for the ILI students relative to the English-medium students. We can express the fully-specified two-level model (column 7) in combined form as

$$\begin{aligned} y_{it} = \hat{\gamma}_0 &+ \hat{\gamma}_1(\text{ILI}_{i.}) + \hat{\gamma}_2(\text{fall 6th z-score}_{i.}) + \hat{\gamma}_3(\text{male}_{i.}) + \hat{\gamma}_4(\text{low SES}_{i.}) \\ &+ \hat{\gamma}_5(\text{ILI}_{i.} \times \text{male}_{i.}) + \hat{\gamma}_6(\text{trimester}_{.t}) + \hat{\gamma}_7(\text{ILI}_{i.} \times \text{trimester}_{.t}) + r_{i.} + e_{it}. \end{aligned} \tag{7a}$$

$$\widehat{\text{Var}}[r_{i.}] = \hat{\tau}; \tag{7b}$$

$$\widehat{\text{Var}}[e_{it}] = \hat{\sigma}^2. \tag{7c}$$

Table 33: Random intercept models: standardized ELA formative test scores at Site 3

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Fixed effects:							
intercept	0.009	−0.047	−0.012	0.124	0.142	0.112	0.087
	(0.096)	(0.115)	(0.050)	(0.083)	(0.089)	(0.093)	(0.095)
ILI		0.191	0.072	0.050	0.005	0.006	0.088
		(0.212)	(0.092)	(0.091)	(0.119)	(0.119)	(0.133)
fall 6th grade z-score			0.832***	0.816***	0.816***	0.816***	0.816***
			(0.042)	(0.042)	(0.042)	(0.042)	(0.042)
male				−0.045	−0.075	−0.075	−0.075
				(0.083)	(0.099)	(0.099)	(0.099)
low SES				−0.200**	−0.204**	−0.204**	−0.203**
				(0.083)	(0.084)	(0.084)	(0.084)
trimester						0.010	0.018*
						(0.009)	(0.010)
ILI × male					0.108	0.107	0.108
					(0.185)	(0.185)	(0.185)
ILI × trimester							−0.028
							(0.019)
Variance components:							
Random Intercept ($\hat{\tau}$)	0.825	0.826	0.131	0.124	0.125	0.125	0.125
Level-1 Residual ($\hat{\sigma}^2$)	0.200	0.200	0.200	0.200	0.200	0.200	0.199
Data structure:							
Number of Students	92	92	92	92	92	92	92
Number of Observations	638	638	638	638	638	638	638
Selection criteria:							
AIC	1102.301	1104.752	959.041	963.586	966.786	975.085	981.120
BIC	1115.676	1122.585	981.333	994.795	1002.453	1015.210	1025.703

Note: *p<0.1; **p<0.05; ***p<0.01

Table 34: Random intercept models: standardized math formative test scores at Site 3

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Fixed effects:							
intercept	0.032	−0.026	−0.001	−0.017	−0.0003	−0.0001	−0.094
	(0.092)	(0.111)	(0.057)	(0.100)	(0.108)	(0.111)	(0.112)
ILI		0.192	0.106	0.104	0.063	0.063	0.377**
		(0.201)	(0.104)	(0.106)	(0.140)	(0.140)	(0.150)
fall 6th grade z-score			0.747***	0.752***	0.752***	0.752***	0.755***
			(0.048)	(0.050)	(0.050)	(0.050)	(0.050)
male				−0.038	−0.066	−0.066	−0.065
				(0.099)	(0.117)	(0.117)	(0.117)
low SES				0.065	0.061	0.061	0.064
				(0.100)	(0.101)	(0.101)	(0.101)
trimester						−0.0001	0.031***
						(0.009)	(0.010)
ILI × male					0.098	0.098	0.098
					(0.216)	(0.216)	(0.216)
ILI × trimester							−0.108***
							(0.019)
Variance components:							
Random Intercept ($\hat{\tau}$)	0.732	0.733	0.173	0.176	0.178	0.178	0.179
Level-1 Residual ($\hat{\sigma}^2$)	0.191	0.191	0.191	0.191	0.191	0.191	0.180
Data structure:							
Number of Students	89	89	89	89	89	89	89
Number of Observations	608	608	608	608	608	608	608
Selection criteria:							
AIC	1019.411	1021.872	911.188	920.126	923.153	932.765	909.733
BIC	1032.641	1039.513	933.239	950.997	958.434	972.457	953.835
Note:							*p<0.1; **p<0.05; ***p<0.01

The one-way ANOVA models shown in column 1 of Tables 33 and 34 give a sense of how much of the total variance in formative test scores is located between students, before considering any other student-level covariates. The total variance is 1.03 in English language arts and 0.92 in math. In both subjects, between-student variance accounts for about 80% of the total variance in the one-way ANOVA models.

When we account for ILI status in the second iteration of the models (see column 2), the coefficients for ILI are positive (0.19 for both subjects) but not statistically significant. Moreover, the between-student variance for English language arts and math is essentially unchanged relative to the one-way ANOVA models in column 1. In other words, ILI program status on its own does not explain any additional variation in the formative test score data for either subject area.

When we add fall 6th grade z-scores to the models in column 3, the estimated coefficients for ILI program status shrink toward zero—from 0.19 to 0.07 in English language arts and from 0.19 to 0.11 in math. As is evident in the preceding OLS analyses, we see that fall 6th grade z-scores are strongly predictive of students' formative test scores at later time points. Fall 6th grade scores explain 84% of the between-student variance in English language arts and 76% of the between-student variance in math relative to the models that accounted for ILI program status alone.

Including gender and SES in the models (column 4) does not add much, if any, explanatory power. The coefficient estimate for ILI status shrinks slightly closer to zero (from 0.07 to 0.05) in English language arts and is essentially unchanged in math, relative to the models in column 3 that account only for ILI program status and fall 6th grade z-score. The between-student variance is unchanged in math and decreases by only a negligible amount (from 0.13 to 0.12) in English language arts.

The ILI $\times$ male interaction term is also small (approximately 0.1 for both subjects) and not statistically significant.

Since the formative tests are administered at fairly regular time intervals over students' middle school careers, we can incorporate a time index into the analysis to estimate fixed slopes for the changes in scores from one formative assessment instance to the next. Interacting the time index with ILI program status allows us to distinguish between the relative starting points and slopes for ILI students relative to English-medium students. The coefficient for trimester in column 6 (without the ILI $\times$ trimester interaction) is precisely zero, which is what we would expect given that the z-scores are standardized—and hence, mean-centered—at each time point. When we include the ILI $\times$ trimester interaction term in column 7, we can make the following inferences:

- The ILI students' expected winter 6th grade z-score, after accounting for fall 6th grade z-score, gender, and SES, is approximately 0.1 standard deviations higher than the Indigenous-heritage English-medium students' expected score in English language arts and 0.4 standard deviations higher in math. The fixed effect estimate for ILI is not statistically significant for the English language arts formative test, but it is statistically significant at the 95% confidence level for math.

- The initial difference in favor of the ILI students attenuates over time as the ILI and Indigenous English-medium students edge toward parity in their expected formative assessment scores at later time periods. The expected change for English-medium students is +0.02 per academic term in English language arts and +0.03 in math, but for ILI students it is $0.02 - 0.03 = -0.01$ in English language arts and $0.03 - 0.1 = -0.07$ in math.

Although we have considered a variety of alternative parametrization for the data from Site 3, the Akaike and Bayesian information criteria (AIC and BIC) suggest that the relatively simple models in column 3 (accounting just for students' ILI program status and fall 6th grade z-score, and omitting gender, SES, and the time index) strike the optimal balance between model fit and parsimony for this particular context.

8.4 Summative Assessment Results

In addition to the formative assessments described previously, students at Site 3 also take end-of-year reading and math summative assessments that are designed in alignment with grade-level Common Core standards. The English-medium students from this cohort were tested in 6th and 7th grades. The ILI students were not tested in 6th grade, but they were tested in 7th grade. No students took summative assessments in 8th grade due to the COVID-19 school closures.

Although both groups were tested in the 7th grade, the English-medium and ILI students took different summative assessments that year. Therefore, it is not reasonable to assume that we can make a direct one-to-one comparison between the two groups, as we did with the formative assessment results. That said, the test that the ILI students took is highly informative in a special way because was administered in the Indigenous language of instruction that the ILI program at Site 3 uses. The results thus provide a glimpse not only of the ILI students' achievement *vis-a-vis* the 7th grade Common Core reading and math content standards, but also shed light on the ILI students' developing proficiency in their Indigenous language. Unlike the formative assessments reviewed in the preceding section—which are administered in English—this summative assessment is ostensibly much more closely aligned with the ILI curriculum used at Site 3.

Despite the differences in the two summative assessments used in 7th grade, we can draw some general comparisons between them. For example, both tests rate students according to four-level proficiency scales. For the English-medium students, the scale ranges from 1) "not met standard" to 2) "nearly met standard" to 3) "met standard" to 4) " exceeded standard." For the ILI students, the scale ranges from 1) "beginning" to 2) "developing" to 3) "proficient" to 4) "distinguished." Students who score in the upper two categories on either test are considered to have demonstrated grade-level-appropriate proficiency in reading or math.

Tables 35 and 36 show the counts of Indigenous-heritage English-medium students from Site 3 at each proficiency level in 6th grade reading and math. Tables 37 and 38 show the counts of students at each proficiency level in 7th grade reading and math.

Table 35: 6th grade summative reading assessment levels by program at Site 3

proficiency level	n
English	
not tested	*18*
1) not met standard	18
2) nearly met standard	20
3) met standard	7
4) exceeded standard	2
total	**65**
ILI	
not tested	*27*
total	**27**

Table 36: 6th grade summative math assessment levels by program at Site 3

proficiency level	n
English	
not tested	*18*
1) not met standard	33
2) nearly met standard	10
3) met standard	3
4) exceeded standard	1
total	**65**
ILI	
not tested	*27*
total	**27**

Table 37: 7th grade summative reading assessment levels by program at Site 3

proficiency level	n
English	
not tested	*8*
1) not met standard	24
2) nearly met standard	16
3) met standard	16
4) exceeded standard	1
total	**65**
ILI	
not tested	*1*
1) beginning	7
2) developing	7
3) proficient	6
4) distinguished	6
total	**27**

Table 38: 7th grade summative math assessment levels by program at Site 3

proficiency level	n
English	
not tested	*8*
1) not met standard	32
2) nearly met standard	18
3) met standard	4
4) exceeded standard	3
total	**65**
ILI	
not tested	*1*
1) beginning	9
2) developing	6
3) proficient	5
4) distinguished	6
total	**27**

A minority of the Indigenous-heritage English-medium students "met" or "exceeded" the grade-level standards in 6th grade—about 19% (9 out of 47 students tested) in reading, and about 9% (4 out of 47) in math. Slightly higher proportions of this group scored in the upper two levels in 7th grade—about 30% in reading (17 out of 57 students tested) in reading and 12% (7 out of 57) in math.

By comparison, larger proportions of the ILI students demonstrated grade-level proficiency in the two subjects—about 46% (12 out of 26) in Indigenous-language reading and about 41% (11 out of 26) in math. Given that the ILI students scored in similar ranges to their Indigenous-heritage peers in the English-medium program on formative assessments administered *in English*, it might be reasonable to infer that a larger proportion of ILI students from this cohort were capable of demonstrating grade-grade-level proficiency in math and language arts *in both languages* than the proportion of Indigenous-heritage English-medium students demonstrated in English alone.

9 Discussion

Parents, school administrators, and other community members sometimes express reservations about Indigenous-language immersion schooling. One of the most common concerns is that conducting instruction in the Indigenous language will cause students' achievement on standard measures of English language arts and math skills to be inhibited relative to a mainstream English-medium curriculum. But is there an empirical basis for that assumption?

The findings from the three case studies detailed here do not provide any compelling evidence that this fear is warranted, at least insofar as the ILI programs at these sites and the students who constituted each of the analytic samples are typical of ILI schooling more broadly. Rather, the ILI students we observed often performed as well—if not better—on mainstream measures of academic achievement administered in English relative to their Indigenous peers who received English-medium instruction. Not only were the ILI students often capable of demonstrating mastery of grade-level content in English, they showed in many cases that they could do so in their Indigenous language as well. The few exceptions we observed where ILI students appeared to score lower than their English-medium peers were concentrated in early elementary grades on low-stakes formative tests when the ILI students were first learning to read, simultaneously, in two highly dissimilar languages with unique orthographic rules that can be difficult to reconcile. As Holm (1996) noted, English is characterized by a uniquely complex entanglement of Romance and Germanic roots and a morpho-phonemic orthography where many high-frequency words are irregularly-spelled. In contrast, many Indigenous orthographies (including the alphabet of the Indigenous language used at Site 1) tend to be phonemic, with regular one-to-one correspondence between letters and sounds. As various educators at the ILI Study's Partner Schools have noted, ILI students often make rapid progress learning to read and write complex sentences in their Indigenous language in early grades. These same students may find it challenging at first to reconcile a highly-regular, phonemic Indigenous orthography with the highly irregular English spellings that they encounter as they begin to learn to read in English. To the extent that this difficulty is apparent, however, it generally seems to disappear as students' multilingualism matures in later grades.

In the present study, we don't observe any widespread negative association between ILI schooling and mainstream English-language academic achievement outcomes among Indigenous students, particularly in relation to relatively high-stakes summative assessments and longer-term outcomes in late elementary or intermediate grades. Naturally, however, this finding falls short of definitively refuting the idea that there could be negative *causal* effects of ILI on outcomes that are important to students, parents, and communities. To clarify the discussion around this point, we can consider how the conventional assumptions for causal inference might be challenged in the present context and engage in some informed speculation about what conditions would need to be true in order for the data in this study to be obscuring actual negative effects of ILI on the student outcomes we examined.

9.1 Conventional Assumptions for Quantitative Causal Inference and their Implications for ILI Research

The potential outcomes model of causality (Holland, 1986; Rubin, 1974, 1986), which is perhaps the most widely-used causal inference framework in the quantitative social sciences, fundamentally entails a set of four assumptions, often referred to as *consistency*, *positivity*, the *stable unit treatment value assumption (SUTVA)*, and *(conditional) ignorability*. To relate them to the present study, we could use the following definitions.

> *Consistency*: The observed outcome for any given individual student corresponds to the outcome that they would exhibit given the schooling type (*i.e.*, ILI or English-medium) that they experienced.
>
> *Positivity*: Any given individual could potentially be exposed to any of the treatment conditions under consideration. In this case, the "treatment conditions" are ILI and English-medium schooling, and we assume that any student in our sample could potentially have attended the ILI program or the corresponding English-medium comparison school that serves their local community.
>
> *Stable Unit Treatment Value Assumption (SUTVA)*: The potential outcomes for any given individual do not depend on how program status (ILI or English-medium) is determined or on the program status of any other individual in the data.
>
> *(Conditional) Ignorability*: Any given individual's potential outcomes (*i.e.*, the outcomes corresponding to each of the potential treatment conditions they could have experienced) are independent of their treatment status. In other words, the data on the unobserved counterfactual potential outcomes are missing completely at random.

The ignorability assumption is conventionally guaranteed in many research settings via randomization, but randomized control trials are often not a realistic option in education studies. In cases such as the present context, it would be unethical to assign students to one or another educational pathway at random. However, if some set of pre-treatment background characteristics correlates with both treatment status and the outcome of interest, and we can observe and account for all of those characteristics (also known as "confounders") in our causal model, then we may be able to argue that the weaker conditional ignorability assumption is plausible.

We might reasonably assume that the consistency and positivity assumptions are satisfied in the three case studies, but SUTVA and the conditional ignorability assumption present special problems in this context.

The stable unit treatment value assumption may be questionable in these case studies for a few reasons. (Indeed, it is hard to sustain a strict form of it in most educational research contexts where classrooms or other group learning environments play some role.) For example, we know from observing how ILI programs often engage with prospective students and their families that ILI educators tend to believe—with good reason—that peer effects are very important to the outcomes they care about. Although we do not have any reason to believe that the ILI programs in these case studies were selectively admitting students who would score especially highly on standardized tests, we do know that ILI educators view their programs' strength and sustainability as depending greatly on the extent of parents' collective dedication to Indigenous language immersion. An example of this is the insistence of some ILI programs that parents commit to learning the Indigenous language and make efforts to use it at home with their children. As in any educational setting where peer effects play an important role, this can create challenges for straightforwardly estimating causal effects of ILI within the potential outcomes framework. In short, the causal effects of ILI as a pedagogy may well be inextricable from the peer effects that students in these programs experience.

Although the interaction between peer effects and ILI's effectiveness may present certain inferential challenges, it also invites questions for future studies around Indigenous language immersion to investigate. By considering variation between schools in their expectations of the families of the students they admit, we might gain insight that could be helpful to ILI educators in the future. For example, one "high expectation" program might require parents who are not proficient in the Indigenous language to enroll in evening classes and demonstrate progress in speaking and writing as a condition of enrollment. Another program (or the same program at an earlier or later time point) might take a more open approach to admissions with relatively lax expectations for students and their families outside of school. A comparison of relevant outcomes between the two scenarios would help us understand how much these sorts of policies matter, and what outcomes they affect.

Students do not experience school walled off from their peers, and ILI programs do not operate in isolation from each other either. So we may also worry that the stable unit treatment value is not realistic when trying to understand the effects of policies or mandates imposed at the school level. ILI educators share knowledge within and across cultures and places. ILI programs, whether emerging or well-established, are likely to draw inspiration from and emulate practices that are perceived as having been successful before in other schools. Any attempt to study the effects of school-level policies would be remiss not to seek some accounting of how the school's experience was influenced or informed by similar situations elsewhere.

So-called "strong" ignorability via randomization is not viable here, as we have already established. But the weaker conditional ignorability assumption is also impossible to guarantee because families who enroll in ILI programs may differ from families who enroll in English-medium schools on unobserved dimensions that correlate with outcomes of interest. This so-called "unobserved confounding" is the fundamental source of omitted variable bias.

One way that omitted variable bias could manifest itself in this study would be if, before choosing which educational pathway to take, the ILI students tended to be more ambitious or motivated, their parents were more likely to encourage their studies, or they tended to exhibit any other number of unobserved factors relative to their English-medium peers that might positively relate to later test scores. Were there a basis for believing that any of those things were true, we might question whether the students who happened to attend the ILI program would have scored higher than their English-medium peers anyway, on average, in the unobserved counterfactual scenario where the ILI students had instead experienced English-medium schooling. Though it may be impossible to disentangle this sort of selection bias from what we can estimate regarding the effects of ILI programs, we can potentially apply methods of sensitivity analysis (Cinelli & Hazlett, 2020; Frank, 2000; Frank et al., 2013, 2023) to bound these concerns within some range of plausibility.

For example, we can consider the coefficient estimates for ILI participation at Site 1 when we regress 3rd grade English language arts and math summative assessment scores on program (ILI relative to English-medium) and five observed student-level covariates: fall kindergarten letter naming fluency or number identification score, gender, special education, free/reduced-price lunch, and English language learner status. (These are the models shown in column 6 in Tables 15 and 16 in Chapter 6.)

In the model where the English language arts summative assessment score is the outcome, the coefficient estimate for ILI is 0.219 with a standard error of 0.148. Although the estimated association between ILI participation and the outcome is positive after accounting for the observed student characteristics, it is not statistically significant. Suppose we want to know how strong the omitted variable bias would need to be in this model for our coefficient estimate to be obscuring what is actually a *negative* causal effect of ILI on the outcome. In this case, we might consider the magnitude of unobserved confounding that would be necessary to swing the estimate from +0.219 to -0.296 while holding the standard error constant. (-0.296 would be a sufficiently large negative effect to be substantively concerning and just large enough to be statistically significant at the 95% level, given a standard error of 0.148.) In order for that to be true, according to the sensitivity analysis framework proposed by Cinelli and Hazlett (2020), the hypothetical omitted variables would need to explain at least 39.8% of the residual variance both of the program and of the outcome that remained over after accounting for the observed covariates in our model. In other words, the actual effect of ILI schooling on the outcome must less extreme than -0.296 unless we can plausibly conceive of one or more omitted variables that would explain an additional 40% of the residual variance both of a) whether students chose ILI versus English-medium schooling and of b) students' 3rd grade ELA summative assessment scores, above and beyond what is already explained by the five observed covariates that were included in the model.

