Refining State Accountability Systems for English Learner Success

Megan Hopkins
Pete Goldschmidt
Julie Sugarman
Delia Pompa
Lorena Mancilla

MPI
MIGRATION POLICY INSTITUTE

MIGRATION POLICY INSTITUTE
NATIONAL CENTER
ON IMMIGRANT
INTEGRATION POLICY

Refining State Accountability Systems for English Learner Success

Megan Hopkins
Pete Goldschmidt
Julie Sugarman
Delia Pompa
Lorena Mancilla

Migration Policy Institute

January 2024

Contents

Executive Summary

The federal *Every Student Succeeds Act* (ESSA) of 2015 requires states to develop accountability systems to hold K-12 schools accountable for the outcomes for all students, with special attention to historically underserved groups such as English Learners (ELs). Changes from previous versions of the law heightened awareness of how well ELs fare on academic assessments and learning English. Accountability systems and their various components need to support valid claims about schools and produce meaningful data that policymakers, practitioners, families, and community members can trust as they engage in informed decision-making about school improvement. ESSA affords states the flexibility to set benchmarks and consequences within the federal accountability framework, but educators and researchers have expressed concern about whether state accountability systems accurately capture and represent EL students' outcomes.

Accountability systems and their various components need to support valid claims about schools and produce meaningful data that policymakers, practitioners, families, and community members can trust.

State accountability systems center on student outcomes from English language arts (ELA) and math assessments. Unlike their non-EL peers, ELs are expected to learn and master this academic content while also learning and becoming proficient in English. Research shows that ELs' development of English language proficiency (ELP) affects their performance and growth on ELA and math content assessments. In other words, content area test scores are influenced both by how well ELs are taught that content and by their ability to demonstrate knowledge and skills on content assessments in English. But too often, the role ELP plays in ELs' scores in a single year and in progress from one year to the next is not taken into account when reporting and interpreting EL outcome data in academic subjects.

The nature and quality of the language services ELs receive at school are important factors that contribute to these students' performance and growth toward achieving ELP. These services vary widely across and within states, adding to the challenges and complexities of interpreting EL outcomes. For example, some schools use bilingual program models while others use English-only models, and schools may or may not have teachers with specialized training in how to provide English language development instruction. Researchers have long discussed the value of adding opportunity-to-learn indicators to accountability systems in order to make the impact of school system resources, contexts, and practices on student outcomes more transparent. However, such information typically is not reported systematically or connected to accountability outcomes in a meaningful way.

In 2022 and 2023, a team of researchers from the Migration Policy Institute (MPI), California State University Northridge, and University of California San Diego collaborated on a project comprised of two distinct research studies, one quantitative and one qualitative, focused on state accountability systems and ELs.

The project's aim was to rethink accountability with an eye to (1) addressing problems with the validity of summative assessment outcomes for ELs in current systems and (2) incorporating students' opportunities to learn and program quality into an accountability system whose value is being questioned in many communities.

The quantitative study used state-level ELA, math, and ELP assessment data from two states (Hawaii and Ohio) to examine potential refinements to the statistical models used in state accountability systems. Results from these analyses revealed statistical modeling approaches that can help to paint a clearer picture of EL students' status (current year performance) and growth (progress over time).

- The first approach, focused on status, establishes adjusted grade-level standards, or numerical benchmarks, in ELA and math that students must meet based on their ELP level. To generate adjusted academic proficiency cut scores, a regression analysis was conducted to determine the gap between a student's score on the ELA or math assessment and their score on these assessments as a function of their ELP and grade level. These adjusted standards are intended purely for accountability purposes to more accurately reflect how schools are contributing to specific academic outcomes. As the findings show, applying adjusted academic proficiency cut scores based on ELP levels offers more explicit insight into how ELs are performing in academic content areas and thus how schools are serving them.
- The second approach, centered on growth, examined the viability of two statistical methods for generating more accurate claims about schools' ability to facilitate ELs' academic growth. One method involves using a time-varying covariate (TVC) model, and the other method uses a multivariate longitudinal model. Results indicate that using the TVC model improved the correlations of estimated school effects between growth on content and progress toward ELP from about .65 to .90. This increase is a substantively meaningful improvement that increases the likelihood that test scores capture schools' true contribution to student learning. Likewise, in the second method using a multivariate longitudinal model, sample analyses resulted in estimates that more accurately reflect schools' ability to foster ELs' academic growth. When applying this second method, differences in these estimates from traditional growth models were moderate but significant.

The qualitative study gathered data through focus groups and interviews conducted across 18 states with state and local education agency staff, representatives of community advocacy organizations, and parents of EL students. The goal was to identify a set of indicators that (1) shed light on aspects of EL programming that are most salient for understanding the nature and quality of language instructional supports that EL students receive; (2) would be useful to diverse interest groups; and (3) could be feasibly collected and reported. This research illuminated a variety of perspectives on data collection and reporting for accountability purposes. Focus group and interview participants uniformly expressed a desire for actionable data that can be used to drive local decision-making. But, they noted, this requires developing infrastructure and systems that integrate different data sources (such as demographic, teacher certification, and outcome data) and make data accessible and digestible to local groups (including educators and families). Study participants identified four sets of EL-specific indicators they thought would be useful for better understanding the effectiveness of the schools in their communities: program models (especially

bilingual versus English-only approaches), access to qualified teachers, student participation in educational opportunities (such as gifted and talented programs or the Seal of Biliteracy), and family engagement. Additionally, study participants expressed the desire to be able to look at disaggregated data, which would require states to develop consistent definitions for EL student groups (e.g., newcomers) for whom separate data analysis is currently not possible in most states.

The findings and implications from these two studies support calls for revisions in federal policy to better support ELs. Recommendations for changes in federal legislation include:

1. requiring states to include one or more program quality indicators in their accountability system that are flexible enough to accommodate diverse contexts but provide critical information to support decision-making about resources, policies, and instruction;
2. creating pilot programs and competitive grants to support states' creation of new ways to analyze academic outcomes using more nuanced statistical models that include ELP;
3. supporting states' development of enhanced data systems to allow deeper data analysis and connection of program quality and outcome data;
4. providing guidance for states on school improvement strategies for ELs that could be used to respond to more nuanced accountability measures for ELs, as described in this report; and
5. creating common definitions of EL subgroups in order to allow disaggregation of student outcome data and comparison of these data across contexts in order to improve understanding of student needs and appropriate targeting of school improvement responses.

Getting accountability right for the United States' growing EL population will lead to improvement in accountability overall.

Accountability systems designed to identify and close achievement gaps have yet to fully account for ELs' performance and growth and to provide sufficient information to enable school leaders to make decisions about targeting resources and implementing school improvement strategies. When it comes to understanding how EL students perform in comparison to their non-EL peers, and how schools contribute to EL outcomes, too many questions remain unanswered under today's federal accountability framework and state accountability systems. Getting accountability right for the United States' growing EL population will lead to improvement in accountability overall.

1 Introduction

U.S. school reform efforts over the last three decades have centered the idea of improving K-12 education by holding schools publicly accountable for student outcomes. For more than 20 years, the federal government has framed this effort through an accountability system that requires states to administer standardized tests, publish those results and other outcomes such as graduation rates, and identify schools needing improvement. This system focuses especially on using data to illuminate how historically marginalized groups of students—including English Learners (ELs), students living in poverty, and students with disabilities—continue to be disadvantaged by systems that fail to meet their unique needs. But for this system's data to be useful for school improvement, policymakers, educators, and community members need to trust that the data accurately identify successes, challenges, and inequities in educational opportunities provided to students.

Between the 2001 and 2015 reauthorizations of the *Elementary and Secondary Education Act of 1965* (ESEA),[1] the federal framework for school accountability has evolved in terms of the requirements and flexibility afforded to states. One reason policymakers pressed for these changes was to better capture schools' contributions to students' success or lack thereof. Additionally, states have been afforded wider latitude to design systems to help identify schools needing improvement. As states have developed and implemented their ESEA accountability systems, practitioners and researchers have been tracking whether these systems accurately capture and represent the outcomes of students identified as ELs. Aiming to add to this important work, this report explores school accountability through an EL lens to contribute to the refinement of state accountability systems as well as federal accountability for ELs more broadly.

ELs are a diverse and steadily growing group of students who, as of 2020, made up 10 percent of the U.S. K-12 population.[2] Unlike their non-EL peers, ELs are expected to learn and master academic content while also becoming proficient in English. By definition, ELs are not English proficient; yet in nearly all cases, their academic performance and progress is measured using standardized summative assessments[3] in English. Given that English language proficiency (ELP) fundamentally influences students' performance on these assessments, accountability systems that do not consider this relationship cannot accurately measure what ELs know and can do, or how schools have contributed to ELs' academic development.

Given that English language proficiency fundamentally influences students' performance on these assessments, accountability systems that do not consider this relationship cannot accurately measure what ELs know and can do, or how schools have contributed to ELs' academic development.

1 The *Elementary and Secondary Education Act* was reauthorized in 2001 as the *No Child Left Behind Act* and in 2015 as the *Every Student Succeeds Act*.

2 National Center for Education Statistics, "English Learners in Public Schools," updated May 2023.

3 Summative assessments evaluate what students know as a result of instruction, compared to set benchmarks. This contrasts with formative assessments, conducted as a regular part of instruction, to understand what concepts students do not yet understand.

Of course, the type and quality of instruction ELs receive plays an important role in their language development and academic progress, yet accountability systems often do not help decisionmakers understand such important contextual factors when interpreting EL outcomes. Although federal civil rights law requires schools to facilitate ELs' development of English proficiency and to ensure they can meaningfully participate in their school's educational programs, the way schools do this varies greatly from state to state and even within states.[4] The delivery of instructional services varies based on program model implementation and local resources available, such as credentialled EL teachers. This variability adds to the complexity of accurately capturing and understanding ELs' performance and progress in academic content areas and ELP[5] on standardized tests—over and above the complexities inherent in understanding the progress of all students in their academic growth.

Federal and state accountability systems do not capture the nature, quantity, or quality of these language services, nor do they typically tease apart the fundamental relationship between ELP and academic outcomes for ELs. Without such information, decisionmakers are left to guess at answers to key questions that should inform pedagogical and systems-level policy, such as the following:

- Do gaps between ELs and their non-EL peers on academic assessments in English language arts (ELA) and math reflect ineffective instruction in those content areas or the fact that ELs are not yet proficient enough in English to demonstrate their content knowledge via an English-medium assessment?
- What, if anything, do data reported for accountability purposes tell us about how schools are contributing to outcomes reported from ELP assessments?
- What do we know about the nature and quality of the support ELs receive at school to develop their proficiency in English?

State accountability systems must be refined to help answer such questions and provide more accurate information to inform efforts to improve schools' support for ELs.

This report presents findings from two distinct research studies conducted as part of a single project (see Box 1). The goals of the project were to identify (1) methods for improving the accuracy of EL test score analyses and (2) useful and feasible indicators of language instructional supports. Both inquiries aim to inform policymaker efforts to improve current accountability systems and make bold changes in future iterations of federal education law. Sections 2 and 3 of this report present the findings, implications, and recommendations from each study separately. Those sections are followed by a conclusion that summarizes the system changes that would be required to put the studies' recommendations into practice.

4 Julie Sugarman, *A Matter of Design: English Learner Program Models in K-12 Education* (Washington, DC: Migration Policy Institute, 2018); U.S. Department of Justice and U.S. Department of Education, *Dear Colleague Letter: English Learner Students and Limited English Proficient Parents* (Washington, DC: U.S. Department of Justice and U.S. Department of Education, 2015).

5 English language proficiency (ELP) assessments measure progress in listening, speaking, reading, and writing in English. These standardized assessments are administered annually—typically in the winter or spring—to English Learners (ELs) and the results are used for accountability purposes as well as student placement in instructional services.

BOX 1
About This Project and Its Methodology

In 2022 and 2023, a team of researchers from the Migration Policy Institute (MPI), California State University Northridge, and University of California San Diego collaborated on a project comprised of two distinct research studies focused on state accountability systems and English Learners (ELs). Though the studies were conducted independently of one another, they shared a goal of addressing how state accountability systems could provide better information to policymakers about school effectiveness for ELs.

One study, conducted by Dr. Pete Goldschmidt, used quantitative methods to examine potential refinements to states' current growth models. The research investigated whether aligning progress on English language proficiency (ELP) and growth on state content assessments would contribute to a more accurate understanding of schools' contributions to EL success. This study used data from Hawaii and Ohio for school years 2016–17, 2017–18, and 2018–19. These data provide insight into student outcomes from the last three years before school closures caused by the pandemic created drastic disruptions in both instruction and assessment. They also represent the two ELP assessments used by most states, with Hawaii using the ACCESS for ELLs, given to students in the 42 states and territories in the WIDA consortium, and Ohio using the ELP assessment from the seven-state ELPA-21 consortium. The data provided scores on ELP, English language arts (ELA), and math assessments for each student from grade 3 to grade 8 and grade 11. Those data were used to examine the relationship between ELP and content assessments and to develop growth models (mathematical formulas) combining those variables. Those new models were then compared to the growth models that the two states typically used.

The qualitative study, conducted by Dr. Megan Hopkins and Dr. Julie Sugarman, sought to identify opportunity-to-learn indicators that would help policymakers, educators, community members, and other interested parties understand the quality and type of services provided to ELs. Research methods included focus groups and interviews conducted with individuals from diverse interest groups across 18 states. Participants included 17 state education agency staff, 37 representatives from community advocacy organizations, 24 EL family members, and 5 regional or local education agency leaders. The community and EL family member focus groups were conducted by staff from state-level immigrant policy and community interest organizations with whom MPI's National Center on Immigrant Integration Policy has collaborated regularly in its wider work. These organizations' familiarity with study participants helped foster candid conversations in the focus groups.

See Appendix A of this report for more details on both studies' research methodologies.

2 Refining Accountability Models to Include English Language Proficiency

The *Every Student Succeeds Act* (ESSA), the 2015 reauthorization of the ESEA, elevated accountability for ELs by including measures of EL outcomes in the calculations states use to identify which schools are in need of improvement. In other words, for the first time in federal law, ELs' performance and growth in academic subjects and their progress toward learning English mattered for determining whether a school was performing well or poorly. In previous reauthorizations of ESEA, reporting of EL language and academic

outcomes fell under Title III, while accountability for school performance more generally fell under Title I. Because Title I accountability came with higher stakes and Title III accountability only applied to districts who chose to receive federal EL funds under that part of the legislation, the result was a lack of oversight for ELs' progress in school. To address this issue, ESSA moved policies related to EL outcomes from Title III to Title I.

At first, these changes in ESSA were received with optimism by groups interested in and advocating for EL education. However, a 2020 MPI analysis of approved state plans found that, despite ESSA's good intentions vis-à-vis ELs, policies pertaining to EL education and accountability remained scattered and disjointed both within and across states, and detailed information about those policies was still largely inaccessible to local groups, such as local education officials, educators, families, and community advocates.[6] Additionally, while ESSA and its predecessors emphasized that states should assess students in the "language and form most likely to yield accurate data,"[7] the use of native language assessments—intended to reduce the confounding effect of testing students in a language they do not yet know well—is not universal, even in states where such tests are available.[8]

BOX 2
Selected ESSA Requirements for ELs

Among the many requirements laid out in the *Every Student Succeeds Act* (ESSA), the following are some of the most notable in terms of states' obligations relating to ELs and accountability:

- States must establish and implement standardized statewide entrance and exit procedures for ELs (that is, procedures for identifying students as ELs and for later exiting them from EL status when they have reached a certain level of English proficiency); this must include an assurance to the federal government that all students who may be ELs are assessed within 30 days of being enrolled in a school.
- States must adopt an annual ELP assessment that will be taken by all ELs in schools served by the state education agency.
- States must include an indicator in their statewide accountability system that measures ELs' annual progress toward achieving proficiency in English using the statewide annual ELP assessment.
- States must define what it means to reach proficiency in English, as measured by the statewide annual ELP assessment and within a state-determined timeline.
- States must establish ambitious long-term goals and measurements of interim progress for increasing the percentage of ELs making progress toward achieving proficiency in English.

Source: *Elementary and Secondary Education Act of 1965*, as amended through the *Every Student Succeeds Act*, Public Law 114–95, 114th Cong., 2d sess. (December 10, 2015): 1802–2192, 1830, 1835–36.

6 Leslie Villegas and Delia Pompa, *The Patchy Landscape of State English Learner Policies under ESSA* (Washington, DC: Migration Policy Institute, 2020).
7 *Every Student Succeeds Act of 2015*, Public Law No. 114–95, *U.S. Statutes at Large* 129 (December 10, 2015): 1826.
8 Julie Sugarman and Leslie Villegas, *Native Language Assessments for K-12 English Learners: Policy Considerations and State Practices* (Washington, DC: Migration Policy Institute, 2020).

A. The Relationship between ELP and Academic Performance

State accountability systems under ESSA include several indicators to assess school performance and guide appropriate action for schools that need improvement.[9] Two such indicators are status and growth. With respect to status, states are required to report annual, statewide test outcomes in ELA[10] as the share of students who score proficient on those tests. In terms of growth (also called progress over time),[11] these indicators, when used by states, shed light on students' academic gains, giving credit for progress even if students have not yet reached proficient status. States are also required to incorporate a measure of growth in English proficiency in their accountability systems based on annual ELP assessment results for students identified as ELs. Given that proficiency scores on academic content assessments[12] can be shaped by students' prior learning and characteristics, some analysts suggest that growth indicators—because they reflect what students learn each year—afford better information than status indicators on schools' contributions to student learning and development.[13]

In writing plans to implement ESSA, states chose statistical models to calculate growth depending on what questions about students' learning progress they wanted to answer.[14] For example, they may have been interested in whether some students progress faster than others, or they may have preferred to simply describe how much growth occurs overall from year to year. Some growth models allow states to predict how much growth a student should make based on their past performance, with states giving more credit to schools for students who meet or exceed those expectations than those whose students grow less than expected. Other models simply describe how much a student's score goes up or down from one year to the next.[15]

Although ESSA's inclusion of these kinds of growth models allows for more accurate assessments of school effectiveness than prior methods that compared cohorts of students (for example, comparing 3rd graders in 2021 to 3rd graders in 2022),[16] there is still room for improvement. For example, there is a need to ensure that the statistical models used produce results that reflect the effects of instruction on growth rather than extraneous variables and provide information on academic progress that is not confounded by students' progress in ELP.

