Handbook for Developing and Monitoring the English Language Proficiency Indicator and English Learner Progress

Pete Goldschmidt, Ph.D.

California State University Northridge

February 2018

Suggested Citation: Goldschmidt, P. (2018*). Handbook for Developing and Monitoring the English Language Proficiency Indicator and English Learner Progress.* Washington DC: Council of Chief State School Officers.

One Massachusetts Avenue, NW, Suite 700 • Washington, DC 20001-1431
Phone (202) 336-7000 • Fax (202) 408-8072 • www.ccsso.org

Acknowledgement

This Handbook is substantively based on facets of the state ESSA ELP Indicator technical support facilitated by CCSSO and conducted by me, Kenji Hakuta, and Delia Pompa. The Handbook was expertly guided to completion by Fen Chou of CCSSO. Important suggestions were offered by many state Title I, Title III, and Accountability Directors attending the technical assistance meetings in Chicago and Denver. Also, a select group of peer reviewers provided in depth feedback and comments that significantly improved the Handbook. The peers consisted of:

Name	State/Affiliation
Samuel Aguirre	Ohio
Migdalia Arthurton	US Virgin Islands
Kilchan Choi	CRESST/UCLA
Shawn Cockrum	Missouri
Walt Drane	Mississippi
Michael Flicek	Wyoming
Anna Furniss	Mississippi
Matt Goodlaw	New Mexico
Tricia Kerr	Arkansas
Evan Kramer	Tennessee
Audrey Lesondak	Wisconsin
Tammy Mckeown	Arizona
Ann-Michelle Neal	Utah
Jose Perez	US Virgin Islands
Siaosi Siaosi	US Virgin Islands
Chandi Wagner	DC
Kerri Whipple	North Dakota
Kate Wright	Arizona

All errors and omission are the sole responsibility of the author.

Contents

List of Tables

List of Figures

Introduction

Around 5 million children in U.S. schools (about one in ten students) are learning English as a new language. English Learners (ELs) represent the fastest-growing student group nationwide. More than ever, education leaders must design policies, EL programs, and instructional practices to give every EL learner the best opportunity possible for strong English language development (ELD) and for academic achievement.

A key change in the Every Student Succeeds Act (ESSA) is that assessment and accountability of ELs moved from Title III to Title I must be included in the state's overall accountability system. The progress of EL students in achieving English language proficiency is one of the five accountability indicators, so the topic of how to meaningfully incorporate EL progress into school-level accountability has received considerable attention among states.

Throughout 2016 and 2017 states have been making a significant effort to meaningfully incorporate the ELP Indicator into state accountability systems. Implementing school-based accountability for this particular facet of the education system has several tradeoffs, many of which are addressed in Goldschmidt & Hakuta (2016). Aside from the particular challenges and benefits for state accountability models, several positive externalities result from this new ESSA requirement: greater interaction and inclusion of state Title I, Title III, and Accountability leads in Title III, accountability, data/reporting, and policy areas; the use of state extant data to develop the ELP Indicator; and meaningful consideration of stakeholder input. Particularly, data-based decision making—a term often used in discussions of setting education policy—was an integral part of the technical assistance provided to states by the Council of Chief State School Officers (CCSSO) for developing the ELP Indicator. This guide builds upon that technical assistance, focusing on the use of data to help inform policy decisions, and it is intended to be a resource that state staff in Title I, Title III, and accountability can use now and in the future while implementing their ESSA plans.

The literature identifies several uses for school-level indicators. These indicators can provide descriptive information about a particular process, form the basis for holding schools accountable, generate results that help states examine program effectiveness, and provide information that can lead to appropriately targeted interventions (Ogawa and Collem, 2002). Hence, this Handbook is organized into three broad sections: the first describes procedures for **developing an ELP** (English Language Proficiency) **Indicator** that can be incorporated into a statewide accountability system; the second section describes procedures for **monitoring functionality** of the ELP Indicator within the

accountability system and its ability to provide useful descriptive information; and the third and final section introduces procedures for **evaluating the impact** of EL programs on school and student success. Each section contains tasks or procedures along with a brief overview of their relevance, and where necessary, an example and brief interpretation is provided (e.g. a table or chart). Procedures in each section should neither be considered necessary nor sufficient to guarantee an infallible indicator or EL program success.

This Handbook focuses on the use of data to assist a State Education Agency (SEA) in making an informed decision, but it is not suggesting that these analyses occur without values, intentions, and constraints brought about by state specific context – such as assessments used, stakeholder input, or a specific Theory of Action. The guiding questions presented in Goldschmidt and Hakuta (2016) are repeated below, and it should be clear that the intent of this Handbook is to use state data to directly or indirectly support decisions about these guiding questions.

It is important to emphasize that a high quality indicator does not exist in a vacuum, and its implementation within the system ultimately dictates its usefulness. Often, overall results implying that an indicator is not working as intended is based on confounded evidence about the indicator and the system in which it operates. This relates to both claims (Kane, 2006) and collecting validity evidence (Messick, 1995).

At the conclusion of each section there is a "CHECK," which is designed to provide analysts with a quick summary of what the takeaway should be from the section.

Part I: Considerations for Developing an ELP Indicator

There are numerous factors to consider when a State Education Agency (SEA) develops an indicator. Two steps need to be undertaken before the analytical data-informed element is incorporated. Steps one and two are likely iterative: 1) develop a theory of action (ToA) and 2) gather stakeholder input. SEAs generally have good processes in place to complete these steps. This handbook focuses on using analytical tools to inform policy, and does not address developing a ToA or engaging stakeholders.

There is considerable literature on using multiple measures for monitoring or accountability purposes (OECD, 2004). Specifically, the past two reauthorizations of the Elementary and Secondary Education Act (ESEA, 1965) have increasingly considered the use of multiple measures. This is often driven by the desire to improve reliability and validity (Baker, 2003) and to bring emphasis to a specific aspect of school process such as a school's ability to facilitate English language development; or as Baker (2003) notes, increasing the breadth of potential claims about a school.

While ESSA continues to call for the use of multiple measures as a more meaningful way to differentiate schools, there is no universal taxonomy of multiple measures that affords an efficient presentation and discussion of developing, monitoring, evaluating, and using the multiple measures to evaluate school success. Although this Handbook focuses on the ELP Indicator, it is also useful to develop a taxonomy that places various elements into context. This is important because state accountability systems consist of "multiple measures" but must include an EL "indicator." Also, some "measures" consist of several "measures" – e.g. the school quality measure might include chronic absenteeism and a student survey.

The following definitions can be used as a starting point:

> **English Learner (EL)** – (specifics vary by state): A student whose home language is other than English and who does not meet a state's English language proficiency standard on the English language proficiency screening assessment.
>
> **On Track:** In this Handbook, On Track refers to EL students who are within the identified timeframe to exit EL status. On Track refers to time in program, not whether the amount of progress made was sufficient to meet a specified target. For example, if the expected time–five years–is set for a student whose initial ELP level is "2" to exit EL status, and that student has not yet reached his or her fifth program year, the student would be considered On Track.

Off Track: In this Handbook, Off Track refers to an EL who is still receiving program services beyond the expected timeframe for exiting EL status. For example, if the expected time–five years–is set for a student whose initial ELP level is "2" to exit EL status, and that student has surpassed his or her fifth program year, the student would be considered Off Track. This is distinct from a student who has not demonstrated sufficient progress in a particular year.

Long Term EL: A Long Term EL has received EL services for more than five years. A student can be Off Track and not be a Long Term EL, but a Long Term EL is always Off Track.

Indicator[1]**:** Measurable variable used as a representation of an associated or related (but non-measured or non-measurable) factor(s) or quantity/quantities. For example, the ESSA School Quality Indicator is not measured directly, rather it is an aggregation of several measures that are directly measured (e.g. student and teacher attendance).

Factor[2]**:** A numerical expression of a value.

Measure[3]**:** A number or quantity that records a directly observable value or performance. All measures have a unit attached to them: inch, centimeter, dollar, etc.

Variable[4]**:** A quantity that can have any one of a set of values (something that varies).

To evaluate a composite indicator, the relationship and functionality of all indicators in the composite must be considered. Similarly, examining an indicator with multiple measures requires studying the relationship and functioning of the measures in the indicator. Evaluating the properties of an individual measure generally relies on a psychometric-centered approach which does not consider the use of the measure in the composite (see below). Keep in mind, however, that consequences and intended use are important elements of validity evidence (Messick, 1995, Kane, 2006, Herman, Heritage, & Goldschmidt, 2011).

[1] Based on: http://www.businessdictionary.com/definition.html
[2] Based on: http://www.businessdictionary.com/definition.html
[3] Based on: http://www.businessdictionary.com/definition.html
[4] Based on: http://www.businessdictionary.com/definition.html

The indicator may simply be a manipulation of the state English Language Proficiency Assessment (ELPA) – in which case much of the technical evaluation of the indicator is the psychometric properties of the ELPA. However, using an ELP Indicator as a basis for making claims requires that those claims are understood. As noted (Baker, 2013), the way an indicator is used necessitates that states emphasize different elements of validity evidence. For example, if the intent of an indicator is primarily as additional information that provides a broader picture of what schools are accomplishing vs. a deeper understanding of a particular process, then different validity evidence such as the correlation among indicators would be utilized differently. Also, if aggregate performance is intended to provide meaningful disaggregated results from which curricular and instructional guidance is gleaned, validity evidence would focus on whether disaggregated scores allow for the intended claims about individual students and the relationship to curriculum and instruction.

It is important to consider the intended level of aggregation vs. the desired level (Henderson-Montero, Julian, Yen, 2003). If the primary intent is to create a broader measure of school performance, then this requires considering carefully how technical aspects of the ELP assessments relate to other assessments in the system. If the primary intent is to provide an aggregate evaluation of individual student performance on a specific facet of school process, then states must look at what the aggregation of student scores to school mean scores indicate and also what school mean scores imply about individual student scores (e.g. English language development.) For example, the school aggregate of a student survey about school climate may not have a disaggregated analog so the reliability (internal consistency) is less important than it is for an ELPA because ELPA results are utilized at the student level to determine whether a student is English language proficient (or making sufficient progress towards proficiency[5]).

It may appear that an ELP Indicator based solely on the state ELPA is a good choice. The ELPA has gone through extensive vetting and is psychometrically sound, and thus any aggregation will maintain the technical soundness. Unfortunately, this is not strictly true because claims based on aggregates suffer from "ecological fallacy" and the introduction of confounding causes. Ecological fallacy exists when stakeholders attempt to make claims about individual students based on aggregate data – even when the aggregation is based on those same individuals. For example, analyses of school aggregates might reveal that the percent of students who are left-handed at a school is highly correlated with the percent of students who are proficient at a school; which suggests that left-handed students are more likely to be proficient. However, these results do not indicate

[5] This also reflects that that ELPA results have higher consequences which requires more attention to the psychometric properties of the assessment.

who specifically is proficient. It may well be that as the percentage of students who are left-handed increases, right-handed students are motivated to perform better to challenge the notion that left-handed students are smarter. Importantly, a policy decision based on the "evidence" that right-handed students are systematically less likely to be proficient, and therefore need an intervention to catch up, results in an ineffective intervention that addresses incorrect students. Another challenge to the notion that left-handed students are more likely to be proficient relates to potential confounding factors. For example, it may be that left-handed students attend more advantaged schools, and this explains why the percent proficient increases with the percent of left-handed students enrolled[6]. Some aggregation procedures are more robust to confounding factors than others (Choi, Goldschmidt, & Yamashiro, 2005, Wilms & Raudenbush, 1989; Aiken and Longford, 1986).

If the intended claim based on the ELP Indicator is that a school with a high score is facilitating better English language development than a school with a lower score on the indicator, then a simple aggregate of ELPA scores of students in the school is obviously confounded with the initial English language proficiency with which EL students entered. A measure of change or growth is generally a better measure because it represents what the school has contributed to EL student language development (Goldschmidt & Hakuta, 2016; Choi, et. al., 2005).

The following questions are reproduced from Goldschmidt & Hakuta (2016) and address the major considerations in developing and incorporating an ELP Indicator into a state accountability system.

A. What are my state's expectations about English language proficiency development with respect to:
    1. ELP standards?
    2. Trajectory of development?
    3. Time to proficiency?
    4. Reclassification?
    5. Individual student factors that influence growth?
    6. Instructional program factors that influence time to proficiency?

[6] Potential confounding factors can occur with disaggregated data as well, but do not result in Simpson's Paradox – which is that the relationship between two variables is changed when data are aggregated over a confounding variable.

B. Which models should my state consider for the ELP indicator? What tools do I need to effectively communicate these considerations with LEAs, schools, and stakeholders?

C. Which factors should be considered in making a selection? Am I concerned with:
    1. Familiarity to stakeholders?
    2. Transparency of the model?
    3. Sensitivity to meaningful variation (not losing meaningful variation between students, between schools, between years)?
    4. Ability to take initial ELP level, time to proficiency, and other variables into account?
    5. Ability to optimize N-size (e.g. address reliability/stability of results while minimizing loss of schools that do not meet minimum N-size)?
    6. "Fairness" across grade bands (elementary, middle, high)?
    7. Year-to-year stability of the model in enabling the state's accountability goals?
    8. Model consistency with the state's academic achievement indicator approach?

D. What is my state accountability system trying to accomplish by including ELP as an indicator receiving substantial weight?

E. What are my state's considerations in choosing N-size? Are we concerned with:
    1. Percent of schools with ELs that are included or excluded from accountability for ELP?
    2. Number of years after reclassification that exited EL students can be included in the academic achievement subgroup (allowable for up to four years)?
    3. Discrepancy between ELP and academic achievement N-sizes that might occur as a result of decisions about including reclassified EL students (i.e. E2)?

F. How do I know if some schools are doing a better job with EL students than other schools? How can the new accountability system help me in determining this?

G. What kind of data modeling will my state consider in moving forward to include ELs in our plan?

The Handbook focuses on question G as a means of directly or indirectly addressing questions A-F. Questions A, B, and C are mainly addressed in Part I; questions D and E are addressed in Part II; and, questions F and A6 are addressed in Part III. The Handbook provides examples and many different approaches to examine a particular issue. Many of the procedures have been previously identified elsewhere and references can be used to find additional detail.

Two important elements in developing an ELP Indicator are beyond the scope of this Handbook: developing a Theory of Action (ToA) and setting ELP standards. It can be the case that a ToA does not directly relate to the indicators (nor help inform their development).

The Dataset

This Handbook uses a dataset that consists of two years of matched data in which each student will correspond with one row of data. Each row consists of both current year and prior year information[7]. The variables required are detailed in Appendix A. Figure 1 presents a sample of the data layout showing nine students and a few variables.

Student	School	District	Grade	ELPA Scale Score	Initial ELD Level	Current Year ELD Level	Prior Year ELD Level	ELPA Listening	ELPA Speaking	ELPA Reading	ELPA Writing
1	A	D	4	628	1	2.8	2	2	4	2	1
2	A	D	12	699	1	4.5	4.3	3	4	3	3
3	A	D	1	655	3	4	3.5	2	3	2	2
4	A	D	1	692	4	4.7	4.2	3	4	3	2
5	A	D	1	663	4	4	4.2	3	3	3	2
6	A	D	2	669	2	4.3	3	3	3	3	2
7	A	D	4	628	1	2.8	1.3	2	2	1	2
8	A	D	4	721	1	5.5	4.7	3	4	4	3
9	A	D	1	638	4	3.3	4.3	2	3	2	2

FIGURE 1: SAMPLE DATA STRUCTURE

In Figure 1, student 1 attends school *A* in district *D*. This student is in 4th grade (or just completed 4th grade – depending when the data are pulled). Student 1 scored 628 on the ELPA assessment that was administered in the winter while the student was in 4th grade. Student 1's initial[8] ELD level was 1, the current level is 2.8, and the prior level was 2. The

[7] As more years of data become available, it is often more efficient to create a "long" file wherein a student appears in multiple rows in the file – where each row corresponds with a student/year combination.

[8] A student's initial ELD level is based on the first result using the annual English Language Proficiency Assessment (ELPA) and not the screener used to determine program eligibility. There are two reasons a screener is not a good choice as a basis for establishing progress requirements: 1) screeners are generally intended to categorize students as English proficient or not and do not have sufficient precision to establish specific non-English proficient classifications; and 2) using the screener for target setting places

final 4 columns for student 1 indicate the domain performance level scores. Not displayed in Figure 1, but essential for analyses is that the dataset also includes the number of years in the program (which could be the number of assessments - excluding the screener - a student completed. As noted, appendix A details which variables are useful for a dataset supporting the analyses described in this handbook. It is important to note that the ELPA results for both the prior year (2014–2015 school year) and the current year (2015–2016 school year) are in the same row for each student; as well as the initial ELPA score or level. It is also important to note that the dataset includes EL, EO (English only) and REL (reclassified EL) students.