We can also test milder assumptions about the extent of omitted variable bias. For example, how severe would the omitted variable bias need to be in order to shift the positive coefficient we obtained to within the 95% confidence interval range of the -0.296 effect size we deemed as problematic? Unless we can reasonably conceive of one or more unobserved confounders that could explain more than 19.1% of the residual variance both of the program status and of the English language arts summative test outcome that remained after accounting for the five observed covariates, that scenario would not be plausible either. Table 39 shows the sensitivity analysis reporting results for the Site 1 ELA summative assessment model in the format adapted from Cinelli and Hazlett (2020).

Table 39: Sensitivity analysis reporting: causal effects of ILI on 3rd grade English language arts summative assessment at Site 1

Outcome: *standardized 3rd grade ELA summative test score*

Treatment:	Est.	S.E.	t-value	$R^2_{Y\sim D\|\mathbf{X}}$	$RV_{q=2.35}$	$RV_{q=2.35,\alpha=0.05}$
ILI	0.219	0.148	3.476	4.5%	39.8%	19.1%
df = 46		*Bound (1x fall K LNF z-score)*: $R^2_{Y\sim Z\|\mathbf{X},D} = 63.9\%$, $R^2_{D\sim Z\|\mathbf{X}} = 2.3\%$				

In the case of the corresponding model where the 3rd grade math summative test was the outcome, we obtained a coefficient estimate for ILI participation of 0.485 with a standard error of 0.184 after accounting for fall kindergarten number identification score, gender, special education, free/reduced lunch, and English learner status. This estimate of the association between ILI participation and 3rd grade math summative assessment scores is positive and statistically significant. But how severe would any omitted variable bias need to be in order for the point estimate we observe to be obscuring what is actually a null effect? For that to be true, the hypothetical unobserved omitted confounders would need to be able to explain at least 32.0% of the residual variance both of the program status and of the outcome to bring the point estimate to precisely zero. In a less extreme scenario, how strong would the omitted variable bias need to be for us to obtain the estimate we observe if the true effect was still positive but right on the verge of being too small to be statistically significant given the standard error? That situation might be plausible if we could reasonably name one or more omitted variables that could explain at least 8.4% of the residual variance both of ILI/English-medium status and of the 3rd grade math summative assessment score. Table 40 shows the sensitivity analysis reporting results for the Site 1 3rd grade math summative assessment model in the format adapted from Cinelli and Hazlett (2020).

Are any of these omitted variable bias conditions likely to be true? There is no conclusive way to prove

Table 40: Sensitivity analysis reporting: causal effects of ILI on 3rd grade math summative assessment at Site 1

Outcome: *standardized 3rd grade math summative test score*						
Treatment:	Est.	S.E.	t-value	$R^2_{Y\sim D\|\mathbf{X}}$	$RV_{q=1}$	$RV_{q=1,\alpha=0.05}$
ILI	0.485	0.184	2.633	13.1%	32%	8.4%
df = 46	*Bound (1x fall K NIM z-score)*: $R^2_{Y\sim Z\|\mathbf{X},D} = 23.1\%$, $R^2_{D\sim Z\|\mathbf{X}} = 6.8\%$					

or disprove the presence or absence of any precise magnitude of unobserved confounding, but we can try to frame the discussion in terms of what might be plausible, given what we can know about how predictive the observed variables are with regard to both the outcome of interest and schooling choice. Take, for example, the fall kindergarten emergent literacy and numeracy scores that we used as key covariates in our models. In the case of the 3rd grade English language arts summative assessment model, omitted confounding as strong as fall kindergarten letter naming fluency score would explain 63.9% of the residual variance of the outcome and 2.3% of the residual variance of program status. In other words, in order for the true causal effect of ILI on 3rd grade ELA summative test scores to be statistically significant and negative, there would need to be unobserved confounding *orthogonal to the set of observed covariates* that was nearly two-thirds as strong as fall kindergarten letter naming fluency at predicting the outcome and simultaneously more than 17 times as strong as the fall kindergarten score at predicting ILI versus English-medium enrollment.

In the case of the 3rd grade math summative assessment model, unobserved confounding as strong as the fall kindergarten number identification score would explain 23.1% of the residual variance of the outcome and 6.8% of the residual variance of program status. Therefore, in order for the true causal effect of ILI on 3rd grade math summative assessment scores to be zero in the face of the effect size we estimated from the observed data, there would need to be unobserved confounding orthogonal to the set of observed covariates that was 1.39 times more predictive of the outcome *and* 4.7 times more predictive of program status than the fall kindergarten number identification measure.

Although we can never definitively rule out the existence of omitted variables that meet these criteria, it may be hard to imagine any realistic set of omitted variables that plausibly could. In various conversations and interviews with staff and educators at the ILI Study's Partner Schools, our study team has often asked whether students in the ILI programs seem to differ from their Indigenous peers who opt for English-medium schooling in ways that might strongly predict later test scores. Generally, they have told us that the ILI and English-medium students come from similar socioeconomic backgrounds and are otherwise alike in most ways when they start kindergarten.

9.2 Other Causal-Inferential Concerns and Avenues for Future Research

Aside from the challenges inherent to satisfying the assumptions outlined above, another concern that arises with using the potential outcomes model to study Indigenous language immersion is clearly defining the counterfactual contrast of interest. In the present case, we investigate comparisons between ILI programs and contemporary English-medium schools/classrooms in the same communities serving similar student bodies. However, we might also be interested in studying the effects of particular ILI programs relative to differently-specified alternatives. For example,

- The modal or prevailing educational pathway for students in the community prior to establishment of the ILI school;
- Different approaches to ILI within or between cultural-linguistic traditions, (*e.g.*, programs that entail relatively higher versus lower family commitment expectations, year-round programs compared to programs that observe a mainstream fall-winter-spring academic calendar, etc.);
- Well-resourced and adequately-staffed ILI programs compared to relatively under-resourced ILI programs, perhaps at different time frames within the same school.

The multi-year structure of Indigenous language immersion entails various other implications that quantitative researchers must consider. For example, any selection effects that play a role in families' initial school choices are liable to accumulate over time, influencing not only enrollment but also persistence in the program. Students may deliberately opt out of the ILI program because they believe that an alternative pathway would suit them better. Or students may leave the program for other reasons that correlate with outcomes of interest (for example, some kinds of families may be more likely than others to relocate to a distant area for economic or other reasons and be unable to continue in the ILI program). The ILI and English-medium cohorts of students examined in this study were stable and persisted in one program or the other from one year to the next. In other contexts, however, attrition or program switching may present challenges for robust research design.

Multi-year programs can entail complex cross-classified and/or multiple membership data structures as students move from teacher to teacher, classroom to classroom, school to school. There are numerous approaches that accommodate these sorts of data structures (Raudenbush, 2008, 2009; Sun & Pan, 2014), but as researchers look at more schools and classrooms across a wider span of years, the model's complexity will generally grow exponentially, in tandem with the number of potential permutations between levels and across grades. This can be a big problem when, as is often the case with ILI programs, sample sizes are relatively small.

Many conventional approaches to longitudinal data analysis rely on an assumption that the treatment of interest is relatively static, or in the case of fixed-effects models, that important omitted variables do not vary over time. Neither of these may be accurate in the case of ILI programs, where the style of instruction may evolve from one year to another due to turnover in school staff or leadership, or the extent to which English is introduced (or not) into daily instruction from one grade to another. There is also good reason to believe that ILI programs can positively affect unmeasured characteristics like students' self-confidence and sense of belonging (Holm & Holm, 1995), and these effects may compound over time. To the extent that such unobserved factors contribute to other observed outcomes, it may be impossible to distinguish their moderating or mediating effects.

Though Indigenous language immersion programs vary enormously in their levels of intensity and linguistic and cultural foundations, they are invariably multifaceted. This means that researchers must acknowledge a certain tension between understanding holistic effects of the program versus isolating or disentangling distinct effects of specific program components that may interact with and moderate each other. For any given ILI program, we may hope to understand the key aspects that underpin its effectiveness. Or looking at ILI more broadly, we may focus our attention on a set of specific characteristics that virtually all ILI programs seem to share. Conversely, we might look for examples of culturally-unique components that may only be present at specific programs and try to understand their effects.

Indigenous language immersion, broadly defined, spans a wide range of sites, geographic regions, linguistic traditions, and cultures that may have much—or very little—in common with each other. And even within any particular region, different programs with certain linguistic and cultural commonalities (*e.g.*, Kaiapuni Hawaiian language immersion programs across the Hawaiian Islands or Diné immersion schools in the Southwest) may take different forms at different schools taking different pedagogical approaches.

Almost all Indigenous language immersion programs, like other schools, routinely collect information about their students to track academic progress, administer their programs efficiently, and comply with federal, state, and local regulations. These administrative records are a tremendously useful resource for studying ILI programs. They are potentially easily accessible to researchers and educators, they tend to be broadly congruent with each other from school to school, and since they are collected and maintained as a matter of course, they entail minimal (if any) extra intrusion upon teachers, students, or their families—provided, of course, that individuals' privacy and security are appropriately respected.

Administrative data typically comprise relatively coarsely-measured demographic characteristics including federal ethnicity categories, gender, free or reduced-price lunch eligibility, special education status, home language, and English learner status. Administrative data generally do not include information on many other background variables of interest including parents' educational attainment, family size/structure (*e.g.*, number of siblings, whether the family is a single-parent or multi-generational household), students' preschool experience, how much parents spoke to or read to their child in early childhood, number of books in the house, etc.

Moreover, although the academic records stored in a school's student information system likely include standardized test scores and course marks in elementary and secondary grades, they often do not include indicators relating to students' academic preparedness prior to (or early in) kindergarten such as vocabulary, letter recognition listening comprehension or speaking ability (in English or an Indigenous language), or numeracy.

Teacher characteristics are another area where readily-accessible administrative data may be sparse. In the case of Indigenous language immersion, we might expect teacher effectiveness to be moderated by years of experience, fluency in the Indigenous language, or other potentially observable characteristics. Although this information may not be recorded in the school's computer database, it may be feasible for researchers to collect it via interviews with knowledgeable school staff.

Many Indigenous language immersion programs (as well as any local English-medium schools that researchers may want to consider as comparison groups) are small schools in rural areas. Therefore, the sample sizes within any given cohort are likely under-powered for detecting effect sizes within a plausible range. One way to augment statistical power is to aggregate multiple cohorts of students together (as we did with the Site 1 and Site 2 case studies here), but doing so involves the risk of introducing time-varying factors into the analysis, many of which may be unobserved. Examples of these might include teacher and administrative turnover, curriculum changes, changes to enrollment requirements, and fluctuations in class sizes or student/teacher ratios. Again, interviews with knowledgeable school staff may be the best way to incorporate relevant qualitative information about these concerns into the analysis.

Another potential shortcoming of administrative academic records is that any outcome measurements recorded in the data are likely to correspond to constructs that are only partially representative of the holistic aims of an ILI program. For example, schools may have readily-available English language arts and math achievement data for all ILI and English-medium comparison students, but Indigenous language proficiency data for ILI students only (if at all). This presents a problem when researchers hope to understand the extent (or lack) of English-medium students' proficiency in the Indigenous language in order to know whether students in English-medium programs start kindergarten with some degree of listening comprehension and speaking ability, and whether that proficiency changes over time without exposure to Indigenous-language classroom instruction. Other important outcomes that may be difficult to measure are unlikely to be represented in the administrative data at all.

Researchers might also be interested in other important outcomes such as student perceptions of school climate or characteristics such as self-confidence, sense of belonging, cultural pride etc. Existing instruments that aim to measure so-called “non-cognitive” traits may not be adequately validated or reliable with regard to young children or Indigenous populations and in any event, assessing students on these constructs may represent an undue and costly imposition. Imagining ways that educators might observe and quantifiably measure outcomes like these in a non-intrusive, low-cost way seems like a particularly worthy area for further inquiry.

Finally, many of the most important outcomes are long-deferred. Some examples include graduating from high school on time as well as college enrollment, persistence, and degree completion. Researchers and educators may also be interested in long-term community-level outcomes such as prevalence of fluent Indigenous-language speakers, availability of qualified ILI teachers (as the pool of program alumni grows), changes in the use of the Indigenous language in everyday life outside of school, etc. Some of these outcomes (like high school graduation) are relatively straightforward to measure and ILI educators are no doubt already engaging in creative ways of observing other outcomes (such as use of the Indigenous language outside of school).

The lessons that Holm and Holm (1995) took away from their experience at Rock Point and Fort Defiance can inform studies of contemporary ILI programs insofar as they relate to readily-apparent school characteristics. For example, observers of contemporary ILI programs can locate schools on the spectrum between “whole-school” and “supplemental” co-located programs as defined by Holm and Holm (1995), as well as between “total” and “partial” programs. It may be harder to discern different schools’ relative degree of curricular selectivity and focus, but interviews with teachers and administrators can shed light on how much autonomy a school has to set its own priorities. Educator interviews can also address whether and how they perceive external mandates as interfering with or distracting from important priorities. These factors are likely to be important school-level moderators of the effects of ILI that warrant further investigation from a quantitative angle.

Holm and Holm (1995) also alluded to some beneficial effects of the Rock Point program that were strongly evident in their observations: “Just as important—though harder to gauge—was that students had considerably more self-confidence and pride. We believe that people now take considerable pride in being from Rock Point” (p. 148). Similar remarks abound in the Indigenous Language Immersion Study’s corpus of qualitative interviews of contemporary ILI educators, parents, and students (McCarty et al., 2016), but these kinds of outcomes are generally unmeasured in existing quantitative data sources.

Something else that has been echoed in many different places in qualitative interviews of ILI educators is that there are often shortages of well-qualified and highly proficient Indigenous language immersion teachers. As the Rock Point program matured and its earliest cohorts of students grew up, Holm and Holm (1995) noted that “Many of the younger teachers [at Rock Point] are graduates of the community school program” (p. 148). This observation suggests a potential long-term community-level outcome that could be useful for program evaluators long into the future. Specifically, it may be useful to look comprehensively at how various ILI programs keep track of the changes they observe in the pool of qualified teacher candidates in their language community over time. This kind of information, recorded in a systematic way, could constitute persuasive evidence of an Indigenous-language immersion program’s growth, strength, and sustainability over many years.

10 Appendix 1: Supplementary Tables for Site 1

Table 41: Standardized fall kindergarten letter naming fluency test scores at Site 1

	(1)	(2)	(3)	(4)	(5)
intercept	−0.177	0.047	0.119	0.469*	0.491*
	(0.149)	(0.166)	(0.168)	(0.279)	(0.281)
ILI	0.336	0.369*	0.352*	0.319	0.304
	(0.206)	(0.199)	(0.196)	(0.196)	(0.197)
male		−0.546***	−0.459**	−0.487**	−0.488**
		(0.200)	(0.203)	(0.202)	(0.202)
SPED			−0.474*	−0.451*	−0.416*
			(0.244)	(0.243)	(0.247)
FRPL				−0.398	−0.400
				(0.254)	(0.255)
ELL					−0.318
					(0.405)
Observations	93	93	93	93	93
R^2	0.028	0.102	0.139	0.162	0.168
Adjusted R^2	0.018	0.083	0.110	0.124	0.120
Residual Std. Error	0.991	0.958	0.944	0.936	0.938
F Statistic	2.658	5.138***	4.782***	4.259***	3.516***
AIC	266.239	260.863	259.017	258.459	259.799
BIC	273.837	270.993	271.680	273.655	277.527

Note: *p<0.1; **p<0.05; ***p<0.01

Table 42: Standardized fall kindergarten letter sounds fluency test scores at Site 1

	(1)	(2)	(3)	(4)	(5)
intercept	−0.075	0.094	0.097	0.327	0.370
	(0.151)	(0.171)	(0.176)	(0.295)	(0.295)
ILI	0.142	0.167	0.166	0.145	0.116
	(0.208)	(0.205)	(0.207)	(0.208)	(0.208)
male		−0.413**	−0.410*	−0.429**	−0.430**
		(0.206)	(0.213)	(0.214)	(0.213)
SPED			−0.015	−0.00003	0.065
			(0.257)	(0.258)	(0.260)
FRPL				−0.262	−0.265
				(0.269)	(0.268)
ELL					−0.602
					(0.426)
Observations	93	93	93	93	93
R^2	0.005	0.047	0.047	0.058	0.079
Adjusted R^2	−0.006	0.026	0.015	0.015	0.026
Residual Std. Error	1.003	0.987	0.992	0.993	0.987
F Statistic	0.464	2.241	1.478	1.345	1.489
AIC	268.445	266.398	268.394	269.399	269.283
BIC	276.042	276.528	281.057	284.595	287.011
Note:			*p<0.1; **p<0.05; ***p<0.01		

Table 43: Standardized fall kindergarten number identification test scores at Site 1

	(1)	(2)	(3)	(4)	(5)
intercept	−0.265*	−0.205	−0.083	0.241	0.279
	(0.145)	(0.166)	(0.161)	(0.260)	(0.260)
ILI	0.509**	0.519**	0.493**	0.469**	0.442**
	(0.201)	(0.202)	(0.191)	(0.190)	(0.189)
male		−0.150	0.006	−0.016	−0.017
		(0.203)	(0.197)	(0.196)	(0.194)
SPED			−0.829***	−0.804***	−0.741***
			(0.238)	(0.237)	(0.239)
FRPL				−0.380	−0.380
				(0.241)	(0.239)
ELL					−0.579
					(0.391)
Observations	94	94	94	94	94
R^2	0.065	0.071	0.181	0.203	0.223
Adjusted R^2	0.055	0.050	0.154	0.168	0.179
Residual Std. Error	0.972	0.974	0.920	0.912	0.906
F Statistic	6.425**	3.468**	6.638***	5.681***	5.045***
AIC	265.409	266.850	256.966	256.378	256.063
BIC	273.039	277.024	269.682	271.638	273.866
Note:			*p<0.1; **p<0.05; ***p<0.01		

Table 44: Standardized fall 1st grade English reading formative test scores at Site 1

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
intercept	0.214	0.304**	0.428***	0.503***	0.442*	0.482**	0.579**
	(0.152)	(0.125)	(0.142)	(0.144)	(0.239)	(0.240)	(0.250)
ILI	−0.397*	−0.579***	−0.553***	−0.558***	−0.554***	−0.574***	−0.763***
	(0.207)	(0.171)	(0.170)	(0.167)	(0.168)	(0.168)	(0.223)
fall K LNF z-score		0.578***	0.534***	0.494***	0.499***	0.488***	0.487***
		(0.085)	(0.088)	(0.088)	(0.090)	(0.090)	(0.089)
male			−0.305*	−0.249	−0.242	−0.250	−0.474*
			(0.175)	(0.173)	(0.176)	(0.175)	(0.247)
SPED				−0.439**	−0.441**	−0.397*	−0.432**
				(0.207)	(0.208)	(0.209)	(0.210)
FRPL					0.069	0.062	0.069
					(0.216)	(0.214)	(0.214)
ELL						−0.460	−0.441
						(0.338)	(0.337)
ILI × male							0.423
							(0.332)
Observations	91	91	91	91	91	91	91
R^2	0.040	0.369	0.390	0.420	0.421	0.434	0.445
Adjusted R^2	0.029	0.354	0.369	0.393	0.387	0.393	0.398
Residual Std. Error	0.986	0.804	0.794	0.779	0.783	0.779	0.776
F Statistic	3.663*	25.688***	18.540***	15.593***	12.365***	10.717***	9.488***
AIC	259.571	223.396	222.262	219.611	221.503	221.512	221.742
BIC	267.104	233.439	234.817	234.676	239.079	241.599	244.340

Note: *p<0.1; **p<0.05; ***p<0.01

Table 45: Standardized winter 1st grade English reading formative test scores at Site 1