---

9 Susan Lyons, Juan D'Brot, and Erika Landl, *State Systems of Identification and Support under ESSA: A Focus on Designing and Revising Systems of School Identification* (Washington, DC: Council of Chief State School Officers, 2017).

10 Most states also include scores from science assessments in their accountability systems, and some include social studies.

11 The terms growth and progress are commonly used interchangeably to refer to the amount of gains a student makes in academic subject areas over time (that is, from year to year).

12 In addition to its use in the domain of language learning, the term proficiency is also used in this report to refer to student performance with regard to meeting grade-level standards or expectations on standardized tests of academic content such as English language arts (ELA) and math.

13 Kilchan Choi, Pete Goldschmidt, and Kyo Yamashiro, "Exploring Models of School Performance: From Theory to Practice," *Yearbook of the National Society for the Study of Education* 104, no. 2 (2005): 119–46; Pete Goldschmidt et al., *Policymakers' Guide to Growth Models for School Accountability: How do Accountability Models Differ?* (Washington DC: Council of Chief State School Officers, 2005).

14 For a high-level explanation of growth data and the types of growth models used in state accountability systems, see Data Quality Campaign, "Growth Data: It Matters, and It's Complicated" (issue brief, January 2019).

15 Juan D'Brot, *Considerations for Including Growth in ESSA State Accountability Systems* (Washington, DC: Council of Chief State School Officers, 2017).

16 Pete Goldschmidt and Kilchan Choi, *The Practical Benefits of Growth Models for Accountability and the Limitations under NCLB* (Los Angeles: UCLA National Center for Research on Evaluation, Standards, and Student Testing, 2007).

While there is still a need to assess how well growth models work to demonstrate school effectiveness in general, there are additional outstanding questions regarding how to model growth for ELs specifically. Modeling growth for ELs is particularly challenging because language development is not a linear process. ELs typically make faster growth in their first year or two of English language development, with growth slowing in later years as learners are immersed in more complex academic discourse and refine elements of grammar and vocabulary.[17] Nonetheless, many state accountability systems set linear progress goals for ELs (for example, that students will improve one proficiency level each year).

Modeling growth for ELs is particularly challenging because language development is not a linear process.

Notably, both school- and student-level factors contribute to ELP, such as the program model schools use to support ELs' language development and students' proficiency in their primary language(s). Therefore, the amount of time it takes to become proficient in English and how much progress is made from one year to the next can vary widely across schools and students.[18] This variation matters not just for the inclusion of ELP growth in accountability models, but it also reverberates in ELA and math scores because students' performance on academic assessments is highly correlated with their ELP levels.[19]

In short, a student's ELP influences their performance on academic content assessments and, importantly, measurements of their academic progress (see Box 3 for more detail).[20] Thus, ELs and non-ELs with the same prior performance score in an academic content area may be on different growth trajectories. A refined accountability model that can illustrate these differences, without reducing performance expectations for ELs, would better capture schools' contributions to student academic performance and growth and suggest domains in which instructional improvements are needed.

17 H. Gary Cook, Carsten Wilmes, Tim Boals, and Martín Santos, "Issues in the Development of Annual Measurable Achievement Objectives for WIDA Consortium States" (WCER Working Paper No. 2008–2, Wisconsin Center for Education Research, Madison, WI, April 2008); Pete Goldschmidt and Kenji Hakuta, *Incorporating English Learner Progress into State Accountability Systems* (Washington, DC: Council of Chief State School Officers, 2017).

18 National Academies of Sciences, Engineering, and Medicine, *Promoting the Educational Success of Children and Youth Learning English: Promising Futures* (Washington, DC: The National Academies Press, 2017).

19 Megan Hopkins et al., "Fully Accounting for English Learner Performance: A Key Issue in ESEA Reauthorization," *Educational Researcher* 42, no. 2 (2013): 101–8.

20 Alison L. Bailey and Patricia E. Carroll, "Assessment of English Language Learners in the Era of New Academic Content Standards," *Review of Research in Education* 39, no. 1 (2015): 253–94; Alison L. Bailey and Becky H. Huang, "Do Current English Language Development/Proficiency Standards Reflect the English Needed for Success in School?" *Language Testing* 28, no. 3 (2011): 343–65; Mikyung Kim Wolf and Seth Leon, "An Investigation of the Language Demands in Content Assessments for English Language Learners," *Educational Assessment* 14, nos. 3–4 (2009): 139–59; Frances A. Butler and Robin Stevens, "Standardized Assessment of the Content Knowledge of English Language Learners K-12: Current Trends and Old Dilemmas," *Language Testing* 18, no. 4 (2001): 409–27.

BOX 3
Illustrating the Relationship between ELP and Academic Performance with State Data

Building on prior work that used district-level data to examine the relationship between ELP and academic performance, the current project's research team explored this relationship using state-level data from Hawaii and Ohio, from school years 2016–17 through 2018–19. EL students demonstrate considerably more limited ELA and math skills and knowledge than non-ELs, and this gap ranges between 0.6 and 1.2 standard deviations. Although the two states use different assessments for ELP, ELA, and math, the gaps and the pattern of the gaps are generally consistent across these content areas and states. These gaps, measured in standard deviations, could also be represented by the percentage of ELs and non-ELs who are proficient in ELA and math, which would demonstrate a similar pattern.

Further analysis indicated that ELP plays a role in both status and growth. In terms of status, between 28 and 53 percent of the variability in ELA and math performance could be accounted for by students' ELP levels. In other words, between one-quarter and half of ELs' performance on these assessments could be explained by their levels of English proficiency. Thus, attributing 100 percent of academic success to a school's ability to teach content places all the accountability on content instruction, when, in fact, one-quarter to half of the success should be attributed to the EL program. Regarding growth, analyses of data from both states showed that ELs at beginning ELP levels make less academic progress compared to non-ELs, but that ELs at more advanced ELP levels make faster year-to-year academic progress than ELs at beginning ELP levels and, in some cases, their non-EL peers. These findings illustrate how accountability systems that do not account for the role ELP plays in ELs' academic performance make it challenging to accurately identify school successes or areas for improvement.

Sources: For further reading, see Gary Cook, Robert Linquanti, Marjorie Chinen, and Hyekyung Jung Cook, *National Evaluation of Title III Implementation Supplemental Report—Exploring Approaches to Setting English Language Proficiency Performance Criteria and Monitoring English Learner Progress* (Washington, DC: U.S. Department of Education, Office of Planning, Evaluation, and Policy Development, 2012); Megan Hopkins et al., "Fully Accounting for English Learner Performance: A Key Issue in ESEA Reauthorization," *Educational Researcher* 42, no. 2 (2013): 101–8.

B. Accountability Models for Addressing Confounded Results for ELs

Recognizing that ELs' academic performance is confounded by their ELP, the results from the quantitative study revealed statistical modeling approaches that can help to paint a clearer, untangled picture of status and growth. The following sections use examples to present and explain these approaches, first for status and then for growth.

Understanding and Untangling Confounding in Status

To more accurately measure ELs' current year performance on ELA and math assessments, this project's researchers analyzed state data to develop a method that better isolates claims regarding ELs' ELA and math status. This approach establishes adjusted grade-level standards[21] in academic subject areas that students must meet based on their ELP level. To generate adjusted academic proficiency cut scores, a regression analysis was conducted to determine the gap between a student's score on the ELA or math assessment and their score on these assessments as a function of their ELP and grade level. These adjusted standards

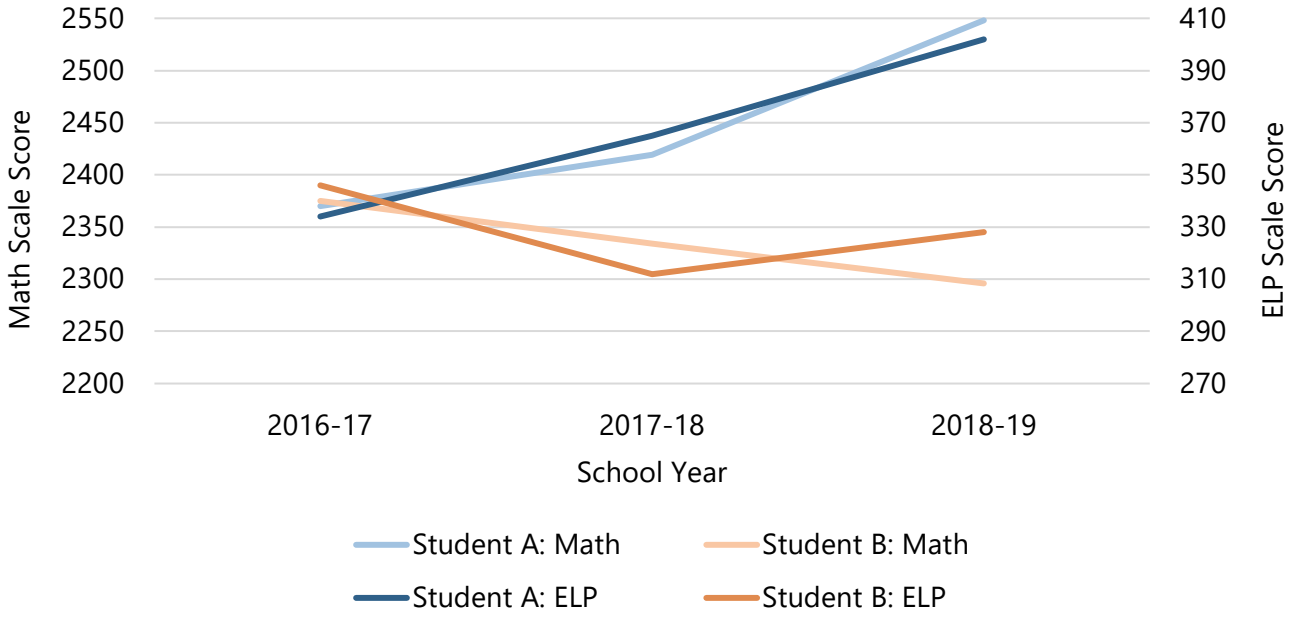
21 The term standard is used in several ways in education research. In this report, the term is used to mean a numerical benchmark or cut score; students scoring at or above that level are thus said to be proficient in the academic content.

are intended purely for accountability purposes to more accurately reflect how schools are contributing to specific academic outcomes. For technical details and additional results, see Appendix B.

Using ELP and math assessment data from two ELs in Hawaii for illustration purposes, it is possible to observe the influence of students' ELP on academic performance. As shown in Figure 1, when Student A and Student B had similar ELP assessment scores in 2016–17, they demonstrated similar performance in math. Then, in 2017–18, Student A scored 2419 while Student B scored 2334 in math—a difference of 85 points. Although there could be many reasons why Student A scored higher than Student B, one reason is that Student A was performing at a more advanced ELP level than Student B that year. In fact, Student A's ELP assessment score was 365 while Student B's score was 312, which is a difference of 53 points.

FIGURE 1

Math and ELP Assessment Scores for Student A and Student B, School Years 2016–17 to 2018–19



Source: Author analysis of Hawaii assessment data from school years 2016–17 through 2018–19, received from the Hawaii State Department of Education in September 2023.

When examining this difference in terms of effect size,[22] the gap between the two students is larger in ELP (1.5) than in math (0.93), meaning that Student B was more limited than Student A in their ability to demonstrate what they know and can do in math due to their ELP level. Examining the relationship between ELP and math performance statewide indicated that ELs were 5.2 times as likely to make gains in math from year to year if they made gains toward ELP.[23] Thus, applying the same math proficiency expectation to all ELs regardless of ELP level does not result in a complete picture of their academic performance.

Using ELP-related estimates to create adjusted cut scores for academic performance in math, results from Hawaii showed that proficiency cut scores in math increased with grade and ELP level (see Figure 2). In other

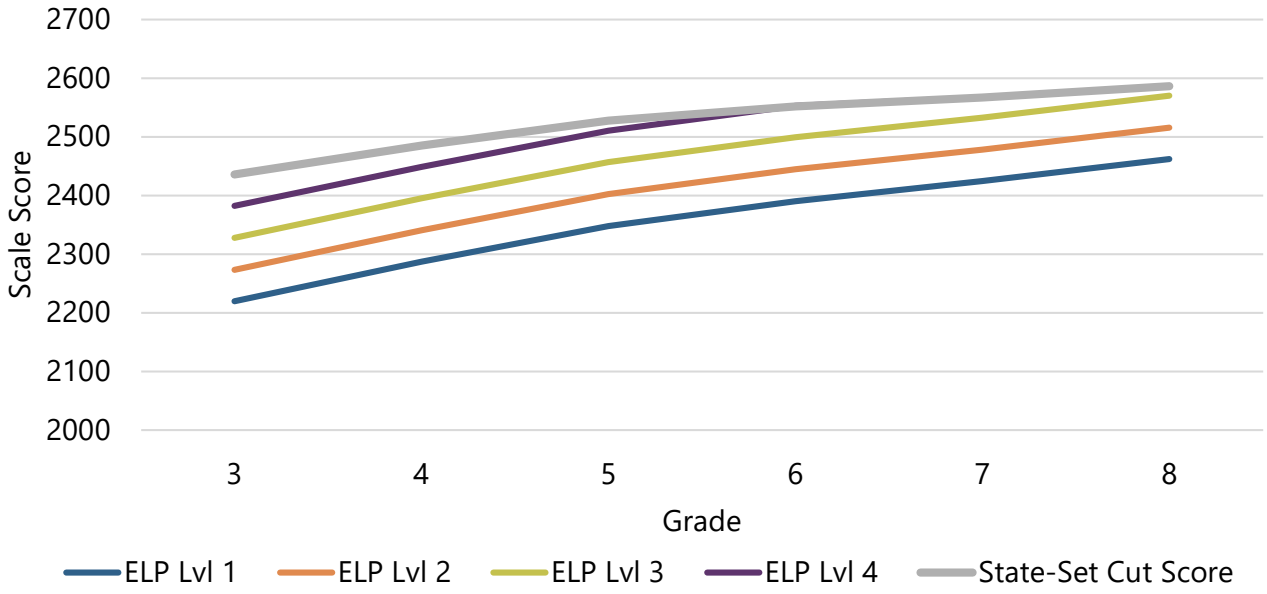
22 Effect size is a statistical way to look at the meaningfulness of a difference in scores. The larger the number, the more meaningful the difference.

23 Similarly, ELs across the state were 4.2 times as likely to make gains in ELA if they made gains toward ELP.

words, as EL students make gains in their language proficiency and achieve higher ELP levels, their expected math performance changes accordingly.[24] For example, a student at ELP Level 1 in 3rd grade would be expected to meet the academic proficiency cut score on the blue target line in Figure 2, whereas the student would be expected to advance one ELP level by 4th grade and thus meet the cut score along the orange target line and so on. In this way, as ELs progress toward proficiency in English, academic content expectations would increase accordingly until they exit EL status, at which point expectations for them would be consistent with those for their non-EL peers.[25]

FIGURE 2

Adjusted Proficiency for Accountability: Math Scale Scores by ELP Level, Hawaii



Source: Author analysis of Hawaii assessment data from school years 2016–17 through 2018–19, received from the Hawaii State Department of Education in September 2023.

Returning to Student A and Student B from Figure 1 to illustrate these trends, Student A's math score in 2016–17 does not meet the originally established math standard, represented by the gray line in Figure 2. However, after applying the adjusted academic proficiency cut scores based on ELP level, Student A's score meets the adjusted standard. In 2017–18, Student A's score does not meet either the original math standard or the adjusted math standard. By 2018–19, Student A was reclassified as English proficient and exited EL status, and their score met both the original and the adjusted math standard. Student B, on the other hand, despite having a larger adjustment in cut scores due to their lower ELP level in most years, did not meet the original or adjusted math standards in any of the three years. As this example reveals, applying adjusted academic proficiency cut scores based on ELP levels offers more nuanced insight into how ELs are performing in academic content areas and thus how schools are serving them.

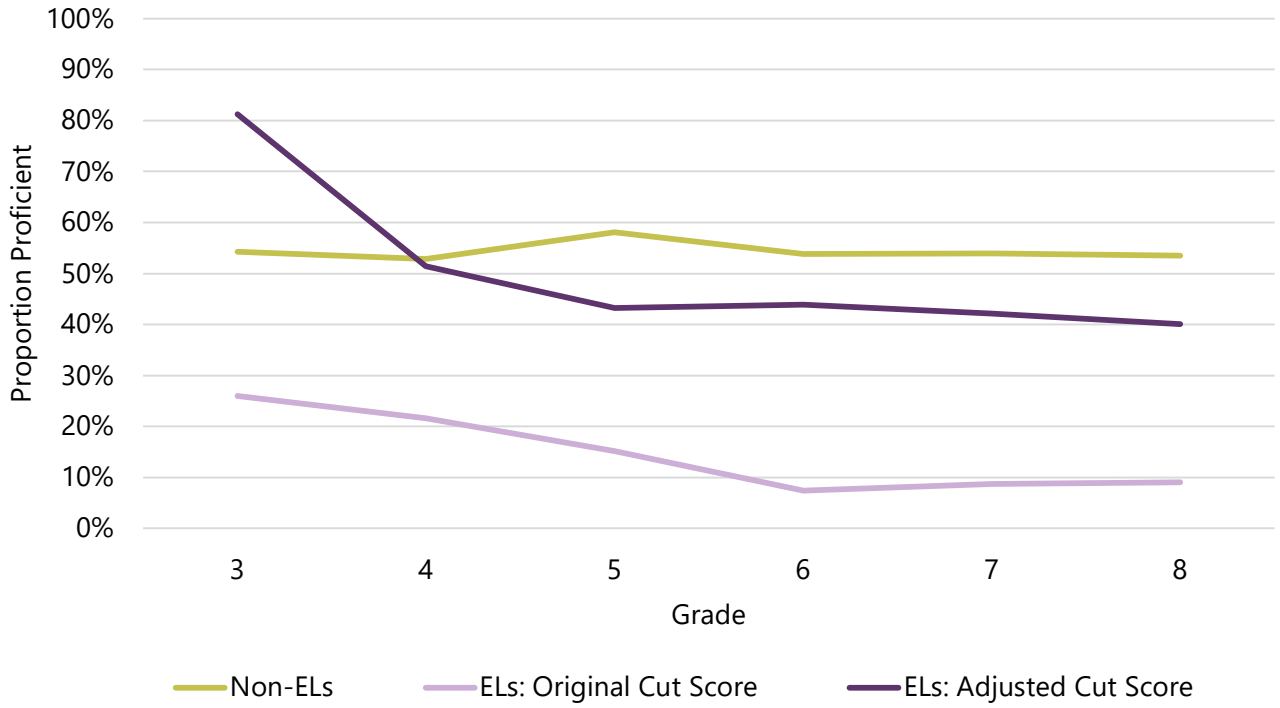
24 It is interesting to note that in math, proficiency cut scores catch up to the established Smarter Balanced Assessment Consortium cut scores in middle school for ELs who have not yet reached ELP. This is consistent with expectations as math is less dependent on a student's language proficiency than ELA. Importantly, this does not imply that demonstrating math skills and knowledge is independent of language.