The dataset includes all students from the current year (2015–2016) as well as ELs who have an ELPA score in 2014–2015, whether or not these ELs have an ELPA score in 2015-16. Specific details about the dataset are presented below.

Details about the ELPA are not critical because consistent with most ELPAs used in the US, the sample ELPA used provides scale scores as well as performance levels for a composite and scores in the four domains of listening, speaking, reading, and writing. The handbook does not rely on whether the scale is vertically equated (across all grades or within grade bands) for any analyses. For the purpose of school accountability, a vertical scale to measure growth is not a necessary condition (Goldschmidt, Choi, Martinez, and Novak, 2010).

State Context

Basic demographic and geographic information about ELs provides the basis for considering the more complex issues needed to develop, monitor, and evaluate schools on their success in providing education to ELs' English language development. The basic descriptive information about ELs in the state may include the following:

Number of ELs. This provides stakeholders context as to the scope of EL issues. The number of ELs should not be construed to imply significance – rather how services might be organized and how this relates to accountability, monitoring, and providing support. This is further clarified by disaggregating and aggregating results in various ways. Explicitly examining the number of ELs in the state is also an opportunity to examine data integrity. For example, students who are ELs may or may not be coded correctly as being EL and then they may or may not have ELP assessment information. Table 1[9] presents the count of EL students.

higher stakes on the assessment that may be detrimental to the screener's primary goal of establishing whether the student is proficient in English.

[9] Data throughout this handbook are based on an anonymous state.

TABLE 1: OVERALL STATE CONTEXT

	N	Percent
State Total K-12 population	291,693	
ELs (denoted by valid ELPA result)	21,286	7.3%

ELs as a percent of all students. This provides ELs in context, beyond simply considering the number of students. However, looking at the state average does not identify potential patterns that may be important and relevant to school-based accountability systems. For example, what is the number and percent of ELs by district, school, regional support center, or other important geographic feature? What is the number and percent of ELs in schools where there are sufficient ELs to meet the state's minimum N size requirement? How many schools have at least one EL? It is useful to maintain a chart to track how many ELs and schools are included in the accountability system at different N-sizes.

The results in Table 2 indicate that although ELs represent about 7.3% of students statewide, if only those schools meeting the minimum N are considered, then the average percent of ELs in a qualifying school increases to 12.4% and 14.6% for minimum Ns of 10 and 20, respectively. This result can inform how to weigh the ELP Indicator (discussed in more detail below).

TABLE 2: PERCENT ELS IN SCHOOLS MEETING MINIMUM N

	Minimum N = 10			Minimum N = 20		
School Level	Percent	N	S.D.	Percent	N	S.D.
K-2	14.3	11	16.1	15.3	10	16.6
Elementary	14.0	105	11.0	16.5	83	11.1
Middle	12.7	74	9.7	14.8	59	9.5
High	9.4	77	9.9	11.5	59	10.3
Statewide	**12.4**	267	10.7	**14.6**	211	10.9

The results in Table 3 indicate that although a large proportion of schools are excluded when setting minimum Ns, most ELs are included in the accountability system. Additionally, the analyst can compute the percent of schools with at least one EL included in the accountability system. For example, using results available in Table 3, a minimum of 10 corresponds to 65.8% (267/[665-259]) of schools, that have at least one EL, included in the accountability system.

TABLE 3: NUMBER AND PERCENT OF STUDENTS INCLUDED BY MINIMUM N

	Total N		Included N		Included Pct	
	Schools	ELs	Schools	ELs	Schools	ELs
All	665	21,286				
No ELs	259	0			38.9%	0.0%
Minimum = 10			267	20,745	40.2%	97.5%
Minimum = 20			211	19,968	31.7%	93.8%

Grade and ELD level at entry. The next level of descriptive analysis focuses specifically on the distribution of initial English language development and when students enter the state school system. Figure 2 includes only students who are ELs in the current year[10]; the underlying data table for Figure 2 is displayed in appendix B. Results for these analyses are useful for developing an accountability model and informing discussions with stakeholders. This chart is particularly useful in supporting conversations related to the time it takes ELs to exit EL status (or reclassification), and different challenges facing different school levels (i.e. elementary, middle, and high). This chart also provides a basis for discussing the extent to which ELs enter in higher grades and how the time to exit will be consistent with the graduation timeline.

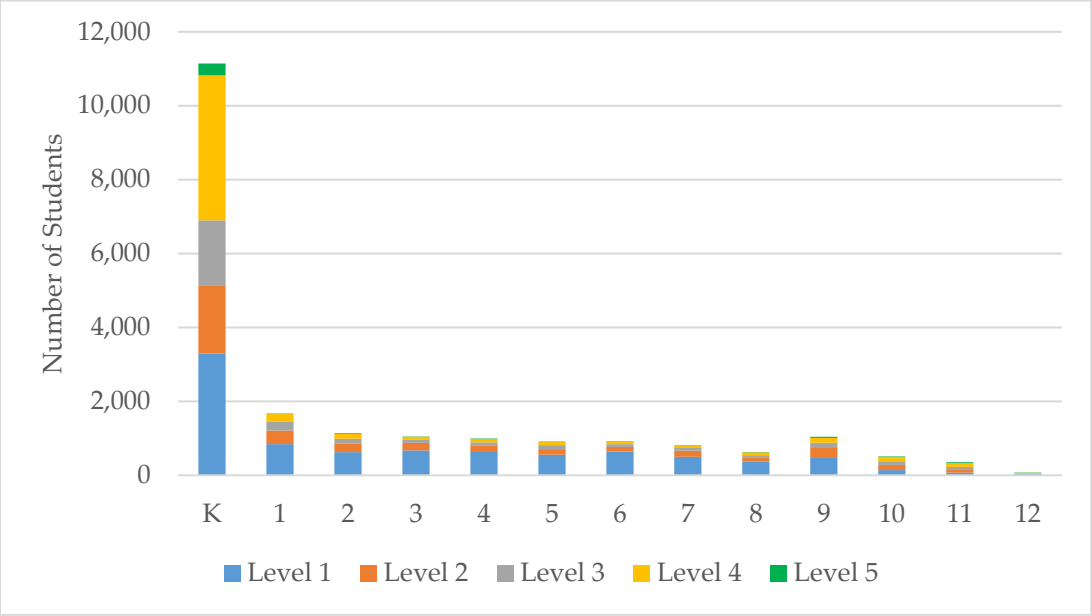


FIGURE 2: INITIAL ELD LEVEL BY INITIAL GRADE

[10] The dataset includes students who have either current and/or previous year's results, but in many instances only current ELs are included in analyses because this set of students are the students used to calculate the ELP Indicator.

The results in Figure 2 indicate that approximately 66% of ELs enter in grades K-2. In most states, the majority of ELs entering in high school are placed into 9th grade (generally related to credits earned). Figure 2 also indicates that there is considerable variability in the ELD level in Kindergarten; this variability decreases over time. A majority of students are at level 1 in later grades.

A state would want to consider both the absolute numbers as well as the distribution over time when developing an ELP Indicator, and consider the pattern as it relates to construction of the indicator and scores on the indicator across grades.

Time in program by grade. Figure 3 displays the number of years in the EL program by grade, further highlighting how the ELP Indicator might be impacted by school level (elementary, middle, high). Addressing specific aspects about the amount of time an EL is in a program are discussed in more detail in the section *Translating Growth Model Results into Individual and School Scores*. The underlying data for constructing Figure 3 are presented in Appendix C. Figure 3 clearly demonstrates that in middle school, we see the lowest concentration of ELs in program years 1 through 3, and the highest concentration in years 7 through 9 (ELs who, in most states, should have exited before program year 7). Given what is known about language development patterns (language acquisition tends to happen rapidly in early years, and then slow over time), middle schools will be most adversely impacted by an ELP Indicator based on linear progress expectations.

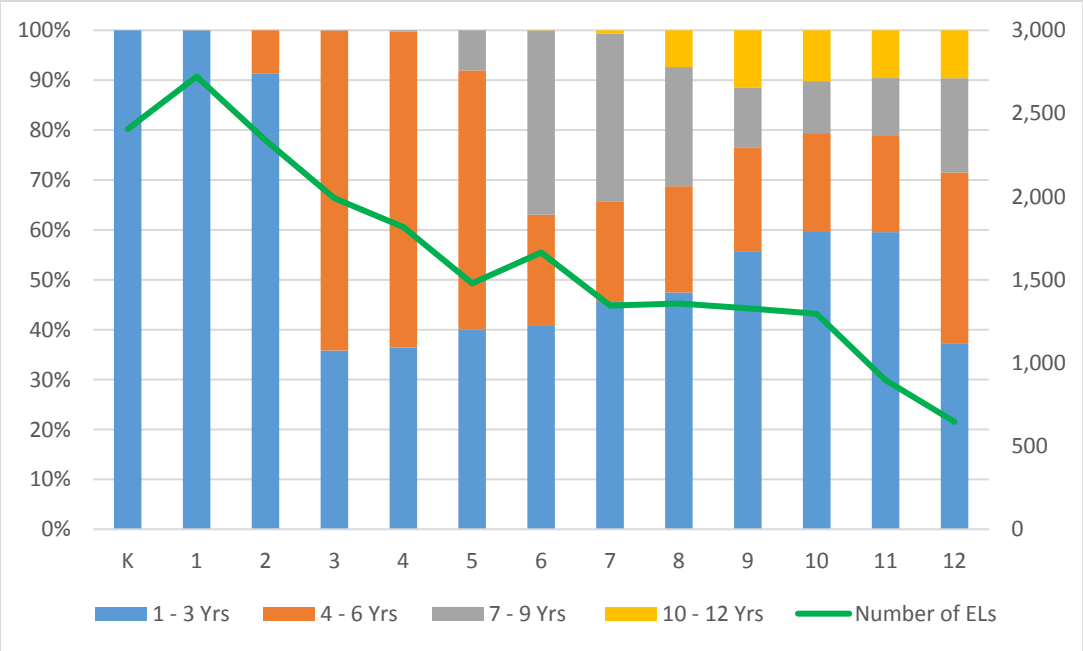


FIGURE 3: DISTRIBUTION OF TIME IN EL PROGRAM BY GRADE

Exit Criteria

The following analyses are somewhat tautological in that exit criteria are needed both to determine the proportion of ELs exiting on time and making progress towards exiting and in modeling the effect of setting the combination of exit criteria, timeframe, and annual progress goals to meet the exit criteria.

Setting exit criteria is a necessary step in developing the ELP indicator, and the Handbook summarizes several approaches as well as provides references with additional extensive guidance. Exit criteria should focus on appropriate academic language requirements necessary to allow students to meaningfully participate in instruction in English, and to have sufficient access to academic content assessments that provide valid results showing what students know and can do (Abedi, 2008: Messick, 1995). This does not imply that keeping students in programs indefinitely is a sound strategy as students begin to lose important opportunities to learn. Many states are finding that with new, more rigorous standards and assessments that exit criteria are lower (in numeric value) than they were previously. For example, some WIDA states have changed the exit criteria from 5.0 to 4.5. In setting exit criteria the state should rely on stakeholder input, English language standards, standards setting, and empirical evidence such as performance on the ELPA, and performance and progress on academic content assessments. When considering the relationship between ELPA performance and ELA, it is useful to examine not only levels of performance, but also the relationship between ELPA performance and growth in ELA. From an accountability perspective this is important because most state systems use some form of ELA growth in the accountability model. An excellent guide to using data for setting exit criteria can be found in Cook, H. G., Linquanti, R., Chinen, M., & Jung, H. (2012), which is available at: https://www2.ed.gov/rschstat/eval/title-iii/implementation-supplemental-report.pdf.

Three empirical methods that can inform exit criteria[11] are:

- Decision consistency analysis
- Logistic regression analysis
- Descriptive box plot analysis (Cook, et. al., 2012)

Decision consistency is the process of maximizing consistent decisions using the ELPA and another criterion (e.g. state ELA assessment) by varying the ELPA exit cut score and examining how this impacts consistent classifications of students on both assessments

[11] In this case, exit criteria refer to setting the ELPA cuts only. Some researchers and EL advocates recommend using an additional (state-wide, uniform) procedure for making exit determinations from the program as a whole.

(Cook, et. al., 2012). Consistent decisions include classifying a student as either pass/pass or not pass/not pass on both criterion, respectively. Figure 4 displays the decision consistency analysis. Because there is no peak from which consistency begins to decrease, states can examine where the consistency gains begin to substantially decrease (i.e. the slope starts to change). In this case there is little gain in consistency beyond Level 5. The data and calculations for Figure 4 are displayed in Appendix D.

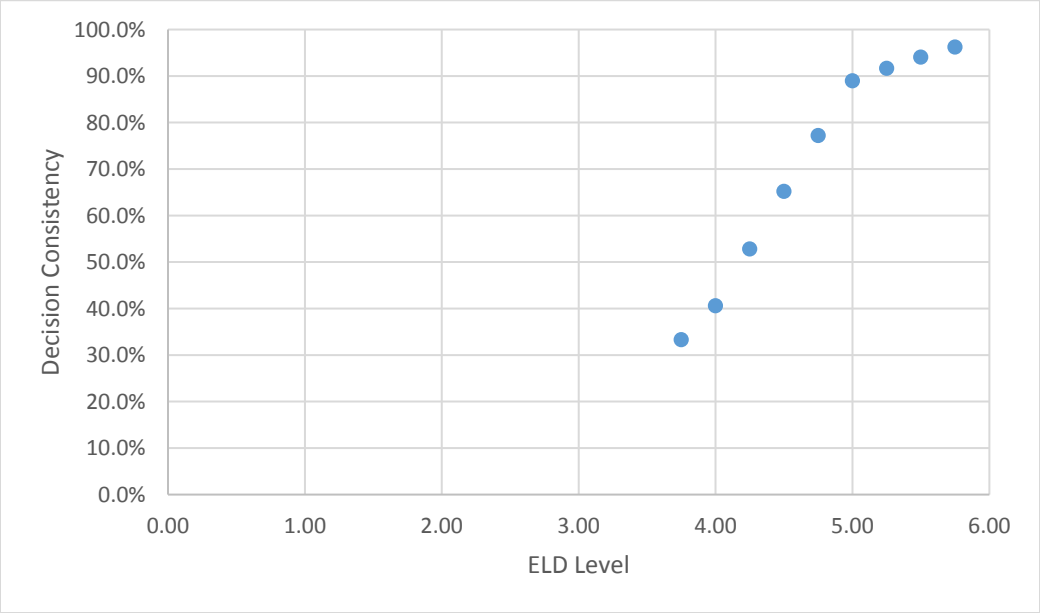


FIGURE 4: DECISION CONSISTENCY PLOTTED OVER ELD LEVELS

Decision consistency can also be applied where the desired exit criteria are based on conjunctive or profile scores, and the state is examining alternatives for monitoring progress. For example, an EL's exit criteria may follow conjunctive rules for four domains, but monitoring progress may be based on an equally weighted sum of the four domain scores[12]. The emphasis is on developing a single score and exit criteria that more easily monitors progress towards English language proficiency for the purposes of the accountability system.

[12] Assuming true weights are unknown and in the absence of purposeful policy weights, equal weighting results in minimum error (in terms of the weights). Two methods states can use to establish weights are: 1) regressing the ELPA scale score on the four domain performance levels, or 2) a logistic regression with the outcome equal to whether or not a student is proficient on the ELA assessment on the four domain performance levels. The latter approach may provide unstable estimates if the percent proficient for ELs is very low. The former approach with the example dataset used in this report results in weights equal to: Listening = .23; Speaking = .27; Reading = .24; Writing = .26. These are quite close to equal weights. The R^2 for the regression model is .94, indicating that the four domains' performance levels account for the variation in the overall ELPA composite scale score.

The following examples use three scores based on the ELPA: domain performance levels (1 to 4); the sum of the four domain performance levels (4 to 16); and the ELPA composite scale score[13]. Figure 5 displays the distribution of scores for a single domain (*listening*) across all possible domain sum scores.