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
intercept	−0.015	0.078	0.186	0.311**	0.429*	0.465*	0.548**
	(0.153)	(0.132)	(0.151)	(0.145)	(0.244)	(0.245)	(0.256)
ILI	0.028	−0.148	−0.121	−0.126	−0.134	−0.152	−0.317
	(0.210)	(0.183)	(0.182)	(0.171)	(0.172)	(0.172)	(0.228)
fall K LNF z-score		0.527***	0.489***	0.422***	0.413***	0.403***	0.403***
		(0.091)	(0.094)	(0.090)	(0.092)	(0.092)	(0.092)
male			−0.274	−0.170	−0.184	−0.190	−0.385
			(0.187)	(0.178)	(0.180)	(0.179)	(0.253)
SPED				−0.775***	−0.772***	−0.730***	−0.760***
				(0.212)	(0.213)	(0.215)	(0.217)
FRPL					−0.134	−0.140	−0.135
					(0.222)	(0.221)	(0.221)
ELL						−0.430	−0.411
						(0.348)	(0.348)
ILI × male							0.373
							(0.341)
Observations	92	92	92	92	92	92	92
R^2	0.0002	0.273	0.291	0.385	0.387	0.398	0.407
Adjusted R^2	−0.011	0.257	0.266	0.357	0.352	0.356	0.357
Residual Std. Error	1.005	0.862	0.857	0.802	0.805	0.803	0.802
F Statistic	0.018	16.726***	12.011***	13.605***	10.878***	9.375***	8.227***
AIC	266.061	238.724	238.504	227.383	228.992	229.352	230.046
BIC	273.626	248.811	251.113	242.514	246.645	249.527	252.742

Note: *p<0.1; **p<0.05; ***p<0.01

Table 46: Standardized spring 1st grade English reading formative test scores at Site 1

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
intercept	0.016	0.100	0.154	0.313**	0.359	0.396	0.521**
	(0.155)	(0.134)	(0.153)	(0.142)	(0.238)	(0.239)	(0.247)
ILI	−0.029	−0.197	−0.182	−0.196	−0.199	−0.218	−0.467**
	(0.211)	(0.184)	(0.186)	(0.167)	(0.168)	(0.168)	(0.220)
fall K LNF z-score		0.523***	0.504***	0.420***	0.417***	0.407***	0.408***
		(0.092)	(0.095)	(0.088)	(0.089)	(0.089)	(0.088)
male			−0.140	−0.003	−0.008	−0.013	−0.318
			(0.190)	(0.174)	(0.176)	(0.176)	(0.247)
SPED				−0.959***	−0.957***	−0.916***	−0.957***
				(0.207)	(0.209)	(0.210)	(0.209)
FRPL					−0.052	−0.058	−0.052
					(0.216)	(0.215)	(0.213)
ELL						−0.434	−0.403
						(0.338)	(0.335)
ILI × male							0.573*
							(0.331)
Observations	91	91	91	91	91	91	91
R^2	0.0002	0.271	0.275	0.420	0.420	0.431	0.451
Adjusted R^2	−0.011	0.254	0.250	0.393	0.386	0.390	0.405
Residual Std. Error	1.005	0.864	0.866	0.779	0.784	0.781	0.772
F Statistic	0.019	16.325***	11.007***	15.537***	12.305***	10.606***	9.736***
AIC	263.222	236.525	237.961	219.749	221.687	221.924	220.691
BIC	270.754	246.568	250.515	234.814	239.263	242.011	243.289

Note: *p<0.1; **p<0.05; ***p<0.01

Table 47: Standardized fall 2nd grade English reading formative test scores at Site 1

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
intercept	−0.014	0.063	0.141	0.277*	0.269	0.310	0.420
	(0.159)	(0.138)	(0.159)	(0.148)	(0.243)	(0.244)	(0.255)
ILI	0.025	−0.141	−0.121	−0.116	−0.116	−0.138	−0.354
	(0.215)	(0.188)	(0.189)	(0.173)	(0.174)	(0.174)	(0.231)
fall K LNF z-score		0.512***	0.485***	0.399***	0.400***	0.389***	0.390***
		(0.093)	(0.097)	(0.091)	(0.092)	(0.092)	(0.092)
male			−0.191	−0.056	−0.055	−0.064	−0.323
			(0.194)	(0.180)	(0.183)	(0.182)	(0.258)
SPED				−0.917***	−0.917***	−0.870***	−0.900***
				(0.217)	(0.218)	(0.220)	(0.220)
FRPL					0.009	0.004	0.009
					(0.220)	(0.219)	(0.218)
ELL						−0.455	−0.433
						(0.345)	(0.343)
ILI × male							0.481
							(0.342)
Observations	88	88	88	88	88	88	88
R^2	0.0002	0.265	0.273	0.402	0.402	0.415	0.429
Adjusted R^2	−0.011	0.248	0.248	0.373	0.366	0.371	0.379
Residual Std. Error	1.006	0.867	0.867	0.792	0.796	0.793	0.788
F Statistic	0.013	15.332***	10.540***	13.952***	11.027***	9.562***	8.579***
AIC	254.714	229.620	230.613	215.474	217.472	217.605	217.450
BIC	262.146	239.530	243.000	230.338	234.814	237.424	239.746

Note: *p<0.1; **p<0.05; ***p<0.01

Table 48: Standardized winter 2nd grade English reading formative test scores at Site 1

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
intercept	0.077	0.138	0.247	0.413***	0.574**	0.603**	0.647**
	(0.159)	(0.141)	(0.159)	(0.147)	(0.242)	(0.243)	(0.257)
ILI	−0.140	−0.276	−0.244	−0.248	−0.257	−0.274	−0.356
	(0.214)	(0.191)	(0.191)	(0.171)	(0.172)	(0.173)	(0.229)
fall K LNF z-score		0.474***	0.435***	0.347***	0.334***	0.327***	0.326***
		(0.095)	(0.098)	(0.090)	(0.091)	(0.091)	(0.092)
male			−0.287	−0.166	−0.186	−0.189	−0.295
			(0.197)	(0.178)	(0.180)	(0.180)	(0.262)
SPED				−1.000***	−0.996***	−0.962***	−0.980***
				(0.214)	(0.214)	(0.217)	(0.220)
FRPL					−0.184	−0.187	−0.186
					(0.218)	(0.218)	(0.219)
ELL						−0.329	−0.317
						(0.344)	(0.347)
ILI × male							0.193
							(0.347)
Observations	89	89	89	89	89	89	89
R^2	0.005	0.230	0.249	0.404	0.409	0.416	0.418
Adjusted R^2	−0.007	0.212	0.222	0.376	0.374	0.373	0.368
Residual Std. Error	1.003	0.888	0.882	0.790	0.791	0.792	0.795
F Statistic	0.427	12.850***	9.393***	14.259***	11.509***	9.733***	8.316***
AIC	257.130	236.295	236.083	217.447	218.690	219.705	221.366
BIC	264.596	246.250	248.526	232.379	236.110	239.614	243.764

Note: *p<0.1; **p<0.05; ***p<0.01

Table 49: Standardized spring 2nd grade English reading formative test scores at Site 1

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
intercept	0.117	0.170	0.272*	0.437***	0.629**	0.651***	0.733***
	(0.160)	(0.140)	(0.159)	(0.148)	(0.240)	(0.243)	(0.256)
ILI	−0.211	−0.343*	−0.318*	−0.327*	−0.337*	−0.350**	−0.501**
	(0.215)	(0.188)	(0.189)	(0.170)	(0.171)	(0.172)	(0.228)
fall K LNF z-score		0.498***	0.461***	0.374***	0.360***	0.354***	0.352***
		(0.093)	(0.097)	(0.090)	(0.091)	(0.091)	(0.091)
male			−0.257	−0.149	−0.172	−0.175	−0.367
			(0.194)	(0.177)	(0.179)	(0.179)	(0.261)
SPED				−0.946***	−0.940***	−0.916***	−0.950***
				(0.212)	(0.212)	(0.216)	(0.218)
FRPL					−0.219	−0.221	−0.219
					(0.216)	(0.217)	(0.217)
ELL						−0.245	−0.225
						(0.342)	(0.343)
ILI × male							0.348
							(0.345)
Observations	88	88	88	88	88	88	88
R^2	0.011	0.260	0.275	0.415	0.422	0.426	0.433
Adjusted R^2	−0.0004	0.242	0.249	0.387	0.387	0.383	0.383
Residual Std. Error	1.000	0.870	0.867	0.783	0.783	0.785	0.785
F Statistic	0.966	14.911***	10.609***	14.704***	11.973***	10.004***	8.722***
AIC	253.744	230.262	230.455	213.587	214.489	215.934	216.819
BIC	261.176	240.172	242.842	228.451	231.831	235.753	239.115

Note: *p<0.1; **p<0.05; ***p<0.01

Table 50: Standardized fall 3rd grade English reading formative test scores at Site 1

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
intercept	0.004	0.075	0.171	0.326**	0.412	0.445*	0.553**
	(0.155)	(0.140)	(0.162)	(0.153)	(0.254)	(0.256)	(0.267)
ILI	−0.007	−0.149	−0.129	−0.139	−0.144	−0.161	−0.371
	(0.211)	(0.193)	(0.193)	(0.177)	(0.178)	(0.179)	(0.238)
fall K LNF z-score		0.451***	0.416***	0.335***	0.328***	0.319***	0.318***
		(0.096)	(0.100)	(0.094)	(0.095)	(0.096)	(0.095)
male			−0.238	−0.121	−0.131	−0.137	−0.386
			(0.199)	(0.184)	(0.187)	(0.187)	(0.264)
SPED				−0.914***	−0.911***	−0.875***	−0.914***
				(0.220)	(0.221)	(0.223)	(0.224)
FRPL					−0.098	−0.103	−0.095
					(0.229)	(0.229)	(0.228)
ELL						−0.377	−0.356
						(0.360)	(0.359)
ILI × male							0.471
							(0.353)
Observations	91	91	91	91	91	91	91
R^2	0.00001	0.201	0.214	0.345	0.347	0.355	0.369
Adjusted R^2	−0.011	0.182	0.186	0.315	0.308	0.309	0.316
Residual Std. Error	1.006	0.904	0.902	0.828	0.832	0.831	0.827
F Statistic	0.001	11.036***	7.874***	11.349***	9.029***	7.715***	6.928***
AIC	263.240	244.876	245.383	230.669	232.475	233.296	233.369
BIC	270.773	254.919	257.937	245.734	250.051	253.383	255.967

Note: *p<0.1; **p<0.05; ***p<0.01

Table 51: Standardized winter 3rd grade English reading formative test scores at Site 1

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
intercept	−0.024	0.084	0.184	0.351**	0.538**	0.576**	0.669***
	(0.153)	(0.134)	(0.154)	(0.142)	(0.234)	(0.235)	(0.246)
ILI	0.044	−0.146	−0.125	−0.136	−0.147	−0.166	−0.348
	(0.210)	(0.184)	(0.184)	(0.164)	(0.164)	(0.164)	(0.219)
fall K LNF z-score		0.522***	0.486***	0.398***	0.384***	0.374***	0.372***
		(0.092)	(0.095)	(0.087)	(0.088)	(0.088)	(0.088)
male			−0.247	−0.121	−0.143	−0.149	−0.365
			(0.189)	(0.171)	(0.172)	(0.172)	(0.243)
SPED				−0.984***	−0.978***	−0.937***	−0.970***
				(0.204)	(0.204)	(0.205)	(0.206)
FRPL					−0.212	−0.218	−0.212
					(0.211)	(0.210)	(0.210)
ELL						−0.432	−0.414
						(0.331)	(0.330)
ILI × male							0.408
							(0.325)
Observations	92	91	91	91	91	91	91
R^2	0.0005	0.269	0.283	0.436	0.443	0.454	0.464
Adjusted R^2	−0.011	0.252	0.258	0.410	0.410	0.415	0.419
Residual Std. Error	1.005	0.864	0.860	0.767	0.767	0.764	0.761
F Statistic	0.044	16.193***	11.450***	16.635***	13.512***	11.637***	10.268***
AIC	266.034	236.523	236.758	216.888	217.812	217.984	218.274
BIC	273.599	246.567	249.312	231.953	235.388	238.071	240.872

Note: *p<0.1; **p<0.05; ***p<0.01

Table 52: Standardized spring 3rd grade English reading formative test scores at Site 1

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
intercept	0.080	0.183	0.271*	0.432***	0.726***	0.748***	0.856***
	(0.153)	(0.138)	(0.159)	(0.149)	(0.244)	(0.247)	(0.258)
ILI	−0.152	−0.317*	−0.298	−0.307*	−0.328*	−0.339*	−0.550**
	(0.211)	(0.190)	(0.191)	(0.173)	(0.172)	(0.173)	(0.231)
fall K LNF z-score		0.467***	0.437***	0.353***	0.329***	0.323***	0.321***
		(0.095)	(0.099)	(0.092)	(0.092)	(0.093)	(0.093)
male			−0.216	−0.095	−0.127	−0.131	−0.378
			(0.196)	(0.180)	(0.180)	(0.180)	(0.254)
SPED				−0.947***	−0.938***	−0.913***	−0.952***
				(0.214)	(0.212)	(0.215)	(0.216)
FRPL					−0.334	−0.337	−0.332
					(0.221)	(0.221)	(0.220)
ELL						−0.256	−0.235
						(0.348)	(0.346)
ILI × male							0.469
							(0.342)
Observations	91	90	90	90	90	90	90
R^2	0.006	0.223	0.234	0.378	0.394	0.398	0.411
Adjusted R^2	−0.005	0.205	0.207	0.348	0.358	0.354	0.361
Residual Std. Error	1.003	0.890	0.889	0.806	0.800	0.802	0.798
F Statistic	0.524	12.482***	8.746***	12.889***	10.925***	9.146***	8.190***
AIC	262.707	239.343	240.082	223.380	222.960	224.372	224.335
BIC	270.240	249.342	252.581	238.379	240.458	244.370	246.833

Note: *p<0.1; **p<0.05; ***p<0.01

Table 53: Standardized fall 4th grade English reading formative test scores at Site 1

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
intercept	−0.132	0.029	0.018	0.186	0.452	0.506	0.754**
	(0.180)	(0.167)	(0.199)	(0.201)	(0.311)	(0.314)	(0.333)
ILI	0.237	−0.027	−0.025	−0.091	−0.088	−0.099	−0.472
	(0.241)	(0.226)	(0.228)	(0.219)	(0.219)	(0.219)	(0.287)
fall K LNF z-score		0.461***	0.463***	0.395***	0.378***	0.359***	0.360***
		(0.111)	(0.115)	(0.113)	(0.114)	(0.115)	(0.112)
male			0.024	0.042	0.019	0.001	−0.447
			(0.227)	(0.217)	(0.218)	(0.218)	(0.314)
SPED				−0.739**	−0.741**	−0.706**	−0.738***
				(0.279)	(0.279)	(0.280)	(0.274)
FRPL					−0.311	−0.326	−0.337
					(0.278)	(0.278)	(0.272)
ELL						−0.470	−0.530
						(0.411)	(0.403)
ILI × male							0.813*
							(0.418)
Observations	70	69	69	69	69	69	69
R^2	0.014	0.215	0.215	0.292	0.306	0.320	0.360
Adjusted R^2	−0.0004	0.191	0.179	0.248	0.251	0.255	0.287
Residual Std. Error	1.000	0.900	0.906	0.867	0.866	0.864	0.845
F Statistic	0.972	9.031***	5.934***	6.610***	5.558***	4.871***	4.904***
AIC	202.650	186.138	188.126	182.967	183.616	184.180	182.023
BIC	209.396	195.074	199.297	196.372	199.255	202.053	202.130

Note: *p<0.1; **p<0.05; ***p<0.01

Table 54: Standardized winter 4th grade English reading formative test scores at Site 1

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
intercept	−0.054	0.100	0.042	0.209	0.552*	0.621*	0.789**
	(0.187)	(0.174)	(0.202)	(0.203)	(0.319)	(0.322)	(0.341)
ILI	0.095	−0.160	−0.157	−0.216	−0.218	−0.234	−0.505*
	(0.247)	(0.231)	(0.232)	(0.223)	(0.222)	(0.221)	(0.290)
fall K LNF z-score		0.452***	0.466***	0.401***	0.378***	0.356***	0.359***
		(0.112)	(0.116)	(0.113)	(0.114)	(0.114)	(0.113)
male			0.132	0.143	0.121	0.104	−0.244
			(0.231)	(0.221)	(0.220)	(0.220)	(0.327)
SPED				−0.757**	−0.732**	−0.687**	−0.721**
				(0.289)	(0.287)	(0.288)	(0.287)
FRPL					−0.402	−0.428	−0.421
					(0.290)	(0.290)	(0.287)
ELL						−0.520	−0.554
						(0.412)	(0.409)
ILI × male							0.615
							(0.431)
Observations	68	67	67	67	67	67	67
R^2	0.002	0.202	0.207	0.286	0.308	0.326	0.348
Adjusted R^2	−0.013	0.178	0.169	0.240	0.251	0.258	0.271
Residual Std. Error	1.006	0.906	0.911	0.871	0.864	0.860	0.853
F Statistic	0.148	8.125***	5.468***	6.206***	5.422***	4.829***	4.501***
AIC	197.816	181.814	183.468	178.412	178.335	178.577	178.304
BIC	204.475	190.633	194.491	191.640	193.768	196.214	198.147

Note: *p<0.1; **p<0.05; ***p<0.01

Table 55: Standardized spring 4th grade English reading formative test scores at Site 1

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
intercept	−0.115	0.035	0.061	0.225	0.365	0.430	0.640*
	(0.186)	(0.175)	(0.208)	(0.206)	(0.324)	(0.327)	(0.347)
ILI	0.205	−0.041	−0.043	−0.099	−0.095	−0.109	−0.439
	(0.249)	(0.235)	(0.237)	(0.225)	(0.227)	(0.226)	(0.301)
fall K LNF z-score		0.435***	0.429***	0.359***	0.353***	0.330***	0.330***
		(0.116)	(0.120)	(0.117)	(0.118)	(0.119)	(0.117)
male			−0.054	−0.011	−0.025	−0.051	−0.450
			(0.237)	(0.225)	(0.228)	(0.228)	(0.332)
SPED				−0.816***	−0.817***	−0.772**	−0.777***
				(0.290)	(0.292)	(0.292)	(0.288)
FRPL					−0.164	−0.180	−0.183
					(0.292)	(0.291)	(0.287)
ELL						−0.523	−0.578
						(0.418)	(0.413)
ILI × male							0.711
							(0.435)
Observations	66	65	65	65	65	65	65
R^2	0.011	0.191	0.192	0.286	0.290	0.309	0.340
Adjusted R^2	−0.005	0.165	0.152	0.239	0.230	0.237	0.259
Residual Std. Error	1.002	0.911	0.918	0.870	0.875	0.871	0.859
F Statistic	0.681	7.336***	4.833***	6.017***	4.821***	4.317***	4.188***
AIC	191.594	177.311	179.256	173.196	174.850	175.118	174.143
BIC	198.163	186.009	190.128	186.243	190.071	192.513	193.712

Note: *p<0.1; **p<0.05; ***p<0.01

Table 56: Standardized fall 5th grade English reading formative test scores at Site 1

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
intercept	−0.130	−0.021	0.153	0.253	0.314	0.329	0.324
	(0.231)	(0.213)	(0.252)	(0.243)	(0.421)	(0.449)	(0.467)
ILI	0.247	−0.029	−0.098	−0.111	−0.115	−0.121	−0.111
	(0.318)	(0.297)	(0.299)	(0.284)	(0.289)	(0.299)	(0.368)
fall K LNF z-score		0.439***	0.412***	0.349**	0.351**	0.348**	0.347**
		(0.140)	(0.141)	(0.136)	(0.139)	(0.144)	(0.146)
male			−0.398	−0.283	−0.300	−0.307	−0.293
			(0.314)	(0.302)	(0.320)	(0.331)	(0.436)
SPED				−0.945**	−0.925**	−0.919*	−0.919*
				(0.425)	(0.445)	(0.455)	(0.462)
FRPL					−0.068	−0.074	−0.075
					(0.382)	(0.391)	(0.398)
ELL						−0.055	−0.055
						(0.490)	(0.498)
ILI × male							−0.030
							(0.621)
Observations	40	39	39	39	39	39	39
R^2	0.016	0.221	0.255	0.350	0.350	0.351	0.351
Adjusted R^2	−0.010	0.178	0.191	0.273	0.252	0.229	0.204
Residual Std. Error	1.005	0.900	0.892	0.846	0.858	0.871	0.885
F Statistic	0.603	5.104**	3.997**	4.570***	3.558**	2.879**	2.391**
AIC	117.872	107.298	107.546	104.255	106.217	108.202	110.199
BIC	122.939	113.953	115.863	114.236	117.862	121.511	125.171