25 Applying the adjusted proficiency for accountability cut score results in different interpretations as to EL performance in math and ELA. Results for ELA can be found in Appendix B.

As another representation of how applying adjusted standards provides more accurate information about ELs' academic status—this time statewide, rather than for individual students—Figure 3 compares ELA proficiency results for students across Hawaii. The over-adjustment in grade 3 notwithstanding,[26] applying adjusted cut scores provides an opportunity to examine how well schools are facilitating ELA learning for ELs. Using the originally established standards, it would be misleading to say that EL students were behind in ELA without considering their ELP level. In Figure 3, the percent of ELs meeting the original 6th grade ELA cut score was less than 10 percent; however, if the cut score is adjusted to reflect students' ELP level, about 45 percent of ELs meet their ELA target in 6th grade. These results help to refine the accountability conversation to one centered on why, when taking ELP into account, only about 45 percent of ELs meet the ELA standards in 6th grade while about 55 percent of their non-EL peers do so.

FIGURE 3

Proportion of EL and Non-EL Students Reaching Proficient Status in ELA, Hawaii



Source: Author analysis of Hawaii assessment data from school years 2016–17 through 2018–19, received from the Hawaii State Department of Education in September 2023.

Understanding and Untangling Confounding in Growth

Like status, when growth in ELA and math are calculated without considering the impact of students' progress toward ELP, this results in biased estimates of growth in ELA and math. Therefore, when holding schools accountable for growth in academic content areas such as ELA and math, it is important for decisionmakers to know whether that growth is due to students' development of content knowledge and skills or to their increasing ELP. For instance, growth in math scores that is due to increased ELP levels should result in different responses than growth in math that is due to improved math skills and knowledge. It is

26 In Figure 3, the adjusted cut score for ELs is higher than that for non-EL students in grade 3. Further analysis is needed to understand why there appears to be such a large overadjustment in the data used for this example.

important to note that it is unlikely that growth results can be perfectly untangled; nonetheless, this study aims to identify approaches that provide information that more closely aligns with accountability-related claims being made about schools (e.g., School X is doing a good job teaching math, or School Y needs support to improve its ELA instruction).

This section presents the results of quantitative analyses using simulated data to examine the viability of two statistical methods for generating more accurate claims (i.e., less biased estimates) about schools' ability to facilitate ELs' academic growth. One approach is to use a time-varying covariate (TVC) model, and the other approach is a multivariate longitudinal model. While states use various growth models as part of their ESSA accountability systems, the models described here are not currently used by any state. Technical details related to each model can be found in Appendix C.

Using a Time-Varying Covariate in Growth Models

Traditional growth models used by states estimate how many scale score points a student improves on academic assessments from one test occasion to another (typically, annually by grade) without considering the student's ELP level. Looking back to Student A and Student B's performance in Figure 1, it is possible to calculate the average of the change in scale scores for each student from 2016–17 to 2017–18 and from 2017–18 to 2018–19, otherwise known as the slope, and infer that this average score represents growth in math. Yet, as noted above, this growth in math is influenced by students' progress toward ELP and thus overstates actual academic progress in math.

One way to capture the influence of ELP is to include annual ELP assessment results (that is, ELP level) as a TVC in academic growth model calculations. A TVC is a variable that varies with time. For example, in general, a student's race or ethnicity is not time-varying because it remains the same across years. However, an EL student's ELP level, as determined using annual ELP assessments administered each winter or spring, varies over time (for example, it might increase from 1.0 to 2.0 as the student moves up in language proficiency).

Adding ELP results as a TVC means that simply calculating the slope, or the year-over-year change in scale scores, is no longer possible; however, the resulting estimates from calculations that include a TVC have the benefit of representing growth in math that is net of progress toward ELP. Growth models that include ELP level as a TVC also have the benefit of addressing potential bias in estimates of growth, and variability in growth, that result from the omission of variables.

An important consideration with using this approach is what to do about missing data, given that non-ELs do not have ELP assessment results. One potential strategy is to estimate separate growth models for ELs and non-ELs; however, this strategy has several limitations for accountability. One, it is likely that the number of ELs is quite small in many schools, which would result in very imprecise estimates for these schools; two, federal accountability does not allow for results to be conditioned on a student characteristic; and three, policymakers generally prefer for transparency and simplicity's sake to create a single estimate of growth to use in accountability calculations and determinations. Additionally, although there is significant literature on how to treat missing data, either through design (such as random assignment) or imputation (that is, using alternative values in place of missing data), these approaches are not viable in this case

because random assignment is not an option and imputing scores makes little practical sense for non-ELs. The present analysis revealed that the dummy variable adjustment (DVA) is a viable strategy[27] to generate less biased estimates of school effects. In each growth model calculated, the DVA equals 1 if the ELP scale score is missing and 0 otherwise.

The final model to come out of this analysis thus uses ELP level as a TVC to better isolate schools' contribution to ELs' academic growth; it also includes a DVA to address missing data as well as the appropriate random student and school effects. Results indicate that this model is significantly less biased than models that exclude the TVC. That is, using the TVC improves the correlations of estimated school effects between growth on content and progress toward ELP from about .65 to .90. This increase is a substantively meaningful improvement, given that a correlation of .90 (rather than .65) would allow researchers and decisionmakers to be far more confident that they are capturing schools' true contribution on students' math learning over time.

Multivariate Longitudinal Model for Academic Content Growth

The second proposed approach to generating less biased estimates about schools' contribution to ELs' academic growth is to simultaneously model growth on content and growth on the ELP assessment. This approach takes advantage of two facts: first, that ELs are progressing toward ELP and growing in academic content knowledge and skills, and second, that ELP assessments are generally administered in the middle of the academic year (winter) while content assessments are administered near the end of the academic year (spring). In addition, this approach models math and ELP for each student simultaneously (rather than as separate lines, as in Figure 1). To accomplish this, each individual assessment result is ordered in the dataset by time (e.g., winter 2018 ELP, spring 2018 math, winter 2019 ELP, spring 2019 math) instead of the traditional way of listing one test at a time with one score for each academic year. These data are then analyzed using the appropriate growth model.

Like the TVC model discussed in the previous section, the statistical model used in this approach estimates growth on academic content with respect to progress toward ELP. Results from simulated data analysis demonstrate that a multivariate longitudinal model also results in estimates that more accurately reflect schools' ability to foster ELs' academic growth. Applying the multivariate, or stacked, model to state-level data, resulted in observed differences in school estimates from traditional growth models.[28] The impact of these changes was moderate but significant.

Consistent with results from the simulated data, results from state data analyses indicated that choice of statistical model can substantially influence the claims made about school effects. Overall, the multivariate or stacked model produced lower estimates of ELs' academic growth than traditional models. It can thus be inferred that the traditional estimates produced using state data are biased upwards, which suggests that ELs are not progressing academically as much as states' current methods imply. Importantly, this analysis found that gain score models[29] were least correlated with models examined that use multiple years of data and should be used with caution.

27 Paul Allen, "Is Dummy Variable Adjustment Ever Good for Missing Data?" Statistical Horizons, September 27, 2022.
28 These include the value-added model and student growth percentile model; see Appendix C for more details.
29 There is considerable literature on using gain scores in school accountability, though this is beyond the scope of these analyses to address.

C. *Implications*

The statistical analyses performed for the quantitative study make it clear that refined accountability models that explicitly account for ELP provide more valid and reliable claims about ELs' academic performance and growth. These results point to the following recommendations:

1. States should consider using a model that bases academic proficiency expectations on ELP levels. For example, for an EL who scores at ELP Level 1, what might be a reasonable expectation for their academic performance in ELA? Clearly it cannot be proficient (i.e., meets or exceeds expectations) in ELA as the student is at the beginning stages of developing English proficiency (Level 1). Thus, states should use a model designed to generate adjusted academic proficiency cut scores for accountability purposes.[30] As the examples using data from Hawaii illustrate, adjusted academic proficiency standards based on ELP can help states more accurately assess how well schools are supporting ELs' academic performance. Ensuring such accuracy is important so that policymakers, educators, families, and community members do not assume language is the sole cause of limited or low academic performance.
2. States should consider refining their accountability models to include an ELP indicator that more explicitly accounts for the role of ELP in academic growth. Without considering ELP progress, school effect estimates co-mingle students' current language skills, progress toward ELP, academic performance status, and academic growth. This obfuscation results in claims about schools that lessen the utility of accountability results as a tool to support school improvement because schools' specific strengths and challenges are not accurately identified.

BOX 4
Example of State-Level Implications

The implications of using refined accountability models can be quite varied and depend on state-specific accountability rules. One requirement under ESSA is for states to identify schools where the performance of any student group (such as ELs) is as poor as the bottom 5 percent of the school in general, as measured by one or more indicators (such as percent meeting grade-level standards). Using the Ohio data analyzed for this report, it is possible to simulate the impact on accountability outcomes for 2018–19 using an approach that includes ELP in the growth model. In one estimate, the authors found that if the state were to use original grade-level standards (i.e., academic proficiency cut scores), 745 schools would be identified as in need of improvement based on their EL students' math performance. However, if adjusted cut scores were used, only 345 schools would be identified. This difference means that EL students' math performance in as many as 400 schools is confounded by their ELP. If the adjusted cut scores were used, then schools would be more accurately held accountable for their students' performance in math.

Source: Author analysis of Ohio assessment data from school years 2016–17 through 2018–19, received from the Ohio Department of Education and Workforce in February 2020.

30 States could use the proposed model, explained in depth in Appendix C, to create their own adjustment for ELs.

3 Using Opportunity-to-Learn Indicators to Contextualize Outcomes

With an accountability system that yields more accurate information on ELs' status and growth, policymakers and educators can turn to the work of improving instruction and educational systems. But to do this, they need information on what is happening in schools and classrooms so that they can identify what practices must change or what resources need to be deployed. One way to provide this information is to expand data collection to include opportunity-to-learn indicators that shed light on students' learning experiences and environments.

A. Beyond Test Scores for School Accountability

Opportunity-to-learn indicators span a wide variety of topics. A recent National Academy of Sciences report on measures related to educational equity proposed 16 indicators that measure, for example, rates of enrollment in preschool and advanced coursework, school climate, and disparities in student discipline.[31] Likewise, research on high-quality instruction has identified practices in instruction, student grouping, teacher qualifications and beliefs, and quality of learning materials, among many other things, that are connected to improved outcomes and may be quantified as educational indicators. These general categories of school quality matter for all students, but it is also possible to identify policies and practices within these categories that must be in place to effectively support ELs.[32] Within school climate, for example, schools serving ELs should demonstrate that ELs and their families feel welcome and that their languages and cultures are valued as opposed to seen as an obstacle to learning. Even more specific rubrics exist for specific types of programs for ELs, such as the *Guiding Principles for Dual Language Education.* For dual language programs that aim to develop bilingualism and biliteracy in English and a partner (non-English) language, rubrics such as the *Guiding Principles* probe specific practices of this educational approach such as how program staff demonstrate the social value of minoritized (non-English) languages and cultures.[33]

There is a long history of school effectiveness research that aims to connect policies and practices at the classroom, school, district, and state level with student outcomes. In addition to expanding scientific knowledge about learning, such research aims to influence decisionmakers in areas such as instructional planning and teacher preparation. School administrators making decisions about resource allocation use similar types of inquiries to connect the dots between, for example, student demographics, instructional practices, and academic achievement in their local context. In the realm of accountability, researchers have long discussed adding opportunity-to-learn indicators to make school system resources, contexts, and

31 National Academies of Sciences, Engineering, and Medicine, *Monitoring Educational Equity*, eds. Christopher Edley Jr., Judith Koenig, Natalie Neilsen, and Constance Citro (Washington, DC: The National Academies Press, 2019).

32 Zenaida Aguirre-Muñoz and Anastasia A. Amabisca, "Defining Opportunity to Learn for English Language Learners: Linguistic and Cultural Dimensions of ELLs' Instructional Contexts," *Journal of Education for Students Placed at Risk (JESPAR)* 15, no. 3 (2010): 259–78; National Academies of Sciences, Engineering, and Medicine, *Promoting the Educational Success*.

33 Elizabeth Howard et al., *Guiding Principles for Dual Language Education,* 3rd ed. (Washington, DC: Center for Applied Linguistics, 2018).

practices more transparent.[34] However, such information typically is not reported systematically, used for high-stakes purposes, or connected to student outcomes.

ESSA requires states to include a measure of school quality or student success—explicitly not tied to assessment results—in their accountability systems. Most states chose to use attendance as the measure used to identify schools for improvement, but many also report indicators such as access to advanced coursework or student discipline. Although these are typically reported for the subgroups of students required by law, including ELs, these indicators are rarely, if ever, EL-specific.[35] Given long-standing inequities in EL education, assessing quality of EL academic and linguistic supports is critical for ensuring that EL students are afforded equitable access and meaningful learning opportunities.[36]

B. *Indicators of Language Instruction Support*

While there are many indicators of instructional quality that apply to all students and are therefore relevant to effective instruction for ELs, the focus for this study was on indicators that illuminate the nature and quality of language instructional supports that EL students receive. As part of school accountability, these indicators could play two roles: one, to contextualize EL student outcomes in order to understand patterns of progress or lack thereof, and two, to gauge whether schools are meeting their civil rights obligations to provide support for EL students that facilitates their ELP and academic achievement.[37]

These indicators could play two roles: one, to contextualize EL student outcomes ... and two, to gauge whether schools are meeting their civil rights obligations to provide support for EL students.

One of the most consequential decisions in designing EL instruction is the program model. This includes whether students will be taught entirely or primarily in English, in students' native languages for a short time before transitioning to English, or in a dual language model that promotes full bilingualism and biliteracy in English and another language. For English-only programs, which are the most common approach in the United States, an important indicator is the delivery of English language development (ELD) instruction, with options including support within the general education classroom or pull-out services that serve EL students in small groups outside of their home classroom.[38] There is evidence that program type is related to EL student outcomes,[39] and strong opinions—but little empirical research—on which ELD approaches are most effective. Other indicators that might

34 Andrew Porter, "Opportunity to Learn" (brief no. 7, Center on Organization and Restructuring of Schools, Madison, WI, 1993).
35 Villegas and Pompa, *The Patchy Landscape*.
36 National Academies of Sciences, Engineering, and Medicine, *Promoting the Educational Success*.
37 Under Title VI of the *Civil Rights Act of 1964* and the *Equal Educational Opportunities Act of 1974*, schools have an obligation to ensure ELs have access to the same rigorous curriculum as all other students. The U.S. Supreme Court ruling *Castaneda v. Pickard* set the standard that instruction for ELs must be based on sound theory, supported with appropriate resources, and shown to be effective. See Julie Sugarman, *Legal Protections for K-12 English Learner and Immigrant-Background Students* (Washington, DC: Migration Policy Institute, 2019).
38 Sugarman, *A Matter of Design.*
39 Jennifer L. Steele et al., "Effects of Dual-Language Immersion Programs on Student Achievement: Evidence from Lottery Data," *American Educational Research Journal* 54, no. 1_suppl (2017): 282S–306S; Ilana M. Umansky and Sean F. Reardon, "Reclassification Patterns among Latino English Learner Students in Bilingual, Dual Immersion, and English Immersion Classrooms," *American Educational Research Journal* 51, no. 5 (2014): 879–912.

be salient for understanding EL program quality include how much time per day or week students receive targeted ELD and whether EL students receive instruction from teachers with specialized credentials, such as English as a second language (ESL) or bilingual certifications. Of course, there are also related indicators that are particularly important for judging the quality of EL instruction but may be collected for all students, including whether EL students have access to advanced academic coursework and the appropriateness of curricular materials.

Given the range of indicators that might be collected to assess programmatic inputs for EL students, the goal for this study was to identify a set of indicators that would be useful to a wide range of interested parties, including leaders from state education agencies (SEAs) and local education agencies (LEAs), community advocates, and ELs' family members, and that could be feasibly collected and reported. Focus groups and interviews with individuals from these groups sought diverse perspectives on what information study participants had about EL programs and instruction and what additional information would be useful and, for system insiders, feasible to collect.

C. Perspectives on Indicators of Program Quality

Although focus group discussions reflected diverse perspectives on the purposes for data—unsurprising, given the participants' diverse priorities—there was considerable agreement on which indicators would be useful for evaluating program quality. Importantly, both system insiders and outsiders acknowledged that there is information that is known to schools and administrators that is not made available publicly, and it was a common theme that sharing information is generally desirable.