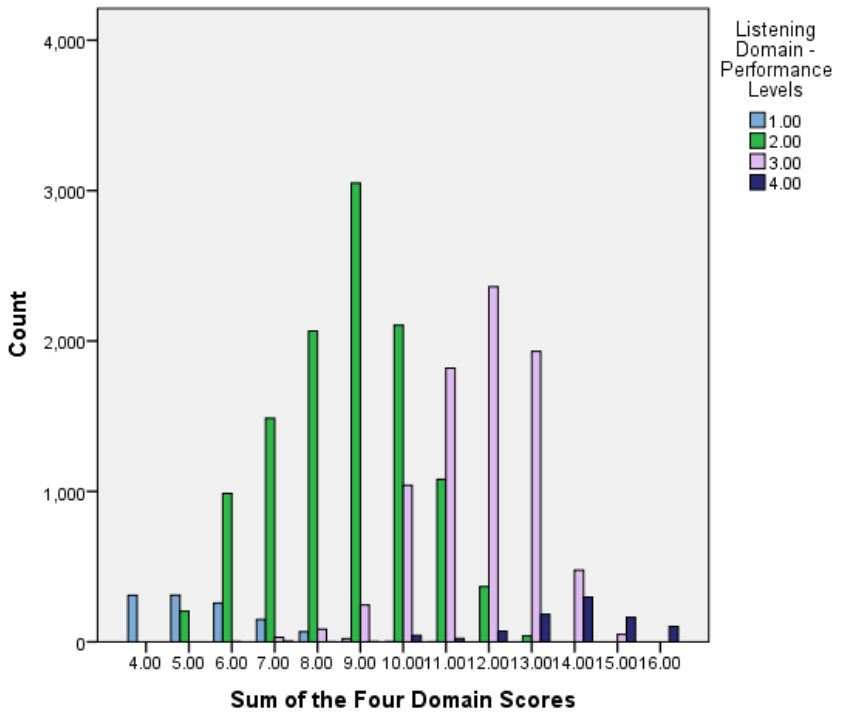


FIGURE 5: DISTRIBUTION OF LISTENING PERFORMANCE ACROSS DOMAIN-SUM SCORES

Figure 5 in conjunction with a similar chart for each of the other three domains provides a basis for setting a cut for the domain sum score, or to determine whether the sum of the four domains is reasonable. For example, if the domain sum score cut were 13, it is very unlikely that an EL would score as low as level 2 on the *listening* domain. Figure 5 can also be used along with similar figures for each domain to set disjunctive rules: for example, a student could be deemed to have met growth if they scored either a 14 or scored at least a four on one domain. If the ELP assessment consisted of only listening and a composite, then students who meet the sum score cut could score as low as 3 on listening and still be considered as having made growth. Students who meet the listening domain cut (4) could score as low as 10 on the composite and still be considered as having made growth. A disjunctive rule with domain scores is likely not applicable. However,

[13] Each state or consortium likely has a different relationship among these three measures.

an application of a disjunctive rule, applied by several states, is whether a student met the growth target, the level target, or exited.

Figure 6 produces results similar to the decision consistency results in Cook et. al., (2012) with the added dimension of two cuts for the *listening* domain. In this case, the consistency is within the ELPA and not between the ELPA and the ELA and/or Math assessment results. The analysis underlying Figure 6 supports the more traditional analysis presented above (consistency between ELPA and ELA) and can be used as a preliminary step. Figure 6 presents the decision consistency between setting the Listening domain passing score at level 3 or 4 and its alignment with the overall domain sum score. That is, if the domain sum cut is set at 12 and the Listening conjunctive cut score is set to 4, 75% of classifications would be consistent, whereas if the Listening conjunctive cut score is set to 3, 82% of classifications would be consistent. The data underlying Figures 5 and 6 are presented in Appendix F.

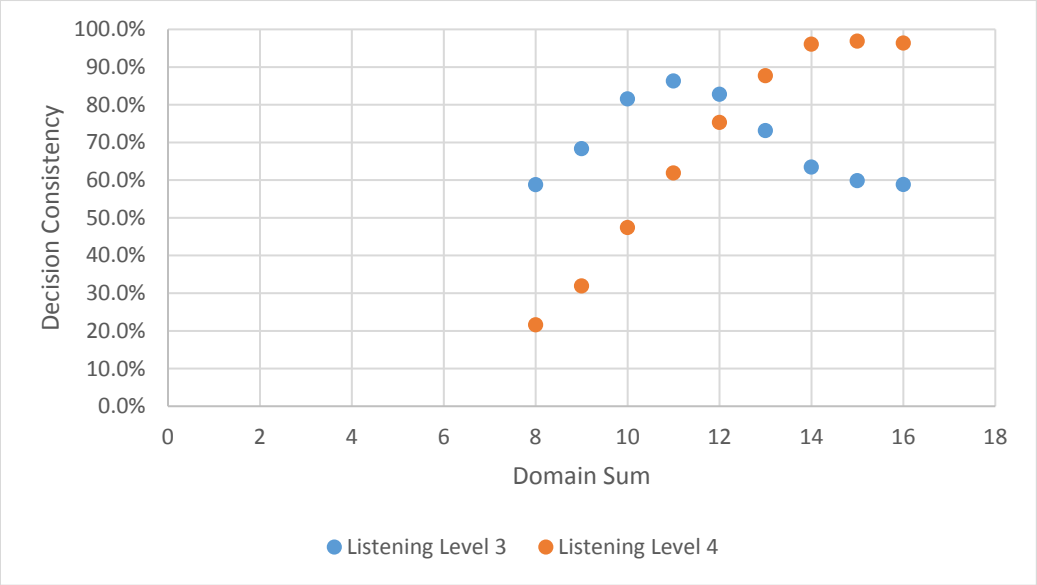


FIGURE 6: DECISION CONSISTENCY BETWEEN THE LISTENING DOMAIN AND DOMAIN SUM SCORE

Figure 7 reproduces results from a logistic regression model of the probability of being proficient in English language arts (ELA) on the ELPA scale score. Cook et. al., (2012) highlight where a student would have equal chances of being proficient or not. This is the scale score associated with a 0.5 probability of proficient on ELA, which is where the Ogive crosses the horizontal (gray) line. This corresponds with a scale score of approximately 740. Another consideration is the probability of meeting ELA proficiency for English Only (EO) students. In this case, the probability is about .40. Given that ELs

should not be held to higher standards than their EO classmates, this second demarcation provides additional insight for setting EL exit criteria. In Figure 7 this corresponds to a scale score of approximately 725.

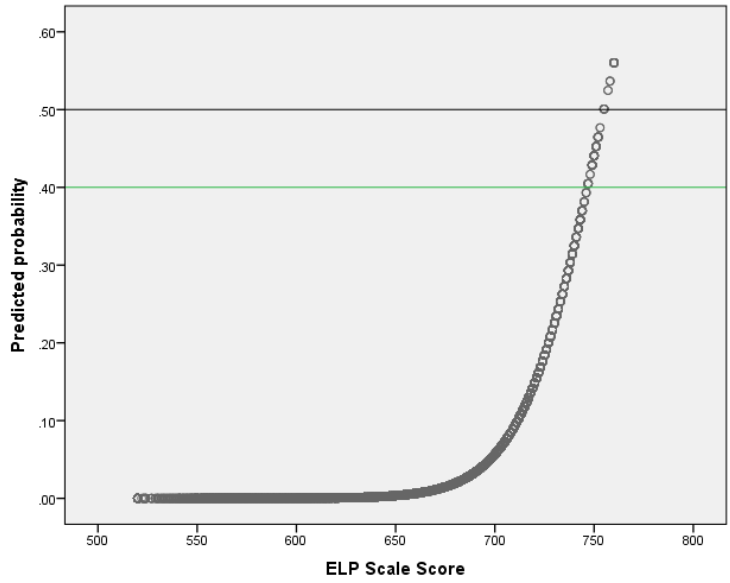


FIGURE 7: PROBABILITY OF MEETING ELA PROFICIENCY – ELEMENTARY GRADES

The logistic model results presented in Figure 7 are created using the code displayed in appendix N.

Figure 8 presents a box plot that demonstrates median performance by ELD level and for EO students. Each box in Figure 8 depicts scores that range from the 25th percentile to the 75th percentile, where each box is consists of three horizontal lines. The middle line (although not necessarily equal distant between the top and bottom line) represents the median score (50th percentile). The bottom line on each box represents the 25th percentile and the top line represents the 75th percentile. It is also useful to indicate the proficiency cut so that EO student median performance can be examined relative to the proficiency cut. ELs at ELD level 5 are performing somewhat lower than EO students, but the range of performance falls within EO performance range. In contrast, ELs at ELD level 4 are performing substantively below EO students (the median performance lies below the 25th percentile of EO students).

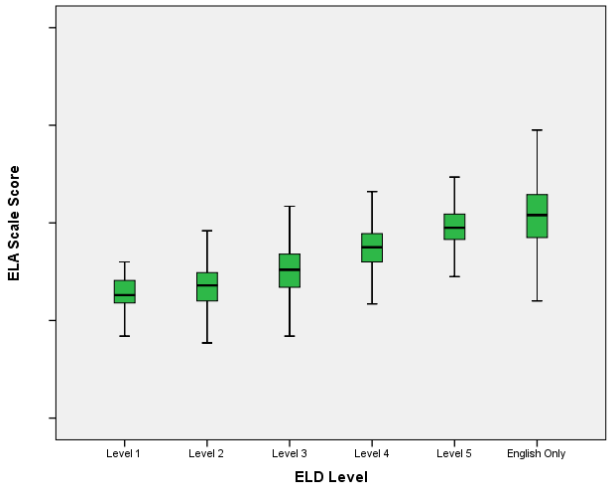


FIGURE 8: BOX PLOT OF ELA SCALE SCORES WITH ELD LEVELS

The data for Figure 8 are presented in Appendix E.

Another, option[14] to consider, when working with a composite, is to use an ordinary least squares (OLS) model to relate ELPA scores to ELA scores more concretely. The coefficients from the model $Y_{ELPA} = b_0 + b_1ELA + b_2grade + e$[15] can be used to provide estimates for the ELPA exit score that corresponds with the proficiency cut score on the ELA assessment. The estimates for b_0, b_1, and b_2 are simply plugged in for each grade and the ELA proficiency cut. For example, if b_0, b_1, and b_2 are estimated to be 20, .50, and 5, respectively, and the ELA cut score for proficient is 100, then the estimated ELPA exit score for grade 3 is computed as: 20 + .50x100 + 5x3 = 65. Analogous to Figure 8, the analyst should create a similar plot replacing ELA performance on the vertical axis with growth in ELA – using the same growth model that the state uses in the accountability system.

If a state is considering using conjunctive exit criteria, it is beneficial to create this type of box plot by individual domain (as in Figure 9) and for the aggregate of the domain scores (particularly if the state is interested in using a single score to monitor progress rather than 4 domain scores). Domain sum score and ELA results are presented in Figure 10. The results in Figure 10 are consistent with those in Figure 8, which is expected given the moderately high correlation between the domain sum score and the ELP scale score composite.

[14] This option was suggested by M. Flicek (WY).

[15] Given that b_0 is the expected score of Y_{ELPA} when ELA and grade are equal to 0, it is advisable to adjust the ELA scores and grade variable so that 0 has real meaning. For example an adjusted grade, grade* = grade -3. By using grade* in the model instead of grade, the interpretation for b_0 is now when ELA = 0 and grade* = 0. Since grade* = 0when grade = 3, b_0 represents the 3rd grade score. It may also be preferable to use a set of indicator variables for grade because this eliminates the the linearity assumption.

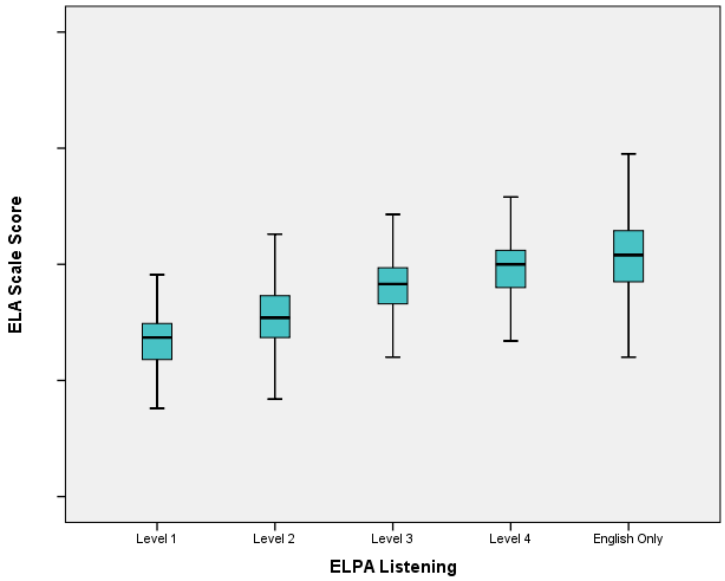
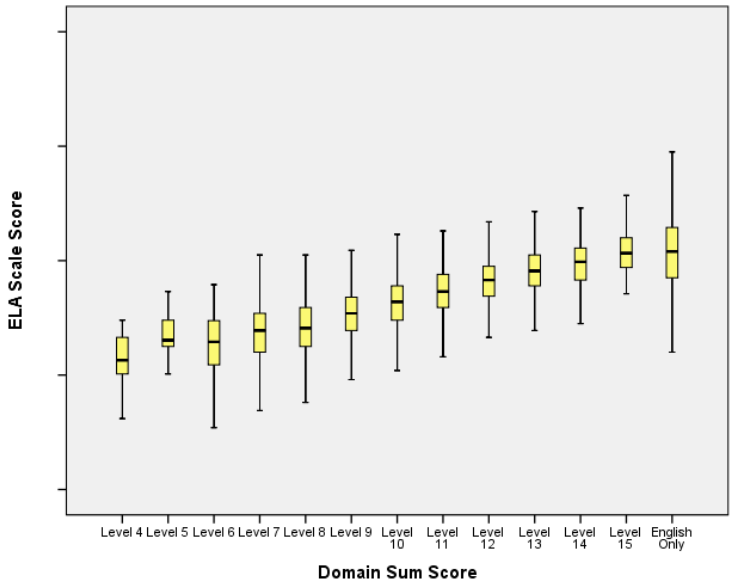


FIGURE 9: BOX PLOT OF ELA SCALE SCORES WITH LISTENING PERFORMANCE LEVELS



*Excludes ELs at sum =16 due to N<20.

FIGURE 10: BOX PLOT OF ELA SCALE SCORES WITH DOMAIN SUM SCORES

Figures 8–10 can be repeated using gains or growth on ELA and Mathematics. An example is presented in Figure 11, which shows gains on ELA. In this case, states should

CHECK

At this stage the SEA should have, at a minimum, a working exit criterion (or criteria if relying on more than the ELPA or ELPA with conjunctive rules). This may need to be revisited after working through the next step of setting an exit timeframe.

examine the results of the method used to calculate growth in the state accountability system.

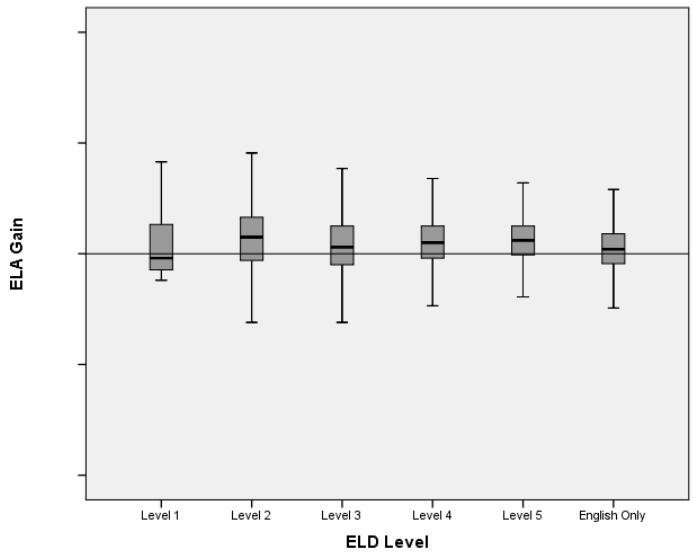


FIGURE 11: BOX PLOT OF ELA GAINS BY ELD LEVEL

Whereas previous figures highlight concurrent status, Figure 11 emphasizes the potential for continued success and/or closing an achievement gap[16] for ELs after exiting. Figure 11 indicates that if the state were using simple annual gains to measure growth on content assessments, ELs who are at least ELD level 3 demonstrate gains in line with their English Only classmates.

- Should the "Check" go here so it's after the end of the final paragraph?

Setting ELs Exit Timeframe

Once exit criteria are established, the state needs to set timelines for ELs to exit EL status. The methods described below are based on existing results and thus provide support for the decision, but in developing an accountability system the SEA needs to appropriately weigh what is realistic with what is ambitious. Exit timeframe is a key component of the EL Progress indicator. The time to exit dictates EL annual progress required to reach proficiency on time.