Note: *p<0.1; **p<0.05; ***p<0.01

Table 57: Standardized winter 5th grade English reading formative test scores at Site 1

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
intercept	−0.100	0.010	0.006	0.111	0.580	0.526	0.528
	(0.225)	(0.212)	(0.254)	(0.241)	(0.421)	(0.446)	(0.462)
ILI	0.195	−0.066	−0.064	−0.072	−0.109	−0.089	−0.094
	(0.315)	(0.301)	(0.309)	(0.288)	(0.286)	(0.294)	(0.359)
fall K LNF z-score		0.396***	0.397**	0.324**	0.337**	0.349**	0.349**
		(0.144)	(0.148)	(0.141)	(0.140)	(0.144)	(0.147)
male			0.011	0.151	0.024	0.047	0.040
			(0.326)	(0.309)	(0.320)	(0.329)	(0.429)
SPED				−1.096**	−0.947**	−0.970**	−0.970**
				(0.439)	(0.447)	(0.456)	(0.463)
FRPL					−0.519	−0.498	−0.497
					(0.384)	(0.392)	(0.399)
ELL						0.200	0.201
						(0.489)	(0.497)
ILI × male							0.017
							(0.618)
Observations	41	40	40	40	40	40	40
R^2	0.010	0.173	0.174	0.299	0.334	0.338	0.338
Adjusted R^2	−0.016	0.129	0.105	0.218	0.236	0.217	0.193
Residual Std. Error	1.008	0.923	0.936	0.874	0.864	0.875	0.888
F Statistic	0.382	3.883**	2.519*	3.725**	3.414**	2.803**	2.330**
AIC	120.941	111.978	113.977	109.414	109.325	111.123	113.122
BIC	126.082	118.734	122.421	119.547	121.148	124.634	128.322

Note: *p<0.1; **p<0.05; ***p<0.01

Table 58: Standardized spring 5th grade English reading formative test scores at Site 1

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
intercept	−0.062	0.058	0.174	0.287	0.355	0.399	0.400
	(0.226)	(0.199)	(0.235)	(0.216)	(0.387)	(0.410)	(0.425)
ILI	0.120	−0.194	−0.236	−0.244	−0.250	−0.266	−0.268
	(0.316)	(0.281)	(0.285)	(0.259)	(0.263)	(0.271)	(0.330)
fall K LNF z-score		0.512***	0.494***	0.416***	0.418***	0.409***	0.409***
		(0.135)	(0.136)	(0.126)	(0.128)	(0.133)	(0.135)
male			−0.281	−0.132	−0.150	−0.169	−0.172
			(0.301)	(0.277)	(0.294)	(0.303)	(0.395)
SPED				−1.170***	−1.148***	−1.130**	−1.130**
				(0.393)	(0.411)	(0.420)	(0.426)
FRPL					−0.076	−0.092	−0.092
					(0.353)	(0.361)	(0.367)
ELL						−0.163	−0.163
						(0.450)	(0.457)
ILI × male							0.008
							(0.568)
Observations	41	40	40	40	40	40	40
R^2	0.004	0.282	0.299	0.440	0.441	0.443	0.443
Adjusted R^2	−0.022	0.243	0.240	0.376	0.359	0.342	0.322
Residual Std. Error	1.011	0.863	0.865	0.784	0.795	0.805	0.817
F Statistic	0.145	7.262***	5.116***	6.885***	5.367***	4.380***	3.641***
AIC	121.189	106.651	107.693	100.678	102.624	104.466	106.466
BIC	126.329	113.407	116.137	110.811	114.446	117.977	121.666

Note: *p<0.1; **p<0.05; ***p<0.01

Table 59: Standardized fall 1st grade math formative test scores at Site 1

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
intercept	−0.189	−0.112	0.013	0.129	0.248	0.268	0.261
	(0.154)	(0.147)	(0.169)	(0.170)	(0.295)	(0.298)	(0.315)
ILI	0.388*	0.202	0.195	0.181	0.181	0.179	0.191
	(0.221)	(0.215)	(0.213)	(0.206)	(0.207)	(0.208)	(0.265)
fall K NIM z-score		0.353***	0.350***	0.256**	0.245**	0.233**	0.234**
		(0.105)	(0.104)	(0.107)	(0.110)	(0.113)	(0.114)
male			−0.303	−0.258	−0.264	−0.259	−0.244
			(0.210)	(0.204)	(0.206)	(0.207)	(0.288)
SPED				−0.709**	−0.716**	−0.686**	−0.684**
				(0.283)	(0.284)	(0.289)	(0.293)
FRPL					−0.137	−0.145	−0.144
					(0.278)	(0.280)	(0.282)
ELL						−0.252	−0.252
						(0.400)	(0.403)
ILI × male							−0.032
							(0.417)
Observations	80	80	80	80	80	80	80
R^2	0.038	0.162	0.184	0.247	0.249	0.254	0.254
Adjusted R^2	0.026	0.140	0.152	0.207	0.199	0.192	0.181
Residual Std. Error	0.987	0.928	0.921	0.891	0.895	0.899	0.905
F Statistic	3.096*	7.416***	5.704***	6.152***	4.920***	4.132***	3.495***
AIC	228.909	219.932	219.777	215.325	217.063	218.631	220.625
BIC	236.055	229.460	231.687	229.617	233.737	237.688	242.063

Note: *p<0.1; **p<0.05; ***p<0.01

Table 60: Standardized winter 1st grade math formative test scores at Site 1

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
intercept	0.169	0.277**	0.375**	0.455***	0.505*	0.547*	0.720**
	(0.151)	(0.137)	(0.160)	(0.164)	(0.287)	(0.287)	(0.292)
ILI	−0.354	−0.604***	−0.611***	−0.627***	−0.627***	−0.630***	−0.950***
	(0.219)	(0.203)	(0.203)	(0.200)	(0.201)	(0.200)	(0.248)
fall K NIM z-score		0.457***	0.456***	0.394***	0.390***	0.364***	0.347***
		(0.100)	(0.099)	(0.104)	(0.106)	(0.108)	(0.106)
male			−0.236	−0.190	−0.193	−0.185	−0.565**
			(0.199)	(0.198)	(0.199)	(0.198)	(0.264)
SPED				−0.489*	−0.491*	−0.435	−0.462*
				(0.267)	(0.269)	(0.271)	(0.265)
FRPL					−0.058	−0.077	−0.090
					(0.271)	(0.270)	(0.264)
ELL						−0.515	−0.517
						(0.386)	(0.378)
ILI × male							0.816**
							(0.386)
Observations	82	82	82	82	82	82	82
R^2	0.032	0.236	0.249	0.280	0.281	0.297	0.338
Adjusted R^2	0.020	0.216	0.220	0.243	0.234	0.241	0.275
Residual Std. Error	0.990	0.885	0.883	0.870	0.875	0.871	0.852
F Statistic	2.620	12.175***	8.627***	7.502***	5.936***	5.293***	5.386***
AIC	235.057	217.668	218.203	216.714	218.665	218.746	215.937
BIC	242.278	227.295	230.236	231.154	235.512	237.999	237.598

Note: *p<0.1; **p<0.05; ***p<0.01

Table 61: Standardized spring 1st grade math formative test scores at Site 1

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
intercept	−0.019	0.061	0.051	0.169	0.073	0.155	0.425
	(0.155)	(0.148)	(0.173)	(0.175)	(0.305)	(0.296)	(0.290)
ILI	0.041	−0.143	−0.143	−0.170	−0.170	−0.176	−0.676***
	(0.225)	(0.220)	(0.221)	(0.214)	(0.216)	(0.208)	(0.247)
fall K NIM z-score		0.351***	0.351***	0.262**	0.271**	0.221*	0.195*
		(0.107)	(0.108)	(0.110)	(0.113)	(0.111)	(0.105)
male			0.024	0.097	0.101	0.119	−0.482*
			(0.218)	(0.213)	(0.215)	(0.207)	(0.267)
SPED				−0.704**	−0.700**	−0.595**	−0.626**
				(0.285)	(0.287)	(0.281)	(0.264)
FRPL					0.111	0.075	0.045
					(0.289)	(0.280)	(0.263)
ELL						−0.998**	−0.995***
						(0.399)	(0.374)
ILI × male							1.275***
							(0.387)
Observations	80	80	80	80	80	80	80
R^2	0.0004	0.123	0.123	0.189	0.191	0.255	0.352
Adjusted R^2	−0.012	0.100	0.089	0.146	0.136	0.193	0.289
Residual Std. Error	1.006	0.949	0.955	0.924	0.930	0.898	0.843
F Statistic	0.033	5.401***	3.559**	4.367***	3.483***	4.154***	5.594***
AIC	231.990	223.521	225.508	221.275	223.116	218.525	209.279
BIC	239.136	233.049	237.418	235.567	239.791	237.582	230.718

Note: *p<0.1; **p<0.05; ***p<0.01

Table 62: Standardized fall 2nd grade math formative test scores at Site 1

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
intercept	−0.041	0.044	0.095	0.138	0.419	0.432	0.510
	(0.159)	(0.145)	(0.179)	(0.188)	(0.349)	(0.357)	(0.318)
ILI	0.124	−0.105	−0.124	−0.136	−0.149	−0.151	−0.730**
	(0.276)	(0.256)	(0.260)	(0.262)	(0.262)	(0.265)	(0.279)
fall K NIM z-score		0.458***	0.452***	0.420***	0.396***	0.389***	0.415***
		(0.121)	(0.122)	(0.129)	(0.131)	(0.136)	(0.121)
male			−0.122	−0.096	−0.132	−0.134	−0.655**
			(0.250)	(0.253)	(0.256)	(0.258)	(0.266)
SPED				−0.268	−0.289	−0.282	−0.286
				(0.329)	(0.330)	(0.334)	(0.297)
FRPL					−0.312	−0.317	−0.158
					(0.327)	(0.331)	(0.297)
ELL						−0.104	0.039
						(0.451)	(0.403)
ILI × male							1.960***
							(0.507)
Observations	60	60	60	60	60	60	60
R^2	0.003	0.204	0.207	0.217	0.230	0.231	0.402
Adjusted R^2	−0.014	0.176	0.165	0.160	0.158	0.143	0.322
Residual Std. Error	1.007	0.908	0.914	0.917	0.917	0.926	0.823
F Statistic	0.203	7.301***	4.882***	3.806***	3.221**	2.646**	5.003***
AIC	175.054	163.580	165.325	166.603	167.603	169.542	156.368
BIC	181.337	171.958	175.796	179.169	182.263	186.297	175.218

Note: *p<0.1; **p<0.05; ***p<0.01

Table 63: Standardized winter 2nd grade math formative test scores at Site 1

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
intercept	0.194	0.259*	0.192	0.235	0.377	0.420	0.493
	(0.156)	(0.147)	(0.179)	(0.187)	(0.353)	(0.359)	(0.335)
ILI	−0.553**	−0.715***	−0.690***	−0.702***	−0.708***	−0.716***	−1.170***
	(0.263)	(0.252)	(0.256)	(0.257)	(0.259)	(0.261)	(0.286)
fall K NIM z-score		0.366***	0.373***	0.342***	0.330**	0.308**	0.320**
		(0.120)	(0.121)	(0.128)	(0.131)	(0.134)	(0.125)
male			0.165	0.185	0.165	0.160	−0.280
			(0.252)	(0.254)	(0.259)	(0.260)	(0.283)
SPED				−0.273	−0.287	−0.258	−0.294
				(0.339)	(0.343)	(0.346)	(0.322)
FRPL					−0.157	−0.175	−0.051
					(0.331)	(0.332)	(0.313)
ELL						−0.356	−0.236
						(0.452)	(0.423)
ILI × male							1.595***
							(0.531)
Observations	60	60	60	60	60	60	60
R^2	0.071	0.201	0.207	0.216	0.220	0.229	0.343
Adjusted R^2	0.055	0.173	0.165	0.159	0.147	0.141	0.254
Residual Std. Error	0.972	0.909	0.914	0.917	0.923	0.927	0.864
F Statistic	4.419**	7.170***	4.875***	3.795***	3.039**	2.617**	3.872***
AIC	170.859	163.800	165.342	166.640	168.389	169.692	162.092
BIC	177.142	172.178	175.814	179.206	183.050	186.447	180.941

Note: *p<0.1; **p<0.05; ***p<0.01

Table 64: Standardized spring 2nd grade math formative test scores at Site 1

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
intercept	0.230	0.293*	0.321*	0.359*	0.634*	0.722*	0.754**
	(0.158)	(0.148)	(0.176)	(0.188)	(0.359)	(0.361)	(0.350)
ILI	−0.636**	−0.798***	−0.808***	−0.817***	−0.829***	−0.845***	−1.172***
	(0.262)	(0.248)	(0.252)	(0.254)	(0.255)	(0.253)	(0.289)
fall K NIM z-score		0.377***	0.374***	0.350***	0.323**	0.284**	0.294**
		(0.118)	(0.119)	(0.127)	(0.130)	(0.132)	(0.128)
male			−0.076	−0.082	−0.114	−0.111	−0.458
			(0.254)	(0.256)	(0.259)	(0.257)	(0.297)
SPED				−0.208	−0.218	−0.158	−0.261
				(0.349)	(0.350)	(0.349)	(0.341)
FRPL					−0.307	−0.354	−0.231
					(0.340)	(0.339)	(0.333)
ELL						−0.607	−0.484
						(0.441)	(0.431)
ILI × male							1.169**
							(0.549)
Observations	58	58	58	58	58	58	58
R^2	0.095	0.236	0.238	0.243	0.254	0.281	0.341
Adjusted R^2	0.079	0.209	0.195	0.186	0.183	0.196	0.249
Residual Std. Error	0.960	0.890	0.897	0.902	0.904	0.896	0.867
F Statistic	5.888**	8.510***	5.609***	4.246***	3.548***	3.323***	3.693***
AIC	163.790	155.950	157.854	159.467	160.562	160.449	157.415
BIC	169.971	164.192	168.157	171.830	174.985	176.933	175.959

Note: *p<0.1; **p<0.05; ***p<0.01

Table 65: Standardized fall 3rd grade math formative test scores at Site 1

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
intercept	−0.107	−0.007	0.079	0.169	0.542	0.587*	0.625*
	(0.154)	(0.141)	(0.173)	(0.180)	(0.335)	(0.340)	(0.333)
ILI	0.331	0.080	0.048	0.021	0.002	−0.001	−0.303
	(0.271)	(0.252)	(0.255)	(0.252)	(0.251)	(0.252)	(0.291)
fall K NIM z-score		0.464***	0.454***	0.390***	0.359***	0.335**	0.349***
		(0.119)	(0.120)	(0.125)	(0.127)	(0.130)	(0.127)
male			−0.208	−0.179	−0.225	−0.232	−0.492*
			(0.241)	(0.239)	(0.240)	(0.240)	(0.270)
SPED				−0.487	−0.514*	−0.497	−0.513*
				(0.307)	(0.305)	(0.307)	(0.300)
FRPL					−0.412	−0.433	−0.347
					(0.313)	(0.315)	(0.311)
ELL						−0.351	−0.274
						(0.428)	(0.420)
ILI × male							1.018*
							(0.528)
Observations	62	62	62	62	62	62	62
R^2	0.024	0.223	0.233	0.266	0.288	0.296	0.342
Adjusted R^2	0.008	0.197	0.194	0.214	0.224	0.220	0.257
Residual Std. Error	0.996	0.896	0.898	0.886	0.881	0.883	0.862
F Statistic	1.500	8.490***	5.884***	5.161***	4.528***	3.863***	4.007***
AIC	179.409	167.259	168.467	167.778	167.890	169.137	167.003
BIC	185.790	175.767	179.103	180.540	182.780	186.154	186.147

Note: *p<0.1; **p<0.05; ***p<0.01

Table 66: Standardized winter 3rd grade math formative test scores at Site 1

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
intercept	−0.034	0.070	0.075	0.127	0.624*	0.706*	0.746**
	(0.157)	(0.137)	(0.175)	(0.181)	(0.353)	(0.359)	(0.357)
ILI	0.152	−0.061	−0.063	−0.057	−0.055	−0.098	−0.328
	(0.331)	(0.287)	(0.293)	(0.293)	(0.288)	(0.289)	(0.335)
fall K NIM z-score		0.547***	0.546***	0.501***	0.440***	0.412***	0.398***
		(0.124)	(0.125)	(0.131)	(0.135)	(0.136)	(0.135)
male			−0.012	0.024	−0.016	−0.019	−0.166
			(0.249)	(0.250)	(0.247)	(0.246)	(0.269)
SPED				−0.346	−0.407	−0.358	−0.390
				(0.314)	(0.311)	(0.312)	(0.311)
FRPL					−0.560	−0.610*	−0.577*
					(0.344)	(0.345)	(0.343)
ELL						−0.555	−0.541
						(0.465)	(0.462)
ILI × male							0.831
							(0.629)
Observations	53	53	53	53	53	53	53
R^2	0.004	0.284	0.284	0.302	0.339	0.359	0.383
Adjusted R^2	−0.015	0.256	0.240	0.244	0.269	0.275	0.287
Residual Std. Error	1.008	0.863	0.872	0.870	0.855	0.851	0.844
F Statistic	0.210	9.925***	6.486***	5.188***	4.823***	4.293***	3.989***
AIC	155.180	139.678	141.676	142.354	141.449	141.833	141.813
BIC	161.091	147.560	151.527	154.176	155.241	157.595	159.546

Note: *p<0.1; **p<0.05; ***p<0.01

Table 67: Standardized spring 3rd grade math formative test scores at Site 1

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
intercept	−0.026	0.076	0.085	0.184	0.775**	0.749**	0.780**
	(0.158)	(0.142)	(0.178)	(0.177)	(0.315)	(0.322)	(0.320)
ILI	0.114	−0.085	−0.088	−0.074	−0.055	−0.039	−0.261
	(0.331)	(0.298)	(0.304)	(0.292)	(0.280)	(0.285)	(0.330)
fall K NIM z-score		0.491***	0.491***	0.401***	0.340**	0.350***	0.339**
		(0.127)	(0.129)	(0.130)	(0.128)	(0.130)	(0.130)
male			−0.023	0.070	0.036	0.035	−0.111
			(0.259)	(0.252)	(0.243)	(0.245)	(0.267)
SPED				−0.709**	−0.758**	−0.779**	−0.806**
				(0.314)	(0.302)	(0.308)	(0.306)
FRPL					−0.698**	−0.684**	−0.647**
					(0.313)	(0.317)	(0.315)
ELL						0.223	0.241
						(0.458)	(0.454)
ILI × male							0.811
							(0.621)
Observations	53	53	53	53	53	53	53
R^2	0.002	0.231	0.231	0.305	0.372	0.375	0.398
Adjusted R^2	−0.017	0.201	0.184	0.248	0.305	0.294	0.304
Residual Std. Error	1.009	0.894	0.903	0.867	0.833	0.840	0.834
F Statistic	0.118	7.525***	4.920***	5.278***	5.571***	4.607***	4.252***
AIC	155.276	143.451	145.443	142.077	138.733	140.459	140.490
BIC	161.187	151.333	155.295	153.899	152.525	156.221	158.223