Purposes for Data Collection

When reflecting on the purposes for collecting information about EL program quality, focus group participants were less interested in the issue of collecting data on EL program quality for traditional accountability purposes (that is, to identify underperforming schools using state-defined metrics) than in having access to a variety of data that would help guide their decision-making. For SEA staff, having reliable data related to the kinds of indicators discussed in the next subsection would assist them in providing timely technical assistance and guidance to the districts and schools that need it most, which would allow state agencies to more effectively use their limited resources. Additionally, having data related to EL program quality would help SEA staff contextualize the student outcome data that is used for state accountability purposes. This additional context would allow staff to understand which kinds of programs are working well, where, and for whom, and even to point out ways that outcome data may not be accurately capturing program quality as compared to more holistic ways of measuring quality or given a particular context. This study's state-level participants referred to this capacity-building approach as "little a accountability," an approach they contrasted with "big A accountability" processes that often carry more negative connotations by focusing on identifying struggling systems versus supporting all schools in differentiated ways. Moreover, SEA staff suggested that requiring schools to simply report on certain

This additional context would allow staff to understand which kinds of programs are working well, where, and for whom.

indicators (without connection to "big A accountability" processes) would elevate their importance. Even without the high stakes of accountability, having the state or federal government require such reporting could increase attention to EL program quality in ways that motivate positive change.

Community advocates and EL families expressed similar sentiments, articulating a desire to have access to data to understand what is happening inside schools. Rather than desiring information to identify "good schools" and "bad schools," they wanted data to help them select schools and programs based on their specific needs. Both advocates and families overwhelmingly felt that they do not get sufficient information from their schools, whether about student outcomes or school policies and practices (of more interest to advocates) or about curricular and extracurricular programs and activities (of more interest to parents). Both sets of community respondents were puzzled and concerned by the enormous variation in curricular and extracurricular offerings between and even within schools (some aspects of which they knew about prior to the focus groups, and others they heard about from other study participants), and many wanted to learn more about why and how schools differed in the programs and services they offered.

Indicators of Program Quality

There was substantial convergence across study participant groups with respect to the data and information they would like to access related to the programs and services available to EL students. Analysis of information gathered in these sessions revealed four key indicators, each of which includes several measures: instructional program, access to qualified teachers, access to opportunities, and family engagement.

Instructional Program

Many study participants said that it would be helpful to know the **program model** each EL student is enrolled in. These programs are either bilingual, including transitional bilingual or dual language education,[40] or are monolingual with academic content and language instruction in English. As part of any program model, schools are required to provide ELD instruction to support EL students' language proficiency. Thus, another facet of the instructional program about which study participants desired information is the **ELD service delivery method** for each EL student, including the mode of delivery (e.g., pull-out classes, push-in support, co-teaching). There were some mixed opinions about the **amount of ELD instructional time**, in minutes, being used as an indicator. Some participants noted that quality mattered more than quantity, but participants also thought such a measure could help identify major problems, such as students receiving no deliberate ELD instruction at all.

Although having more detailed information about how language and content instruction is provided would reveal whether and how schools are meeting their civil rights obligations, SEA and LEA staff noted that it would be difficult to collect reliable information on the program model and ELD service delivery method

40 Although participants were most interested in a measure indicating either bilingual or monolingual instruction, there are also important differences among bilingual models. Transitional bilingual education uses the student's home language to bridge to English, usually for one to three years. In contrast, dual language programs have the maintenance of a partner (non-English) language as a goal, and thus last for at least five years and use the partner language for at least half of instructional time. Some dual language programs enroll only ELs or only non-speakers of the partner language, whereas two-way immersion programs enroll an equal number of each group. See Sugarman, *A Matter of Design*.

offered to each EL student. For one thing, the instructional program in which EL students are enrolled might change as students' needs and ELP levels change, even within a given school year, and some students receive multiple types of services simultaneously. Additionally, some programs are implemented in unusual ways, and the personnel responsible for data collection sometimes do not have sufficient knowledge of the different approaches to enter accurate information. As one example, several SEA staff members described how some schools have reported three students enrolled in a bilingual program at their sites; however, three students is not a sufficient number to constitute a program, and it is more likely that these students receive some bilingual supports (such as assistance from a bilingual paraprofessional) rather than being part of a bilingual program.

Access to Qualified Teachers

Beyond instructional program, all study participants were keenly interested in having more information related to EL students' access to qualified teachers. Participants identified three possible measures to assess this indicator within schools. First, having data on **EL-related teacher certification** was deemed highly important, specifically the number of teachers in each school who hold an ESL or bilingual certification or endorsement. These data would include both EL-only teachers (e.g., ELD teachers) as well as content-area teachers to provide information on all teachers' preparation to support EL students' academic and linguistic development.

A second measure of teachers' qualifications to work with EL students is the amount of **EL-related professional development** teachers participate in each year, which could be reported in hours or minutes. This professional development could be provided by the school or sought out by individual teachers. Study participants also expressed interest in collecting information about the specific topics addressed in professional development sessions, but they acknowledged that this is not especially feasible. Third, participants described the utility of reporting information on **EL teacher to EL student ratios** by school. These ratios would offer a general sense of each school's capacity to serve its EL students, irrespective of the instructional program offered.

With respect to the feasibility of collecting data on these measures, local leaders involved in the focus groups noted that they already collect information on teacher certification and professional development hours, but that they rarely, if ever, report such data to the state. At the state level, data related to teacher certification is often collected and housed separately from student demographic and program data, and SEA staff often do not have the capability to merge these datasets to examine EL-related teacher qualifications for instructors within specific programs in a school.

Access to Opportunities

The third indicator of most importance to study participants was the extent to which EL students have access to and successfully engage in the same opportunities as their non-EL peers. A first key measure within this indicator is **access to advanced coursework** at the secondary level. State and local EL program administrators are likely especially aware of this issue because a robust body of research shows that EL students are often excluded from advanced courses that would make them eligible for graduation and/

or college enrollment.[41] As such, access to coursework is generally considered an important measure of EL program effectiveness,[42] and some districts and states already collect data related to this measure.

Also at the secondary level, many study participants indicated that tracking EL students' **completion of the Seal of Biliteracy**[43] during high school, and the associated milestones at the end of elementary and middle school, would help them assess the quality of an EL program, particularly the extent to which EL students' linguistic strengths are valued and nurtured.

At the elementary level, where course enrollment is not a relevant measure, participants noted that **access to gifted and talented programs** would be a helpful measure of quality, as it would indicate whether EL students are equitably included in advanced instructional opportunities. While these measures are often collected at the local level (and sometimes at the state level), EL students' participation tends not to be reported publicly.

Beyond coursework, **participation in extracurricular activities** was noted as a key measure for understanding EL students' inclusion in opportunities outside of the classroom, as is required by civil rights law; these activities include sports, clubs, afterschool programs, and tutoring. However, local leaders noted that it would be very difficult to gather reliable information related to this measure, given fluctuations in when extracurricular activities are offered during the year, how often students can attend, and the staff who run them.

Family Engagement

The fourth indicator of interest to this study's participants was family engagement, as they acknowledged that providing meaningful opportunities for ELs' families to partner with schools is a key aspect of program quality (and is required by law) but is often missing in the schools they work with. Nonetheless, few participants had concrete ideas related to the specific measures that could be used to assess authentic EL family engagement. Further, it did not seem feasible to state and local leaders to collect data showing whether schools are consistently translating materials and events into the languages represented at each school.

Use of Existing Data

In addition to the four indicators described above, all of the interest groups represented in this study expressed a need for more nuanced reporting of outcome data to better assess EL program quality and

41 Rebecca M. Callahan, Lindsey Wilkinson, and Chandra Muller, "Academic Achievement and Course Taking among Language Minority Youth in U.S. Schools: Effects of ESL Placement," *Educational Evaluation and Policy Analysis* 32, no. 1 (2010): 84–117; Angela Johnson, "The Effects of English Learner Classification on High School Graduation and College Attendance," *AERA Open* 5, no. 2 (2019); Ilana M. Umansky, "To Be or Not to Be EL: An Examination of the Impact of Classifying Students as English Learners," *Educational Evaluation and Policy Analysis* 38, no. 4 (December 1, 2016): 714–37.

42 Rebecca M. Callahan and Megan Hopkins, "Policy Brief: Using ESSA to Improve Secondary English Learners' Opportunities to Learn through Course Taking," *Journal of School Leadership* 27, no. 5 (2017): 755–66; Rebecca M. Callahan and Dara Shifrer, "Equitable Access for Secondary English Learner Students: Course Taking as Evidence of EL Program Effectiveness," *Educational Administration Quarterly* 52, no. 3 (2016): 463–96.

43 The Seal of Biliteracy is a program available in nearly all states that awards a special recognition to high school graduates who show proficiency and literacy in English plus another language.

effectiveness. Specifically, they would like to see achievement and ELP assessment as well as graduation rate **data disaggregated by EL student group**. These groups include:

- newcomers (that is, EL students who have been in U.S. schools for less than three years);
- students with limited or interrupted formal education who are also EL classified;
- long-term ELs (students who have been EL classified for more than five years);
- EL students who are also identified for special education services; and
- former EL students (students who have been reclassified as English proficient).

Having access to outcome data by EL student group would provide information on the schools that are serving specific groups well (or not well), which would be valuable for families as they make decisions for their children and for leaders as they work to identify the kinds of EL programs that are most effective for particular students. However, data disaggregation by EL student group presents challenges. For instance, many states do not have concrete definitions for some of the groups listed above (e.g., newcomers, students with limited or interrupted formal education); such definitions would need to be developed to ensure consistency in reporting. Moreover, some student groups may be too small to ensure confidentiality in school- or district-level reporting. Nonetheless, local leaders often look at student performance data in disaggregated ways to inform program enrollment decisions and to examine student progress. But because of a lack of flexibility in state and district data systems, these data are often housed in a dataset developed by an EL director or other staff member and are not shared with others.

D. Implications

While this study's findings offer several lessons about the kinds of information that would be important to policymakers, educators, families, and community members, they also challenge thinking about EL accountability in general. The study's overarching research question concerned what opportunity-to-learn indicators could be added to statewide accountability systems. Adding indicators to the ESSA accountability system would seem advantageous, as some states are reluctant to add new data collection requirements beyond what is in ESSA. Additionally, indicators required by law tend to take on disproportionate importance—for better or worse—in terms of where schools and districts focus their attention and resources.[44]

While this study's findings offer several lessons about the kinds of information that would be important to policymakers, educators, families, and community members, they also challenge thinking about EL accountability in general.

However, the distinction that study participants articulated between "big A" and "little a" accountability suggests that the study's initial idea of how to frame new indicators would need additional nuance.

44 David Figlio and Susanna Loeb, "School Accountability," in *Handbook of the Economics of Education*, eds. Eric A. Hanushek, Stephen Machin, and Ludger Woessmann (North Holland, The Netherlands: Elsevier, 2011), 383–421.

ESSA does require schools, districts, and states to collect and report more information than is used for identifying schools for improvement. But it is not clear whether study participants considered that data to be accountability, or whether that word tends to be associated with the metrics that are used to compute school ratings and identify the lowest performers (i.e., test score outcomes, graduation rates, and one student success indicator, which in most states is attendance).[45] Rather than collecting data solely to identify schools for improvement, study participants articulated a strong desire to require data collection and reporting that would be used for targeted technical assistance and capacity building.

In order to implement new data collection related to EL programs, this study's findings point to the following recommendations:

1. Whereas most of the attention around accountability focuses on outcome measures and school ratings, federal and state accountability policy could lift up the importance of data collection and reporting related to programmatic inputs. Requiring data to be collected on the opportunities afforded to EL students, as well as the resources allocated to EL programming, would provide a more accurate picture of EL equity and signal to schools the inputs they need to consider.

2. SEAs should develop more robust data infrastructures to ensure alignment in data collection and reporting between the local and state levels and to support the integration of different data sources (e.g., demographic, teacher certification, and outcome data). To support these state-level infrastructures, federal financial support and guidance would be needed, including but not limited to a federal technical assistance center focused on data system development as well as support for the development of statewide data dashboards that can display opportunity-to-learn indicators in ways that are useful and comprehensible to a variety of interested audiences. Such systems would support LEAs with flexibly combining their outcome-based data with other data sources focused on program quality.

3. Federal and state policy and associated guidance could elevate attention to specific indicators of EL program quality that this study's participants deemed both desirable and feasible to collect and report: access to qualified teachers and access to opportunities. Specific measures associated with these indicators include: (1) counts of teachers with EL-related certifications, (2) EL students' Seal of Biliteracy completion, and (3) ELs' enrollment in gifted and talented programs. Many districts and states already collect this information, although it is not often examined in the context of assessing EL program quality. If these indicators were specified in federal and state policies, and data infrastructures put in place to support associated data collection and reporting on them, then LEA and SEA staff may be more likely to use them for decision-making.

4. To support the data disaggregation that so many of the study's participants desired, SEAs should consider developing consistent definitions for EL student groups and examining outcomes for each group by district and school, where possible given the size of the group. This includes definitions that most states have yet to incorporate into statewide data systems, such as newcomers, students with limited and interrupted formal education, and long-term ELs. Some states have already developed

45 See for example, The Education Trust, "New School Accountability Systems in the States: Both Opportunities and Peril," accessed November 3, 2023.

definitions of these groups and collect data on them; these states could serve as models for other states as they engage in similar work.

4 Conclusion

As part of a systemic school reform effort over several decades, the federal government's system of school accountability was designed to ensure that students from historically underserved groups are given access to high-quality instruction so that all students are prepared for college and careers. For ELs, accountability has undoubtedly raised the profile of persistent achievement gaps, but the system has yet to be refined to ensure accurate and useful information for decision-making, as this report illustrates.

For ELs, accountability has undoubtedly raised the profile of persistent achievement gaps, but the system has yet to be refined to ensure accurate and useful information for decision-making.

The findings presented here support the use of a statistical model that incorporates ELP levels when interpreting student outcomes in ELA and math. If states set expectations for EL performance and growth in academic content areas informed by ELP levels, they will have more actionable information to determine whether and how schools are supporting students in progressing toward and meeting grade-level expectations. Community members, educators, families, and policymakers, meanwhile, expressed a desire for better access to data collected using current indicators along with data that capture indicators of program quality to provide context for and boost understanding of student outcomes. Together, more nuanced student outcome findings and more information on whether and how schools provide language instructional support could help individual schools target resources and improvement efforts, while also contributing to a better understanding of EL program effectiveness more generally.

These conclusions signal the need for deeper and more transparent use of data in interpreting academic outcomes and improvement strategies for ELs. Much of the work would happen at the state level where SEA personnel conduct applied research that closely analyzes data collected from schools and districts. This report proposes statistical methods that states could apply to their current assessment outcomes to develop a more nuanced understanding of EL performance as well as new indicators to provide more data on program quality. States could revise their current ESSA plans with the statistical adjustments in order to begin generating better data without waiting for changes to the accountability provisions in the law. And with focus group and interview participants indicating there is considerable practitioner and public interest in program quality data for program improvement ("little a" accountability) rather than as a component of the accountability models that identify underperforming schools ("big A" accountability), states could proceed with adding new indicators to their data dashboards even without revisions to the federal accountability system.

A key assumption that bears mention in discussion of these recommendations is that refining accountability models should not signal lower expectations for EL students. Rather, these recommendations, and recommendations being considered in other arenas, should reflect a need to better interpret test scores and

what information decisionmakers need to provide appropriate instruction for EL students in both content area subjects and in language development.

While more research is needed on existing state-level accountability practices, the findings and implications reported here highlight the need for revisions in federal policy to better support ELs. Additionally, new assessments and alternative accountability approaches not covered in this report, such as testing students more frequently during the school year, must be analyzed for their impact on ELs. Reauthorization of ESEA is not imminent, and the political climate calls for flexibility in state level decisions about accountability. However, it is not too early for states and advocates to begin putting together the evidence necessary to support changes in federal legislation. Such changes could include the following:

1. ESSA reauthorization could require states to include one or more program quality indicators in their accountability systems, such as those suggested in this report. This legislation could set parameters for quality indicators that are broad enough to accommodate state context and community priorities but also have evidence-based links to school effectiveness. By doing so, federal legislation can help states generate additional information for better instructional decisions while maintaining state flexibility.

2. The federal government could also create incentives for states to conduct deeper analysis of state-level outcomes using more nuanced statistical models that include ELP. Study participants, including state education agency personnel, expressed a desire for finer-grained information from accountability systems. However, most states are challenged by insufficient knowledge about how to create statistical processes that yield accurate data about all subgroups and too many other competing priorities for limited data analysis funding. Federal investment in pilot programs and competitive grants would provide the resources states need to investigate how to incorporate the new approaches suggested in this report into their unique contexts.

3. Federal grants could support states' development of enhanced data systems. As this study unfolded, researchers found that states often have the data necessary to make better decisions about EL outcomes and instructional responses, but that this information is located in different datasets, making it difficult for states to undertake deeper analyses. Additionally, these datasets are often not easily available to community advocates and education researchers who would wish to conduct deeper analyses. Federal funding could facilitate development of the architecture to combine a broader array of data in more accessible formats.

4. While ESSA required states to include ELs as one of the subgroups that could trigger eligibility for federal school improvement funding, initial state ESSA plans fell short in describing how they would implement improvement strategies responsive to EL needs.[46] Even if states developed more nuanced accountability systems for ELs, the outcomes generated would need to trigger specific responses. Reauthorized ESSA legislation could require more intentional improvement plans tailored to unique subgroups. At the same time, the federal government could issue guidance that provides models of customized strategies and highlights states that have been successful in developing actionable and specific improvement strategies.

46 Villegas and Pompa, *The Patchy Landscape*.

5 Numerous interest groups have called for the federal government to create a common definition of newcomer students to facilitate data collection.[47] The current federal definition of newcomers is applied in different ways in different states and, sometimes, districts within the same state. The lack of a consistent definition often masks the performance of newcomers, resulting in overly broad instructional responses for their needs. The same may be said for other groups of ELs, such as students with limited or interrupted formal education or long-term ELs. Common definitions would provide better information about how each of these groups are faring academically and would encourage appropriate instructional responses.

These steps would help states maintain flexibility while establishing more specific parameters that reflect this research and best practices from leading states. Crucially, getting accountability right for the United States' growing EL population will lead to improvement in accountability overall.