Shorter timeframes lead to greater progress requirements and fewer students meeting those targets. This relates to lower long term goals because the baseline data will show a lower proportion of students meeting targets – resulting in long term goals that appear

[16] The inferences are bounded by the limitations of the growth model used for comparison.

weak. Hence, by setting high expectations by requiring shorter timeframes to exit for individual students, the overall state EL plan does not appear to be rigorous. Moreover, shorter timeframes, if schools do not meet the challenge, will result in a greater proportion of Long Term ELs.

Setting exit timeframes too long results in more students initially meeting progress targets, higher long term goals, but also results in timeframes that extend into middle, high, or beyond high school. These targets may not represent a meaningful improvement from one year to the next. This results in a false sense of progress in early years, only to have students "suddenly" fall behind because early performance masked the need for intervention. This results in middle and high schools facing difficult challenges and in an accountability system where the EL Progress indicator is biased against them.

Hence, exit timelines should be developed based on a clear ToA that articulates how the state envisions progress and how the ELP Progress indicator supports monitoring of this progress. These timelines should be informed by:

- Existing Research
- Stakeholder recommendations
- Empirical evidence

Research indicates that time to language proficiency can vary from four to seven years (Linquanti & Hakuta, 2012; Hakuta, Butler, & Witt, 2000); establishing an appropriate timeframe is not an entirely open-ended process given what is known about language development. ESSA requires that at least one student factor (at least initial ELD level) be included in developing exit timeframes. Cook, et. al. (2012) outline approaches to establishing exit timeframes that include calculating the percentage of students exiting by initial performance level and year, as well as using event history analysis to inform the decision (Cook, et. al., 2012). Cook, et. al. (2012) has access to five years of data while this Handbook generally applies two assessment occasions from which to draw inferences[17]. This Handbook demonstrates what Cook, et. al. (2012) recommends but applied two years of data. The SEAs can also produce these results and determine whether they are tenable as a basis for the setting exit timeframes. The SEA should continue to revisit these analyses at least until the analyses can be conducted with 5 years

[17] Using five or more years of data to examine the likelihood of students exiting or the cumulative probability of exiting vs. some of the analyses suggested in this Handbook also points to the substantive difference between using data to develop the ELP Indicator and how the accountability system will use data. As a first step, the accountability system will likely depend on students with valid ELPA scores. Event history analysis begins with all students who are initially EL and follows them through to exit, or until their data is missing or data collection has ended.

of data using the new assessment. Monitoring strategies are presented in more detail in Part II.

Table 4 presents the data used to determine the cumulative probability of exiting EL status. Only those students whose initial ELD level was Level 1 are presented; the remaining initial ELD levels are treated similarly.

TABLE 4: EXIT STATUS DATA

Exit Status over time

		Most Recent ELPA Score							
Initial ELD Level	Year in Program	(A) Score in 15 Missing 16	(B) Exit in 15	Level 1 in 16	Level 2 in 16	Level 3 in 16	Level 4 in 16	(C) Exit in 16	(D) Total N
Level 1	1	247	0	1059	0	0	0	0	1306
	2	169	54	144	607	556	241	26	1797
	3	162	134	82	296	416	406	72	1568
	4	132	178	32	146	328	359	79	1254
	5	117	194	36	150	341	588	106	1532
	6	108	177	24	82	260	442	68	1161
	7	74	101	12	45	186	406	79	903
	8	44	60	8	37	117	295	50	611
	9	47	38	9	22	56	211	22	405
	10	36	36	3	15	64	151	12	317
	11	0	0	0	11	42	91	6	150
	Total	1136	972	1409	1411	2366	3190	520	11004

The data in Table 4 represent only two years of matched student data, while this type of analysis generally uses 5 or more years. The data in Table 4 are backwards looking instead of starting from a specific cohort and following their progress forward, and the table makes explicit use of the number of years an EL has completed up to the current year. Hence, Table 4 creates an artificial longitudinal dataset by taking a snapshot of where students are and how long it took them to get there. Every survival analysis suffers from attrition, which in this instance can only occur once – that is, students who were in the dataset in 2015 and missing in 2016, but did not exit in 2015. These students are represented in column A. These students are included in the calculations since at the beginning of 2015 they had an opportunity to either exit, continue, or attrite. Column B displays the number of students who exited in the prior year. Column C displays the number of students who exited in the current year (2016), and column D displays the total number of students who have data in either 2015 or 2016 at each year of program participation. These data form the basis for Table 5 and Figure 12.

TABLE 5: CUMULATIVE PROBABILITY OF EXIT

(D)	(F)	(G)	(H)	(I)
N	Year	Exit	p(not Exit)	p(exit)
1797	2	80	96%	4%
1568	3	206	87%	17%
1254	4	257	80%	34%
1532	5	300	80%	47%
1161	6	245	79%	58%
903	7	180	80%	66%
611	8	110	82%	73%
405	9	60	85%	77%
317	10	48	85%	80%
150	11	6	96%	81%

In Table 5, column D repeats from Table 4, Column F shows the year in program, and Column G shows the sum of columns B and C from Table 4. We see that in program year 2, 80 (54+26) ELs exited. Column H shows the probability that a student does not exit in the given year. We see that 1797 students in their second year who could have exited and 80 (column G) did exit. Hence, the probability, p(not exit) is H = (D-G)/D, or (1797-80)/1797 = .955 or 96%. The probability of exit, p(exit) is I = 1 – H_2 (H_2 refers to column H, year 2, or 1 - .96 = .04 or 4%.) Each year has its own probability of not exiting; however, the cumulative probability in column I includes all previous probabilities of not exiting. As we see in year 4: I = 1 – H_2*H_3*H_4, or 1 – (.96*.87*.80*.80) = 1 - .53 = .47, or 47%. The results for all four initial ELD levels are displayed graphically in Figure 7. A short YouTube video at: https://youtu.be/ArWM0jnKQ6A provides a step guide to conduct this analysis.

Given the paucity of data, a table of mean time to exit provides one element of empirical evidence for setting the exit timeframe. Table 6 presents the mean time to exit for students who have exited but also the mean time for students currently in the program, who have not yet exited.

As noted above, mean time to exit is an underestimate because it is based only on students who have exited. By including the mean times of students currently in the program,

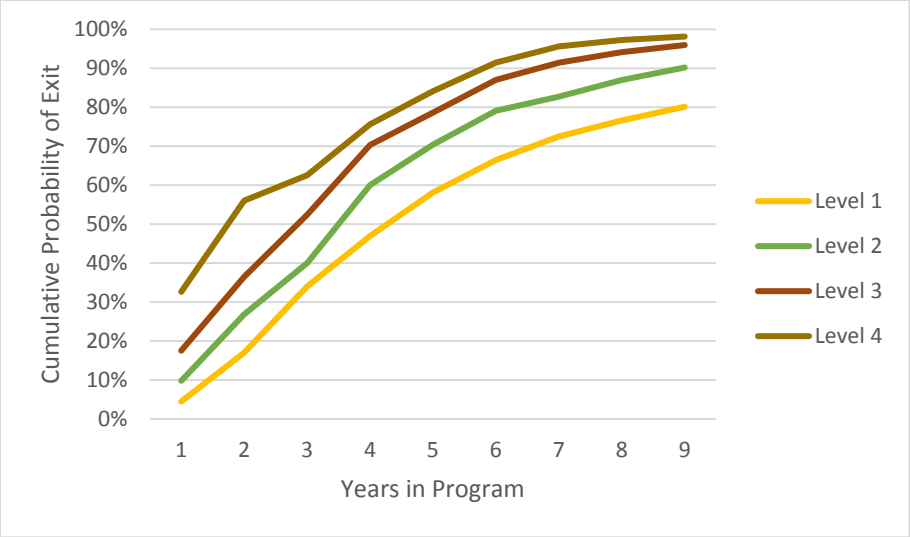


FIGURE 12: CUMULATIVE PROBABILITY OF EXIT BY INITIAL ELD LEVEL

it becomes apparent that some students are in the program longer than the successfully exited students.

Table 6 shows that of students who entered the program at level 1 and were still in the program at the beginning of 2016, 520 out 8,896 (about 6%) exited at the end of 2016. These students were in the program an average of 5.5 years. Of the initial ELD level 1 students who have not exited, they have been in the program for 4.3 years. Of the 8,376 initial ELD level 1 students who have not exited, about 61% [18] are more than 1 performance level from reaching the exit criteria.

[18] Supplementary analyses are accomplished by selecting only ELs who have not exited and creating a cross tabulation of initial ELD level by current ELD level.

TABLE 6: MEAN NUMBER OF YEARS IN THE PROGRAM

	Have not Exited			Exited		
Initial ELD Level	Mean	N	SD	Mean	N	SD
Level 1	4.3	8376	2.6	5.5	520	2.0
Level 2	2.8	3632	2.2	4.5	196	2.0
Level 3	2.8	2741	2.1	4.4	213	2.2
Level 4	2.2	4591	1.4	2.8	515	1.3
Level 5			0.7	1.0	500	0.2
Total	3.3	19340	2.4	3.4	1944	2.3

States can also calculate the mean gains by initial ELD level and year in program, as displayed in Table 7. This does not follow the same students over time, but rather shows the gains in performance from the previous to the current year by students who happen to be at different points in the time continuum. Table 7 presents results only for seven years and only for students whose initial ELD level was level 1. The results indicate that it would take an EL approximately until the end of year 5 to accumulate sufficient gains to meet the exit criteria (if set at Level 5). A similar table can be constructed for each domain and the domain sum score.

TABLE 7: PERFORMANCE LEVEL GAINS BY YEAR IN PROGRAM

Initial ELD Level	Years in Program	Mean Gain	N	S.D.	Cumulative Gain
Level 1	2	1.52	1459	0.95	1.5
	3	0.91	1193	0.96	2.4
	4	0.64	882	0.90	3.1
	5	0.52	1155	0.83	3.6
	6	0.42	831	0.82	4.0
	7	0.36	701	0.81	4.4
	8	0.32	472	0.81	4.7

[1]A complete version for all initial ELD levels is in Appendix M.

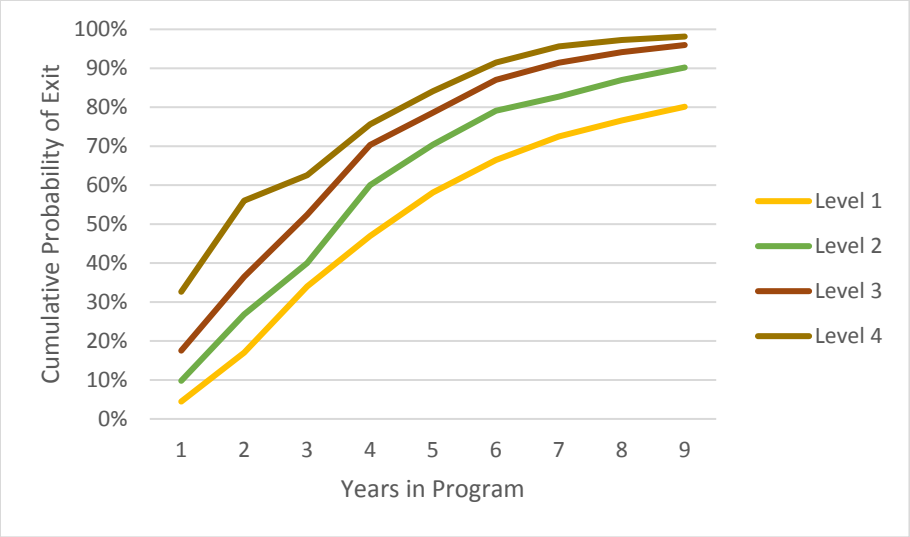


FIGURE 13: CUMULATIVE PROBABILITY OF EXIT BY INITIAL ELD LEVEL

The preceding analysis examples presented several methods to examine time to exit using only two years of robust accurate data[19]. Results developed in this manner should be treated with caution, and considered against research results and stakeholder input. The results are consistent with research indicating a likely time to exit is about four to seven years. In developing the ELP Indicator, it is important to consider both current progress and desired progress to form realistic expectations for EL language development.

One ubiquitous issue arises when states change assessments. In 2017, most states administered a relatively new ELPA but must develop an EL progress indicator based on only two or three years of results. Growth models that form primarily normative results (Student Growth Percentile (SGP) and Value Added Model (VAM)) are the most robust to changing assessments (Goldschmidt & Hakuta, 2016).

ELPA performance is fundamental for developing an EL *progress* indicator. Ideally, multiple years of data are utilized to examine trajectories over time. In fact, it is not unreasonable to expect changes in performance between the first and second administration of a new assessment (or a significantly revised version of an old assessment), or to demonstrate greater gains[20] than will be observed in subsequent years (Linn, 1998). As instruction improves so should performance, but this likely lags behind the increased rigor of assessments. Using simple linear gains to monitor performance

[20] Depressed scores on the first administration of a new assessment are generally due to students and teachers not being familiar with the new standards, coverage, rigor, test format, and administration (e.g. computer vs. paper pencil).

over time is potentially problematic due to both the nature of performance gains and the nature of language development[21]. The key consideration for developing an indicator is to identify a pattern and develop appropriate growth expectations. This is discussed in more detail below.

A single year of new ELPA results. Using two years of data presents inferential challenges, but a state may need to set time expectations based on a single administration of a new assessment or be able to make an adjustment to the accountability system during the transition from one assessment to another. In this instance, scores (performance levels) from the old assessment need to be placed on the new assessment scale – minimally through equipercentile equating[22]. For indicators based on some measure of growth, it is important to correlate old and new scores and compare that to the correlation of old and older scores because a significant change in the correlations foreshadows changes in accountability results. These correlations represent the ordering of students from one year to the next and do not account for shifts in performance due to different standards. Also, it is important to note that given the non-linearity of language development, correlations will understate the relationship between performance and time.

Having only a single year of new assessment results is obviously limiting, but the SEA can reproduce the analyses in Figures 7 through 11 (as well as the OLS analyses described in that section) using the old assessment and the new assessment as a way to incorporate a fixed criterion (e.g. ELA) to compare performance. Also, SEAs using three or more assessment occasions (e.g. a longitudinal growth model, VAM, or SGP) can use two prior (old) years and a current (new) year to estimate the fixed effect of the assessment change. It would likely be prudent to evaluate whether the change in the current year score is related to either (or both) initial ELD level and number of occasions in the program.

These methods are at best approximations because it is likely that the relationship between the last year of an old assessment and the first year of the new assessment will change; given that the first administration of a new assessment tends to result in lower initial performance and faster initial improvement over the first few years (Linn, 1998). Thus, estimating long term trajectories with a single new data point is not recommended. This is particularly problematic for an indicator based on progress where initial progress will be confounded with facets associated with changing assessments. Using the initial

[21] This is discussed in more detail in Goldschmidt and Hakuta, 2016 and Castellano and Ho 2013.
[22] For example see Livingston, S. (2004). Equating Test Scores (without IRT) at: http://www.ets.org/Media/Research/pdf/LIVINGSTON.pdf.

results to calculate a temporary adjustment for the accountability system is reasonable. Figure 14 emphasizes this point as the trajectory based on the first administration of a new assessment presents a more optimistic picture with respect to changes in expected growth over time than does the subsequent administration.

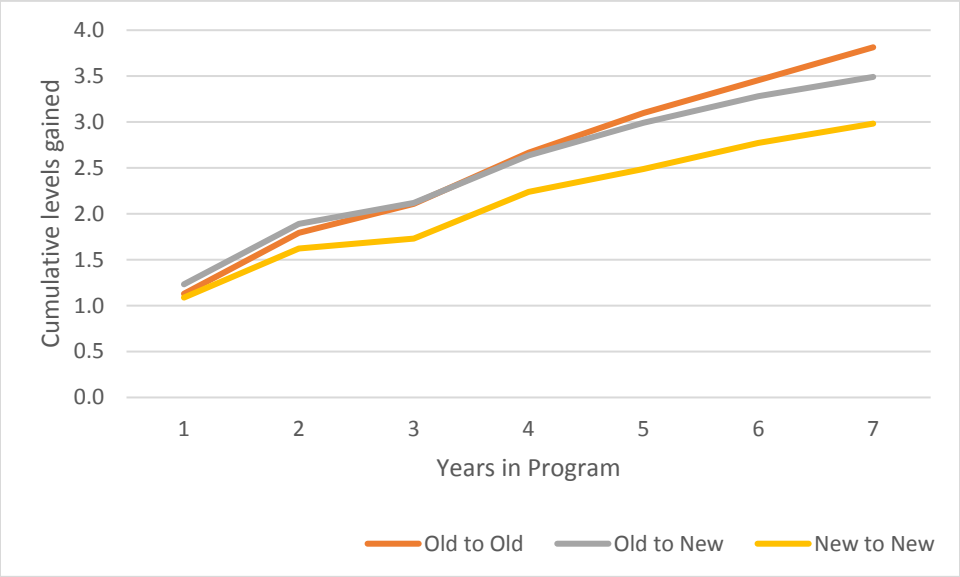


FIGURE 14: CUMULATIVE GROWTH BY YEAR IN PROGRAM - OLD AND NEW ELPA RESULTS

Figure 14 compares cumulative growth for students whose initial ELD level is 2. Three trajectories are represented and provide some indication on the impact of the new assessment on time to exit. The "Old to Older" line represents cumulative gain scores that are calculated by subtracting 2013 scores from 2014 scores (both reflect scores on the old ELPA). The "Old to New" trajectory subtracts 2014 scores from 2015 scores (2015 reflects scores on the new ELPA). "New to New" is calculated by subtracting 2015 from 2016 (2016 is the second year of the new ELPA). The results demonstrate a consistent pattern across assessments. However, initial results imply that time to exit using the new assessment is approximately two years longer than it was using the old assessment. This can be seen by comparing at which year the trajectories approach 3.0 levels gained; given the ELs in Figure 13 started at level 2, and exit at level 5.