Note: *p<0.1; **p<0.05; ***p<0.01

Table 68: Standardized fall 4th grade math formative test scores at Site 1

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
intercept	−0.220	−0.079	−0.084	−0.021	0.426	0.462	0.575
	(0.170)	(0.163)	(0.215)	(0.222)	(0.355)	(0.370)	(0.354)
ILI	0.788**	0.558*	0.561*	0.565*	0.587*	0.570*	0.131
	(0.321)	(0.305)	(0.317)	(0.316)	(0.310)	(0.316)	(0.355)
fall K NIM z-score		0.449***	0.448***	0.386**	0.356**	0.345**	0.325**
		(0.153)	(0.155)	(0.164)	(0.162)	(0.166)	(0.157)
male			0.011	0.043	0.016	0.007	−0.309
			(0.279)	(0.279)	(0.274)	(0.278)	(0.297)
SPED				−0.381	−0.408	−0.409	−0.488
				(0.338)	(0.332)	(0.336)	(0.319)
FRPL					−0.541	−0.559	−0.486
					(0.340)	(0.347)	(0.329)
ELL						−0.213	−0.259
						(0.541)	(0.512)
ILI × male							1.445**
							(0.630)
Observations	43	43	43	43	43	43	43
R^2	0.128	0.283	0.283	0.306	0.351	0.354	0.438
Adjusted R^2	0.107	0.247	0.228	0.233	0.263	0.246	0.326
Residual Std. Error	0.945	0.868	0.879	0.876	0.858	0.868	0.821
F Statistic	6.016**	7.895***	5.132***	4.194***	3.999***	3.282**	3.898***
AIC	121.130	114.711	116.709	117.293	116.440	118.256	114.231
BIC	126.414	121.756	125.515	127.860	128.769	132.345	130.082

Note: *p<0.1; **p<0.05; ***p<0.01

Table 69: 3rd grade summative English language arts assessment levels by program at Site 1

proficiency level	n
English	
1) minimally proficient	18
2) partially proficient	1
3) proficient	5
4) highly proficient	1
total	**25**
ILI	
1) minimally proficient	12
2) partially proficient	11
3) proficient	3
4) highly proficient	2
total	**28**

Table 70: 3rd grade summative math assessment levels by program at Site 1

proficiency level	n
English	
1) minimally proficient	10
2) partially proficient	9
3) proficient	6
4) highly proficient	0
total	**25**
ILI	
1) minimally proficient	5
2) partially proficient	10
3) proficient	11
4) highly proficient	2
total	**28**

Table 71: 4th grade summative English language arts assessment levels by program at Site 1

proficiency level	n
English	
not tested	*4*
1) minimally proficient	10
2) partially proficient	5
3) proficient	4
4) highly proficient	0
total	**23**
ILI	
not tested	*1*
1) minimally proficient	11
2) partially proficient	9
3) proficient	5
4) highly proficient	1
total	**27**

Table 72: 4th grade summative math assessment levels by program at Site 1

proficiency level	n
English	
not tested	*4*
1) minimally proficient	8
2) partially proficient	7
3) proficient	4
4) highly proficient	0
total	**23**
ILI	
not tested	*1*
1) minimally proficient	9
2) partially proficient	8
3) proficient	7
4) highly proficient	2
total	**27**

Table 73: 5th grade summative English language arts assessment levels by program at Site 1

proficiency level	n
English	
1) minimally proficient	9
2) partially proficient	5
3) proficient	6
4) highly proficient	0
total	**20**
ILI	
1) minimally proficient	7
2) partially proficient	6
3) proficient	7
4) highly proficient	1
total	**21**

Table 74: 5th grade summative math assessment levels by program at Site 1

proficiency level	n
English	
1) minimally proficient	12
2) partially proficient	6
3) proficient	2
4) highly proficient	0
total	**20**
ILI	
1) minimally proficient	8
2) partially proficient	5
3) proficient	7
4) highly proficient	1
total	**21**

Table 75: Standardized 3rd grade English language arts summative test scores at Site 1

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
intercept	−0.535***	−0.402***	−0.318**	−0.139	−0.007	−0.034	0.143
	(0.167)	(0.127)	(0.143)	(0.121)	(0.207)	(0.210)	(0.215)
ILI	0.301	0.173	0.239	0.240	0.226	0.219	−0.110
	(0.230)	(0.174)	(0.181)	(0.146)	(0.148)	(0.148)	(0.202)
fall K LNF z-score		0.584***	0.545***	0.465***	0.444***	0.453***	0.431***
		(0.092)	(0.097)	(0.080)	(0.085)	(0.086)	(0.083)
male			−0.234	−0.147	−0.150	−0.126	−0.466**
			(0.187)	(0.152)	(0.153)	(0.156)	(0.211)
SPED				−0.828***	−0.840***	−0.792***	−0.863***
				(0.159)	(0.160)	(0.169)	(0.165)
FRPL					−0.150	−0.126	−0.161
					(0.191)	(0.193)	(0.185)
ELL						−0.361	−0.256
						(0.394)	(0.380)
ILI × male							0.642**
							(0.282)
Observations	53	53	53	53	53	53	53
R^2	0.033	0.464	0.480	0.668	0.672	0.678	0.711
Adjusted R^2	0.014	0.442	0.449	0.640	0.637	0.636	0.666
Residual Std. Error	0.834	0.627	0.624	0.504	0.506	0.507	0.485
F Statistic	1.722	21.631***	15.099***	24.133***	19.273***	16.146***	15.838***
AIC	135.161	105.881	106.225	84.498	85.812	86.854	83.076
BIC	141.072	113.762	116.076	96.320	99.604	102.616	100.809

Note: *p<0.1; **p<0.05; ***p<0.01

Table 76: Standardized 4th grade English language arts summative test scores at Site 1

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
intercept	−0.479***	−0.361**	−0.345*	−0.178	0.465	0.423	0.705**
	(0.167)	(0.163)	(0.192)	(0.196)	(0.287)	(0.288)	(0.292)
ILI	0.139	−0.050	−0.048	−0.119	−0.150	−0.095	−0.549**
	(0.219)	(0.216)	(0.219)	(0.210)	(0.193)	(0.199)	(0.258)
fall K LNF z-score		0.291***	0.286**	0.231**	0.174*	0.167	0.194**
		(0.105)	(0.112)	(0.109)	(0.102)	(0.102)	(0.096)
male			−0.036	−0.062	−0.075	−0.077	−0.578**
			(0.218)	(0.207)	(0.190)	(0.189)	(0.266)
SPED				−0.616**	−0.632**	−0.574**	−0.640***
				(0.267)	(0.245)	(0.250)	(0.235)
FRPL					−0.731***	−0.703***	−0.709***
					(0.253)	(0.254)	(0.237)
ELL						−0.506	−0.476
						(0.455)	(0.425)
ILI × male							0.879**
							(0.348)
Observations	45	44	44	44	44	44	44
R^2	0.009	0.162	0.163	0.263	0.396	0.415	0.503
Adjusted R^2	−0.014	0.121	0.100	0.188	0.316	0.320	0.407
Residual Std. Error	0.726	0.679	0.687	0.652	0.599	0.597	0.558
F Statistic	0.405	3.966**	2.590*	3.484**	4.977***	4.380***	5.212***
AIC	102.825	95.635	97.605	93.973	87.253	87.805	82.624
BIC	108.245	102.772	106.526	104.678	99.742	102.079	98.682

Note: *p<0.1; **p<0.05; ***p<0.01

Table 77: Standardized 5th grade English language arts summative test scores at Site 1

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
intercept	−0.518***	−0.455***	−0.449**	−0.360**	−0.079	−0.031	−0.114
	(0.158)	(0.157)	(0.188)	(0.174)	(0.307)	(0.325)	(0.327)
ILI	0.231	0.100	0.098	0.092	0.069	0.051	0.238
	(0.220)	(0.223)	(0.229)	(0.209)	(0.209)	(0.214)	(0.254)
fall K LNF z-score		0.179	0.178	0.117	0.125	0.115	0.105
		(0.107)	(0.109)	(0.102)	(0.102)	(0.105)	(0.104)
male			−0.016	0.101	0.025	0.004	0.257
			(0.241)	(0.224)	(0.233)	(0.240)	(0.304)
SPED				−0.919***	−0.830**	−0.810**	−0.814**
				(0.317)	(0.326)	(0.332)	(0.329)
FRPL					−0.311	−0.329	−0.355
					(0.280)	(0.286)	(0.283)
ELL						−0.179	−0.191
						(0.356)	(0.352)
ILI × male							−0.579
							(0.438)
Observations	41	40	40	40	40	40	40
R^2	0.028	0.088	0.088	0.265	0.290	0.296	0.332
Adjusted R^2	0.003	0.039	0.012	0.181	0.186	0.168	0.186
Residual Std. Error	0.705	0.685	0.694	0.632	0.630	0.637	0.630
F Statistic	1.104	1.781	1.157	3.148**	2.781**	2.309*	2.273*
AIC	91.590	88.077	90.072	83.462	84.038	85.735	85.609
BIC	96.731	94.832	98.516	93.595	95.860	99.246	100.809

Note: *p<0.1; **p<0.05; ***p<0.01

Table 78: Standardized 3rd grade math summative test scores at Site 1

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
intercept	−0.722***	−0.581***	−0.512***	−0.385**	−0.083	−0.101	−0.001
	(0.168)	(0.139)	(0.158)	(0.149)	(0.247)	(0.250)	(0.267)
ILI	0.668***	0.410**	0.460**	0.516***	0.490**	0.485**	0.293
	(0.231)	(0.194)	(0.202)	(0.185)	(0.183)	(0.184)	(0.259)
fall K NIM z-score		0.513***	0.500***	0.358***	0.314***	0.315***	0.305***
		(0.098)	(0.099)	(0.100)	(0.103)	(0.103)	(0.104)
male			−0.180	−0.102	−0.086	−0.069	−0.263
			(0.195)	(0.180)	(0.178)	(0.181)	(0.257)
SPED				−0.705***	−0.755***	−0.715***	−0.761***
				(0.214)	(0.213)	(0.223)	(0.227)
FRPL					−0.349	−0.334	−0.350
					(0.228)	(0.231)	(0.231)
ELL						−0.310	−0.254
						(0.477)	(0.479)
ILI × male							0.378
							(0.357)
Observations	53	53	53	53	53	53	53
R^2	0.141	0.445	0.455	0.556	0.577	0.580	0.591
Adjusted R^2	0.124	0.423	0.421	0.519	0.532	0.526	0.527
Residual Std. Error	0.840	0.681	0.682	0.622	0.614	0.618	0.617
F Statistic	8.368***	20.081***	13.629***	15.002***	12.800***	10.606***	9.274***
AIC	135.833	114.638	115.729	106.901	106.337	107.853	108.552
BIC	141.744	122.519	125.580	118.723	120.129	123.615	126.284

Note: *p<0.1; **p<0.05; ***p<0.01

Table 79: Standardized 4th grade math summative test scores at Site 1

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
intercept	−0.555***	−0.432**	−0.375*	−0.244	0.153	0.158	0.109
	(0.194)	(0.178)	(0.212)	(0.213)	(0.305)	(0.312)	(0.339)
ILI	0.306	0.024	0.018	−0.022	−0.020	−0.029	0.060
	(0.256)	(0.244)	(0.246)	(0.238)	(0.231)	(0.245)	(0.334)
fall K NIM z-score		0.423***	0.429***	0.348***	0.322**	0.326**	0.321**
		(0.124)	(0.126)	(0.127)	(0.125)	(0.130)	(0.132)
male			−0.117	−0.088	−0.051	−0.052	0.052
			(0.230)	(0.222)	(0.217)	(0.220)	(0.344)
SPED				−0.639**	−0.630**	−0.636**	−0.625*
				(0.308)	(0.300)	(0.307)	(0.311)
FRPL					−0.505*	−0.507*	−0.512*
					(0.283)	(0.288)	(0.291)
ELL						0.066	0.059
						(0.569)	(0.576)
ILI × male							−0.178
							(0.450)
Observations	45	45	45	45	45	45	45
R^2	0.032	0.241	0.246	0.319	0.371	0.371	0.374
Adjusted R^2	0.010	0.205	0.191	0.251	0.290	0.272	0.255
Residual Std. Error	0.847	0.759	0.766	0.737	0.717	0.727	0.735
F Statistic	1.431	6.674***	4.458***	4.693***	4.594***	3.734***	3.152**
AIC	116.745	107.800	109.515	106.902	105.380	107.364	109.173
BIC	122.165	115.027	118.548	117.742	118.027	121.817	125.433

Note: *p<0.1; **p<0.05; ***p<0.01

Table 80: Standardized 5th grade math summative test scores at Site 1

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
intercept	−0.780***	−0.712***	−0.859***	−0.813***	−0.533*	−0.471	−0.445
	(0.159)	(0.152)	(0.173)	(0.168)	(0.272)	(0.283)	(0.291)
ILI	0.546**	0.407*	0.467**	0.472**	0.471**	0.450**	0.378
	(0.222)	(0.216)	(0.214)	(0.206)	(0.204)	(0.206)	(0.250)
fall K NIM z-score		0.269**	0.268**	0.222**	0.218**	0.196*	0.201*
		(0.108)	(0.105)	(0.104)	(0.103)	(0.106)	(0.108)
male			0.366	0.466**	0.396*	0.373	0.274
			(0.223)	(0.219)	(0.224)	(0.226)	(0.296)
SPED				−0.646**	−0.552*	−0.523	−0.522
				(0.315)	(0.320)	(0.324)	(0.327)
FRPL					−0.334	−0.357	−0.341
					(0.257)	(0.260)	(0.264)
ELL						−0.292	−0.283
						(0.350)	(0.354)
ILI × male							0.229
							(0.438)
Observations	41	41	41	41	41	41	41
R^2	0.134	0.256	0.307	0.379	0.408	0.420	0.424
Adjusted R^2	0.112	0.217	0.251	0.310	0.323	0.317	0.302
Residual Std. Error	0.711	0.668	0.653	0.627	0.621	0.624	0.630
F Statistic	6.032**	6.539***	5.458***	5.497***	4.819***	4.097***	3.475***
AIC	92.378	88.147	87.252	84.730	84.799	85.970	87.631
BIC	97.519	95.002	95.820	95.012	96.794	99.678	103.053

Note: *p<0.1; **p<0.05; ***p<0.01

Table 81: Log odds of scoring "proficient" or higher on 3rd grade English language arts summative test at Site 1

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
intercept	−1.153**	−0.993*	−0.826	−0.378	−0.383	−0.383	0.024
	(0.468)	(0.555)	(0.598)	(0.643)	(1.400)	(1.400)	(1.416)
ILI	−0.373	−0.922	−0.732	−0.841	−0.840	−0.840	−2.080
	(0.680)	(0.858)	(0.899)	(0.932)	(0.980)	(0.980)	(1.294)
fall K LNF z-score		1.367***	1.242***	1.130**	1.131**	1.131**	1.157**
		(0.471)	(0.474)	(0.484)	(0.507)	(0.507)	(0.581)
male			−0.620	−0.609	−0.609	−0.609	−19.379
			(0.906)	(0.917)	(0.922)	(0.922)	(5,160.041)
SPED				−18.070	−18.069	−18.084	−19.336
				(2,597.202)	(2,597.041)	(2,783.626)	(4,487.653)
FRPL					0.005	0.005	0.090
					(1.276)	(1.276)	(1.288)
ELL						0.114	1.506
						(7,757.900)	(12,016.520)
ILI × male							20.262
							(5,160.041)
Observations	53	53	53	53	53	53	53
Log Likelihood	−26.915	−20.374	−20.140	−17.530	−17.530	−17.530	−15.385
AIC	57.830	46.748	48.279	45.060	47.060	49.060	46.769
BIC	61.771	52.659	56.160	54.912	58.882	62.852	62.532

Note: *p<0.1; **p<0.05; ***p<0.01

Table 82: Log odds of scoring "proficient" or higher on 4th grade English language arts summative test at Site 1

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
intercept	−1.322**	−1.191**	−1.061	−1.148	0.493	0.377	0.932
	(0.563)	(0.575)	(0.659)	(0.720)	(1.089)	(1.107)	(1.195)
ILI	0.118	−0.085	−0.073	−0.037	−0.154	−0.016	−0.864
	(0.730)	(0.770)	(0.773)	(0.785)	(0.838)	(0.855)	(1.125)
fall K LNF z-score		0.224	0.181	0.209	0.051	0.034	0.077
		(0.356)	(0.374)	(0.386)	(0.403)	(0.401)	(0.413)
male			−0.297	−0.283	−0.363	−0.347	−1.492
			(0.766)	(0.770)	(0.801)	(0.808)	(1.363)
SPED				0.302	0.272	0.402	0.305
				(0.967)	(1.043)	(1.072)	(1.126)
FRPL					−1.936**	−1.861**	−1.950**
					(0.921)	(0.922)	(0.954)
ELL						−16.083	−16.063
						(2,797.291)	(2,764.446)
ILI × male							1.925
							(1.732)
Observations	45	44	44	44	44	44	44
Log Likelihood	−23.824	−23.385	−23.310	−23.262	−20.998	−20.621	−19.956
AIC	51.647	52.771	54.620	56.524	53.996	55.242	55.912
BIC	55.261	58.123	61.756	65.445	64.701	67.732	70.185

Note: *p<0.1; **p<0.05; ***p<0.01

Table 83: Log odds of scoring "proficient" or higher on 5th grade English language arts summative test at Site 1

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
intercept	−0.847*	−0.754	−0.707	−0.553	0.399	1.386	1.243
	(0.488)	(0.501)	(0.589)	(0.602)	(1.041)	(1.298)	(1.441)
ILI	0.362	0.116	0.097	0.104	0.024	−0.268	0.237
	(0.663)	(0.695)	(0.706)	(0.724)	(0.738)	(0.786)	(0.934)
fall K LNF z-score		0.363	0.355	0.247	0.287	0.180	0.167
		(0.326)	(0.330)	(0.332)	(0.339)	(0.347)	(0.358)
male			−0.114	0.082	−0.240	−0.777	0.006
			(0.755)	(0.787)	(0.860)	(0.955)	(1.214)
SPED				−16.978	−16.645	−17.399	−17.371
				(1,763.377)	(1,753.687)	(2,618.111)	(2,593.455)
FRPL					−1.049	−1.615	−1.795
					(0.933)	(1.110)	(1.240)
ELL						−18.438	−18.381
						(2,756.969)	(2,832.649)
ILI × male							−1.876
							(1.821)
Observations	41	40	40	40	40	40	40
Log Likelihood	−26.172	−25.179	−25.168	−23.238	−22.589	−20.612	−20.041
AIC	56.345	56.359	58.336	56.476	57.178	55.224	56.082
BIC	59.772	61.426	65.092	64.920	67.311	67.046	69.593
Note:							*p<0.1; **p<0.05; ***p<0.01

Table 84: Log odds of scoring "proficient" or higher on 3rd grade math summative test at Site 1

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
intercept	−1.153**	−1.017**	−0.856	−0.661	18.346	18.382	18.913
	(0.468)	(0.510)	(0.561)	(0.571)	(2,409.947)	(2,550.157)	(2,514.209)
ILI	1.010*	0.675	0.796	1.003	0.754	0.754	−0.286
	(0.602)	(0.663)	(0.698)	(0.711)	(0.795)	(0.795)	(1.109)
fall K NIM z-score		0.882**	0.842**	0.552	0.311	0.311	0.281
		(0.352)	(0.352)	(0.362)	(0.433)	(0.433)	(0.439)
male			−0.440	−0.264	−0.135	−0.135	−1.324
			(0.673)	(0.692)	(0.764)	(0.764)	(1.246)
SPED				−2.102*	−19.019	−19.055	−19.353
				(1.159)	(2,409.947)	(2,550.157)	(2,514.209)
FRPL					−19.279	−19.315	−19.441
					(2,409.947)	(2,550.157)	(2,514.209)
ELL						0.384	0.596
						(7,867.086)	(7,403.362)
ILI × male							2.271
							(1.667)
Observations	53	53	53	53	53	53	53
Log Likelihood	−33.114	−29.302	−29.086	−26.807	−21.027	−21.027	−20.015
AIC	70.227	64.603	66.173	63.615	54.055	56.055	56.031
BIC	74.168	70.514	74.054	73.466	65.876	69.847	71.793
Note:							*p<0.1; **p<0.05; ***p<0.01