47 Kara Arundel, "Advocates Seek More Resources for Newcomer Students from Ed Dept," K–12 Dive, September 7, 2022.

Appendices

Appendix A. Methodology

This report addresses findings from two independent, but related, studies. Using quantitative and qualitative methods, the project sought to investigate how states might refine their school accountability models to provide more accurate data on English Learner (EL) growth. The quantitative side of the project was directed by Dr. Pete Goldschmidt of California State University Northridge, and the qualitative side was directed by Dr. Megan Hopkins of the University of California, San Diego and Dr. Julie Sugarman of the Migration Policy Institute (MPI).

Quantitative Methods

Following on prior research on the connection between English language proficiency (ELP) and academic achievement, the quantitative side of the project aimed to propose refinements to current growth models used to judge the effectiveness of academic instruction for all students. The proposed refinements would more accurately model EL progress on English language arts (ELA) and math and explicitly consider ELP in interpretation of academic performance. Refined models may produce results that are more sensitive to variability in schools' contribution to EL student success. The following three questions guided this research:

1. Does aligning EL progress toward ELP and growth on state content assessments result in more coherent understanding of schools' contribution to EL success?
2. Does incorporating EL performance (either ELP levels or progress) into monitoring EL performance and progress on content generate a more coherent model that supports valid claims about schools?
3. Incorporating results from questions 1 and 2, what are the benefits and challenges related to (a) specifically monitoring current and former ELs as separate subgroups, and (b) the impact of minimum N on inclusion, accuracy, and claims about schools?

The researcher used state-level datasets from Hawaii and Ohio that included data from school years 2016–17, 2017–18, and 2018–19, the last three years before COVID-19 disrupted instruction and the administration of state standardized tests. Using those two states allowed comparisons of students who take the two most common ELP assessments: ACCESS for ELLs and ELPA-21. Hawaii is one of 42 states and territories belonging to the consortium using the former test, and Ohio is one of seven states using the latter. With these datasets, the researcher applied a three-step process to develop and test alternative methodological options that potentially enhance coherence and validity:

- **Step one** examines state operational data—that is, test scores from actual administrations of ELA, math, and ELP tests—to identify the structure of EL progress, the structure of content progress, and the relationship between the two.
- **Step two** develops a simulated dataset based on the analyses from step one. The simulated data make it possible to create known school effects that the refined models attempt to capture. Specifically, these models move away from simply looking at student gains or growth-to-target to more

dynamic models that better reflect EL progress toward English proficiency. For content performance (demonstrated status and growth), the refined models move away from simply using percent proficient, or change in percent proficient, to conditional models that not only capture content performance but also explicitly consider EL language ability.

- **Step three** applies models developed in step two to state operational data to examine the similarities and differences in claims about schools based on the state accountability model and the developed refined model.

Qualitative Methods

The qualitative study focused on aspects of EL programming that are most salient for understanding the nature and quality of language instructional supports that EL students receive. The overarching research question guiding this study was: What information about EL services would be useful for diverse audiences to assess program quality in an accountability system? Specifically, the study sought to understand:

1. What information about language instructional services would be of use to diverse interest groups (i.e., state and local leaders, parents and families, and community groups) in an accountability system?
2. What indicators would help to gauge whether schools are meeting their civil rights obligations to provide support for ELs that facilitates ELP and academic achievement?
3. How can data related to these new indicators be feasibly collected and reported by local and state leaders?
4. How could parents, families, and community groups use this information to make sense of student outcomes?

This study's research methods include focus groups and interviews conducted with individuals from diverse interest groups from 18 states. States were purposively selected to include distinct geographic and policy contexts, from states with large and stable EL populations to states with small and growing populations as well as states that offer explicit support for dual language and/or bilingual education programs and states that do not. Including these states made it possible to identify common challenges and opportunities across contexts.

Participants included 17 state education agency (SEA) staff, 37 representatives from community advocacy organizations, 24 EL family members, and 5 regional or local education agency (LEA) leaders. The advocate and EL family member focus groups were conducted by staff from state-level immigrant policy and community interest organizations with whom MPI's National Center on Immigrant Integration Policy has collaborated regularly in its broader work. The EL family member focus groups were conducted in Spanish, and all participants were Spanish-speaking mothers of students who were previously or currently identified as ELs.

Data were collected and analyzed across two phases. The first phase focused on identifying a set of new accountability indicators. The second phase assessed the feasibility of collecting data from districts and schools related to the indicators identified in the first phase.

Phase 1

The first phase of the study aimed to collect information on what indicators of program quality would be interesting and useful to a variety of interested groups. It proceeded along two lines:

SEA staff. First, researchers identified SEA staff in 15 states with distinct contexts, including states with large and small populations of ELs, as well as those that explicitly support bilingual education and those that do not (see Table A–1 for which states were represented in which aspects of the data collection). Each administrator participated in a 45–60 minute virtual focus group held using Zoom, and the groups included three to five participants, for a total of 17 participants.

SEA staff were asked to describe their state's current approaches to assessing EL program quality and to discuss the utility of collecting data on a range of potential program quality measures. Potential measures included: language(s) of instruction, program type, English language development (ELD) instructional delivery (including mode and amount of time), access to advanced coursework, teacher certification, curricular materials, and targeted interventions. Participants also had the opportunity to identify any additional measures, to discuss how they might use information related to these measures, and to describe any challenges with collecting these data.

Community members. MPI's National Center on Immigrant Integration Policy has frequently collaborated with state-level immigrant policy and community interest organizations in the course of its work on issues related to EL and immigrant education. For the community member focus groups, MPI asked five of its partner organizations to facilitate focus groups within their communities. This approach allowed the researchers to attend to differential power dynamics between participants and the research team that could have affected participants' comfort and limited their candid responses. The partner organization staff recruited participants, were trained by MPI on procedures and using the focus group protocol, conducted the focus groups, and submitted recordings and summaries of the sessions.

The groups conducting community advocate focus groups were Californians Together (who did two such focus groups), the New York Immigration Coalition, and Intercultural Development Research Association (IDRA), based in Texas. The focus groups ranged in size from 4 to 18 participants, for a total of 37 participants. Three were conducted virtually using Zoom, while one was held in person.

Additionally, three organizations conducted focus groups in Spanish with EL family members, Conexión Américas (Tennessee), IDRA (Texas), and the Latino Community Fund (Georgia). All of the family members were Latina mothers with one or more students who were currently or had at one point been identified as ELs and received services from their elementary, middle, and/or high school. All of the family focus groups were done on Zoom, and they ranged from 6 to 11 participants.

Community advocates and EL family members were asked similar sets of questions. The questions concerned what they know about the services that ELs receive in their school or community, what more they would like to know about those services, how knowing this information might inform their decision-making or advocacy, and what kind of data reports would be helpful. They were asked about the potential utility of specific measures, including teacher certification, amount of time for ELD, use of native language for instruction, how much ELs are integrated with non-ELs for instruction, and class sizes.

Analysis. All focus groups were audio recorded with participants' permission, and extensive notes were taken during each discussion. After completing focus groups for each group, notes were analyzed to identify measures that were consistently raised as important. The research team reviewed the audio recordings if clarification questions arose or more detail was needed. Then, the researchers grouped the measures identified by each group into categories representing overarching indicators of EL program quality. Finally, the researchers compared findings across participant groups to generate a list of potential indicators for review in Phase 2.

Phase 2

In the study's second phase, the researchers interviewed five regional or local education agency leaders from three different states. First, the researchers asked these leaders to describe their current EL programs and how they assess program quality across schools. Participants then reviewed the list of indicators developed in Phase 1 and reflected on whether and how they would collect this information and how this information would (or would not) contribute to understanding program quality in their region or district. As with the focus groups, each interview was audio recorded with participants' permission, and detailed notes were taken to capture participants' responses. These notes were then analyzed to identify indicators that could be reasonably collected across districts and to assess common challenges to data collection.

TABLE A–1
States Represented in the Qualitative Study

State	State Education Agency Staff	Community Advocates	EL Family Members	Regional or Local Education Agency Staff
California*	X	X		
Connecticut	X			
Delaware	X			
Georgia			X	
Indiana	X			
Louisiana	X			
Massachusetts*	X			
Michigan	X			
Minnesota*	X			
Mississippi	X			X
New Jersey*	X			X
New York*		X		
Ohio	X			X
Oregon*	X			
Tennessee			X	
Texas*	X	X	X	
Utah*	X			
Virginia*	X			

* indicates a state where bilingual or dual language programs are explicitly supported in law or educational policy.

Appendix B. Statistical Compendium for Understanding and Untangling Confounding in Status

Technical Details

The researchers first ran a simple regression analysis to estimate model parameters. The model is given as:

$$Gap = b_0 + b_1(ELP\_level) + b_2(current\_grade) + e \quad (1)$$

In this model, *Gap* represents the difference between a student's scale score and the state grade-level proficiency cut score. *ELP_level* is a students' ELP level, *current_grade* is the same student's current grade level, and *e* is the residual. The parameters to estimate are b_0, b_1, and b_2, which represent the intercept, the adjustment due to ELP level, and the adjustment due to grade, respectively.

Then, the researchers used the difference between a student's current ELP level and the state exit score, or an ELP level of 5,[48] as well as the model parameters from above (a function of ELP level and current grade level) to generate adjusted proficiency cut scores for accountability purposes.[49] If a student's ELP level is less than the exit score of 5, then the model is given as:

$$Adj\_Cut = Cut - b_1(ELP\_Exit - ELP\_level) + b_2{*}current\_grade \quad (2)$$

Else

$$Adj\_Cut = Cut.$$

In this model, *Adj_Cut* is the adjusted cut score for a given ELP level and grade. *Cut* is the state's grade-level cut score for proficiency. *ELP_Exit* is the ELP level required for a student to exit EL status (5 in this example). *ELP_level* is the student's current ELP level. All other elements of (2) are as defined in (1).

It is important to emphasize that only the parameter estimates for ELP level and grade level are used to create adjusted cut scores, as the accountability adjustment creates appropriate expectations given ELP and is not attempting to equalize proficiency rates between non-ELs and ELs. These estimates are consistent with the original proficiency cut scores, they increase over grade levels (because Hawaii has a vertical scale),[50] and they increase as ELs near English proficiency.

Additional Results

Results from analysis of math data from Hawaii are presented in the body of the report, in Figure 2 within Section 2.B. Figure B–1 displays analogous results for ELA and shows that cut scores would increase with grade and ELP level. In other words, as ELs score at higher ELP levels, their expected ELA performance would change accordingly.

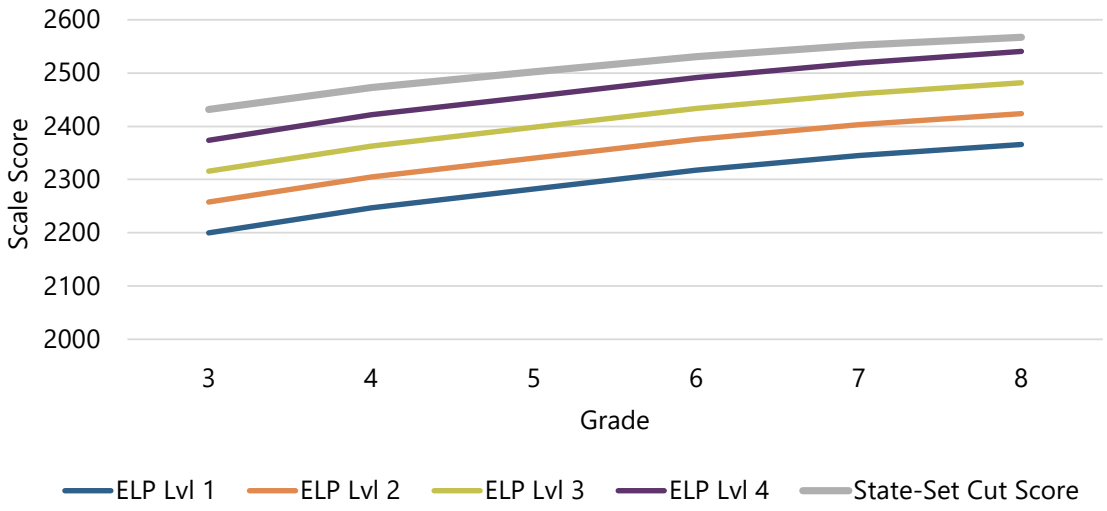
48 This example is based on Hawaii. The state uses the WIDA Access for ELLs 2.0 assessment where 5 was the determined ELP cut score. For content assessments, Hawaii uses the Smarter Balance Assessment System. Of course, any meaningful cut scores can be used.

49 Such an analysis could be run by states to create their own adjustment for ELs.

50 A vertical scale places scores for every grade level on a common scale to allow comparison over time. The process described here is also possible without a vertical scale in either or both content assessments or ELP assessment. ELP levels or domain level sum scores can be used as an input, and given that the model accounts for grade, content does not need to be on a vertical scale.

FIGURE B–1

Adjusted Proficiency for Accountability: ELA Scale Scores by ELP Level, Hawaii

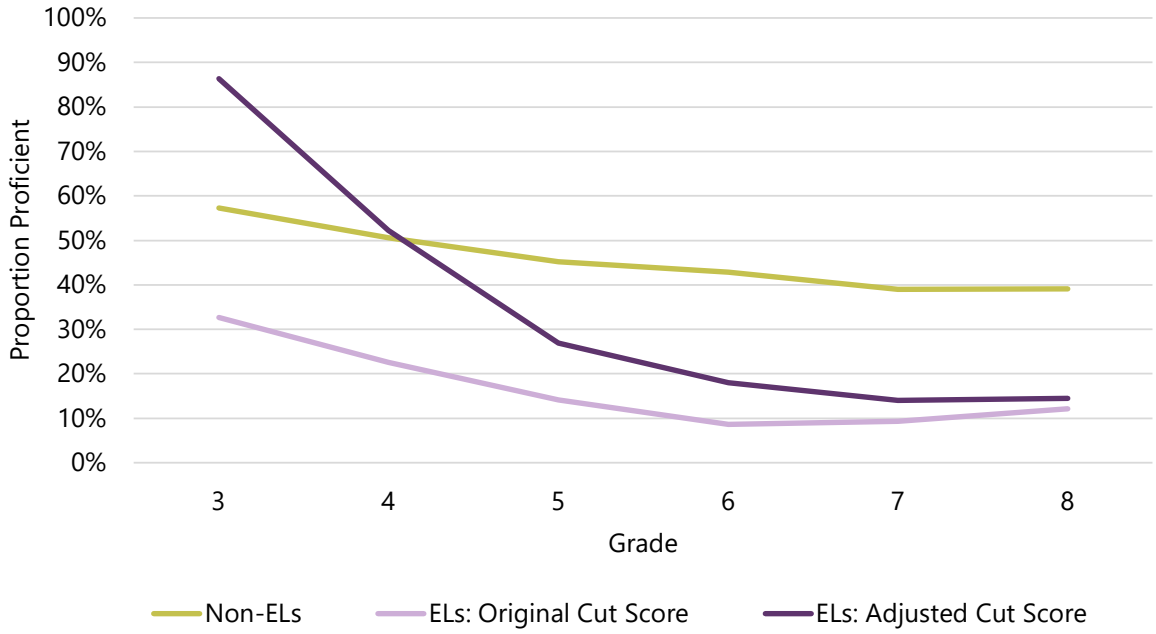


Source: Author analysis of Hawaii assessment data from school years 2016–17 through 2018–19, received from the Hawaii State Department of Education in September 2023.

Next, while the impact of applying adjusted cut scores for the purposes of accountability on proficiency results in ELA for students in Hawaii is presented in Figure 3 within Section 2.B of the report, the results for math are displayed in Figure B–2. The comparison between original and adjusted cut scores is particularly striking here, as ELs in Hawaii quickly fall behind non-ELs despite consideration for their ELP level.

FIGURE B–2

Proportion of EL and Non-EL Students Reaching Proficient Status in Math, Hawaii

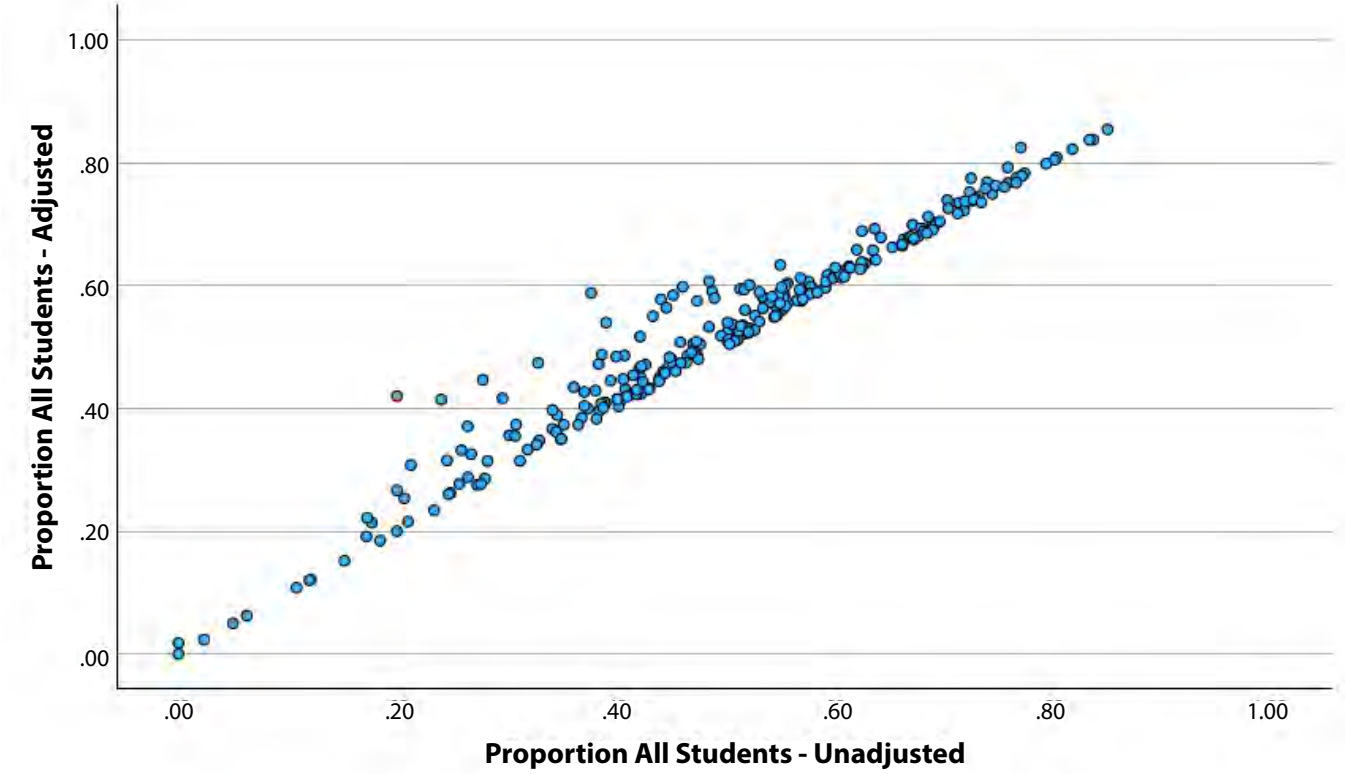


Source: Author analysis of Hawaii assessment data from school years 2016–17 through 2018–19, received from the Hawaii State Department of Education in September 2023.