Developing a Model that Reproduces Growth Trajectories

An important step in developing the ELP Indicator is to develop a model that describes the observed growth as closely as possible, resulting in a model that reproduces the existing patterns. A progress model that applies incompatible business rules results in a poorly functioning ELP Indicator. For example, if the ELP Indicator is based solely on the percent of students exiting on time, then middle and high schools are likely to be disadvantaged because ELs in middle and high school generally are not exiting on time. Another example is if expectations are either too low or too high, then the ELP Indicator will not differentiate performance among schools very well (an extreme example would be if the growth expectation is set to 0, then all schools would meet this expectation and schools would all perform equally well).

Part of this step is to compare model results against actual growth (as depicted in Figure 13). For example, assuming

CHECK

At this stage the SEA should have a working exit timeframe. This includes having created a table/chart of growth over time disaggregated by initial ELD level. This table/chart should be based on the metric the state intends to use to monitor progress – e.g. ELD levels, scale scores, domain scores, sums of domain scores, etc.

a state uses a version of a Growth-To-Target model with expected annual growth targets presented in Table 8, Figure 15 represents a comparison of actual growth and model-based expected growth.

TABLE 8: ANNUAL GROWTH EXPECTATIONS BY INITIAL ELD LEVEL

Initial ELD Level	Year 2	Year 3	Year 4	Year 5
Level 1	1	1	1	1
Level 2	1	1	1	
Level 3	1	1		
Level 4	1			

Table 8 presents a simple linear model that sets the exit criteria at level 5, and expects ELs to gain one ELD level each year. This trajectory can be graphed against the gains observed by initial ELD and years in program from two assessment occasions.

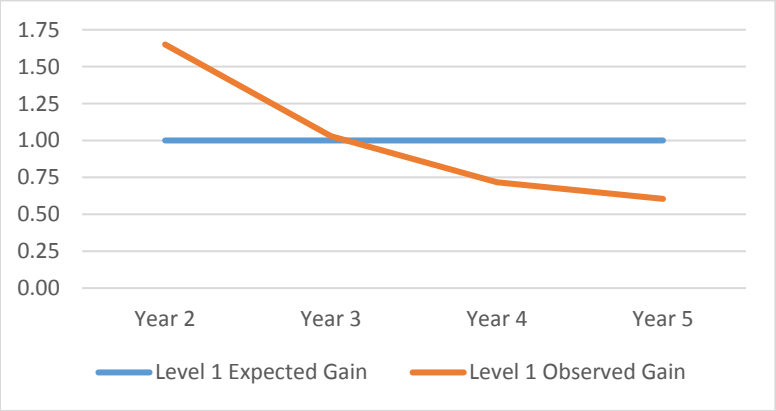


FIGURE 15: COMPARISON OF EXPECTED AND OBSERVED GAINS

Figure 15 displays results for students whose initial ELD is level 1. Figure 15 demonstrates that on average EL gains exceed growth expectations in year 2, equal expectations in year 3, and are below expectations in years 4 and 5. By design, this model results in students easily meeting expectations at first but then finding it more difficult over time to meet expectations. Moreover, meeting the expectation in year 1 does not provide a good estimate of whether a student is likely to remain on track with respect to expectations or to exiting on time. This is clear when plotting the cumulative gain (or change from baseline) against the cumulative expectations based on Table 8.

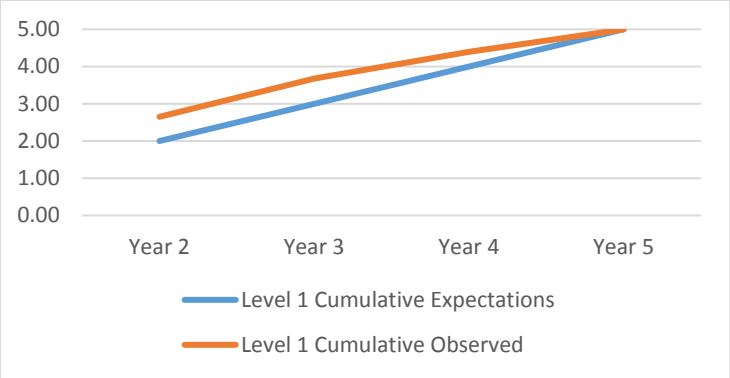


FIGURE 16: COMPARISON OF CUMULATIVE EXPECTED AND OBSERVED GROWTH

Figure 16 is based on the same parameters as Figure 15 and demonstrates that on average, while ELs are meeting exit criteria in the expected timeframe, schools are not receiving credit for progress towards on-time exit because the growth expectations and observed growth deviate sufficiently to cause a disconnect in the accountability system. Results of the ELP Indicator based strictly on a linear growth model[23] will not allow for consistent claims.

Consistent with Figures 15 and 16 it is possible to build a table that provides the probability of meeting growth expectations by year. Table 9 summarizes the results for growth expectations presented in Table 8. Consistent with Figure 14, there is a .84 probability of meeting the year 2 growth target for students whose initial ELD level is Level 1. The individual cells representing level-by-year provide guidance on selecting an approach and timeframe that meet the expectations of the SEA. It is also useful to notice the Level totals. There is a substantial decrease in the probability of meeting growth expectations (for any year in program) by initial ELD level. This may be meaningful if schools have systematic differences in the proportion of initial ELD level students enrolled. This will be addressed further in STEP 4 of developing the ELP Indicator.

[23] It is important to note that this result is based on the sample state's data. Previous evidence suggests that the same model applied in different states can have substantively different results (Goldschmidt, Choi and Beaudoin, 2012).

TABLE 9: PROBABILITY OF MEETING GROWTH TARGET BY YEAR IN PROGRAM AND INITIAL ELD LEVEL

Initial ELD Level	Year	p(Meet)	N	S.D.
Level 1	2	0.84	1459	0.36
	3	0.58	1193	0.49
	4	0.41	882	0.49
	5	0.36	1155	0.48
	Total	0.58	4689	0.49
Level 2	2	0.56	769	0.50
	3	0.32	508	0.47
	4	0.16	422	0.37
	Total	0.39	1699	0.49
Level 3	2	0.31	575	0.46
	3	0.29	442	0.46
	Total	0.30	1017	0.46
Level 4	2	0.08	1678	0.27
	Total	0.08	1678	0.27

Table 9 can be reproduced based on any growth model to ascertain whether an EL met the growth expectation or not. For example, the proportion of students whose scores based on a value-added model (VAM) are above 0[24], or students whose SGP is above 50 can be placed into a table like Table 9. Figure 17 reproduces Figures 15 and 16 based on a VAM. The VAM model (equation 1 below) appears to be more closely aligned with both annual changes in ELD levels as well as cumulative growth than the simple linear model.

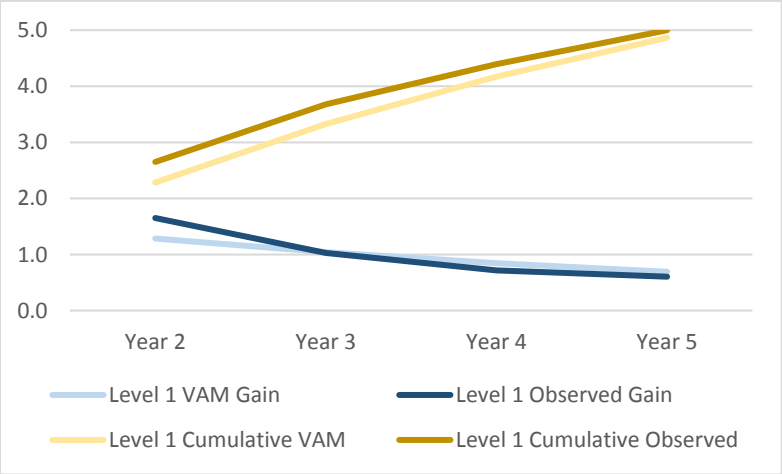


FIGURE 17: VAM COMPARISON OF MODEL AND OBSERVED RESULTS

[24] Whether a Value Added of 0 or a SGP of 50 is the appropriate benchmark requires additional consideration.

Table 10 shows that, consistent with expectations, the VAM results are somewhat more egalitarian across initial ELD levels.

TABLE 10: PROBABILITY OF MEETING GROWTH TARGET BY YEAR IN PROGRAM AND INITIAL ELD LEVEL BASED ON VAM MODEL

Initial ELD Level	Program Year	P(Meet)	N	S.D.
Level 1	2	0.86	1,574	0.35
	3	0.40	1,193	0.49
	4	0.26	882	0.44
	5	0.27	1,155	0.45
	Total	0.49	4,804	0.50
Level 2	2	0.66	797	0.47
	3	0.26	508	0.44
	4	0.13	422	0.34
	Total	0.42	1,727	0.49
Level 3	2	0.48	600	0.50
	3	0.25	442	0.43
	Total	0.38	1,042	0.49
Level 4	2	0.16	1,726	0.37
	Total	0.16	1,726	0.37
Level 5	1	1.00	499	0.00
	Total	1.00	499	0.00
	Total	0.43	9,798	0.50

The previous set of analyses compared results of a VAM to observed growth and a simple Linear Gain Model. It is important to understand the potential strengths and limitations of the state's desired growth model beyond the differences noted above. These are detailed in Goldschmidt and Hakuta (2016). As of summer of 2017, four types of growth models appear to be most prevalent among state ESSA plans:

- Transition Matrix/Value Table
- Growth to Target
- Value Added Model
- Student Growth Percentile (including Percentile Rank of the Residual).

Each of these has strengths and limitations. Several factors need to be considered when choosing a growth model:

- Assessment to which the model is applied.

    - What is the metric of the exit criterion? Is the exit criterion set in composite ELP levels, a profile of domain ELP levels, composite scale scores, sum of domain ELP levels, or another metric?
    - How does the metric used to model EL English language development align with the exit criteria (i.e., does monitoring lead to consistent results in terms of exiting)? For example, if the state monitors a composite scale score, but uses four domain ELP levels to exit ELs, do these two metrics align and result in ELs actually exiting?
- Transparency and communication potential.
    - Generally, growth based on more data tends to be more robust to confounding factors, but is more difficult to communicate to stakeholders in the field. A panel growth model is technically superior to a simple gain model, but the former is both more difficult to calculate and more difficult to present results to stakeholders than a year over year gain approach.
    - Another consideration is whether to match a state's existing growth model used for academic content. A state that has already developed strong understanding of SGPs or VAM might want to build on this infrastructure and use the same model for monitoring EL progress.
    - However, although there might be general understanding of an existing model, it is likely that EL coordinators and teachers are less adept at using SGP or VAM results. Further, stakeholders in the field are unlikely to readily translate school results into program meaning—and conversely, meaningfully aggregate individual student results into school results.
- Technical qualities
    - Are results of the model robust to factors outside of a school's control? Generally, school input characteristics such as the percentage of students who qualify for free/reduced lunch (FRL) or other student characteristics are correlated to a model's results to determine whether any unwanted influences exist. A rule of thumb is if the R^2 between a school input factor and model results is less than or equal to .05, the factor is not substantively biasing model results. It should be noted that summing each school input to determine a summed, or joint R^2 , is inappropriate, because the impact of inputs are not independent. In order to examine the joint effect of several input factors, states should conduct an ordinary least squares (OLS) regression.
- Capacity
    - Does the SEA have the staff to manipulate, develop, implement, monitor, interpret, and adjust the model as required?

    - For example, if the SEA introduces a new assessment administration option that is not universally adopted, does the SEA have capacity to adjust results to take all of the options into account?
- Does the SEA have staff to work with a vendor to monitor, interpret, and understand the need for potential adjustments. A vendor may not be aware of policy changes that might necessitate a model adjustment.
    - For example, if the SEA introduces a new assessment administration option that is not universally adopted, does the SEA have capacity to recognize the need to inform the vendor?

Growth models can either explicitly or implicitly incorporate other factors as such initial grade (e.g., this can be a variable in a VAM), or the model can be applied separately by grade span or school level, for example.

The preceding growth examples have disaggregated growth by initial ELD level. It is very likely that this is the most significant factor in determining how long an EL takes to exit. Initial ELP may impact the shape of the growth trajectory and actual growth over time (i.e., not simply different starting levels running in parallel). This variation is likely not so great as to warrant a joint effect in the growth model. To test whether growth varies by starting ELD level, the desired model can include the joint, or interaction, effects between initial ELD level and years in the program. For example if the VAM[25] is:

$$\Delta Y_{it} = b_0 + b_1(\text{TIME}) + b_2(\text{TIME}^2) + b_3(\text{initial Level 2}) + b_4(\text{initial Level 3}) + b_5(\text{initial Level 4}) + r_{it} \quad (1)$$

In (1) ΔY_{it} = Change in ELPA score between the current year and the previous year;
b_0 = is the intercept and represents the initial change in scores for students whose initial ELD is Level 1;
b_1 = the average change in scores over TIME;
TIME = the number of years in the program;
b_2 = the deviation from linear growth (usually deceleration[26];
b_3, b_4, b_5 = the incremental score due to having an initial ELD level different from Level 1;

[25] There are many different approaches to specifying a VAM, this specification is used to mirror the use of simple gains. Also, VAM often incorporate school fixed or random effects (rather than simply aggregating individual residuals, the details of which are beyond the scope of this Handbook. Additional information is available in Bryk & Raudenbush (2002).
[26] Goldschmidt, Choi, and Martinez (2010) discuss the impact modeling non-linear growth in accountability settings.

Level 2, Level 3, Level 4 = initial ELD Level (Level 1 is the left-out category and Level 5 is excluded because students whose initial score is Level 5 are exited and do not have a growth score); and,
r_{it} = the residual = model-based estimate of ΔY_{it} – observed $\Delta Y_{it.}$

The residuals form the basis of comparison (e.g., in Figure 17 and Ttable 9 the test was whether the residual is greater than 0 or not). An important check for SEAs using a normative model is to relate the norms (e.g. a VAM result = 0, or an SGP = 50) back to a criterion in order to understand what a trajectory of a series of 0s or 50s portends for the student. Consistent across all models (simple gains, Growth to Target, VAM, etc.) is the implicit assumption that the mean trajectory (or point estimate) is normally distributed around the estimate. A SEA could determine that progress is only met if growth is equal to the mean + 1 SD, for example. This reduces type I error, increases the likelihood that students meeting annual targets are, in fact, proficient in the pre-determined time, but increases the type II error and decreases the percent of students meeting targets. This, of course, is not a problem if long term goals and school expectations are developed accordingly. The SEA or stakeholders may consider that language development occurs not only over different timeframes but also follows different trajectories. In order to test whether growth trajectories differ by initial ELD Level, equation (1) needs to be expanded to include:

$$\Delta Y_{it} = b_0 + b_1(\text{TIME}) + b_2(\text{TIME}^2) + b_3(\text{initial Level 2}) + b_4(\text{initial Level 3}) + b_5(\text{initial Level 4}) + b_6(\text{TIME*initial Level 2}) + b_7(\text{TIME*initial Level 3}) + b_8(\text{TIME*initial Level 4}) + r_{it} \quad (2)$$

In this case, interpretation of the TIME coefficients changes; b_1 is no longer the average growth of ELs over TIME, rather is the growth of initial Level 1 ELs. Growth for the other initial ELD levels is now $b_1 + b_6$ for initial Level 2 ELs, $b_1 + b_7$ for initial Level 4 ELs, and $b_1 + b_8$ for initial Level 4 ELs. Overall, if the change in R^2 is significant between the model without joint effects and the model with joint effects, including the joint effects statistically improves the model. Whether this improvement is substantively relevant requires additional considerations. The significance of b_6 through b_8 directly tests whether the growth trajectory of Level 2 through Level 4 (initial levels) ELs differs significantly from the growth trajectory of Level 1 ELs. Even if some or all of the b_6, b_7, b_8 are statistically significant, substantively the effect may be minimal, especially given that the timeframe is relatively short (likely four to eight years). Also, one can correlate the residuals to determine whether adding joint effects creates important changes in the residuals. A model without joint effects is likely sufficient, and it may be beneficial to demonstrate this to various stakeholders.