Table 85: Log odds of scoring "proficient" or higher on 4th grade math summative test at Site 1

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
intercept	-1.322^{**}	-1.202^{**}	-0.700	-0.577	1.384	1.340	0.936
	(0.563)	(0.574)	(0.654)	(0.689)	(1.078)	(1.082)	(1.200)
ILI	0.686	0.314	0.233	0.185	0.299	0.405	1.357
	(0.698)	(0.754)	(0.789)	(0.795)	(0.892)	(0.902)	(1.219)
fall K NIM z-score		0.553	0.635^{*}	0.577	0.542	0.475	0.449
		(0.365)	(0.381)	(0.394)	(0.427)	(0.438)	(0.459)
male			-1.126	-1.115	-1.167	-1.179	0.302
			(0.747)	(0.750)	(0.855)	(0.855)	(1.359)
SPED				-0.650	-0.856	-0.839	-0.757
				(1.207)	(1.415)	(1.429)	(1.369)
FRPL					-2.572^{***}	-2.522^{**}	-2.782^{**}
					(0.984)	(0.984)	(1.082)
ELL						-15.626	-15.741
						(2,630.636)	(2,482.122)
ILI × male							-2.448
							(1.854)
Observations	45	45	45	45	45	45	45
Log Likelihood	-26.549	-25.350	-24.124	-23.967	-19.853	-19.586	-18.650
AIC	57.099	56.700	56.249	57.933	51.707	53.172	53.300
BIC	60.712	62.120	63.476	66.967	62.547	65.818	67.753
Note:							*p<0.1; **p<0.05; ***p<0.01

Table 86: Log odds of scoring "proficient" or higher on 5th grade math summative test at Site 1

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
intercept	-2.197^{***}	-2.184^{***}	-2.747^{***}	-2.683^{***}	-2.399^{*}	-2.202	-1.785
	(0.745)	(0.764)	(0.965)	(0.987)	(1.323)	(1.365)	(1.437)
ILI	1.712^{**}	1.534^{*}	1.815^{*}	1.916^{*}	1.905^{*}	1.857^{*}	1.336
	(0.870)	(0.891)	(0.958)	(1.004)	(1.003)	(1.011)	(1.217)
fall K NIM z-score		0.531	0.585	0.437	0.427	0.338	0.350
		(0.396)	(0.418)	(0.425)	(0.426)	(0.435)	(0.436)
male			1.074	1.361	1.244	1.110	0.339
			(0.907)	(0.971)	(1.034)	(1.038)	(1.599)
SPED				-17.530	-17.414	-17.343	-18.487
				(2,652.616)	(2,661.728)	(2,573.826)	(4,150.513)
FRPL					-0.318	-0.351	-0.351
					(1.022)	(1.034)	(1.065)
ELL						-16.269	-17.349
						(2,936.325)	(4,876.863)
ILI × male							1.259
							(2.019)
Observations	41	41	41	41	41	41	41
Log Likelihood	-20.457	-19.545	-18.820	-17.467	-17.420	-17.023	-16.826
AIC	44.914	45.090	45.641	44.935	46.839	48.045	49.652
BIC	48.341	50.231	52.495	53.503	57.121	60.040	63.360
Note:							*p<0.1; **p<0.05; ***p<0.01

11 Appendix 2: Supplementary Tables for Site 2

Table 87: Standardized fall 4th grade English reading test scores at Site 2

	(1)	(2)
intercept	0.191	0.359**
	(0.142)	(0.169)
ILI	−0.549**	−0.466*
	(0.236)	(0.237)
male		−0.406*
		(0.228)
Observations	72	72
R^2	0.071	0.112
Adjusted R^2	0.058	0.086
Residual Std. Error	0.964	0.949
F Statistic	5.384**	4.361**
AIC	202.967	201.732
BIC	209.797	210.839
Note:	*p<0.1; **p<0.05; ***p<0.01	

Table 88: Standardized fall 4th grade math test scores at Site 2

	(1)	(2)
intercept	0.294**	0.405**
	(0.139)	(0.169)
ILI	−0.700***	−0.650***
	(0.233)	(0.236)
male		−0.261
		(0.226)
Observations	73	73
R^2	0.113	0.129
Adjusted R^2	0.100	0.104
Residual Std. Error	0.953	0.951
F Statistic	9.018***	5.196***
AIC	204.120	204.745
BIC	210.992	213.907
Note:	*p<0.1; **p<0.05; ***p<0.01	

Table 89: Standardized winter 4th grade English reading formative test scores at Site 2

	(1)	(2)	(3)	(4)
intercept	0.012	−0.086	−0.072	−0.027
	(0.140)	(0.087)	(0.110)	(0.120)
ILI	−0.353	0.118	0.124	−0.051
	(0.236)	(0.154)	(0.157)	(0.243)
fall 4th ELA z-score		0.788***	0.785***	0.776***
		(0.078)	(0.079)	(0.080)
male			−0.032	−0.130
			(0.145)	(0.179)
ILI × male				0.292
				(0.309)
Observations	66	66	66	66
R^2	0.034	0.633	0.633	0.638
Adjusted R^2	0.019	0.621	0.615	0.614
Residual Std. Error	0.915	0.569	0.573	0.573
F Statistic	2.233	54.231***	35.622***	26.892***
AIC	179.528	117.711	119.661	120.704
BIC	186.097	126.470	130.609	133.842
Note:		*p<0.1; **p<0.05; ***p<0.01		

Table 90: Standardized spring 4th grade English reading formative test scores at Site 2

	(1)	(2)	(3)	(4)
intercept	0.151	0.085	0.218*	0.224
	(0.161)	(0.102)	(0.125)	(0.139)
ILI	−0.644**	−0.206	−0.144	−0.168
	(0.264)	(0.174)	(0.173)	(0.294)
fall 4th ELA z-score		0.716***	0.676***	0.674***
		(0.085)	(0.086)	(0.089)
male			−0.298*	−0.312
			(0.170)	(0.215)
ILI × male				0.037
				(0.362)
Observations	48	48	48	48
R^2	0.115	0.658	0.680	0.680
Adjusted R^2	0.096	0.643	0.658	0.650
Residual Std. Error	0.884	0.556	0.543	0.550
F Statistic	5.977**	43.247***	31.186***	22.866***
AIC	128.325	84.718	83.475	85.463
BIC	133.938	92.203	92.831	96.690
Note:		*p<0.1; **p<0.05; ***p<0.01		

Table 91: Standardized fall 5th grade English reading formative test scores at Site 2

	(1)	(2)	(3)	(4)
intercept	0.177	0.173	0.459**	0.521**
	(0.245)	(0.189)	(0.181)	(0.207)
ILI	−0.624*	−0.233	0.030	−0.153
	(0.332)	(0.269)	(0.242)	(0.374)
fall 4th ELA z-score		0.745***	0.612***	0.579***
		(0.158)	(0.141)	(0.151)
male			−0.854***	−1.042**
			(0.246)	(0.382)
ILI × male				0.330
				(0.510)
Observations	33	33	33	33
R^2	0.102	0.484	0.635	0.641
Adjusted R^2	0.074	0.450	0.598	0.590
Residual Std. Error	0.948	0.731	0.625	0.631
F Statistic	3.539*	14.096***	16.852***	12.489***
AIC	94.082	77.786	68.347	69.859
BIC	98.572	83.772	75.830	78.838
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 92: Standardized winter 5th grade English reading formative test scores at Site 2

	(1)	(2)	(3)	(4)
intercept	0.044	0.086	0.194	0.162
	(0.189)	(0.115)	(0.137)	(0.146)
ILI	−0.833***	−0.401**	−0.269	0.008
	(0.284)	(0.185)	(0.205)	(0.465)
fall 4th ELA z-score		0.690***	0.625***	0.645***
		(0.108)	(0.116)	(0.121)
male			−0.312	−0.219
			(0.226)	(0.267)
ILI × male				−0.347
				(0.521)
Observations	25	25	25	25
R^2	0.272	0.744	0.765	0.770
Adjusted R^2	0.240	0.720	0.731	0.724
Residual Std. Error	0.706	0.428	0.420	0.425
F Statistic	8.594***	31.909***	22.783***	16.746***
AIC	57.428	33.335	33.164	34.615
BIC	61.085	38.210	39.258	41.928
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 93: Standardized spring 5th grade English reading formative test scores at Site 2

	(1)	(2)	(3)	(4)
intercept	0.098	0.158	0.277	0.385**
	(0.254)	(0.139)	(0.167)	(0.165)
ILI	−0.618	−0.073	0.074	−0.864
	(0.394)	(0.228)	(0.254)	(0.525)
fall 4th ELA z-score		0.970***	0.897***	0.823***
		(0.134)	(0.144)	(0.140)
male			−0.345	−0.662**
			(0.277)	(0.303)
ILI × male				1.187*
				(0.593)
Observations	24	24	24	24
R^2	0.101	0.742	0.761	0.803
Adjusted R^2	0.060	0.718	0.725	0.761
Residual Std. Error	0.951	0.521	0.514	0.479
F Statistic	2.459	30.266***	21.229***	19.315***
AIC	69.621	41.608	41.812	39.221
BIC	73.155	46.321	47.702	46.290
Note:		*p<0.1; **p<0.05; ***p<0.01		

Table 94: Standardized fall 6th grade English reading formative test scores at Site 2

	(1)	(2)	(3)	(4)
intercept	−0.017	−0.120	−0.134	−0.232
	(0.282)	(0.182)	(0.207)	(0.233)
ILI	−0.250	0.088	0.079	0.289
	(0.367)	(0.243)	(0.255)	(0.343)
fall 4th ELA z-score		0.836***	0.837***	0.804***
		(0.138)	(0.142)	(0.147)
male			0.039	0.319
			(0.245)	(0.392)
ILI × male				−0.475
				(0.518)
Observations	27	27	27	27
R^2	0.018	0.610	0.611	0.625
Adjusted R^2	−0.021	0.578	0.560	0.557
Residual Std. Error	0.937	0.602	0.615	0.617
F Statistic	0.463	18.779***	12.019***	9.163***
AIC	77.018	54.082	56.053	57.037
BIC	80.906	59.265	62.532	64.812
Note:		*p<0.1; **p<0.05; ***p<0.01		

Table 95: Standardized winter 6th grade English reading formative test scores at Site 2

	(1)	(2)	(3)	(4)
intercept	−0.140	−0.237	−0.311	−0.393
	(0.326)	(0.247)	(0.295)	(0.322)
ILI	0.608	0.662	0.678	0.879
	(0.523)	(0.395)	(0.406)	(0.501)
fall 4th ELA z-score		0.791***	0.781***	0.750***
		(0.218)	(0.225)	(0.233)
male			0.206	0.444
			(0.422)	(0.544)
ILI × male				−0.648
				(0.907)
Observations	18	18	18	18
R^2	0.078	0.508	0.517	0.535
Adjusted R^2	0.020	0.443	0.413	0.392
Residual Std. Error	1.082	0.816	0.837	0.852
F Statistic	1.353	7.752***	4.986**	3.736**
AIC	57.783	48.467	50.162	51.470
BIC	60.454	52.029	54.614	56.812
Note:		*p<0.1; **p<0.05; ***p<0.01		

Table 96: Standardized spring 6th grade English reading formative test scores at Site 2

	(1)	(2)	(3)	(4)
intercept	−0.226	−0.308	−0.434	−0.366
	(0.324)	(0.274)	(0.323)	(0.355)
ILI	0.563	0.609	0.635	0.469
	(0.519)	(0.436)	(0.444)	(0.552)
fall 4th ELA z-score		0.669**	0.652**	0.678**
		(0.241)	(0.246)	(0.257)
male			0.353	0.156
			(0.461)	(0.599)
ILI × male				0.536
				(1.000)
Observations	18	18	18	18
R^2	0.068	0.384	0.409	0.421
Adjusted R^2	0.010	0.302	0.282	0.243
Residual Std. Error	1.074	0.902	0.915	0.939
F Statistic	1.176	4.675**	3.226*	2.368
AIC	57.527	52.084	53.346	54.953
BIC	60.198	55.645	57.798	60.295
Note:		*p<0.1; **p<0.05; ***p<0.01		

Table 97: Standardized fall 7th grade English reading formative test scores at Site 2

	(1)	(2)	(3)	(4)
intercept	−0.040	−0.277	−0.180	−0.226
	(0.274)	(0.186)	(0.244)	(0.272)
ILI	−0.387	0.199	0.159	0.284
	(0.501)	(0.350)	(0.362)	(0.473)
fall 4th ELA z-score		0.706***	0.681***	0.695***
		(0.142)	(0.150)	(0.157)
male			−0.206	−0.110
			(0.326)	(0.403)
ILI × male				−0.315
				(0.735)
Observations	20	20	20	20
R^2	0.032	0.605	0.615	0.619
Adjusted R^2	−0.022	0.559	0.542	0.518
Residual Std. Error	1.026	0.674	0.686	0.705
F Statistic	0.597	13.023***	8.509***	6.102***
AIC	61.671	45.742	47.248	49.005
BIC	64.658	49.725	52.227	54.979
Note:		*p<0.1; **p<0.05; ***p<0.01		

Table 98: Standardized winter 7th grade English reading formative test scores at Site 2

	(1)	(2)	(3)	(4)
intercept	−0.038	−0.246	−0.268	−0.371
	(0.278)	(0.185)	(0.238)	(0.261)
ILI	−0.325	0.220	0.228	0.506
	(0.496)	(0.339)	(0.354)	(0.453)
fall 4th ELA z-score		0.681***	0.688***	0.725***
		(0.138)	(0.149)	(0.154)
male			0.051	0.290
			(0.335)	(0.414)
ILI × male				−0.707
				(0.718)
Observations	19	19	19	19
R^2	0.025	0.615	0.616	0.640
Adjusted R^2	−0.033	0.567	0.539	0.538
Residual Std. Error	1.004	0.650	0.671	0.672
F Statistic	0.431	12.779***	8.007***	6.235***
AIC	57.960	42.301	44.271	44.999
BIC	60.794	46.078	48.993	50.666
Note:		*p<0.1; **p<0.05; ***p<0.01		

Table 99: Standardized spring 7th grade English reading formative test scores at Site 2

	(1)	(2)	(3)	(4)
intercept	−0.157	−0.284	−0.288	−0.296
	(0.289)	(0.204)	(0.269)	(0.304)
ILI	−0.207	0.163	0.164	0.189
	(0.533)	(0.383)	(0.399)	(0.548)
fall 4th ELA z-score		0.635***	0.636***	0.638***
		(0.155)	(0.169)	(0.179)
male			0.007	0.026
			(0.375)	(0.471)
ILI × male				−0.060
				(0.834)
Observations	17	17	17	17
R^2	0.010	0.548	0.548	0.548
Adjusted R^2	−0.056	0.484	0.444	0.398
Residual Std. Error	1.001	0.700	0.726	0.756
F Statistic	0.151	8.494***	5.259**	3.644**
AIC	52.150	40.813	42.813	44.805
BIC	54.650	44.146	46.979	49.805
Note:		*p<0.1; **p<0.05; ***p<0.01		

Table 100: Standardized fall 8th grade English reading formative test scores at Site 2

	(1)	(2)	(3)	(4)
intercept	0.119	0.214	0.198	0.215
	(0.284)	(0.220)	(0.280)	(0.309)
ILI	−0.083	0.401	0.386	0.255
	(0.480)	(0.393)	(0.431)	(0.942)
fall 4th ELA z-score		0.897***	0.910***	0.895**
		(0.245)	(0.286)	(0.310)
male			0.046	−0.002
			(0.463)	(0.568)
ILI × male				0.170
				(1.078)
Observations	20	20	20	20
R^2	0.002	0.441	0.441	0.442
Adjusted R^2	−0.054	0.375	0.336	0.293
Residual Std. Error	1.025	0.789	0.813	0.839
F Statistic	0.030	6.699***	4.209**	2.971*
AIC	61.628	52.038	54.026	55.993
BIC	64.615	56.021	59.004	61.967
Note:		*p<0.1; **p<0.05; ***p<0.01		

Table 101: Standardized winter 8th grade English reading formative test scores at Site 2

	(1)	(2)	(3)	(4)
intercept	0.126	0.273	0.292	0.437
	(0.307)	(0.276)	(0.348)	(0.365)
ILI	−0.184	0.161	0.177	−0.922
	(0.506)	(0.465)	(0.510)	(1.069)
fall 4th ELA z-score		0.761**	0.747*	0.623
		(0.308)	(0.352)	(0.364)
male			−0.051	−0.458
			(0.549)	(0.645)
ILI × male				1.426
				(1.222)
Observations	19	19	19	19
R^2	0.008	0.282	0.282	0.346
Adjusted R^2	−0.051	0.192	0.138	0.159
Residual Std. Error	1.063	0.932	0.963	0.951
F Statistic	0.132	3.136*	1.964	1.849
AIC	60.133	55.995	57.984	58.222
BIC	62.966	59.773	62.706	63.889
Note:		*p<0.1; **p<0.05; ***p<0.01		

Table 102: Standardized spring 8th grade English reading formative test scores at Site 2

	(1)	(2)	(3)	(4)
intercept	0.310	0.423**	0.628***	0.745***
	(0.280)	(0.172)	(0.200)	(0.203)
ILI	−0.679	−0.201	−0.016	−0.900
	(0.499)	(0.317)	(0.316)	(0.620)
fall 4th ELA z-score		1.061***	0.895***	0.779***
		(0.195)	(0.207)	(0.209)
male			−0.579	−0.916**
			(0.330)	(0.376)
ILI × male				1.179
				(0.724)
Observations	19	19	19	19
R^2	0.098	0.684	0.738	0.779
Adjusted R^2	0.045	0.644	0.685	0.716
Residual Std. Error	1.010	0.617	0.580	0.551
F Statistic	1.853	17.307***	14.056***	12.367***
AIC	58.202	40.287	38.748	37.450
BIC	61.036	44.065	43.470	43.117
Note:		*p<0.1; **p<0.05; ***p<0.01		

Table 103: Standardized winter 4th grade math formative test scores at Site 2

	(1)	(2)	(3)	(4)
intercept	0.074	−0.167**	−0.140	−0.142
	(0.147)	(0.074)	(0.093)	(0.102)
ILI	−0.413	0.304**	0.313**	0.322
	(0.256)	(0.135)	(0.137)	(0.209)
fall 4th math z-score		0.899***	0.896***	0.896***
		(0.063)	(0.064)	(0.064)
male			−0.059	−0.055
			(0.121)	(0.149)
ILI × male				−0.015
				(0.264)
Observations	67	67	67	67
R^2	0.038	0.771	0.772	0.772
Adjusted R^2	0.024	0.764	0.761	0.758
Residual Std. Error	0.983	0.483	0.486	0.490
F Statistic	2.600	108.015***	71.235***	52.581***
AIC	191.873	97.608	99.354	101.350
BIC	198.488	106.427	110.377	114.578
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 104: Standardized spring 4th grade math formative test scores at Site 2

	(1)	(2)	(3)	(4)
intercept	0.080	−0.102	−0.036	−0.011
	(0.178)	(0.093)	(0.119)	(0.133)
ILI	−0.438	0.177	0.205	0.103
	(0.297)	(0.162)	(0.166)	(0.283)
fall 4th math z-score		0.874***	0.856***	0.847***
		(0.076)	(0.079)	(0.082)
male			−0.143	−0.195
			(0.159)	(0.199)
ILI × male				0.154
				(0.344)
Observations	50	50	50	50
R^2	0.043	0.750	0.754	0.755
Adjusted R^2	0.024	0.739	0.738	0.733
Residual Std. Error	1.007	0.520	0.522	0.526
F Statistic	2.182	70.412***	47.011***	34.694***
AIC	146.540	81.495	82.630	84.410
BIC	152.276	89.143	92.191	95.882
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 105: Standardized fall 5th grade math formative test scores at Site 2