The overall results do not drastically change when aggregated to the school level. Figure B–3 plots the original unadjusted proportion of all students in Hawaii schools who are meeting academic proficiency in ELA on the horizontal axis against the adjusted proportion on the vertical axis. The correlation between the results is .98, meaning that the adjustment does not change overall claims about schools in terms of status. This result is expected because in most schools, most students are not identified as ELs. The results for math were very similar and are not presented graphically here.

FIGURE B–3

Impact of Unadjusted and Adjusted Proficiency Cut Scores on the Proportion of Students within Schools Meeting Standards in ELA, Hawaii

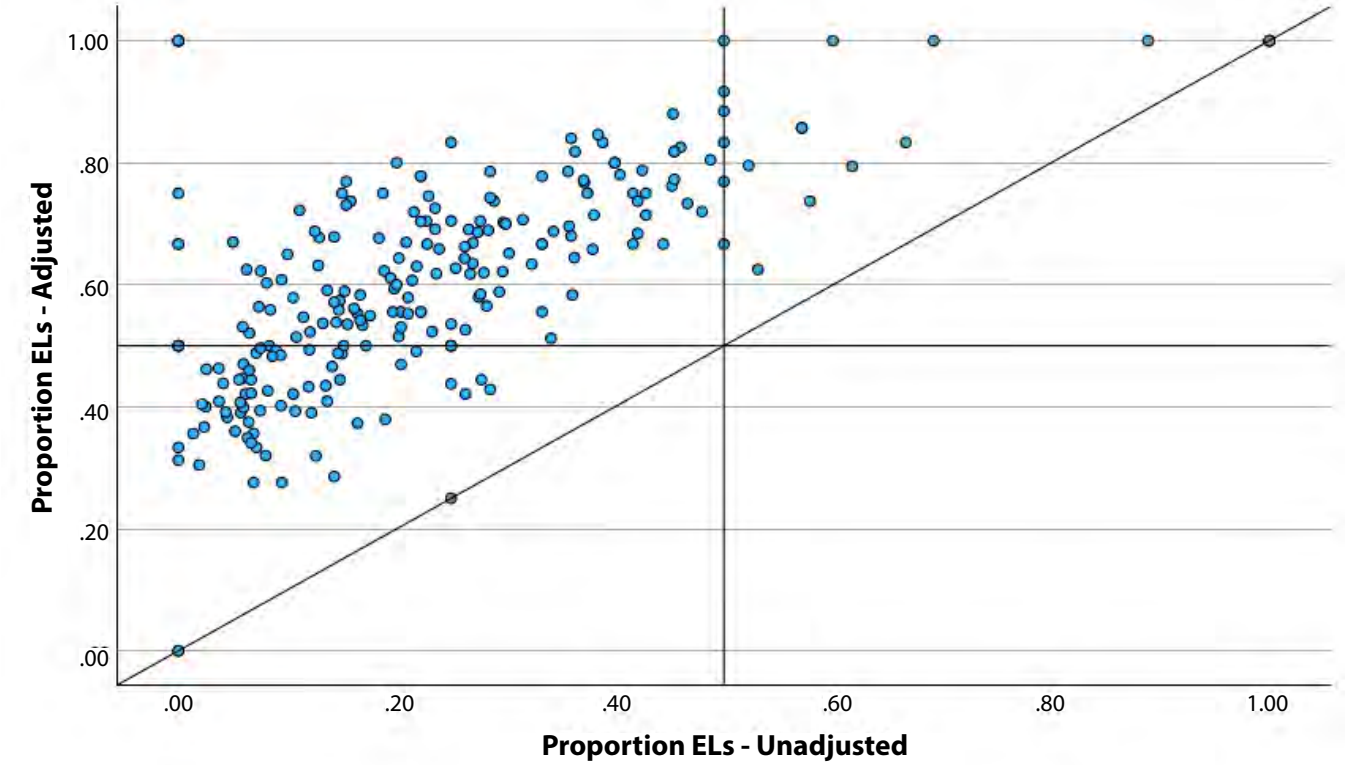


Source: Author analysis of Hawaii assessment data from school years 2016–17 through 2018–19, received from the Hawaii State Department of Education in September 2023.

Examining EL students specifically, Figure B–4 shows the impact of unadjusted and adjusted academic proficiency cut scores in ELA on ELs within Hawaii schools. For reference, in Hawaii the overall average percentage of students meeting grade-level standards in ELA is about 50 percent. With this reference point in mind, the results in Figure B–4 indicate that rather than virtually every school reporting that fewer than 50 percent of their ELs met the ELA proficiency cut score, the distribution of schools whose ELs achieve proficiency in ELA is substantially more evenly distributed around 50 percent when using the adjusted cut scores. As the example presented here shows, there is still a fairly strong relationship between the traditional and adjusted approaches; however, the adjusted cut scores afford opportunities to consider more nuanced reasons for why ELs are not meeting targets.

FIGURE B–4

Impact of Unadjusted and Adjusted Proficiency Cut Scores on the Proportion of ELs within Schools Meeting Standards in ELA, Hawaii



Source: Author analysis of Hawaii assessment data from school years 2016–17 through 2018–19, received from the Hawaii State Department of Education in September 2023.

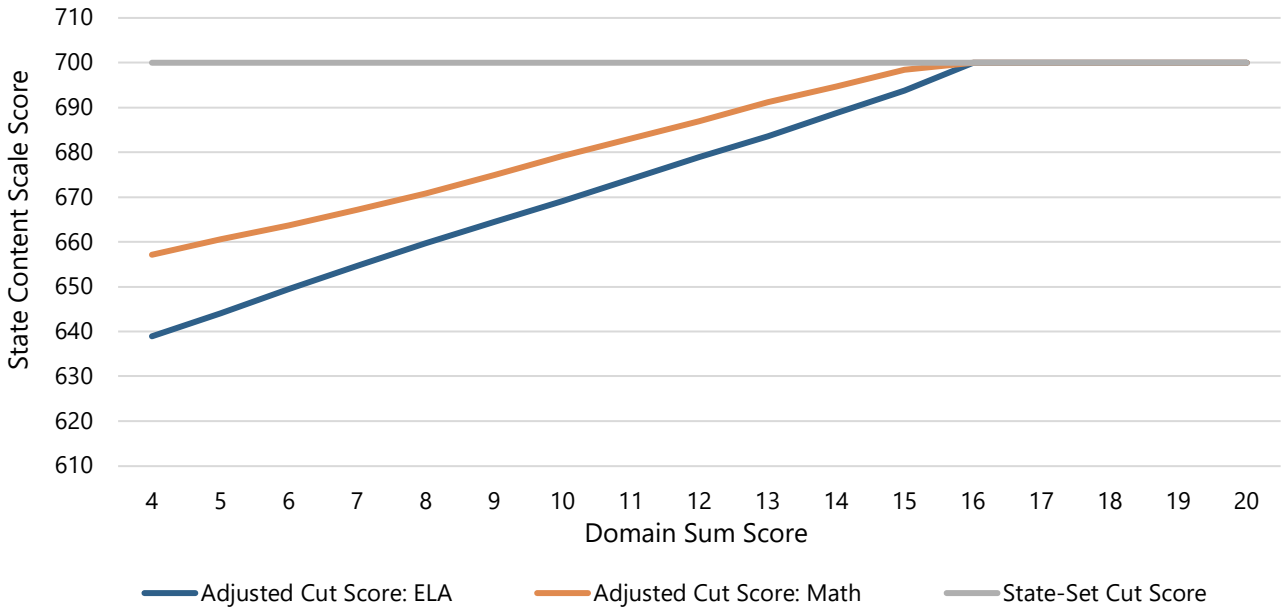
The researchers repeated the above analyses using data from Ohio. Ohio provides a distinct example of the impact of adjusting proficiency levels for accountability in that Ohio's ELA, math, and ELP assessments are not on a vertical scale like Hawaii's are. However, the process used for Hawaii data can still be applied to Ohio data in order to make the same sorts of inferences.

Figure B–5 presents the adjusted proficiency cut scores for Ohio ELs in ELA and math by ELP domain sum scores.[51] Given that Ohio does not use a vertical scale, the proficiency cut scores for ELA and math are 700 across all assessed grades. Figure B–5 shows how accountability proficiency cut scores increase with ELP domain sum scores until reaching the standard cut score at a domain sum score of 16, which is the minimum required score for reaching ELP.

51 For summative ELP assessment, Ohio uses the ELPA-21 assessment that provides domain levels (reading, writing, listening, and speaking) that can be summed to create a domain sum score. Each domain has five performance levels. Domain sum scores are highly correlated with overall composite scores ($r > .90$) and can be used as composite levels as was done for Hawaii to create adjusted state assessment proficiency cut scores for accountability.

FIGURE B–5

Adjusted Proficiency for Accountability: ELA and Math Scale Scores by ELP Domain Sum Scores, Ohio

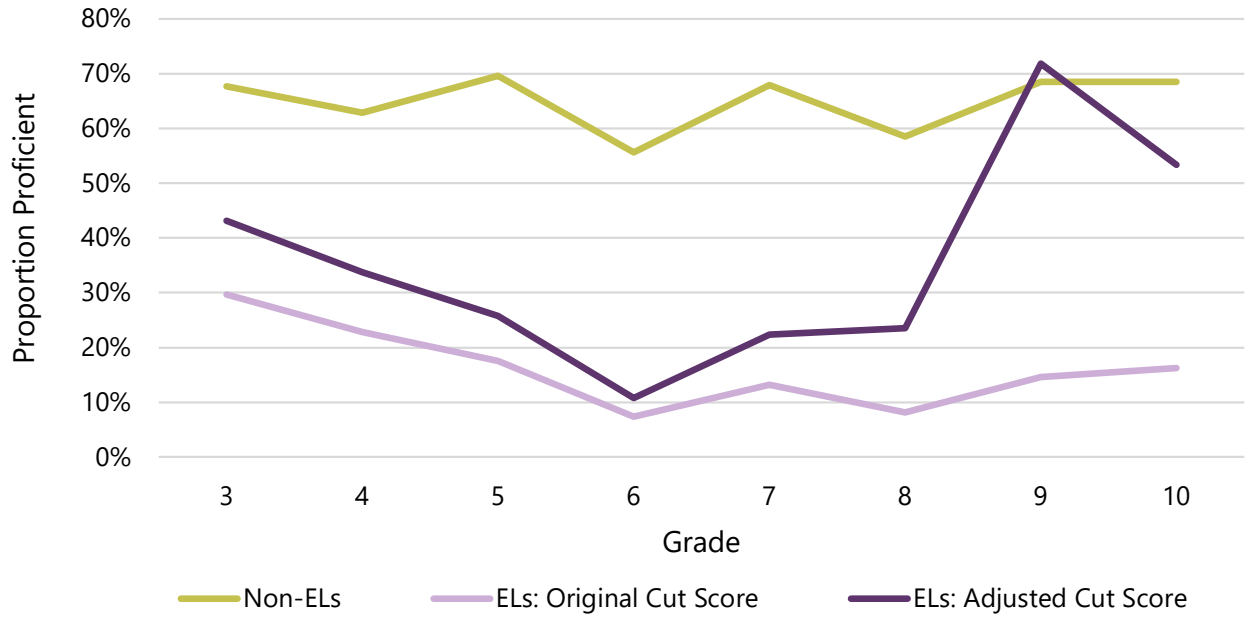


Source: Author analysis of Ohio assessment data from school years 2016–17 through 2018–19, received from the Ohio Department of Education and Workforce in February 2020.

Figure B–6 summarizes the impact of applying adjusted cut scores for the purposes of accountability on proficiency results in ELA for EL and Non-EL students in Ohio. Figure B–7 presents these results for math. Overall, these figures show that, using the original benchmarks, it would be misleading to say that ELs in Ohio are behind in ELA without considering where they should be given their ELP level at the time of the assessment. For example, in Figure B–6, the percentage of ELs meeting the 4th grade cut score is less than 25 percent; however, if the cut score is adjusted to reflect their ELP, about 35 percent of ELs meet their 4th grade ELA target. The data then change the conversation to one that centers on why, when taking ELP into account, only 35 percent of ELs meet the proficiency cut score compared to nearly 65 percent of non-ELs. The adjusted results are particularly striking in Ohio as there are only minimal changes for most grade levels. Because only the ELP level and grade parameters are used to generate adjusted proficiency cut scores, these adjustments are specifically addressing the role of language and are not attempting to equalize proficiency rates between ELs and non-ELs.

FIGURE B–6

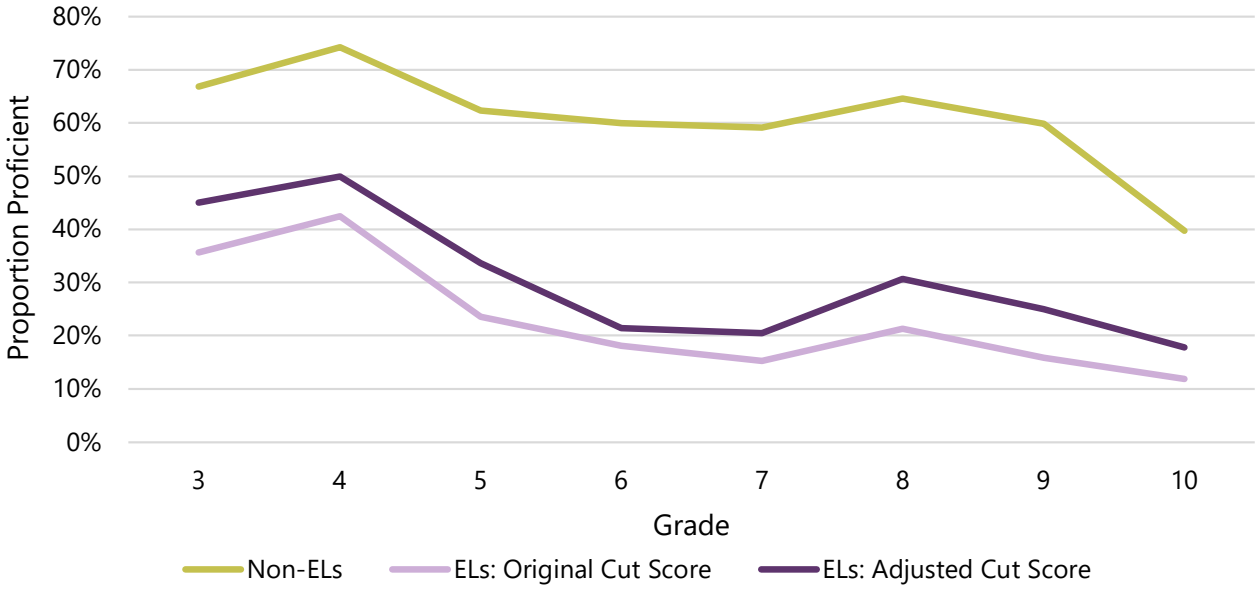
Proportion of EL and Non-EL Students Reaching Proficient Status in ELA, Ohio



Source: Author analysis of Ohio assessment data from school years 2016–17 through 2018–19, received from the Ohio Department of Education and Workforce in February 2020.

FIGURE B–7

Proportion of EL and Non-EL Students Reaching Proficient Status in Math, Ohio

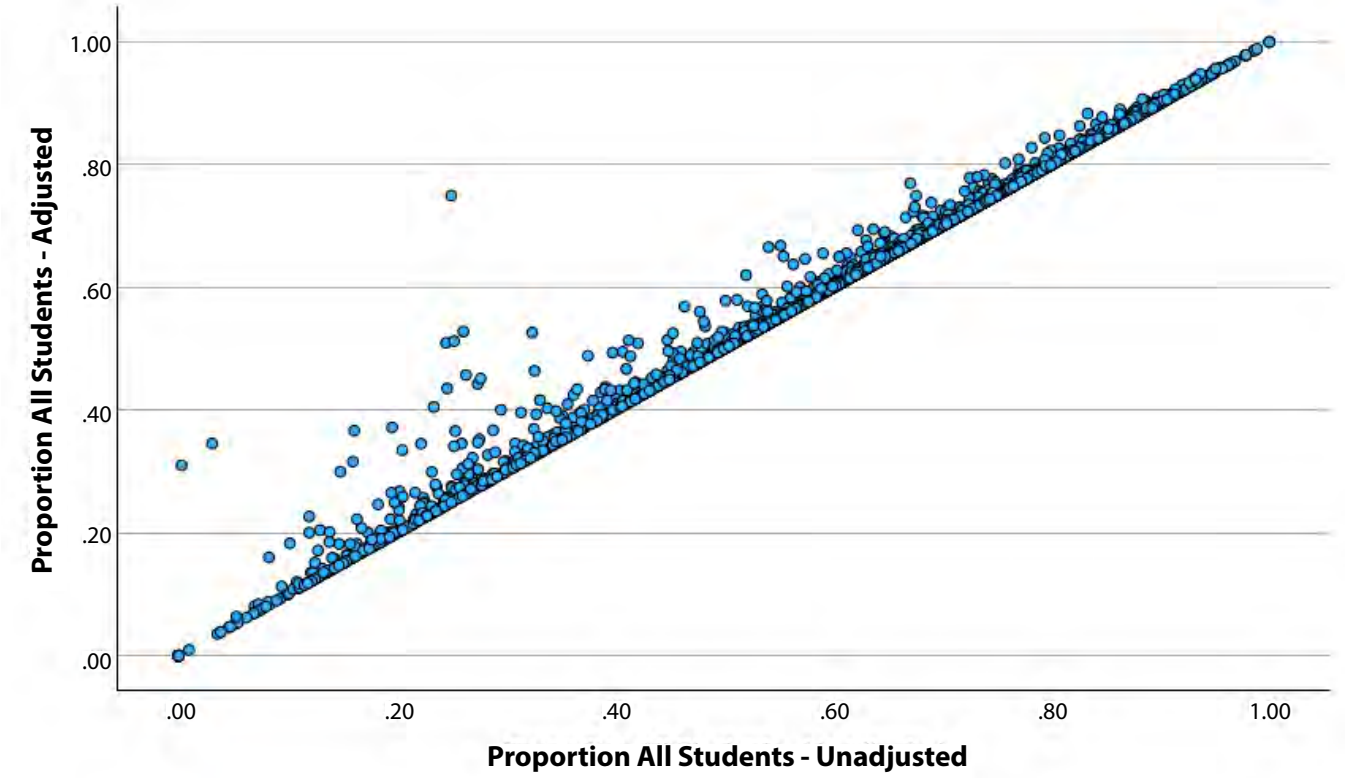


Source: Author analysis of Ohio assessment data from school years 2016–17 through 2018–19, received from the Ohio Department of Education and Workforce in February 2020.