Other EL-specific variables that should be examined are Recently Arrived English Learner (RAEL) and Students with Interrupted Formal Education (SIFE) status. RAEL and/or SIFE can be included in equation (1) and the impact of either of these characteristics can be explicitly tested. An example of these analyses are presented in Part III.

A particularly useful statistic is the percentage of students who *should* exit in the given year compared to those that do, using the expected exit criteria of the new ELPA[27]. Given the deviation from linearity the remaining examples will use the annual growth expectations presented in Table 11 that is based on the gains depicted in Table 7.

TABLE 11: NON-LINEAR GROWTH EXPECTATIONS

Initial ELD Level	Year 2	Year 3	Year 4	Year 5
Level 1	1.2	1	1	0.80
Level 2	1.2	1	0.80	
Level 3	1	1		
Level 4	1			

Using the expectations in Table 11, it is possible to calculate the percentage of students who exit "on time"—before or during the year indicated based on the initial ELD level[28].

Using the growth expectations and the timeframe presented in Table 11, Table 12 indicates that between 42% and 54% of ELs exit on time. Table 12 provides another option for the SEA to consider whether these percentages are in line with expectations.

CHECK

At this stage the SEA should have selected both the metric to monitor EL progress and the growth model. The SEA should know which variables will be included in the growth model and the probabilities of achieving expected growth, disaggregated by the variables selected to be in the model.

[27] This analysis only considers assessment results since other exit criteria may not be comparable.

[28] Assuming this is the only factor that is used to create growth expectations.

TABLE 12: PERCENT OF EXITING ELS EXITING ON TIME

Initial ELD Level	Percent Exit on Time
Level 1	54.4%
Level 2	52.6%
Level 3	41.8%
Level 4	54.0%

Translating Growth Model Results into Individual and School Scores

The next step in developing an ELP Indicator involves translating the results of the growth model into an individual[29] student score, and aggregating this score to the school. Aggregating individual student scores to the school level is a transparent way to create a school summary measure, but implicitly assumes that school effects are merely the sum of individual student performance. Ignoring the clustering of students in schools ignores school context (Burstein, 1980) and potentially produces biased estimates of school effects (Willms & Raudenbush, 1989). When the intra-class correlation is greater than zero, simple aggregates of student performance ignore the fact that students attend specific schools and ignore the mixes within and between school variations in performance (Aitkin & Longford, 1986).

If the SEA intends to create an ELP Indicator that is amenable to aggregation and disaggregation of results between students, schools, districts, and the state, it needs to establish business rules governing progress. The basic distinction for describing student progress is: met/not met target and how much progress. The dichotomous met/not met criteria can be further refined into partial and extra credit, depending on the degree of progress toward the target. Awarding partial/extra credit describes how much progress was made. Goldschmidt and Hakuta (2016) provide an example detailing the tradeoffs of the met/not met approach as it relates to individual student progress.

One option is to use Met/Not Met (e.g., creating a single Met Target variable coded as Met =1 and Not Met =0). This coding is useful because the average of Met Target indicates the percentage of students that met the progress target in the given year. As noted, Met

[29] It may be the case that a state uses a growth model that provides a direct measure of school performance without needing to aggregate individual scores. For example, a SEA might use the Empirical Bayes (EB) residuals from a random effect VAM as a basis for the ELP Indicator. While EB residuals have several desirable properties, there is a disconnect between school EB residuals and individual student performance that – especially as school Ns become small require additional communication in order for schools to understand and meaningfully utilize results. Hence, the school EB residual is another example of a metric that does not really have an individual analog.

Target can be based on the growth expectations in Table 9. An EL whose initial level is Level 2 and is in year 3 of the program is expected to make one level of gain. If the student made one or more levels of gain then Met Target is coded as 1. Similarly, results from the VAM can be coded as 0/1 by setting a target (e.g., a student's value added score is greater than or equal to 0).

Dichotomous determinations (0 or 1) based on VAM, or SGP need demonstrate that a student meeting the target annually is likely to exit in the prescribed timeframe. The transition Matrix/Value Table/Growth-to-Target approach presented in Table 11 is explicitly based on a set of annual targets that lead to reaching the exit criteria in the prescribed time. However, normative models like VAM and SGP may not result in annual targets that meet the exit criteria in the prescribed timeframe because the norms are based on students who are not exited on the set timeframe. A state using a VAM or SGP model needs to consider how the relative results relate to actual progress, given that the ELP Indicator is intended to monitor progress towards a specific goal in a specific timeframe. For example, a VAM provides school contributions to progress that deviates from the overall mean trajectory (in most cases this is a residualized gain[30]). If the mean trajectory (by initial ELP level) does not result in students meeting the exit criteria in the prescribed timeframe, then a state must adjust business rules to ensure that inadequate ELD progress is not deemed acceptable. This is also true when using SGPs. Results for nine students are presented in Table 13.

The second option is that the SEA may want to incorporate a standard error in setting targets. For example, a range in SGPs from 45 to 55 accounts for potential sources of error around the target. The SEA needs to consider whether this additional complexity is a reasonable tradeoff for transparency.

The third option is to award partial/extra credit. We know that targets can be based on table of values, VAM, or SGP. Rather than earning either 0 or 1 point towards a school's aggregate score, each score might be awarded as a fraction of a score[31].

The fourth option is to award credit based on actual model results. Table 13 presents all of these options for a sample of nine students in the dataset. Table 13 displays both the initial ELD level of each student as well as year in the program. Based on these two

[30] See Goldschmidt and Hakuta, 2016).

[31] Under ESSA, fractions greater than 1 for better than targeted performance cannot offset lower performance – e.g., "advanced" growth cannot earn 1.25 points if "nearly meets" growth earns 0.75. points. States have proposed awarding "advanced" points only to the net difference in the number of students who are "advanced" compared to those who are "nearly meets." For example, if there are 25 ELs at "advanced" and 20 ELs and "nearly meets," the 25-20 = 5 students would receive 1.25 points.

parameters, each student has a growth target. Student 1006 had an initial ELD level of 1 and is in year 2 of the program. These values in Table 13 show a growth target of 1.2 for student 1006. Student 1006's growth is 2.5 and so this meets the established growth target (Met Growth Value Table =1). The VAM model presented in (1) results in a residual score of 1.23, which is above expectations because actual growth is significantly above expected growth (the growth target from Table 11 was informed by average growth in that initial ELD level/year in program combination). The business rule for Met Growth (VAM) is that a student meets his or her growth target if the residual is greater than or equal to 0; which, for student 1006 it is. The last column of Table 13 is based on Table 11 values and awards partial credit based on the ratio of growth to growth target (for student 1 this is 2.5/1.2 > 2). Credit is awarded as follows: a ratio less than .67 receives 0 credits; a ratio between .67 and 1 receives 0.5 credits; a ratio of 1 to 1.249 receives 1 credit; and, a ratio of 1.25 and greater receives 1.25 credits. These rules are flexible and should align with state, district, and school contexts.

TABLE 13: INDIVIDUAL STUDENT RESULTS BASED ON VARIOUS APPROACHES FOR AWARDING CREDIT

Student	Initial ELD Level	Program Year	Growth Target	Growth	Met Growth (Value Table)	VAM (residual)	Met Growth (VAM)	Partial/Extra Credit
1006	1	2	1.2	2.5	1	1.23	1	1.25
1013	1	3	1	1.5	1	0.44	1	1.25
1028	4	2	1	-0.2	0	-0.58	0	0
1035	1	5	0.8	0.2	0	-0.43	0	0
1038	1	4	1	1	1	0.16	1	1
1062	4	1	.	.	.	.	.	.
1063	2	7	.	-0.4	.	-0.33	0	.
1067	1	1	.	.	.	.	.	.
Mean	1.88	3.13	1.00	0.77	0.60	0.08	0.50	0.70

*Met Growth: Yes =1, No=0.

When the results are compared across the approaches in Table 13, it is clear that they are fairly similar. Overall, 60% (0.60) of students met their value table growth targets. However, growth in terms of the VAM (residual) is slightly above 0. The VAM (residual) result is consistent when comparing the mean growth of 0.08 to the mean growth target of 1.0. This points to a tradeoff of using a 0/1 counting rule: that the magnitude of growth is less readily apparent. For this set of students the partial/extra credit approach only marginally changed overall results, consistent with expectations.

Creating the ELP indicator

In this last step, the SEA must determine the final set of business rules that calculate the ELP Indicator, giving special consideration to 1) whether to add other measures into the ELP Indicator, and/or 2) whether the ELP Indicator should explicitly hold students and schools accountable when students fall behind the initial ELD level/year in program combinations in Table 11 (i.e. how to treat Off Track ELs in the accountability system).

A SEA may want to monitor progress based on meeting growth targets and include an additional measure such as the percentage of students exiting on time (similar to the graduation rate indicator). The SEA can also capture idiosyncratic growth by not only assessing growth in the current year against a target, but by tracking change from the baseline against a cumulative target, an approach which parallels the cumulative nature of language development.

If the state intends to use multiple measures to create the EL index, it is advisable to normalize each component to avoid different measurement properties of the various measures (OECD, 2008). This is akin to thinking about the effect size of each component of the indicator with respect to the total indicator score. That is, a one standard deviation change in the first measure results in how many units of change in the indicator compared with a one standard deviation change in the second measure[32]? If all the indicators are expressed in the same measure, then normalization is unnecessary (OECD, 2008).

Specifically, some state may include both the proportion of students meeting a growth target and whether or not a student exits. For example, the school score (Score = $(N_{meet\ progress} + N_{exit})/(N_{assessed} + N_{exit})$) includes the number exited in the numerator and the denominator. It should be clear that the impact of this is not uniform and depends on both the percentage of students exiting and the probability of meeting progress targets. The lower the probability of meeting progress targets, the greater the impact of including number exiting; the greater the number exiting, the larger the impact this component has.

Considerations for Monitoring Off Track EL Students

According to ESSA, state accountability systems must monitor the progress of all ELs. Setting a fixed timeframe for exiting is required and helps set reasonable and rigorous expectations for students and schools. An important question that arises is what happens to students once they have exceeded their timeframe? This question applies particularly to long term ELs (ELs who have been in the program more than five years). It also applies

[32] This is not to say that different effect sizes are not warranted. For example, if the state wants the measure of the index to receive more weight, then this is one method to operationalize this weighting scheme.

to ELs whose initial ELD level was 2 or higher and had less than five years to exit. For example, an initial ELD level 3 student is expected to exit after two years, but is in year 3. This student is not a long term EL, but is not on the value table combination. Students no longer on the exit timeframe (whether long term or not) are considered Off Track. Some schools then will receive students who are past the exit timeframe but not included in the ELP Indicator. This may be problematic, especially if the only means of earning credit for these students is through exiting, because any progress towards exiting is not acknowledged. The SEA will want a system that discourages becoming an OFF Track student, encourages schools receiving Off Track students to continue to work for their progress, and does not reduce the rigorous expectations for students to become English proficient in a timely manner.

This is addressed by considering the three malleable components that impact how student progress is translated into a school score. One, the definition of a track; two, the growth model and its amenability to monitor progress past a specific time point; and three, the business rules that govern how individual progress is translated into school performance.

The SEA may define the timeframe with additional sophistication to what has been presented above – in this way a student's On/Off Track status may be derived differently that the business rules described in the preceding paragraph. For example, it may be that a student's timeframe is adjusted as s/he changes school levels. For example, an EL matriculating from elementary to middle school would also transition to a middle school transition matrix/value table where time (or Track) is based on the most recent ELD level and set accordingly. Another option is to account for mobility by addressing its impact on expected time to completion. Adjustments such as these may impact when a student is considered Off Track and may capture the vast majority of cases such that extensive business rules are not necessary.

The SEA should also examine its growth model to ensure it is amenable to including Off Track students. If the SEA is using a statistical model (growth model, VAM, or SGP) that does not explicitly rely on the timeframe to monitor progress, then this allows students to be evaluated for progress irrespective of the number of occasions the EL has been in the program.

However, if the SEA utilizes a model that is explicitly based on an end point (such as a Transition Matrix/Value Table, or Growth to Target model), the SEA may want to create a set of growth targets for students who are Off Track separate for those who are On Track. Rules to include Off Track students in state accountability measures should not lower expectations for the exit timeframe (i.e., lengthen it). Students and schools must

still be held to the growth targets based on a fixed exit timeframe. However, by creating a second set of growth targets for Off Track students, schools will have incentives to continue to make progress with these students. Also, schools who do not exit students on time will receive low scores because they are being held accountable to a fixed standard, but can potentially earn credit for growth towards exiting and not only upon exit

Finally, the SEA can adjust business rules such that schools can earn credit for Off Track student progress. An example of adding an Off Track growth target to the ELP Indicator is for a student/school to earn credit (e.g., counted as 1) if the student either grows or exits. Including this additional element to the ELP Indicator will likely slightly reduce school performance on the ELP Indicator because schools will be held accountable not just for the most successful ELs, but for all ELs. The SEA should examine the impact of the business rules on school performance on the indicator; particularly if it favors one school type over another. Monitoring the performance of the indicator is addressed in Part II.

Exhibit I summarizes tradeoffs between awarding credit for exiting and awarding credit for progress and whether the credit a school earns should reflect extra credit or partial credit. In general, the tradeoff of using exit vs. progress is that exit rewards schools once for exiting, but does not monitor progress towards exiting along the way. Awarding credit for progress ensures that progress is continually monitored and that students "count" at all stages of language development (which appears to be consistent with the intent of ESSA). Awarding Extra vs. partial credit has both response and appearance consequences. Schools may react to the different incentives provided by extra or partial credit – e.g. working with students to exit as quickly as possible, delaying exit, or not worrying about exiting at all. Needless to say, credit for progress or exiting of Off Track ELs could also be the same as it is for on Track ELs. In any case, the SEA should develop business rules defining when a student is classified as Off Track, how the growth model measures Off Track student progress, and decide the amount credit a school earns by meeting progress or exit goals.

Exhibit I: School Credit for Off Track ELs		
	Extra	Partial
Upon Exit	Provides incentives to focus on exiting EL.	Provides incentives to focus on exiting EL.
	Appearance of benefit for late exit.	Clarifies that on-time exit is goal.
	Does not monitor annual progress.	Does not monitor annual progress.
	May provide incentives for ES (Elementary School) to delay exit until EL is Off Track	Reduces incentives for ES to delay exit until EL is Off Track
	Increases incentive to work with Off Track ELs, especially if far from Exit = greater impact for MS/HS (Middle/High School)	Reduce incentive to work with Off Track ELs, especially if far from Exit = greater impact for MS/HS
Progress	Monitor Progress annually.	Monitor Progress annually.
	Provides incentive for meeting annual progress goals.	Provides incentive for meeting annual progress goals.
	Reduces incentive to accelerate progress.	May reduce incentive to attempt to meet progress goals.
	Appearance of benefit for late exit.	Clarifies that on-time exit is goal.
	Maintains incentive to work with all ELD levels.	Maintains incentive to work with all ELD levels.
	Progress might be too slow to exit in timely manner.	Progress might be too slow to exit in timely manner.
	Provides incentive for meeting annual progress goals.	Provides incentive for meeting annual progress goals.

Example 1: Using the year-to-year gain

Gain scores can be based on scale scores, domain sum scores, or performance levels (generally partial levels are more amenable than simply whole levels which may be too restrictive). Three examples of how to use gain scores are shown in Table 13.

1) One option is to compare the gain score to a target (for example, initial ELD level and year in program) and count a student as either 0 = not met or 1 = met. The average of the 0s and 1s represent the ELP Indicator for the school. If the nine students in Table 13 formed a school, the value would be .60. If the SEA accountability system is based on a 100 point scale and if the ELP Indicator is worth 10 points, then the school in Table 13 would earn 6.0 (.60*10) points.
2) There are other options for including the .60 into the accountability system. For example, the SEA can set performance cuts around the average meeting the target. These could be based on the overall state average (perhaps by grade band) and thus directly related to long term goals. For example, if the state average of meeting targets is .50, then .60 could be scored a 'B,' or green, etc.
3) Another option is to base school points on the average gain of students in the school. The average gain could either be multiplied by a factor to earn points, or average points could be used to classify schools into performance categories. The school's mean growth could include a confidence interval to reflect uncertainty in line with the N-size for the particular school.