	(1)	(2)	(3)	(4)
intercept	−0.120	−0.405**	−0.262	−0.138
	(0.252)	(0.182)	(0.204)	(0.237)
ILI	−0.262	0.386	0.492*	0.168
	(0.341)	(0.262)	(0.268)	(0.413)
fall 4th math z-score		0.735***	0.690***	0.640***
		(0.126)	(0.128)	(0.137)
male			−0.375	−0.690*
			(0.261)	(0.402)
ILI × male				0.558
				(0.541)
Observations	33	33	33	33
R^2	0.019	0.540	0.571	0.586
Adjusted R^2	−0.013	0.510	0.526	0.527
Residual Std. Error	0.975	0.678	0.666	0.666
F Statistic	0.593	17.626***	12.854***	9.928***
AIC	95.893	72.875	72.609	73.378
BIC	100.382	78.861	80.091	82.357
Note:		*p<0.1; **p<0.05; ***p<0.01		

Table 106: Standardized winter 5th grade math formative test scores at Site 2

	(1)	(2)	(3)	(4)
intercept	0.026	−0.263**	−0.233	−0.165
	(0.231)	(0.122)	(0.154)	(0.171)
ILI	−0.552	0.053	0.077	−0.209
	(0.341)	(0.187)	(0.205)	(0.374)
fall 4th math z-score		0.699***	0.689***	0.663***
		(0.083)	(0.091)	(0.095)
male			−0.070	−0.232
			(0.216)	(0.280)
ILI × male				0.410
				(0.447)
Observations	26	26	26	26
R^2	0.099	0.779	0.780	0.789
Adjusted R^2	0.061	0.760	0.750	0.748
Residual Std. Error	0.866	0.438	0.447	0.448
F Statistic	2.624	40.578***	26.034***	19.595***
AIC	70.213	35.640	37.517	38.496
BIC	73.987	40.673	43.807	46.044
Note:		*p<0.1; **p<0.05; ***p<0.01		

Table 107: Standardized spring 5th grade math formative test scores at Site 2

	(1)	(2)	(3)	(4)
intercept	0.045	−0.266*	−0.121	−0.064
	(0.256)	(0.137)	(0.169)	(0.187)
ILI	−0.507	0.118	0.271	−0.100
	(0.396)	(0.218)	(0.239)	(0.554)
fall 4th math z-score		0.756***	0.709***	0.679***
		(0.095)	(0.099)	(0.108)
male			−0.353	−0.479
			(0.251)	(0.305)
ILI × male				0.461
				(0.621)
Observations	24	24	24	24
R^2	0.069	0.767	0.788	0.794
Adjusted R^2	0.027	0.745	0.757	0.751
Residual Std. Error	0.957	0.490	0.479	0.484
F Statistic	1.639	34.626***	24.829***	18.344***
AIC	69.907	38.638	38.366	39.677
BIC	73.441	43.350	44.256	46.746

Note: *p<0.1; **p<0.05; ***p<0.01

Table 108: Standardized fall 6th grade math formative test scores at Site 2

	(1)	(2)	(3)	(4)
intercept	−0.509**	−0.652***	−0.674***	−0.663***
	(0.218)	(0.149)	(0.167)	(0.191)
ILI	0.443	0.770***	0.750***	0.727**
	(0.283)	(0.199)	(0.212)	(0.284)
fall 4th math z-score		0.572***	0.568***	0.570***
		(0.103)	(0.105)	(0.109)
male			0.065	0.033
			(0.199)	(0.324)
ILI × male				0.053
				(0.417)
Observations	27	27	27	27
R^2	0.089	0.603	0.605	0.605
Adjusted R^2	0.053	0.570	0.553	0.533
Residual Std. Error	0.722	0.486	0.496	0.507
F Statistic	2.456	18.222***	11.732***	8.427***
AIC	62.921	42.512	44.386	46.366
BIC	66.808	47.695	50.865	54.141

Note: *p<0.1; **p<0.05; ***p<0.01

Table 109: Standardized winter 6th grade math formative test scores at Site 2

	(1)	(2)	(3)	(4)
intercept	−0.475*	−0.661***	−0.646***	−0.569***
	(0.227)	(0.136)	(0.162)	(0.165)
ILI	0.259	0.706***	0.703***	0.526*
	(0.363)	(0.226)	(0.234)	(0.257)
fall 4th math z-score		0.749***	0.751***	0.829***
		(0.133)	(0.137)	(0.143)
male			−0.044	−0.310
			(0.228)	(0.287)
ILI × male				0.712
				(0.493)
Observations	18	18	18	18
R^2	0.031	0.690	0.691	0.734
Adjusted R^2	−0.030	0.649	0.625	0.652
Residual Std. Error	0.751	0.439	0.453	0.437
F Statistic	0.509	16.710***	10.437***	8.956***
AIC	44.668	26.138	28.091	27.412
BIC	47.340	29.699	32.543	32.755
Note:		*p<0.1; **p<0.05; ***p<0.01		

Table 110: Standardized spring 6th grade math formative test scores at Site 2

	(1)	(2)	(3)	(4)
intercept	−0.604**	−0.745***	−0.814***	−0.635***
	(0.220)	(0.180)	(0.211)	(0.173)
ILI	0.942**	1.280***	1.290***	0.878***
	(0.353)	(0.298)	(0.304)	(0.269)
fall 4th math z-score		0.568***	0.559***	0.740***
		(0.175)	(0.178)	(0.149)
male			0.197	−0.420
			(0.296)	(0.300)
ILI × male				1.653***
				(0.515)
Observations	18	18	18	18
R^2	0.308	0.594	0.606	0.780
Adjusted R^2	0.265	0.540	0.522	0.713
Residual Std. Error	0.730	0.578	0.589	0.456
F Statistic	7.121**	10.974***	7.191***	11.543***
AIC	43.644	36.044	37.486	28.992
BIC	46.315	39.606	41.937	34.334
Note:		*p<0.1; **p<0.05; ***p<0.01		

Table 111: Standardized fall 7th grade math formative test scores at Site 2

	(1)	(2)	(3)	(4)
intercept	−0.230	−0.274**	−0.206	−0.138
	(0.253)	(0.123)	(0.157)	(0.166)
ILI	−0.066	0.442*	0.415	0.212
	(0.462)	(0.233)	(0.240)	(0.296)
fall 4th math z-score		0.862***	0.842***	0.829***
		(0.111)	(0.116)	(0.116)
male			−0.157	−0.314
			(0.219)	(0.257)
ILI × male				0.543
				(0.473)
Observations	20	20	20	20
R^2	0.001	0.779	0.786	0.803
Adjusted R^2	−0.054	0.753	0.745	0.750
Residual Std. Error	0.947	0.459	0.466	0.461
F Statistic	0.020	29.913***	19.536***	15.271***
AIC	58.489	30.345	31.719	32.037
BIC	61.477	34.328	36.698	38.011
Note:		*p<0.1; **p<0.05; ***p<0.01		

Table 112: Standardized winter 7th grade math formative test scores at Site 2

	(1)	(2)	(3)	(4)
intercept	−0.088	−0.153	−0.166	−0.062
	(0.322)	(0.147)	(0.185)	(0.189)
ILI	0.022	0.649**	0.653**	0.343
	(0.574)	(0.273)	(0.284)	(0.337)
fall 4th math z-score		1.044***	1.048***	1.027***
		(0.129)	(0.137)	(0.132)
male			0.033	−0.234
			(0.268)	(0.309)
ILI × male				0.850
				(0.546)
Observations	19	19	19	19
R^2	0.0001	0.804	0.804	0.833
Adjusted R^2	−0.059	0.779	0.765	0.785
Residual Std. Error	1.162	0.531	0.548	0.524
F Statistic	0.001	32.728***	20.480***	17.423***
AIC	63.521	34.601	36.582	35.549
BIC	66.354	38.378	41.304	41.215
Note:		*p<0.1; **p<0.05; ***p<0.01		

Table 113: Standardized spring 7th grade math formative test scores at Site 2

	(1)	(2)	(3)	(4)
intercept	−0.071	−0.049	−0.131	−0.016
	(0.350)	(0.148)	(0.186)	(0.183)
ILI	−0.097	0.305	0.318	−0.067
	(0.645)	(0.277)	(0.282)	(0.335)
fall 4th math z-score		1.100***	1.125***	1.122***
		(0.132)	(0.138)	(0.127)
male			0.198	−0.079
			(0.264)	(0.288)
ILI × male				0.947*
				(0.521)
Observations	17	17	17	17
R^2	0.001	0.833	0.840	0.875
Adjusted R^2	−0.065	0.809	0.803	0.833
Residual Std. Error	1.211	0.513	0.521	0.480
F Statistic	0.022	34.950***	22.760***	20.931***
AIC	58.638	30.223	31.503	29.361
BIC	61.137	33.556	35.669	34.360

Note: *p<0.1; **p<0.05; ***p<0.01

Table 114: Standardized fall 8th grade math formative test scores at Site 2

	(1)	(2)	(3)	(4)
intercept	0.219	0.023	0.590**	0.583*
	(0.300)	(0.244)	(0.254)	(0.299)
ILI	−0.277	0.256	0.611*	0.645
	(0.494)	(0.422)	(0.345)	(0.814)
fall 4th math z-score		0.718***	0.481**	0.485**
		(0.214)	(0.181)	(0.211)
male			−1.205***	−1.192**
			(0.357)	(0.465)
ILI × male				−0.042
				(0.915)
Observations	19	19	19	19
R^2	0.018	0.423	0.672	0.672
Adjusted R^2	−0.040	0.351	0.606	0.578
Residual Std. Error	1.039	0.821	0.639	0.661
F Statistic	0.315	5.861**	10.246***	7.174***
AIC	59.249	51.155	42.415	44.412
BIC	62.083	54.933	47.137	50.079

Note: *p<0.1; **p<0.05; ***p<0.01

Table 115: Standardized winter 8th grade math formative test scores at Site 2

	(1)	(2)	(3)	(4)
intercept	0.179	0.025	0.418	0.614*
	(0.304)	(0.255)	(0.310)	(0.340)
ILI	−0.239	0.230	0.503	−0.571
	(0.501)	(0.441)	(0.430)	(0.945)
fall 4th math z-score		0.670***	0.519**	0.389
		(0.223)	(0.219)	(0.238)
male			−0.861*	−1.259**
			(0.440)	(0.533)
ILI × male				1.352
				(1.065)
Observations	19	19	19	19
R^2	0.013	0.370	0.498	0.550
Adjusted R^2	−0.045	0.291	0.398	0.421
Residual Std. Error	1.052	0.867	0.799	0.783
F Statistic	0.228	4.698**	4.963**	4.277**
AIC	59.749	53.224	50.903	50.833
BIC	62.582	57.002	55.625	56.500
Note:		*p<0.1; **p<0.05; ***p<0.01		

Table 116: Standardized spring 8th grade math formative test scores at Site 2

	(1)	(2)	(3)	(4)
intercept	0.221	0.051	0.385	0.453
	(0.289)	(0.211)	(0.242)	(0.274)
ILI	−0.445	0.060	0.331	−0.086
	(0.513)	(0.389)	(0.370)	(0.812)
fall 4th math z-score		0.792***	0.658***	0.607**
		(0.192)	(0.183)	(0.206)
male			−0.794**	−0.942*
			(0.359)	(0.448)
ILI × male				0.542
				(0.934)
Observations	19	19	19	19
R^2	0.042	0.535	0.649	0.657
Adjusted R^2	−0.014	0.477	0.579	0.560
Residual Std. Error	1.040	0.747	0.670	0.686
F Statistic	0.751	9.205***	9.254***	6.717***
AIC	59.307	47.579	44.224	45.773
BIC	62.141	51.357	48.947	51.440
Note:		*p<0.1; **p<0.05; ***p<0.01		

12 Appendix 3: Supplementary Tables for Site 3

Table 117: Standardized fall 6th grade ELA formative test scores at Site 3

	(1)	(2)	(3)	(4)
intercept	−0.042	0.060	0.250	0.248
	(0.124)	(0.167)	(0.211)	(0.226)
ILI	0.142	0.113	0.081	0.087
	(0.230)	(0.232)	(0.232)	(0.305)
male		−0.194	−0.223	−0.220
		(0.212)	(0.211)	(0.251)
low SES			−0.307	−0.307
			(0.211)	(0.212)
ILI × male				−0.013
				(0.473)
Observations	92	92	92	92
R^2	0.004	0.014	0.037	0.037
Adjusted R^2	−0.007	−0.009	0.004	−0.007
Residual Std. Error	1.003	1.004	0.998	1.004
F Statistic	0.384	0.610	1.122	0.832
AIC	265.688	266.827	266.625	268.624
BIC	273.253	276.914	279.234	283.755
Note:		*p<0.1; **p<0.05; ***p<0.01		

Table 118: Standardized winter 6th grade ELA formative test scores at Site 3

	(1)	(2)	(3)	(4)	(5)
intercept	−0.104	−0.068	−0.004	0.109	0.108
	(0.123)	(0.062)	(0.083)	(0.106)	(0.113)
ILI	0.354	0.232**	0.214*	0.197*	0.200
	(0.227)	(0.115)	(0.116)	(0.115)	(0.152)
fall 6th grade z-score		0.854***	0.848***	0.834***	0.834***
		(0.053)	(0.053)	(0.053)	(0.053)
male			−0.124	−0.144	−0.142
			(0.106)	(0.106)	(0.125)
low SES				−0.181*	−0.181*
				(0.106)	(0.107)
ILI × male					−0.007
					(0.235)
Observations	92	92	92	92	92
R^2	0.026	0.753	0.757	0.765	0.765
Adjusted R^2	0.015	0.747	0.748	0.754	0.751
Residual Std. Error	0.992	0.503	0.502	0.496	0.499
F Statistic	2.426	135.619***	91.230***	70.651***	55.872***
AIC	263.632	139.451	140.042	138.999	140.998
BIC	271.198	149.539	152.651	154.130	158.651
Note:			*p<0.1; **p<0.05; ***p<0.01		

Table 119: Standardized spring 6th grade ELA formative test scores at Site 3

	(1)	(2)	(3)	(4)	(5)
intercept	−0.088	−0.052	−0.060	−0.010	0.020
	(0.124)	(0.057)	(0.077)	(0.099)	(0.106)
ILI	0.301	0.176	0.178	0.171	0.096
	(0.228)	(0.106)	(0.108)	(0.108)	(0.142)
fall 6th grade z-score		0.880***	0.881***	0.875***	0.875***
		(0.049)	(0.049)	(0.050)	(0.050)
male			0.015	0.006	−0.044
			(0.098)	(0.099)	(0.117)
low SES				−0.080	−0.086
				(0.099)	(0.100)
ILI × male					0.179
					(0.220)
Observations	92	92	92	92	92
R^2	0.019	0.791	0.791	0.793	0.794
Adjusted R^2	0.008	0.786	0.784	0.783	0.782
Residual Std. Error	0.996	0.462	0.465	0.466	0.467
F Statistic	1.744	168.361***	111.017***	83.091***	66.347***
AIC	264.313	124.086	126.061	127.379	128.673
BIC	271.879	134.173	138.670	142.510	146.325

Note: *p<0.1; **p<0.05; ***p<0.01

Table 120: Standardized fall 7th grade ELA formative test scores at Site 3

	(1)	(2)	(3)	(4)	(5)
intercept	0.012	0.009	0.025	0.213**	0.204*
	(0.123)	(0.065)	(0.086)	(0.106)	(0.114)
ILI	0.039	−0.046	−0.050	−0.078	−0.054
	(0.226)	(0.119)	(0.120)	(0.116)	(0.153)
fall 6th grade z-score		0.873***	0.872***	0.852***	0.852***
		(0.057)	(0.058)	(0.056)	(0.056)
male			−0.031	−0.066	−0.050
			(0.110)	(0.107)	(0.127)
low SES				−0.302***	−0.300***
				(0.107)	(0.108)
ILI × male					−0.056
					(0.237)
Observations	91	91	91	91	91
R^2	0.0003	0.727	0.727	0.750	0.751
Adjusted R^2	−0.011	0.721	0.718	0.739	0.736
Residual Std. Error	0.983	0.517	0.520	0.500	0.503
F Statistic	0.030	117.153***	77.311***	64.633***	51.150***
AIC	259.182	143.080	144.997	138.919	140.859
BIC	266.714	153.123	157.551	153.984	158.435

Note: *p<0.1; **p<0.05; ***p<0.01

Table 121: Standardized winter 7th grade ELA formative test scores at Site 3

	(1)	(2)	(3)	(4)	(5)
intercept	−0.028	−0.031	−0.011	0.068	0.085
	(0.124)	(0.083)	(0.110)	(0.142)	(0.152)
ILI	0.107	0.032	0.026	0.014	−0.029
	(0.228)	(0.152)	(0.154)	(0.155)	(0.203)
fall 6th grade z-score		0.778***	0.776***	0.768***	0.768***
		(0.073)	(0.073)	(0.074)	(0.075)
male			−0.039	−0.053	−0.083
			(0.141)	(0.142)	(0.169)
low SES				−0.127	−0.131
				(0.142)	(0.143)
ILI × male					0.104
					(0.316)
Observations	91	91	91	91	91
R^2	0.002	0.565	0.565	0.569	0.570
Adjusted R^2	−0.009	0.555	0.550	0.549	0.544
Residual Std. Error	0.994	0.660	0.664	0.665	0.668
F Statistic	0.219	57.111***	37.699***	28.406***	22.511***
AIC	261.152	187.661	189.582	190.745	192.629
BIC	268.685	197.705	202.136	205.811	210.205

Note: *p<0.1; **p<0.05; ***p<0.01

Table 122: Standardized spring 7th grade ELA formative test scores at Site 3

	(1)	(2)	(3)	(4)	(5)
intercept	0.003	0.0004	−0.030	0.096	0.094
	(0.125)	(0.078)	(0.104)	(0.133)	(0.143)
ILI	−0.016	−0.095	−0.087	−0.105	−0.101
	(0.229)	(0.144)	(0.146)	(0.145)	(0.191)
fall 6th grade z-score		0.812***	0.815***	0.801***	0.801***
		(0.069)	(0.070)	(0.070)	(0.070)
male			0.059	0.036	0.038
			(0.133)	(0.133)	(0.159)
low SES				−0.202	−0.202
				(0.133)	(0.135)
ILI × male					−0.009
					(0.296)
Observations	91	91	91	91	91
R^2	0.0001	0.611	0.612	0.622	0.622
Adjusted R^2	−0.011	0.603	0.599	0.605	0.600
Residual Std. Error	0.997	0.625	0.628	0.624	0.627
F Statistic	0.005	69.237***	45.801***	35.436***	28.019***
AIC	261.760	177.743	179.540	179.143	181.142
BIC	269.293	187.787	192.094	194.208	198.718

Note: *p<0.1; **p<0.05; ***p<0.01

Table 123: Standardized fall 8th grade ELA formative test scores at Site 3

	(1)	(2)	(3)	(4)	(5)
intercept	0.016	0.050	0.008	0.130	0.172
	(0.121)	(0.068)	(0.092)	(0.117)	(0.125)
ILI	0.289	0.171	0.181	0.159	0.054
	(0.227)	(0.128)	(0.130)	(0.129)	(0.171)
fall 6th grade z-score		0.801***	0.805***	0.790***	0.790***
		(0.058)	(0.058)	(0.059)	(0.059)
male			0.080	0.060	−0.009
			(0.118)	(0.117)	(0.138)
low SES				−0.197*	−0.208*
				(0.118)	(0.118)
ILI × male					0.247
					(0.261)
Observations	91	91	91	91	91
R^2	0.018	0.690	0.692	0.702	0.705
Adjusted R^2	0.007	0.683	0.681	0.688	0.687
Residual Std. Error	0.977	0.552	0.553	0.548	0.548
F Statistic	1.628	98.054***	65.128***	50.573***	40.590***
AIC	257.988	154.984	156.497	155.561	156.605
BIC	265.520	165.027	169.052	170.626	174.181
Note:			*p<0.1; **p<0.05; ***p<0.01		