Like Hawaii, aggregation to the school level leaves the results virtually unchanged because the proportion of ELs in Ohio schools is generally very small. The correlation of overall school percent proficient before and after adjustment is approximately .99 (see Figure B–8).

FIGURE B–8

Impact of Unadjusted and Adjusted Proficiency Cut Scores on the Proportion of Students within Schools Meeting Standards in ELA, Ohio

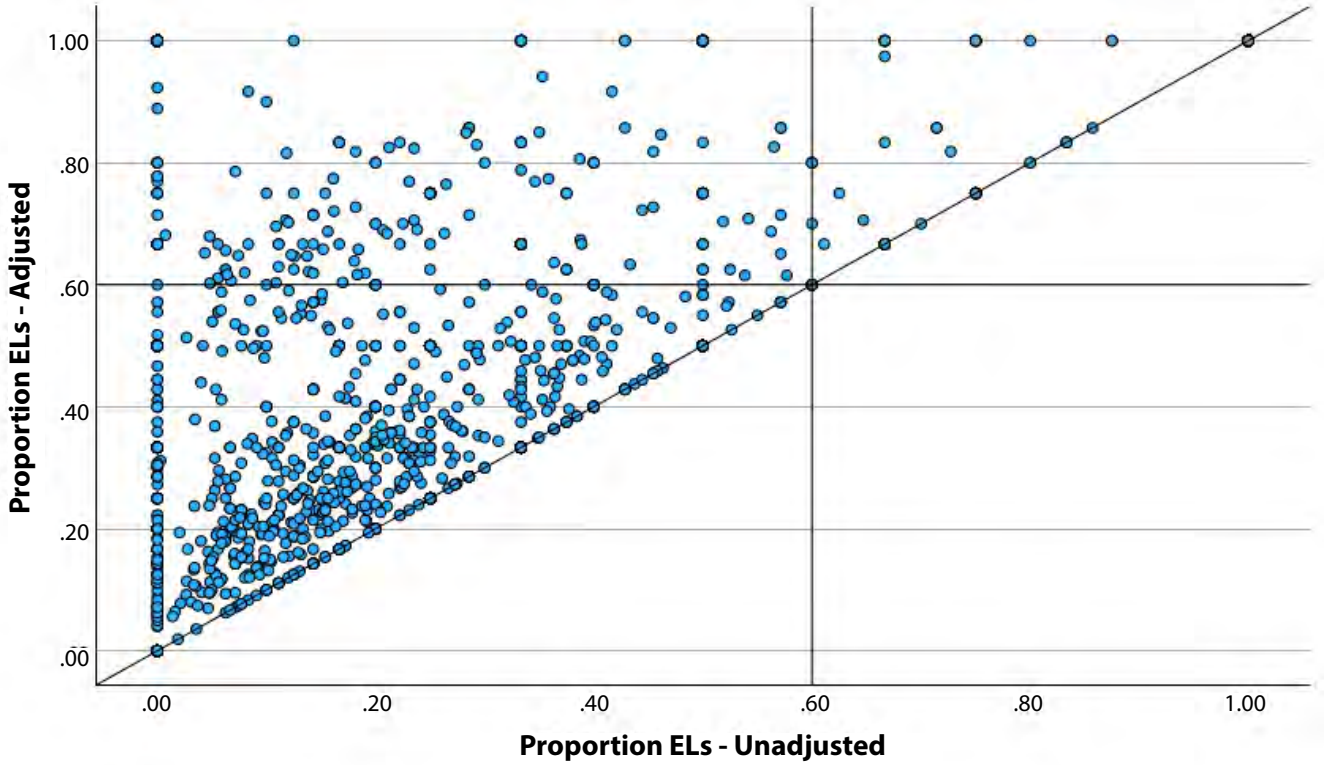


Source: Author analysis of Ohio assessment data from school years 2016–17 through 2018–19, received from the Ohio Department of Education and Workforce in February 2020.

However, the results for Ohio ELs vary substantively for ELA, as highlighted in Figure B–9. These results indicate that rather than virtually every school achieving less than 60 percent of ELs meeting the proficiency cut score (which is the overall average percentage of students meeting proficiency across Ohio schools), the distribution is substantially more evenly distributed around 60 percent using the adjusted proficiency cut.

FIGURE B–9

Impact of Unadjusted and Adjusted Proficiency Cut Scores on the Proportion of ELs within Schools Meeting Standards in ELA, Ohio



Source: Author analysis of Ohio assessment data from school years 2016–17 through 2018–19, received from the Ohio Department of Education and Workforce in February 2020.

Appendix C. Statistical Compendium for Understanding and Untangling Confounding in Growth

Technical Details: Using a Time-Varying Covariate in the Model

A growth model that does not include a variable for ELP assessment results is given as:

$$Y_{tij} = b0_{ij} + b1Grade_{tij} + e1_{tij} \quad (1)$$

In this model, Y_{tij} is the academic assessment results at grade *t*, for student *i*, attending school *j*. $b0_{ij}$ is the expected initial score in the first assessed grade.[52] *Grade*[53] is the assessed grade level at time *t* for student *i*, attending school *j*. In Model (1), *b1* captures the effect of grade (time) on Y_{tij} and represents academic progress. $e1_{tij}$ represents the random measurement error at time *t* for student i, attending school *j*. This

52 In order for this interpretation to be true, *Grade* is recoded to equal *Grade* – 3, such that the interpretation of *b0* is the expected score when the recoded *Grade* equals 0.

53 Generally, growth models are a function of both a linear and a quadratic time marker, in order to capture the deceleration of academic growth of time. Equation (1) is a simplified example to demonstrate the potential bias without loss of generalizability. Equation (1) is not used in subsequent analyses of either simulated or state operational data.

type[54] of model is commonly used in state accountability systems. Since it does not include a variable for ELP, ELP becomes an omitted variable and results in omitted variable bias—in particular, it upwardly[55] biases variable *b1* and increases the variability of $e1_{tij}$. In other words, the growth estimate is biased due to the omitted language proficiency of an EL as indicated by their ELP assessment results. This results in a less precise and less reliable estimation of *b1*, which in turn results in less accurate estimates of a school's contribution to student progress and growth.

A refined accountability model, marked as (2), that includes a time-varying covariate (TVC) for ELP assessment results, is given as:

$$Y_{tij} = b0_{ij} + b1Grade_{tij} + b2ELP_{tij} + e2_{tij} \quad (2)$$

In Model (2), the variables and parameters are defined as in Model (1) above but with the addition of ELP_{tij}, which is the ELP level at time *t*, for student *i*, attending school *j*. Here, *b2* captures the effect of ELP on student content assessment results, Y_{tij}. The addition of ELP_{tij} addresses the omitted variable bias issue present in Model (1) and takes into account the connection between ELP and academic progress and growth over time. It is important to note that $b2ELP_{tij}$ is time (grade) dependent and is not a fixed indicator of a student characteristic, but rather a time-varying indicator of a school-facilitated input (annual ELP assessments).

Including additional assessment information appears to be a viable option to address the omitted variable problem identified above. As noted previously, ELP assessment results are time dependent[56] so that including ELP assessment results is, in fact, including a TVC. The estimated effect of time (*Grade*), *b1*, represents the change in content scores associated with *Grade*, that is uncorrelated with other changes in content results due to changes in ELP assessment results.

The following presents a stylized example of the dummy variable adjustment (DVA) as an option to generate less biased estimates of school effects.

The true model that the data are generated by is given as:

$$Y_{tij} = 50*initial\_level_{tij} + 50*Grade_{tij} - 2*Grade2_{tij} - 30*EL_{tij} + .15*ELPSS_{tij}*EL_{tij} + e_{tij} \quad (3)$$

In this model, *Y* is the academic content scale score, *initial_level* is a random indicator of student preparedness to learn in 3rd grade, and *Grade* is the current grade level. In Model (3), subscript *t* refers to time (grade), *i* refers to student, and *j* refers to school. In this model, the numeric values are scale score points; that is, *50*Grade* indicates that a student grows at 50 scale score points per grade. *Grade2* is grade level squared and captures the deceleration in growth over time, *EL* is an indicator for whether the student

54 Type is used broadly to describe growth models that are based on statistical modeling, such as growth models, value-added models (VAM), or student growth percentile (SGP) models (as opposed to a value table that compares prior performance levels with current performance levels).

55 Omitted variables can create both positive and negative bias in estimates. Given that language proficiency is positively related to grade and content, the bias is positive.

56 ELP assessments are typically administered in the winter or spring.

is an EL, *ELPSS* is an EL student's score on the ELP assessment (which is 0 for non-EL students), and *e* is the residual.

The results of using Model (3) for ELs to estimate the parameters are presented in Table C–1. Consistent with expectations, the model reproduces the true parameter estimates quite well. The emphasis is on *Grade* and *Grade2*, as these together represent the growth in academic content over grades (time). The estimate for *Grade* demonstrates about .02 percent of bias, while the estimate of *Grade2* is unbiased (to two decimal places). The estimates in Table C–1 are intended to demonstrate that the correct model will capture the "true" results accounting for a random residual that is included in generating the data.

TABLE C–1

Results Using Known Model

	B	Std. Error	Beta	t	Sig.
(Constant)	-0.20	0.28		-0.69	0.488
initial_level	50.08	0.05	0.51	1078.20	<.001
Grade	50.01	0.07	1.43	710.80	<.001
Grade2	-2.00	0.01	-0.77	-406.98	<.001
EL	-30.03	0.41	-0.12	-72.95	<.001
ELPSS*EL	0.15	0.00	0.07	43.35	<.001

Source: Authors' simulated data.

Table C–2 presents results of a naïve model estimating growth in academic content for ELs. The naïve model is given as:

$$Y_{tij} = b0_{ij} + b1Grade_{tij} + b2_{tij}Grade2 + r1_{tij} \quad (4)$$

The known parameters are 50 for *Grade* and -2 for *Grade2*. The results in Table C–2 indicate that if Model (4) is utilized to estimate growth in academic content, the growth estimates will be biased. In this case, the linear growth term is biased by about 12 percent and the deceleration term is biased by about 6 percent.

TABLE C–2

Results Using Naïve Model for ELs

	Unstandardized Coefficients				
	B	Std. Error	Pct Bias	t	Sig.
(Constant)	-101.29	1.27		-80.04	<.001
Initial_level	59.96	0.19		320.17	<.001
Grade	55.89	0.46	12	121.28	<.001
Grade2	-1.88	0.04	-6	-45.29	<.001

Standard error of the estimate = 25.2
Source: Authors' simulated data.

If, however, a naïve model attempting to capture growth over time assumes that ELs and non-ELs demonstrate academic content growth at the same rate, and excludes the effect of ELP, the model estimates will be biased. The results in Table C–3 summarize the effects of using the naïve model on the entire sample (i.e., both ELs and non-ELs). These results indicate substantial bias, as the linear growth term is biased by about 5 percent and the deceleration term is biased by approximately 6 percent. The standard error of the estimate (SEE) is, as expected, slightly larger than the actual standard deviation of the residual. This outcome is typical when there are omitted variables that are correlated with variables in the model and with the outcome. The impact of the omitted variable is attenuated because a vast majority of the students are not ELs (as would be the case in any state).

TABLE C–3

Results of Naïve Model for All Students

	Unstandardized Coefficients				
	B	Std. Error	Pct Bias	t	Sig.
(Constant)	-14.45	0.46		-31.69	<.001
Initial_level	51.48	0.05		1109.37	<.001
Grade	52.28	0.14	5	367.13	<.001
Grade2	-2.12	0.01	6	-198.14	<.001

Standard error of the estimate = 25.4
Source: Authors' simulated data.

A model applying the DVA approach to data for all students is given as:

$$Y_{tij} = b0 + b1Grade_{tij} - b2Grade2_{tij} - b3ELP\_IND_{tij} + b4ELPSS_{tij} + e_{tij} \quad (5)$$

In this model, $ELPSS_{tij}$ is equal to the ELP scale score for students with an ELP scale score and is equal 0 if it is missing. ELP_IND_{tij} is the DVA and equals 1 if $ELPSS_{tij}$ is missing and 0 otherwise. The results, presented in Table C–4, show that bias is eliminated for the parameter estimates for growth using the DVA model.

TABLE C–4

Results of Dummy Variable Adjustment Model for All Students

	Unstandardized Coefficients				
	B	Std. Error	Pct Bias	t	Sig.
(Constant)	-0.15	0.47		-0.31	-0.755
Initial_level	50.08	0.05		1040.73	<.001
Grade	49.96	0.14	0.00	351.11	<.001
Grade2	-2.00	0.01	0.00	-187.85	<.001
ELP_IND	-30.33	0.48		-63.81	<.001
ELPSS*EL	0.15	0.00		37.97	<.001

Standard error of the estimate = 25.
Source: Authors' simulated data.

It is difficult to predict the effect of omitted variables in more complex models.[57] Complexity is mainly due to the operational nature of state assessment data where students are enrolled in schools. The complexity arises from the additional random effects that are included in such longitudinal data structures. Besides the random residual (measurement error, e_{tij}) for student *i* at time *t* attending school *j*, there are student residuals *r0* and *r1*. These residuals are unique to each student but constant over time, with random effects in initial performance (*r0*) and growth (*r1*).[58] Capturing school effects is accomplished by introducing two additional random effects *U0* and *U1*, where *U0* represents initial random differences in academic performance among schools and *U1* represents a school's contribution to student academic growth.[59] School effects are generally considered to represent context and are not simply the aggregate of student effects.[60]

To examine the potential of alternative approaches for capturing true school effects, the data-generating Model (3) is expanded to include the appropriate random student and school effects.[61] The new model is given as:

$$Y_{tij} = \gamma_{010}*INITIAL_{ij} + \gamma_{100}*Grade_{tij} + \gamma_{200}*ELP\_IND_{tij} + \gamma_{300}*ELPSS_{tij}*Grade_{tij} + \gamma_{400}*Grade2_{tij} + r_{0ij} + r_{1ij}*Grade_{tij} + u_{00j} + u_{10j}*Grade_{tij} + e_{tij} \quad (6a)$$

Where Y_{tij} is the academic content scale score for student *i* at time *t* in school *j*, and

$$Y_{tij} = 50*INITIAL_{ij} + 50*Grade_{tij} + -80*EL\_E2_{tij} + .05*ELP\_ED2_{tij} + -2*Grade2_{tij} + r_{0ij} + r_{1ij}*Grade_{tij} + u_{00j} + u_{10j}*Grade_{tij} + e_{tij} \quad (6b)$$

$$Y_{tij} = 50*INITIAL_{ij} + 50*Grade_{tij} + -80*EL\_E2_{tij} + .05*ELP\_ED2_{tij}*Grade_{tij} + -2*Grade2_{tij} + r_{0ij} + r_{1ij}*Grade_{tij} + u_{00j} + u_{10j}*Grade_{tij} + e_{tij} \quad (6c)$$

Data are generated using both (6a) and (6b). Generated academic content scores based on (6a) represent static growth as the effect of ELP, γ_{300}, which is constant over grades,[62] while content scores based on (6b) are considered dynamic as the effect of ELP changes over grades.

Given that student and school random effects have been included in the data generation, the DVA model in (5) is mis-specified because student and school effects should be captured either through fixed or random effects. Otherwise, estimated parameters will be confounded representations of within school and between school effects on growth, and estimated standard errors will biased.[63]

57 Jee-Seon Kim and Edward W. Frees, "Omitted Variables in Multilevel Models," *Psychometrika* 71, no. 4 (2006): 659–90.

58 Stephen W. Raudenbush and Anthony S. Bryk, *Hierarchical Linear Models: Applications and Data Analysis Methods,* 2nd ed. (Newbury Park, CA: Sage Publications, 2002).

59 In this case, only the linear term for growth (grade) is subject to a random effect. It is also possible to include a random effect for deceleration, but this additional complexity likely provides no additional substantive guidance for developing alternative approaches to accurately capturing schools' contributions to student academic growth.

60 Leigh Burstein, "The Analysis of Multilevel Data in Educational Research and Evaluation," *Review of Research in Education* 8 (1980): 158–233; Raudenbush and Bryk, *Hierarchical Linear Models*.

61 In this case, data generation includes the random effects *r0* (intercept) and *r1* (Grade), which are drawn from a joint distribution with both *r0* and *r1* N(0,5) and r(*r0*,*r1*) = -.2. The random effects *U0* (intercept) and *U1* (Grade) are drawn from a joint distribution with both *U0* and *U1* N(0,5) and r(*U0*,*U1*) = -.2.