Example 2: Using the VAM results

VAM model results are a residual that indicates the extent to which a student's current ELPA score deviates from the expectation (target) that is based on the student's initial ELD level and year in program[33]. The residual is generally normalized (meaning it has a mean of 0 and a standard deviation of 1). There are many options for converting the residual into the ELP Indicator. Two options are presented in Table 13.

1) One is to award credit (e.g., 1) for any student who has a residual value greater than or equal to 0[34]. The ELP Indicator is then simply the average (percent) of each school's 0s and 1s. In Table 13 (if the nine students formed a school) this is equal to .50. If the SEA accountability model uses a 100 point scale, and if the weight of the ELP Indicator is 10 points, then the school would earn 5.0 (.50*10) points. Options presented for the gain model can be applied to the VAM results as well.
2) If the SEA wants to use the residual as opposed to percent meeting target, there are a number of ways to rescale from the VAM results onto the 0 to 10 ELP indicator scale. Equation (3) shows a simple linear conversion anchored at residuals that range from -3 to 3:

$$\text{ELP Indicator} = 1.67(\text{residual}) + 5. \qquad (3)$$

[33] This is somewhat different from a traditional VAM which uses conditions on prior scores, i.e. the previous year's score.

[34] Keeping the previous discussion in mind whether 0 is the appropriate cut.

It is important to note that Equation (3) is based on finding the slope and intercept that have a minimum and maximum of -3 and 3, respectively. These values are a policy choice, and could be based on the empirical minimum and maximum. Using a policy minimum and maximum can result in truncation of scores (e.g., no school earns 10 points if no school reached 3. Basing Equation (3) on the empirical minimum and maximum each year resolves this issue, but subsequently changes how performance is converted into accountability points each year. Using the results in Table 13, the ELP Indicator score = 5.13 (1.67*0.08 + 5). One difference between a gain score approach and a VAM approach is that instead of applying a confidence interval, the SEA could use the EB residual, which takes the reliability of the estimated school mean into account. The disadvantage of this is that individual student performance does not aggregate up to school performance. This is another example where the school aggregate may have desirable properties (i.e. is BLUP[35]) but loses meaning when disaggregated. Converting any percentage-based measure (e.g., percent meeting target or percent demonstrating having a VAM residual above 0) includes a transformation such as Equation (3), but excludes any shift in the intercept (i.e., the percentage is simply multiplied by the maximum number of points possible for the indicator).

Example 3: Extending business rules to capture the SEA Theory of Action

The preceding examples are relatively simple but may not fully reflect a state's Theory of Action. Example 3 is similar to Example 1, but includes additional elements that the SEA may believe both represents the language development process and provides the correct incentives to schools to continue to work with ELs irrespective of initial ELD level or time in program, but holding rigorous expectations for success. The SEA may recognize that annual growth is varied for an individual student and that cumulative growth is also important in that this provides an indication of whether the student is on track, overall, and not just in the most recent year. Therefore, the ELP Indicator might be based on a series of steps that capture both current year and cumulative growth. States are applying this concept is different ways – including evaluating current and two year growth, as well as current and cumulative growth over all occasions. There are several steps required to expand the definition of growth:

1) Determine whether the student met her most recent annual target or whether the student is cumulatively on track. For example, a student whose initial ELD level is 1 is expected to grow 1.2 levels in Year One and 1 level in Year Two. This implies that the student is at level 2.2 at the end of Year Two and 3.2 at the end of Year Three. If the student grows 1.8 levels in Year One and 0.6 levels in Year Two, this student will meet the Year One target but not meet the Year Two target. However, in Year Two,

[35] BLUP = Best Linear Unbiased Predictor.

this student is at level 3.4, which is above the expected level of 3.2. Considering cumulative growth recognizes that this student has progressed to (beyond) the expected level given the initial ELD level and the number of years in the program. In this way, a student meeting the exit criteria is recognized for exiting, even if the last year of growth was less than expected. In Table 13, an additional column would need to include what the current ELD level is to determine whether (in this case) any of the students had reached an appropriate level in order to warrant a "met" designation under the cumulative growth scenario. Student 1013, for example, would need to be at ELD level 3.2 (Year Three, initial level 1 = 1 + 1.2 + 1) to meet the cumulative target.

2) Consistent with earning credit for exiting, an EL meeting the exit criteria in Year One would count as having met the growth target. This business rule expands students included in the system because growth is not reflected until the second assessment result is available. In the sample data, this rule increases the number of students included in the annual accountability decision by about 600 students. The overall impact increases the percentage of students meeting the target by about 2%. In this state the impact mostly occurs in grades K-2.
3) While adhering to the initial timeframe, the SEA might add business rules for students who exceed the timeframe (Off Track ELs). Excluding these ELs from the ELP Indicator creates a disincentive for schools to ambitiously move these students to English language proficiency. For example, a school can earn credit if the student met a growth target or met the exit criteria. The growth target could be set as the last target from the growth expectations table (e.g., Table 10). For example, for an initial ELD level 1 student, the growth target is 0.8.

This series of expansions to the simple value table of comparing gains to targets may more fully capture the intent of the SEA Theory of Action: all ELs must progress towards English proficiency as efficiently as possible and schools are held accountable for faceting progress for all ELs, irrespective of initial level ELD level or year in program. The business rules might simply be:

If ELD Level = 5, Met_Target[1] = 1; or
If year >= 2 and gain >= Growth Target[2], Met_Target = 1; or
If year >= 2 and ELD Level >= Level Target[3], Met_Target = 1; or
Otherwise, If year >= 2, Met_Target = 0.

(1) Met_Target 1 = yes, 0 = No.

(2) Growth target depends on Year in Program and Initial ELD Level.

(3) For On Track ELs, Level Target depends on Initial ELD Level and Year in Program; for Off Track ELs, Level Target = 5.

These rules result in the probabilities of meeting the growth or cumulative level targets presented in Table 14 (the complete results are displayed in Appendix F).

TABLE 14: PROBABILITY OF MEETING GROWTH OR CUMULATIVE LEVEL TARGETS

Initial_ELD		P(Meet)	N	S.D.
1	2	0.86	1,574	0.35
	3	0.62	1,216	0.49
	4	0.33	891	0.47
	5	0.34	1,159	0.47
	Total	0.46	7,534	0.50
2	2	0.66	797	0.47
	3	0.33	511	0.47
	4	0.18	423	0.39
	Total	0.39	2,423	0.49
3	2	0.48	600	0.50
	3	0.33	446	0.47
	Total	0.31	1,944	0.46
4	2	0.16	1,726	0.37
	Total	1.00	500	0.00
5	1	1.00	499	0.00
Overall	Total	0.39	15,805	0.49

Setting Long Term Goals

Setting long term goals that are closely align with the ELP Indicator provides another tool with which to monitor performance over time. Close alignment means that the ELP Indicator and the long term goals are in the same metric. In fact, a state's long term goal may be an aggregate of school performance on the ELP Indicator. Long term goals are not limited to aggregates of an indicator. For example, a long term goal may be based on the percentage of students exiting on time, or the percentage of students meeting growth expectations each year. If the ELP Indicator is based on a value table that assigns points (values) to different combinations of initial ELD level and growth in a particular year (e.g. 50 points for an initial ELD level 1 student who progressed 0.5 levels in year three) then these values will be less aligned with

CHECK

CHECK: At this stage the SEA should have business rules defining how students and schools will earn points on the ELP Indicator. Included in these business rules are any conjunctive or composite rules for multiple measures and any business rules to address other issues the SEA determines to require attention (i.e. rules that align the functioning of the ELP Indicator with the SEA's Theory of Action) such as how Off Track ELs will be monitored

percent meeting a growth target (even though they are essentially identifying the same phenomenon). Long term goals need to be based on extant performance. Although the language of accountability suggests that indicators and school results should be aligned with long term goals, it is likely a more coherent exercise to develop long term goals from the "bottom up," i.e. by following the steps presented in Part I and ending with a long term goal.

Generally there are four elements to consider in developing long term goals: 1) What is the measure? 2) What is the baseline? 3) What is the target (and when will it be reached)? 4) What are the annual increments?

Beginning with question 1 and taking the most aligned approach, a long term goal, based on the examples above, is measured by the percent of students meeting either growth or cumulative growth expectations. That is, the state aggregate of students meeting their progress expectations is the measure of the long term goal and indicates the extent to which students are progressing as targeted. The SEA must decide what the long term goal in this metric ought to be. This can be a stakeholder decision, a policy decision, or a decision informed by data. For example, if the SEA is considering a graded or color-coded system for each indicator, the long term goal can be aligned with cut values. The long term goal could also be based on the percent of "green" schools. Hence, a school that is classified as green is also meeting long term goals. The SEA may desire a simpler conversion of measure to metric by simply using the percent of students meeting progress targets at each school as the metric for the long term goal. The SEA should be careful in how the data are aggregated to form the state average – especially if progress and school Ns vary considerably. Unweighted school means may not equal the simple aggregate of all students in the state. Using the simple aggregate can produce results that are not representative of the distribution of school performance.

The previous discussion focused on means, but in explicitly addressing question 2, the SEA may want to consider various options for forming the baseline for long term goals. One option is to choose a baseline using a given percentile of the Met Progress Target (ELP Indicator) (Linquanti, 2017). For example, the 25th percentile of Met Progress Target in the sample state is 28% of students meeting the target. The baseline could be based on the mean percentage of students meeting progress targets - currently about 40%. Using a lower percentile results in more schools initially meeting the target, may signal that lower percentages of EL making progress is reasonable, and will require steeper growth to reach an "optically" reasonable goal; resulting in an increasing number of schools (and the state) not meeting long term goals – which may present communication issues. Using the mean results in more schools initially not meeting targets, can result in

communication challenges that imply the targets are too difficult, but also requires slower growth towards a reasonable long term goal.

CHECK

The SEA should have Long Term Goal metric, baseline, end target, and MIPs for reaching the Long Term Goal.

Question three addresses what the long term goal ought to be. Again these decisions are informed by data but are also based on recommendations of stakeholders and SEA policy staff. The simplest option is to select a value – e.g. 75% of students are making progress. Goal setting may also be based on reducing the gap between the long term goal and the baseline by some fraction. In this instance the SEA must consider both the long term goal and the timeframe in which the goal will be reached. Another option is to use extant data (if available) and determine what sort of improvements have occurred in EL Progress and project that into the future. Or the SEA can (as noted above) use the distribution of mean school progress to set long term goals. For example, if the mean school has 40% of students meeting progress targets, and if a school in the 75th or maybe 90th percentile is achieving 80% of its EL students meeting growth targets, then this can be a long term goal as well.

Question 4 then simply addresses a process of developing annual targets. Setting the annual targets could be calculated as the difference between the long term goal and the baseline, divided by the number of years to reach the goal, times the fraction reduction. Another option is to begin with the baseline and consider what would be a reasonable amount of progress over the next set of years. This latter method can be informed by using recent results to determine how much improvement has taken place in meeting the target. In this case the focus is on the percent of students meeting the growth targets in the same initial ELD and year in program combination in the two years of data[36]. For example, in Table 14, 46% of initial ELD level 1 students met the target in program year 4 in 2016, whereas

[36] The extent to which including EL progress into Title I accountability improves meeting targets is not captured by this analysis.

the 2015 cohort with the same combination of initial ELD level and year in program met the target at 44%. This represents a 2% improvement. The weighted (by initial ELD level and year in program) average of all ELs for both years provides an estimate of improvement. However, as noted, it is likely that improvement from the first and second year of a new assessment (and standards etc.) may not reflect long term growth exactly, but it at least provides some basis for thinking about establishing annual targets towards the long term goal.

A long term goal built from the bottom up has the advantage of allowing the SEA to consider how the state level long term goal relates to school and district progress. If the long term goal is that all schools will be classified as "green," and if it is understood that a "green" school is one in which at least two thirds of students (for example) are meeting growth targets or exiting on time, then it is straight forward for stakeholders to understand what it means for a school to be "green" or what it means when the district has 4 out of 10 "green" schools. This is a coherent system where at every level the meaning remains constant and the response would be consistent.

Part II: Monitoring the intended functioning of the ELP Indicator

Ideally, results based on the ELP indicator allow for inferences about how well a school is facilitating English language development. This is akin to making causal claims about schools and their impact on ELs' English language development. The degree to which causal inferences can be made about school effects has received substantial attention in the literature (Burstein, 1980; Aitkin & Longford, 1986; Wilms and Raudenbush, 1989; Goldschmidt, et. al, 2010). This handbook takes a pragmatic approach, recognizing that causal claims are limited, but this should not prevent stakeholders from holding schools accountable for the success of ELs. It is unlikely that conditions for unfettered causal claims exist; it is, nevertheless, worthwhile to evaluate the extent to which the ELP Indicator is related to confounding factors that blatantly impact the validity of inferences. Researchers agree that among the various attributes to examine in considering the quality of an indicator, validity is critically important (Porter, 1991; Profit, et. al., 2010; Lyons & Dadey, 2017). Validity is not related to the indicator (i.e. an indicator is not valid nor invalid), rather validity is a collection of evidence supporting the desired claims based on the results of the indicator (Messick, 1995). For example, an ELP Indicator based on the percent of students exiting the program is neither valid nor invalid; rather it is the claim based on the Indicator that is subject to validity evidence. A claim that ELs are making adequate annual progress in learning English based on this Indicator would likely not garner much validity evidence because this Indicator summaries an endpoint but provides little information on the path towards exiting. This is a traditional view of validity from the assessment literature, and it sets a reasonable framework for examining an indicator. The concept might be modified to focus on importance, usability, and support for inferences based on results (Profit, et. al., 2010), reliability (Goldschmidt et. al., 2012), and precision (Porter, 1991). As noted in Lyons and Dadey (2017), there are both policy and technical elements to consider.

The intended purpose of the ELP Indicator is two-fold. First, the state must consider how the indicator functions as a stand-alone indicator of EL progress. Secondly is how it provides additional information about schools as part of the overall accountability system. The potential impact of the ELP Indicator as part of the overall accountability system depends on the SEA Theory of Action (ToA) and relates to the discussion in Part I with respect to whether the indicator is intended to add breadth or depth to the accountability system (Baker, 2003). If it is the former, the SEA would expect overall school classifications to change based on how well schools facilitate language

development, and this impact will be governed by the relationship between how well schools can accomplish the various elements of the system. If the ELP Indicator is intended to add depth to the accountability system, the SEA is assuming that school processes are correlated and school classifications will not change substantially but that results will be more precise. Important first steps for integrating the ELP Indicator into the overall accountability system include deciding when the ELP Indicator is included (minimum N) and how it is incorporated (weighting). The extent to which combining indicators is functional depends on the method of combining[37] (Martinez, et. al., 2016). Including the ELP Indicator in the overall accountability system is considered first, followed by several steps to check the intended functioning of the ELP Indicator.

Including the ELP Indicator into the Overall Accountability System

There are many methods for including the ELP Indicator into the overall accountability system. One consideration, as noted, is whether the measures provide greater breadth (Baker, 2003). Chester (2003) argues that combining different measures should be done using a conjunctive approach. Combining different measures using a (linear) composite assumes that the two measures are substitutes and that more of one compensates for less of another. Given the high stakes nature of SEA accountability systems, the idea is often to allow different measures to compensate because schools are likely to have different strengths and weaknesses. Conjunctive rules can be applied very differently (e.g. disjunctive) and they will lead to different classifications than composite rules (Martinez, et. al., 2016). A state combining different measures using a composite should first normalize scores (OECD, 2008), placing them on the same scale. As noted in Part I, converting measures that are percentages is simply a linear transformation and keeps indicators on the same scale, but this does not guarantee that each indicator behaves similarly. Deciding how the indicator functions relates to both the minimum N for inclusion of the ELP Indicator and how the ELP Indicator is included in the overall accountability system.