Table 124: Standardized winter 8th grade ELA formative test scores at Site 3

	(1)	(2)	(3)	(4)	(5)
intercept	−0.007	0.030	0.082	0.282**	0.318**
	(0.124)	(0.083)	(0.111)	(0.141)	(0.150)
ILI	0.226	0.035	0.024	0.006	−0.088
	(0.230)	(0.156)	(0.157)	(0.154)	(0.203)
fall 6th grade z-score		0.759***	0.752***	0.718***	0.720***
		(0.073)	(0.074)	(0.074)	(0.074)
male			−0.103	−0.148	−0.210
			(0.143)	(0.141)	(0.166)
low SES				−0.316**	−0.323**
				(0.142)	(0.143)
ILI × male					0.222
					(0.311)
Observations	90	90	90	90	90
R^2	0.011	0.559	0.562	0.586	0.588
Adjusted R^2	−0.0004	0.549	0.546	0.566	0.564
Residual Std. Error	0.990	0.665	0.666	0.652	0.654
F Statistic	0.962	55.157***	36.739***	30.063***	24.013***
AIC	257.545	186.824	188.285	185.182	186.637
BIC	265.044	196.823	200.784	200.180	204.136
Note:			*p<0.1; **p<0.05; ***p<0.01		

Table 125: Standardized fall 6th grade math formative test scores at Site 3

	(1)	(2)	(3)	(4)
intercept	−0.034	0.108	0.404*	0.386*
	(0.128)	(0.171)	(0.212)	(0.229)
ILI	0.113	0.070	0.022	0.063
	(0.232)	(0.233)	(0.229)	(0.303)
male		−0.268	−0.321	−0.292
		(0.215)	(0.211)	(0.253)
low SES			−0.474**	−0.470**
			(0.210)	(0.212)
ILI × male				−0.098
				(0.467)
Observations	89	89	89	89
R^2	0.003	0.020	0.076	0.076
Adjusted R^2	−0.009	−0.002	0.043	0.032
Residual Std. Error	1.004	1.001	0.978	0.984
F Statistic	0.239	0.899	2.326*	1.736
AIC	257.321	257.724	254.543	256.496
BIC	264.787	267.678	266.986	271.428
Note:		*p<0.1; **p<0.05; ***p<0.01		

Table 126: Standardized winter 6th grade math formative test scores at Site 3

	(1)	(2)	(3)	(4)	(5)
intercept	−0.094	−0.010	0.072	0.122	0.139
	(0.123)	(0.058)	(0.077)	(0.100)	(0.107)
ILI	0.320	0.172	0.148	0.141	0.100
	(0.228)	(0.105)	(0.105)	(0.106)	(0.140)
fall 6th grade z-score		0.813***	0.802***	0.793***	0.794***
		(0.048)	(0.048)	(0.050)	(0.050)
male			−0.156	−0.167*	−0.195
			(0.097)	(0.099)	(0.117)
low SES				−0.078	−0.082
				(0.100)	(0.100)
ILI × male					0.097
					(0.215)
Observations	92	89	89	89	89
R^2	0.021	0.770	0.777	0.778	0.779
Adjusted R^2	0.011	0.765	0.769	0.768	0.765
Residual Std. Error	0.995	0.454	0.450	0.451	0.453
F Statistic	1.974	143.852***	98.482***	73.676***	58.422***
AIC	264.084	116.947	116.314	117.669	119.453
BIC	271.649	126.901	128.758	132.601	136.873
Note:			*p<0.1; **p<0.05; ***p<0.01		

Table 127: Standardized spring 6th grade math formative test scores at Site 3

	(1)	(2)	(3)	(4)	(5)
intercept	−0.137	−0.045	−0.059	−0.052	−0.030
	(0.122)	(0.067)	(0.091)	(0.119)	(0.128)
ILI	0.478**	0.353***	0.356***	0.355***	0.303*
	(0.228)	(0.124)	(0.125)	(0.127)	(0.169)
fall 6th grade z-score		0.748***	0.750***	0.748***	0.749***
		(0.057)	(0.057)	(0.060)	(0.060)
male			0.026	0.024	−0.011
			(0.115)	(0.117)	(0.139)
low SES				−0.011	−0.017
				(0.119)	(0.120)
ILI × male					0.121
					(0.258)
Observations	91	88	88	88	88
R^2	0.047	0.686	0.686	0.686	0.687
Adjusted R^2	0.036	0.678	0.675	0.671	0.668
Residual Std. Error	0.982	0.529	0.532	0.535	0.537
F Statistic	4.405**	92.661***	61.100***	45.286***	35.932***
AIC	258.845	142.521	144.469	146.460	148.225
BIC	266.378	152.430	156.856	161.324	165.566

Note: *p<0.1; **p<0.05; ***p<0.01

Table 128: Standardized fall 7th grade math formative test scores at Site 3

	(1)	(2)	(3)	(4)	(5)
intercept	−0.143	−0.116	−0.039	−0.205	−0.137
	(0.123)	(0.080)	(0.107)	(0.137)	(0.146)
ILI	0.425*	0.337**	0.316**	0.342**	0.179
	(0.225)	(0.145)	(0.146)	(0.145)	(0.189)
fall 6th grade z-score		0.762***	0.752***	0.784***	0.786***
		(0.067)	(0.067)	(0.068)	(0.068)
male			−0.147	−0.112	−0.226
			(0.136)	(0.135)	(0.160)
low SES				0.259*	0.245*
				(0.137)	(0.137)
ILI × male					0.387
					(0.292)
Observations	90	88	88	88	88
R^2	0.039	0.619	0.624	0.639	0.647
Adjusted R^2	0.028	0.610	0.610	0.622	0.625
Residual Std. Error	0.980	0.626	0.626	0.617	0.614
F Statistic	3.549*	68.945***	46.439***	36.777***	30.038***
AIC	255.789	172.364	173.154	171.465	171.604
BIC	263.289	182.273	185.541	186.329	188.945

Note: *p<0.1; **p<0.05; ***p<0.01

Table 129: Standardized winter 7th grade math formative test scores at Site 3

	(1)	(2)	(3)	(4)	(5)
intercept	−0.082	−0.063	−0.065	−0.115	−0.116
	(0.122)	(0.084)	(0.113)	(0.147)	(0.159)
ILI	0.330	0.253*	0.254	0.261*	0.263
	(0.224)	(0.152)	(0.155)	(0.156)	(0.206)
fall 6th grade z-score		0.736***	0.736***	0.746***	0.746***
		(0.070)	(0.071)	(0.074)	(0.074)
male			0.004	0.015	0.017
			(0.143)	(0.146)	(0.173)
low SES				0.078	0.078
				(0.147)	(0.148)
ILI × male					−0.006
					(0.318)
Observations	91	89	89	89	89
R^2	0.024	0.571	0.571	0.573	0.573
Adjusted R^2	0.013	0.561	0.556	0.552	0.547
Residual Std. Error	0.977	0.658	0.662	0.665	0.669
F Statistic	2.166	57.295***	37.753***	28.146***	22.249***
AIC	257.942	183.124	185.124	186.825	188.825
BIC	265.475	193.079	197.567	201.757	206.245
Note:					*p<0.1; **p<0.05; ***p<0.01

Table 130: Standardized spring 7th grade math formative test scores at Site 3

	(1)	(2)	(3)	(4)	(5)
intercept	−0.081	−0.049	−0.062	−0.187	−0.120
	(0.125)	(0.084)	(0.114)	(0.147)	(0.156)
ILI	0.222	0.147	0.150	0.165	−0.003
	(0.233)	(0.154)	(0.156)	(0.156)	(0.206)
fall 6th grade z-score		0.760***	0.761***	0.784***	0.786***
		(0.070)	(0.071)	(0.073)	(0.073)
male			0.023	0.053	−0.059
			(0.144)	(0.145)	(0.170)
low SES				0.196	0.183
				(0.147)	(0.146)
ILI × male					0.391
					(0.314)
Observations	90	88	88	88	88
R^2	0.010	0.582	0.582	0.591	0.598
Adjusted R^2	−0.001	0.572	0.567	0.571	0.574
Residual Std. Error	1.000	0.659	0.663	0.660	0.658
F Statistic	0.908	59.090***	38.951***	29.935***	24.415***
AIC	259.338	181.284	183.256	183.378	183.735
BIC	266.837	191.193	195.643	198.242	201.076
Note:					*p<0.1; **p<0.05; ***p<0.01

Table 131: Standardized fall 8th grade math formative test scores at Site 3

	(1)	(2)	(3)	(4)	(5)
intercept	0.045	0.135*	0.112	0.076	0.103
	(0.122)	(0.075)	(0.103)	(0.141)	(0.146)
ILI	0.002	−0.324**	−0.317*	−0.313*	−0.416*
	(0.266)	(0.162)	(0.164)	(0.166)	(0.211)
fall 6th grade z-score		0.707***	0.710***	0.717***	0.718***
		(0.067)	(0.068)	(0.071)	(0.072)
male			0.044	0.054	0.001
			(0.135)	(0.139)	(0.155)
low SES				0.054	0.057
				(0.143)	(0.143)
ILI × male					0.266
					(0.337)
Observations	81	78	78	78	78
R^2	0.00000	0.596	0.596	0.597	0.601
Adjusted R^2	−0.013	0.585	0.580	0.575	0.573
Residual Std. Error	0.975	0.584	0.587	0.591	0.592
F Statistic	0.0001	55.283***	36.451***	27.057***	21.658***
AIC	229.686	142.331	144.219	146.068	147.397
BIC	236.869	151.757	156.003	160.208	163.894
Note:			*p<0.1; **p<0.05; ***p<0.01		

Table 132: Standardized winter 8th grade math formative test scores at Site 3

	(1)	(2)	(3)	(4)	(5)
intercept	0.104	0.131	0.192*	0.160	0.091
	(0.120)	(0.079)	(0.107)	(0.141)	(0.149)
ILI	−0.242	−0.363**	−0.378**	−0.375**	−0.200
	(0.223)	(0.146)	(0.147)	(0.149)	(0.196)
fall 6th grade z-score		0.743***	0.734***	0.741***	0.736***
		(0.067)	(0.068)	(0.071)	(0.070)
male			−0.116	−0.108	0.009
			(0.136)	(0.138)	(0.162)
low SES				0.049	0.061
				(0.140)	(0.140)
ILI × male					−0.407
					(0.299)
Observations	90	88	88	88	88
R^2	0.013	0.595	0.599	0.599	0.608
Adjusted R^2	0.002	0.586	0.585	0.580	0.584
Residual Std. Error	0.958	0.623	0.624	0.628	0.624
F Statistic	1.184	62.549***	41.805***	31.057***	25.474***
AIC	251.658	171.490	172.735	174.604	174.634
BIC	259.157	181.399	185.122	189.468	191.975
Note:			*p<0.1; **p<0.05; ***p<0.01		

References

Barik, H. C., & Swain, M. (1976). A longitudinal study of bilingual and cognitive development. *International Journal of Psychology*, *11*(4), 251–263.

Brayboy, B. M. J., Gough, H. R., Leonard, B., Roehl II, R. F., & Solyom, J. A. (2012). Reclaiming scholarship: Critical Indigenous research methodologies. In F. J. Riemer, M. T. Quartaroli, & S. D. Lapan (Eds.), *Qualitative Research: An Introduction to Methods and Designs* (pp. 423–450). Jossey-Bass.

Campbell, D. T. (1957). Factors relevant to the validity of experiments in social settings [Publisher: US: American Psychological Association]. *Psychological Bulletin*, *54*(4), 297. https://doi.org/10.1037/h0040950

Center for Applied Linguistics. (1999). *Rating scale for CAL Oral Proficiency Exam (COPE) and Student Oral Proficiency Assessment (SOPA)*. Center for Applied Linguistics. Washington, DC. Retrieved May 6, 2024, from https://www.fhps.net/wp-content/uploads/2015/09/SOPA_Rubric.pdf

Cinelli, C., & Hazlett, C. (2020). Making sense of sensitivity: Extending omitted variable bias. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *82*(1), 39–67. https://doi.org/10.1111/rssb.12348

Cram, F., Kennedy, V., Paipa, K., Pipi, K., & Wehipeihana, N. (2015). Being culturally responsive through Kaupapa Maōri evaluation. In S. Hood, R. K. Hopson, & H. Frierson (Eds.), *Continuing the Journey to Reposition Culture and Cultural Context in Evaluation Theory and Practice* (pp. 289–311). Information Age Publishing, Inc.

Every Student Succeeds Act (2015).

Frank, K. A. (2000). Impact of a confounding variable on a regression coefficient. *Sociological Methods & Research*, *29*(2), 147–194. https://doi.org/10.1177/0049124100029002001

Frank, K. A., Lin, Q., Xu, R., Maroulis, S., & Mueller, A. (2023). Quantifying the robustness of causal inferences: Sensitivity analysis for pragmatic social science. *Social Science Research*, *110*, 102815. https://doi.org/10.1016/j.ssresearch.2022.102815

Frank, K. A., Maroulis, S. J., Duong, M. Q., & Kelcey, B. M. (2013). What would it take to change an inference? Using Rubin's causal model to interpret the robustness of causal inferences. *Educational Evaluation and Policy Analysis*, *35*(4), 437–460. https://doi.org/10.3102/0162373713493129

Frazier-Anderson, P., Hood, S., & Hopson, R. K. (2012). Preliminary considerations of an African American culturally responsive evaluation system. In F. J. Riemer, M. T. Quartaroli, & S. D. Lapan (Eds.), *Qualitative Research: An Introduction to Methods and Designs* (pp. 347–372). Jossey-Bass.

Genesee, F. (1994). *Integrating language and content: Lessons from immersion* (Educational Practice Report No. 11). National Center for Research on Cultural Diversity and Second Language Learning. Washington, DC.

Greene, E. B. (1941). *Measurements of human behavior.* [Pages: xxi, 777]. Odyssey Press. https://doi.org/10.1037/12221-000

Hermes, M., & Kawaiʻaeʻa, K. (2014). Revitalizing Indigenous languages through Indigenous immersion education. *Journal of Immersion and Content-Based Language Education*, *2*(2), 303–322. https://doi.org/https://doi.org/10.1075/jicb.2.2.10her

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, *81*(396), 945–960.

Holm, A., & Holm, W. (1990). Rock Point, a Navajo way to go to school: A valediction. *The Annals of the American Academy of Political and Social Science*, *508*(1), 170–184. https://doi.org/10.1177/0002716290508001014

Holm, A., & Holm, W. (1995). Navajo language education: Retrospect and prospects. *Bilingual Research Journal*, *19*(1), 141–167. https://doi.org/10.1080/15235882.1995.10668595

Holm, W. (1996, November). On some of the advantages of Navajo language literacy.

Hood, S., Hopson, R. K., & Kirkhart, K. E. (2015, October 14). Culturally responsive evaluation. In K. E. Newcomer, H. P. Hatry, & J. S. Wholey (Eds.), *Handbook of Practical Program Evaluation* (pp. 281–317). John Wiley & Sons, Inc. https://doi.org/10.1002/9781119171386.ch12

Lee, T. S., McCarty, T. L., Halle-Erby, K., Jacobson, T., Seltzer, M. H., McKenzie, J., & Nicholas, S. E. (2024). Indigenous language immersion schools—The link to sustainable and healthy Indigenous community futures. *Wicazo Sa Review*, in press.

Leonard, W. Y. (2023). Refusing "endangered languages" narratives. *Daedalus*, *152*(3), 69–83. https://doi.org/10.1162/daed_a_02018

May, S., Hill, R., & Tiakiwai, S. (2004). *Bilingual/immersion education: indicators of good practice : final report to the Ministry of Education* [OCLC: 156748446]. Ministry of Education.

McCarty, T. L. (2021). The holistic benefits of education for Indigenous language revitalisation and reclamation (ELR ^{2}). *Journal of Multilingual and Multicultural Development*, *42*(10), 927–940. https://doi.org/10.1080/01434632.2020.1827647

McCarty, T. L., Lee, T. S., Nicholas, S. E., Seltzer, M. H., Jacobson, T. A., & McKenzie, J. (2016). Indigenous-language immersion and Native American student achievement study, Spencer Foundation Award 201700054 (2016-2023).

McCarty, T. L., Noguera, J., Lee, T. S., & Nicholas, S. E. (2021). "A viable path for education"—Indigenous-language immersion and sustainable self-determination. *Journal of Language, Identity & Education*, *20*(5), 340–354. https://doi.org/10.1080/15348458.2021.1957681

Nowell, A., & Hedges, L. V. (1998). Trends in gender differences in academic achievement from 1960 to 1994: An analysis of differences in mean, variance, and extreme scores. *Sex Roles*, *39*(1), 21–43. https://doi.org/https://doi.org/10.1023/A:1018873615316

Paipa, K., Cram, F., Kennedy, V., & Pipi, K. (2015). Culturally responsive methods for family centered evaluation. In S. Hood, R. K. Hopson, & H. Frierson (Eds.), *Continuing the Journey to Reposition Culture and Cultural Context in Evaluation Theory and Practice* (pp. 313–334). Information Age Publishing, Inc.

Raudenbush, S. W. (2008). Advancing educational policy by advancing research on instruction. *American Educational Research Journal*, *45*(1), 206–230. https://doi.org/10.3102/0002831207312905

Raudenbush, S. W. (2009). Adaptive centering with random effects: An alternative to the fixed effects model for studying time-varying treatments in school settings. *Education Finance and Policy*, *4*(4), 468–491. https://doi.org/10.1162/edfp.2009.4.4.468

Rosier, P., & Farella, M. (1976). Bilingual education at Rock Point: Some early results. *TESOL Quarterly*, *10*(4), 379. https://doi.org/10.2307/3585519

Rosier, P., & Holm, W. (1980, September). *The Rock Point experience: A longitudinal study of a Navajo school program (saad naaki bee na'nitin)* (Bilingual Education Series No. 8). Center for Applied Linguistics. Washington, DC.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*(5), 688–701. https://doi.org/10.1037/h0037350

Rubin, D. B. (1986). Statistics and causal inference: Comment: Which ifs have causal answers. *Journal of the American Statistical Association*, *81*(396), 961. https://doi.org/10.2307/2289065

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Houghton Mifflin.

Shepard, L. A. (2017, September 25). Formative assessment: Caveat emptor. In C. A. Dwyer (Ed.), *The Future of Assessment: Shaping Teaching and Learning* (1st ed., pp. 279–303). Routledge. https://doi.org/10.4324/9781315086545

Smith, L. T. (2021). *Decolonizing Methodologies: Research and Indigenous Peoples* (3rd). Bloomsbury Publishing.

Sun, S., & Pan, W. (2014). A methodological review of statistical methods for handling multilevel non-nested longitudinal data in educational research. *International Journal of Research & Method in Education*, *37*(3), 285–308. https://doi.org/10.1080/1743727X.2014.885012

Swender, E., Conrad, D. J., & Vicars, R. (2012). *ACTFL proficiency guidelines 2012*. American Council on the Teaching of Foreign Languages. Alexandria, VA. Retrieved May 6, 2024, from https://www.actfl.org/uploads/files/general/ACTFLProficiencyGuidelines2012.pdf

United Nations Department of Economic and Social Affairs. (n.d.). *International Decade of Indigenous Languages 2022 – 2032 | Division for Inclusive Social Development (DISD)*. Retrieved May 9, 2024, from https://social.desa.un.org/issues/indigenous-peoples/international-decade-of-indigenous-languages-2022-2032

Usborne, E., Peck, J., Smith, D.-L., & Taylor, D. M. (2011). Learning through an Aboriginal language: The impact on students' English and Aboriginal language skills. *Canadian Journal of Education*, *34*(4), 200–215.

Voyer, D., & Voyer, S. D. (2014). Gender differences in scholastic achievement: A meta-analysis. *Psychological Bulletin*, *140*(4), 1174–1204. https://doi.org/10.1037/a0036620

Wilson, W. H., DeCaire, R., Gonzalez, B. N., & McCarty, T. L. (2022). Progress, challenges, and trajectories for Indigenous language content-based instruction in the United States and Canada. *Journal of Immersion and Content-Based Language Education*, *10*(2), 343–373. https://doi.org/10.1075/jicb.21023.wil

Wing, H. (1980). Practice effects with traditional mental test items. *Applied Psychological Measurement*, *4*(2), 141–155. https://doi.org/10.1177/014662168000400201