62 This does not mean that the language proficiency scores are constant over time (grades), just that the effect is constant across the grades.

63 Raudenbush and Bryk, *Hierarchical Linear Models.*

Data generation by (6a) and (6b) is crossed with three different assumptions about school random effects. The first assumption is that there is no correlation between school random effects for academic growth and ELP progress, which implies that a school's facilitation of academic growth is unrelated to that same school's facilitation of progress toward ELP. The second assumption is that there is a negative correlation between school random effects for academic growth and ELP progress, implying that schools that do a "good" job of facilitating academic growth do a "poor" job of facilitating ELP progress. The third assumption is that there is a positive correlation between school random effects for academic growth and ELP progress, which implies that schools are "good" at facilitating both academic growth and progress toward ELP. Given the additional random effects, the correlations among random effects change considerably. The correlation between true school effects for growth (i.e., a school's contribution to growth) and estimated school effects based on the DVA model and the naïve model are presented in Table C–5. The results in Table C–5 are robust to the relationship of school effects between academic growth and progress toward ELP. The results also indicate that naïve models do less well than DVA models in capturing true school effects.

TABLE C–5

Dummy Variable Adjustment and Naïve Growth Model Results

	DVA Pct Bias		Correlation with True School Effect	Naïve Bias		Correlation with True School Effect
Content/ELP Corr.	Linear	Deceleration		Linear	Deceleration	
Static effect						
None	2.0	2.0	0.92	42.0	75.0	0.66
Negative	1.8	1.5	0.92	40.0	75.0	0.59
Positive	2.0	2.0	0.91	42.0	75.0	0.69
Dynamic effect						
None	0.1	1.5	0.90	12.0	71.6	0.65
Negative	0.1	1.5	0.90	12.0	71.6	0.65
Positive	0.1	1.5	0.90	12.0	71.6	0.65

Source: Authors' simulated data.

The DVA approach also works well when applied to a growth model, which is consistent with expectations because the DVA model closely approximates the data generation process. To evaluate the robustness of the DVA approach, DVA is applied to value-added models (VAM) and student growth percentile (SGP) models. The results of including a DVA in VAM and SGP models estimating school effects are summarized in Table C–6. These results indicate that for both VAM and SGP models, there is improvement in the correlations between estimated school effects and true school effects. This finding is consistent across school academic growth/ELP progress effect relationships and whether the effect of ELP is assumed static or dynamic in its influence on content growth. Improvement ranges from about 3 percent to 16 percent.

TABLE C–6

Model Change by Including Dummy Variable Adjustment

	Percent Improvement	
	Value-Added Model	Student Growth Percentile Model
ELP Effect Is Static		
No	9.2%	4.9%
Negative	7.6%	2.7%
Positive	10.4%	6.6%
ELP Effect Is Dynamic		
No	9.3%	12.6%
Negative	12.3%	16.0%
Positive	2.8%	8.2%

Source: Authors' simulated data.

Another common option is to use gains as a basis for measuring growth. Results based on the simulated data indicate that gains capture the true school effect fairly well when multiple years are used to estimate the gains. The correlations range from about .5 to .9. The highest correlations are observed when school content effects and school ELP assessment effects are positively correlated.[64]

Technical Details: Multivariate Longitudinal Model for Academic Content Growth

A model for academic growth that simultaneously models growth on academic content and ELP assessment is given as:

$$Y_{tij} = \gamma_{000} + \gamma_{100}*Grade_{tij} + \gamma_{200}*IND_{tij} + \gamma_{300}*GradeSQR_{tij} + r_{0ij} + r_{1ij}*Grade_{tij} + u_{00j} + u_{10j}*Grade_{tij} + e_{tij} \quad (7)$$

In this model, Y_{tij} is the assessment at time *t* for student *i* attending school *j*. For example, at *Grade* equal to 3, the assessment score represents academic content and at *Grade* equal to 2.5, the assessment score represents ELP.[65] *IND* is used to keep track of which assessment is academic content and which is ELP. All other variables, parameters, and subscripts in Model (7) are the same as in Model (6). In Model (7), the linear component of growth on content is γ_{100} and the linear component on progress on ELP assessment is $\gamma_{100} + \gamma_{200}$. This approach creates only a linear offset in slopes, which could be modeled with a more sophisticated relationship. However, for accountability purposes, this may be the desired solution as it forces the academic growth estimate to be interpreted as net of progress in ELP. Additional functional forms were evaluated on the simulated data, and these did not improve the ability of the model to reproduce true school effects. That is, the researchers modeled different versions of (7) to examine whether more complex representations

64 Additional study is warranted to examine the extent to which gains reproduce accurate school effects as a function of the correlation of content and ELP assessment effects, as well as the magnitude and reliability of school effects.

65 This approach is best suited for assessment results reported on a vertical scale; however, it is possible to simulate growth results without a vertical scale if the assessment scale was developed in grade pairs or grade bands. See Pete Goldschmidt, "The Impact of English Language Proficiency Assessments Claims about Progress" (paper presented at the annual meeting of the American Education Research Association, Chicago, 2023).

of growth substantively changed interpretations about schools, and they did not meaningfully change interpretations.

Applying Model (7) to the simulated data results in robust estimates across different conditions, where conditions refer to assumptions about the relationship between schools' ability to facilitate ELs' progress toward ELP and schools' ability to foster ELs' academic growth. The correlation between estimated school effects and true school effects for academic content are approximately .90 in all instances tested. Table C–7 presents the results for the different approaches in ELA in Hawaii.

TABLE C–7

Correlations among School Effect Estimates: ELA for Students in Hawaii

	EL Count	HI_R_ DVA	HI_R_ Trad	NHI_R_ DV	NHI_R_ Tr	HI_R_ Trad	HI_R_ Stack
GTT	-.143*	.262**	.264**	.245**	.246**	.303**	.372**
EL Count	1	-0.045	-0.063	-0.032	-0.049	0.023	-.120*
HI_R_DVA	-0.045	1	.998**	.991**	.991**	.668**	.623**
HI_R_Trad	-0.063	.998**	1	.987**	.991**	.665**	.625**
NHI_R_DV	-0.032	.991**	.987**	1	.997**	.661**	.609**
NHI_R_Tr	-0.049	.991**	.991**	.997**	1	.657**	.610**
HI_R_Trad	0.023	.668**	.665**	.661**	.657**	1	.867**
HI_R_Stack	-.120*	.623**	.625**	.609**	.610**	.867**	1

* Correlation is significant at the 0.05 level (2-tailed).
** Correlation is significant at the 0.01 level (2-tailed).
Source: Author analysis of Hawaii assessment data from school years 2016–17 through 2018–19, received from the Hawaii State Department of Education in September 2023.

Results based on the simulated data indicated that the multivariate longitudinal model is best able to reproduce true school effects and is robust to assumptions about the growth process. Hence, the researchers use this model as a basis of comparison. *HI_R_Stack* includes the school effects estimated in this way, and these results appear to be the most robust across different approaches. The state's approach, however, most closely resembles the *NHI_R_Tr* approach that is based on an SGP model. Based on the results in Table C–7, the researchers observe that the relationships among the models are moderate, implying that the claims about schools will depend on the model utilized. Further, the DVA approach does not move the SGP estimates closer to the hypothesized optimal model.

Next, Table C–8 presents estimated school effects results for math in Hawaii, and the results are consistent with those presented in Table C–7. However, it is noteworthy that estimated school effects are generally more robust to model choice in math than they are in ELA. Consistent across both Tables C–7 and C–8 is that the mixed effects longitudinal models, whether multivariate or traditional, are highly correlated, indicating robustness in this approach.

TABLE C–8

Correlations among School Effect Estimates: Math for Students in Hawaii

	EL Count	HI_M_ DVA	HI_M_ Trad	NHI_M_ DV	NHI_M_ Tr	HI_M_ Trad	HI_M_ Stack
GTT	-.143*	.244**	.258**	.227**	.237**	.348**	.415**
EL Count	1	-0.012	-0.008	-0.015	-0.010	0.003	-0.069
HI_M_DVA	-0.012	1	.997**	.983**	.979**	.769**	.750**
HI_M_Trad	-0.008	.997**	1	.983**	.984**	.778**	.762**
NHI_M_DV	-0.015	.983**	.983**	1	.998**	.764**	.744**
NHI_M_Tr	-0.010	.979**	.984**	.998**	1	.768**	.751**
HI_M_Trad	0.003	.769**	.778**	.764**	.768**	1	.964**
HI_M_Stack	-0.069	.750**	.762**	.744**	.751**	.964**	1

* Correlation is significant at the 0.05 level (2-tailed).
** Correlation is significant at the 0.01 level (2-tailed).
Source: Author analysis of Hawaii assessment data from school years 2016–17 through 2018–19, received from the Hawaii State Department of Education in September 2023.

Turning to Ohio's data, the results in Table C–9 summarize the various approaches for estimating school effects in ELA in Ohio. These results indicate that model choice substantially influences claims about school effects. If *OH_R_Stack* is the model that most likely captures true school effects, the other options are quite dissimilar. Including DVA applied to VAM results (*R_DVA_S*) and to SGP results (*NR_DVA_S*) results in slight improvements.[66]

TABLE C–9

Correlations among School Effect Estimates: ELA for Students in Ohio

	R_DVA_S	R_Trad	NR_ DVA_S	NR_Trad	OH_R_ Trad	OH_R_ Stack
MET_TARGET Pct	0.040	0.041	.044*	.047*	-0.020	-0.040
R_DVA_S	1	.997**	.879**	.875**	.285**	.268**
R_Trad	.997**	1	.875**	.878**	.280**	.256**
NR_DVA_S	.879**	.875**	1	.995**	.231**	.227**
NR_Trad	.875**	.878**	.995**	1	.225**	.212**
OH_R_Trad	.285**	.280**	.231**	.225**	1	.930**
OH_R_Stack	.268**	.256**	.227**	.212**	.930**	1

* Correlation is significant at the 0.05 level (2-tailed).
** Correlation is significant at the 0.01 level (2-tailed).
Source: Author analysis of Ohio assessment data from school years 2016–17 through 2018–19, received from the Ohio Department of Education and Workforce in February 2020.

66 Given the lack of vertical scales in Ohio, additional adjustments are likely required to bring results into greater alignment.

Finally, the results in Table C–10 summarize the various approaches applied to math in Ohio. These results are consistent with those in Table C–9 with one key exception. Applying DVA to VAM (*M_DVA_S*) and SGP (*NM_DVA_S*) results significantly improves the relationship of estimated school effects based on the reference model results (*OH_M_Stack*). Consistent with results from Hawaii, the mixed effects longitudinal modeling approach is robust to either the traditional approach or the multivariate approach, although some variability remains.

TABLE C–10

Correlations among School Effect Estimates: Math for Students in Ohio

	M_ DVA_S	M_Trad	NM_DVA_S	NM_ Trad	OH_M_ Trad	OH_M_ Stack
MET_TARGET Pct	.138**	.170**	.147**	.182**	-0.015	-0.018
M_DVA_S	1	.708**	.911**	.697**	.381**	.346**
M_Trad	.708**	1	.737**	.963**	.182**	.153**
NM_DVA_S	.911**	.737**	1	.757**	.404**	.364**
NM_Trad	.697**	.963**	.757**	1	.169**	.138**
OH_M_Trad	.381**	.182**	.404**	.169**	1	.942**
OH_M_Stack	.346**	.153**	.364**	.138**	.942**	1

* Correlation is significant at the 0.05 level (2-tailed).
** Correlation is significant at the 0.01 level (2-tailed).
Source: Author analysis of Ohio assessment data from school years 2016–17 through 2018–19, received from the Ohio Department of Education and Workforce in February 2020.

About the Authors

MEGAN HOPKINS

Megan Hopkins is Associate Professor in the Department of Education Studies at the University of California, San Diego. Her research focuses on policy and leadership with a specific emphasis on the education of multilingual learners in the K-12 education system. Over the last decade, she has engaged in numerous applied research projects that examine the implementation of state policies related to bilingual education programming and the preparation of teachers of multilingual learners.

Dr. Hopkins co-leads a research-practice partnership that leverages and conducts research to build state education agency leaders' capacity to advance equity for multilingual learners in their unique contexts. She also serves as co-advisor to the English Learners Collaborative of the Council of Chief State School Officers, the largest professional association for state education agency leaders in the country. Her work has been published widely in the form of academic journal articles, books, and policy reports and briefs.

PETE GOLDSCHMIDT

Pete Goldschmidt is a Professor in the College of Education at California State University Northridge, where he teaches graduate courses in statistics, research methods, and program evaluation. His expertise includes advanced methods in quasi-experimental analyses and longitudinal modeling, which he applies to program and school evaluations, as well as state accountability.

Dr. Goldschmidt provided technical support to 35 states as they developed their state accountability plans under the *Every Student Succeeds Act* (ESSA). He serves on several state technical advisory committees that provide guidance to support state assessment and school accountability systems. He is also a principal investigator on a project funded by the U.S. Department of Education's Competitive Grants for State Assessment that is developing innovative approaches to monitoring English Learners' progress toward English language proficiency. He has published numerous articles in peer reviewed journals such as the *American Educational Research Journal*, *Educational Evaluation and Policy Analysis*, and *School Effectiveness and School Improvement*.

JULIE SUGARMAN

@julie_sugarman

Julie Sugarman is Associate Director for K-12 Education Research at the Migration Policy Institute (MPI) National Center on Immigrant Integration Policy, where she focuses on issues related to immigrant and English Learner students. Among her areas of focus: policies, funding mechanisms, and district- and school-level practices that support high-quality instructional services for these youth, as well as the particular needs of immigrant and refugee students who first enter U.S. schools at the middle and high school levels.

Dr. Sugarman came to MPI from the Center for Applied Linguistics, where she specialized in the evaluation of educational programs for language learners and in dual language/two-way immersion programs. She earned a BA in anthropology and French from Bryn Mawr College, an MA in anthropology from the University of Virginia, and a PhD in second language education and culture from the University of Maryland, College Park.

DELIA POMPA

Delia Pompa is Senior Fellow for Education Policy at MPI's National Center on Immigrant Integration Policy, where her work focuses on research and policy analysis related to improving educational services for immigrant students and English Learners.

Ms. Pompa came to MPI from the National Council of La Raza, where she was Senior Vice President for Programs, overseeing its education, health, housing, workforce development, and immigrant integration work, and where she previously served as Vice President of Education. She has had a key role in shaping federal education policy through her positions as Director of the Office of Bilingual Education and Minority Languages Affairs in the U.S. Department of Education, and as Executive Director of the National Association for Bilingual Education.

Ms. Pompa came to Washington, DC, to serve as Director of Education for the Children's Defense Fund after serving as Assistant Commissioner for Program Development at the Texas Education Agency. Her previous experience as Executive Director for Bilingual and Migrant Education in the Houston Independent School District and as a bilingual classroom teacher and instructor to prospective teachers at the graduate level has anchored her work.

LORENA MANCILLA @LaDraMancilla

Lorena Mancilla is Associate Director for K-12 Partnerships and Policy at MPI's National Center on Immigrant Integration Policy, where she oversees a range of program activities that assist stakeholders at state and local levels in understanding and addressing policy challenges and opportunities affecting English Learners and immigrant-background students.

Dr. Mancilla previously worked at WIDA at the Wisconsin Center for Education Research at the University of Wisconsin-Madison, including as Director of WIDA Early Years, and at Hope Chicago, a Chicago-based nonprofit where she served as Parent Program Director. Dr. Mancilla holds a BS in human computer interaction and a MS in education from DePaul University. She earned her PhD in curriculum and instruction at the University of Wisconsin-Madison with a concentration in multicultural education.

Acknowledgments

This report is part of a series of research on innovative assessment approaches and alternative accountability models coordinated under The K12 Research for Equity Hub (www.edudream.org/thehub). The Hub is managed by EduDream and funded by the Bill & Melinda Gates Foundation and the Walton Family Foundation. No personnel from the Bill & Melinda Gates Foundation or the Walton Family Foundation participated in the creation of Hub research. The findings and conclusions contained in this report are those of the authors and do not necessarily reflect the positions and/or policies of the Bill & Melinda Gates Foundation or the Walton Family Foundation.

The authors are grateful to the team at EduDream—Monica Martinez, Michelle Oliva, and Sophia Velez—for their support for this project, and to Sophia, Michelle, and their colleagues Gloria Corral, Lindsay Fryer, and Ellen Sherratt for their valuable feedback. This research was also made possible by data partnerships through the U.S Department of Education's Competitive Grants for State Assessment.

The authors would like to thank all the focus group and interview participants who shared valuable insights that enriched this research. They extend a particularly warm thank-you to the following partners who organized and conducted focus groups with community advocates and EL family members: Xilonin Cruz-Gonzalez and Shelly Spiegel-Coleman of Californians Together; Tara Lentz and Maria Paula Zapata of Conexión Américas; Morgan Craven, Aurelio Montemayor, and Lizdelia Piñón of the Intercultural Development Research Association (IDRA); Gilda (Gigi) Pedraza of the Latino Community Fund; and Andrea Ortiz, Liza Schwartzwald, and Kim Sykes of the New York Immigration Coalition. Finally, the authors thank Lauren Shaw for her expert editing.

The Migration Policy Institute (MPI) is an independent, nonpartisan policy research organization that adheres to the highest standard of rigor and integrity in its work. All analysis, recommendations, and policy ideas advanced by MPI are solely determined by its researchers.

Design: Sara Staedicke, MPI
Layout: Liz Hall
Cover Photo: Allison Shelley/EDUimages

DOI: https://doi.org/10.62137/NYYA6627

Suggested citation: Hopkins, Megan, Pete Goldschmidt, Julie Sugarman, Delia Pompa, and Lorena Mancilla. 2024. *Refining State Accountability Systems for English Learner Success*. Washington, DC: Migration Policy Institute.

The Migration Policy Institute is an independent, nonpartisan think tank that seeks to improve immigration and integration policies through authoritative research and analysis, opportunities for learning and dialogue, and the development of new ideas to address complex policy questions.

www.migrationpolicy.org

1275 K St. NW, Suite 800, Washington, DC 20005
202-266-1940