Minimum N

The minimum number of ELs in a school before the ELP Indicator is included in a school's overall accountability determination has several ramifications. One, minimum N impacts how many schools and EL students will be held accountable for EL progress. Two, minimum N impacts the reliability of a school's estimate. And three, minimum N

[37] Martinez et., al., (2016) examine multiple measures with respect to teacher evaluation, but the analyses on different means of combining apply here as well, particularly as states consider different weights as part of a composite and (in some instances) consider conjunctive (dashboard-like) rules for combining indicators.

interacts with the underlying growth model selected to monitor progress. Figure 18 displays the relationship between minimum N and students, schools, and the percent of ELs in included schools. The percent of ELs represented in the state accountability system is fairly robust to variation in minimum N; however, the percentage of schools decreases rapidly as minimum N increases. The SEA must decide how much school representation is reasonable in a school-based accountability system. Figure 18 presents three distinct percentages and how they relate to minimum N. The (blue) line, "percent of schools," represents the percentage of schools in the state that would include an ELP indicator at the corresponding minimum N. For example, at a minimum of 20 about 33% of schools, that have at least one EL student, would be eligible to include the ELP indicator.

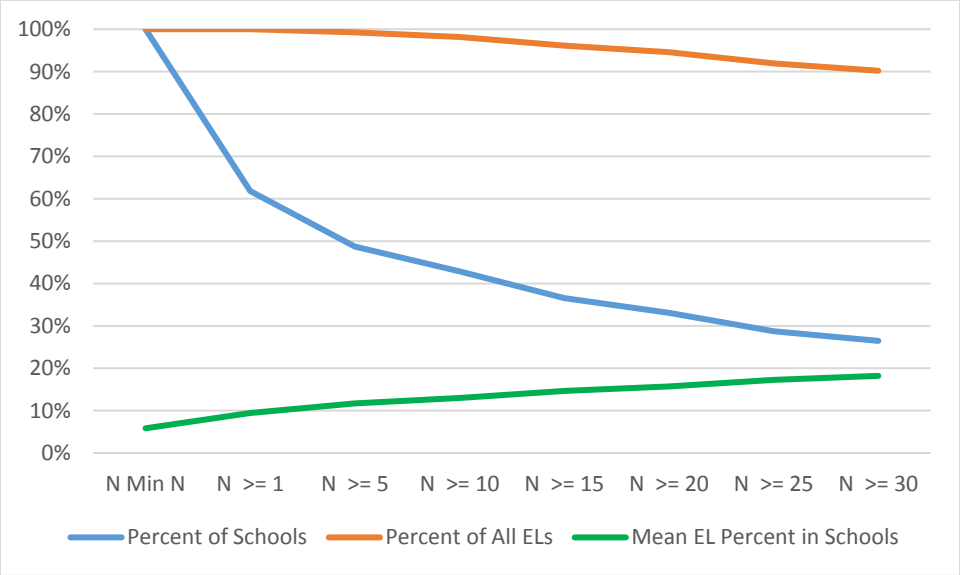


FIGURE 18: EFFECT OF MINIMUM N ON STUDENT AND SCHOOL REPRESENTATION

The percentage of schools eligible to include the ELP indicator generally decreases very rapidly with increasing minimum N. The (orange) line, "Percent of All ELs," represents the percentage of all of the state's EL who would be represented by the ELP indicator at a given minimum N. In contrast to schools, the overall percentage of ELs included in the system decreases very slowly as the minimum N is increased. For example, at a minimum of 20, about 93% of the state's ELs would still be included in accountability system for EL progress. The (green) line, represents the average percent of ELs in a school among schools that meet a particular minimum N. For example, if the minimum N is set to 20, then the average percent ELs in schools meeting this minimum N is 16% - compared with an overall state average of 6%. Figure 18 is based on the number of ELs in the current accountability year. The percentage of schools meeting the minimum N is also affected by the requirements of the growth model. For example, if minimum N is set to 20 and the state uses gain scores, then this likely reduces the number of schools meeting the minimum N because fewer will have a current and a prior score than simply a current

score. More sophisticated growth models requiring more prior years of data will further reduce the number of schools meeting the minimum N requirement. An ELP indicator that uses several criteria can mitigate this reduction in participating schools by tracking both annual growth and target levels. The benefit of annual target levels is that students with a missing prior score can still be included.

Minimum N and school inclusion is driven by state specific context and is likely one of the biggest challenges faced by states. The absolute number of ELs in a state impacts what Figure 18 might look like as well as the distribution of ELs. The issue becomes how to handle the majority of schools serving ELs in schools not meeting the minimum N. Several options exist[38]. One option is to use an aggregation rule until the minimum N is reached (within a certain timeframe; for example, two or three years). Using multiple years changes the inference about the school—instead of asking how the school impacts English language development in the previous year, the question is instead how the school has impacted English language development over the past three years. This may seem like an obvious distinction but it should be noted that the former is based on many different students coming and going, while the latter will be more dependent on the particular characteristics of a cohort. This is a good option for an SEA to consider accountability for small schools. Another option is to include district level accountability and either report the district level or assign schools district-level results. A third option is to report separate EL results consistent with the minimum N, and create business rules that allow the ELP Indicator results to be included as long as N is equal to one. This latter option is discussed in the section on weighting.

Additional options include aggregating scores back to the district or regional service area. Hence, the business rules might be that a school is first checked to see whether it meets the minimum N and if not, secondary rules such as aggregating over years, aggregating to all elementary schools in the district, aggregating to all schools in the district, aggregating among feeder patterns, or aggregating to regional service areas are all options that could either be explicitly included in the accountability model or part of the reporting system.

Additional considerations relate to reliability and validity. Reliability is detailed in *Properties of the ELP Indicator at the School Level.* It is important to reiterate that validity when it comes to an ELP Indicator is not a question of valid or invalid; the focus is on the evidence available to support whether or not claims based on indicator results are consistent with intentions. Minimum N relates to claims in that the number of students

[38] Although the extent to which they are allowable under ESSA or whether a waiver is possible is not clear.

contributes to the functioning of the indicator. For example, a ten percentage point increase in the number of students meeting growth targets might be viewed differently depending on whether it is based on ten students or 100 students. The former change represents one student improving while the latter represents ten students. Conversely, if any specific classification is based on such changes, then a school may fall into a category on the basis of a single student. Smaller minimum Ns introduce the potential for rival hypotheses to reduce the internal validity of claims regarding schools' systematic impact on the outcome.

This latter notion, of reducing internal validity, is also linked to a state's view of whether accountability results are based on a population or on a representative sample (Seastrom, 2017). If the accountability system is assumed to be based on the population, then a state is not using any measures of uncertainty around point estimates of performance, and meaningful differences, gains, or growth are strictly policy decisions. If a state views each year's accountability results as a sample (each year is a sample of students from all students over time who form that specific group as a basis for accountability in a school), then confidence intervals inform uncertainty related to sampling[39] and provide guidance for establishing meaningful differences, gains, or growth[40]. A population or sampling approach not only affects whether uncertainty in results is considered in the system, but also substantively affects how likely a small school is to demonstrate meaningful differences. Assuming the level of a meaningful difference is held constant, the chances that a school meets a threshold are less likely under a sampling framework than a population framework. This is true until the confidence interval is less than the value set as the meaningful difference (Seastrom, 2017). The IES handbook on setting minimum Ns (Seastrom, 2017) for accountability details additional considerations and steps to selecting a minimum N.

Weighting the ELP Indicator

Determining how to include the ELP Indicator generally consists of two steps: one, deciding on how the ELP Indicator will be included in the overall accountability system; and two, how much weight the ELP Indicator should receive. As noted earlier there are different means for including an indicator in a multi-indicator system. How the indicator is included depends on the state's Theory of Action. Weights are generally a reflection of importance, and importance can be expressed in the case of the ELP Indicator by the aligning it with the proportion of students a school is held accountable for on the ELP

[39] Additional uncertainty also exists due to measurement error in assessments.

[40] A state could base this on statistically significant differences and that determination is based on setting type I error rates (α), and whether it is based on a directional or none-directional hypothesis.

Indicator. Another view of weighting, however, is to weight according to the inverse of the correlation among the indicators in order to reduce double counting (OECD, 2008). Reducing the weight of highly correlated indicators attempts to reduce redundancy and is consistent with Baker's (2003) notion of multiple indicators providing breadth of information. However, relying on only a single indicator of a set of highly correlated indicators potentially results in the surrogate fallacy (Kane and Case, 2004). For example, language arts and mathematics results are highly correlated, and it is reasonable to use only one indicator of academic performance in the accountability system. However, if mathematics is excluded, then this may cause less emphasis on mathematics and more emphasis on language arts – likely resulting in a looser coupling between language arts and mathematics. Weighting considerations are important because they reflect the importance stakeholders place on the components of the accountability system. This importance is operationalized through the ToA.

The ELP Indicator may be made up of multiple measures. Combining these measures into a single ELP Indicator is similar to the discussion of combining the indicators into an overall accountability system. One option is particularly germane to the ELP Indicator - combining both progress and exiting into a single value. This can be accomplished in different ways – including calculating each component separately and combining into a composite score. Consistent with any composite index, it is advisable to normalize results before combining in order to reduce level and variance effects.

The option to combine both progress and exiting into a single calculation results in equation (4), which several states are considering:

$$\text{Score} = (N_{\text{meet progress}} + N_{\text{exit}})/(N_{\text{assessed}} + N_{\text{exit}}). \quad (4)$$

Equation (4) counts the number of students meeting the progress target divided by the number of students who were assessed (specifically who have two assessment occasions). The number of students who exited is added to both the numerator and denominator, which will increase the overall resulting score (percentage). The impact of this is not uniform and depends on both the percent of students meeting the progress target and the proportion of students of the total who exited. For example,

$\text{Score}_1 = (50+5)/(100+5) = 55/105 = .524$, whereas
$\text{Score}_2 = (50+10)/(100+10) = 60/110 = .545$.

Figure 19 presents a range of impact across various probabilities of meeting the progress target. The denominator in Figure 19 is the number of who have two assessment occasions plus the number of exited students. For example, for Score_1 and Score_2 the

percentage of exited students equals 4.8% (5/105) and 9.1% (10/110), respectively. Figure 19 indicates that when lower levels of students meet targets the impact of including exited students will be higher. Figure 19 presents five scenarios relating to the percent (p) of students meeting progress targets. The results indicate that for lower p, the greater the impact of including exited students.

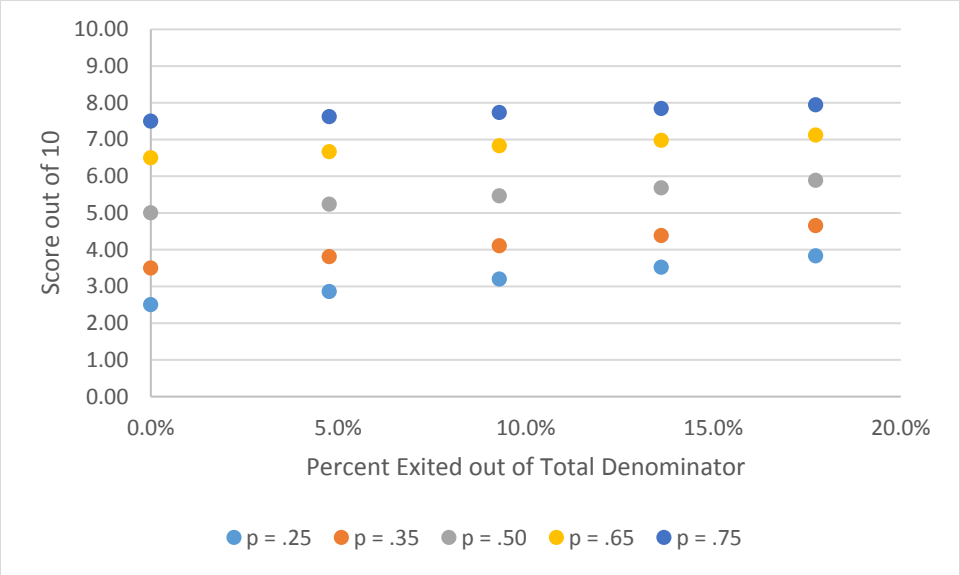


FIGURE 19: POTENTIAL ELP INDICATOR POINTS WHEN INCLUDING EXIT STATUS TO PROGRESS

Including exited students in this way also impacts the relationship between school input factors and the ELP Indicator result. The discussion in Part I, related to whether Off Track[41] ELs should count towards partial or extra credit for a school, is implicit in Equation 4, in that all exited students, whether early, on time, or late represent equal credit for a school.

Some SEAs intend to use a dashboard approach which implies a conjunctive model. The SEA is taking the view that each indicator provides additional breadth about school processes (Baker, 2003). Using a conjunctive model does not automatically mean that each indicator contributes equally to success or failure because the distribution of performance of each indicator is likely different. For example, if the accountability system contains two indicators (percent proficient and EL progress), then the probability of success (passing, etc.) is equal to the joint probability of meeting the percent proficient standard and the ELP Indicator standard. If the probability of meeting the percent proficient standard is 0.5 and the probability of meeting the ELP indicator standard is 0.9, then percent proficient has a greater impact on the joint result because it is more difficult

[41] Off Track refers to students who have exceeded the expected time to exit.

to meet the percent proficient standard (this may well reflect a state's Theory of Action). The same is true when business rules focus on not meeting a standard. For example, schools that do not meet the standards associated with both Indicators are eligible for comprehensive support (e.g. CSI). The SEA can set these probabilities for the indicators in such a way that the probability of not meeting all standards is .05, which would result in approximately 5% of schools failing to meet all the standards. Importantly, this is not accomplished by setting the standard for each indicator at 5% and identifying schools that conjunctively meet the 5% threshold, because this does not result in an overall 5% probability. The following paragraph provides an example of how conjunctive probabilities work.

A state using this approach would want to define the joint conjunctive or disjunctive probabilities of success or failure to determine whether these are in line with expectations. For example, if a state intends to have three classifications: superior, average, and needs support, and if schools are held accountable for performance on four indicators (percent proficient, growth, EL progress, and student OTL), and if the probability of meeting the benchmark on each of the four indicators is 0.5, then the probability of a school being classified as superior is 0.0625 (0.5 x 0.5 x 0.5 x 0.5). Conversely a disjointed model identifying only schools eligible for support because they did not meet all of the standards is 1-p (1 - percent meeting the standard), which in this case results in a 0.0625 probability of missing all of the standards. It should be noted that it will not always be the case that the two rules will result in equal probabilities. For example, if the probabilities of meeting the indicator standards are 0.5, 0.6, 0.4, and 0.7, then the probability of being classified as superior is equal to 0.084; while the probability of being classified as needs support is equal to 0.036.

Another approach to incorporating the ELP Indicator into the overall accountability system is to form a composite index. States are likely to use a linear weighting scheme, which implies that performance on one indicator can compensate for performance on another[42]. The weights indicate the desired rate of substitution. There are several options for determining the weight of the ELP Indicator, many of which are discussed in (Martinez, et al., 2016). One option is to use a policy-based weight. This can be derived from stakeholders and also based on the SEA's Theory of Action. Policy weights can be informed by the representation of ELs in schools. Figure 18 provides some guidance. Note that the percent of ELs, on average, in schools that meet the minimum N

[42] Using geometric mean weighting provides an alternative to linear weighting that has the advantage of creating an incentive for schools to improve on the indicator that is performing most poorly—similar to incentives for a conjunctive model. However, this method is less transparent than linear weighting. More detail is available in the OECD Handbook on Composite Indicators (2008).

requirement is more appropriate to consider than the overall percent of ELs in the state. Another policy, common weighting scheme, is to weight each indicator equally; although this is not possible under ESSA. Weights can also be derived empirically. Detailed guidance for more advanced empirical weight derivations are found in the OECD Handbook on Composite Indicators (2008). Correlations among the indicators can also serve as a guide, with some argument given for allocating lower weights to indicators that are highly correlated (Saisana, Saltelli, & Tarantola, 2005).

A third approach is to weight according to representation. In this instance, the weight itself is not fixed numerically, but rather it is fixed through an equation: e.g., the weight is equal to N_{ELj}/N_{totalj}, where N_{ELj} = the number of ELs in school j and N_{totalj} = the total number of students in school j. The rationale behind this approach are twofold: one, it allows small N sizes (down to a single EL student) to be included in accountability; and two, it likely improves the validity of inferences based on EL progress results. For example, Seastrom (2017) explains how inferences based on performance changes may be suspect when the changes are not considered in context. Specifically, if there is an 11 percentage point drop in performance, but this drop is based on only two fewer students meeting the target, the notion that the drop in performance meaningful is may not be justified. Similarly, if EL students make up 5% of the school's population while the indicator is weighted 10%, then each student's performance is over-weighted. If ELs represent 30% of the school's population and the ELP Indicator weight is 10%, then EL student progress is under-weighted.

An extension of using a formula to weight rather than a value is through business rules that aggregate the various indicators. Generally indicators are aggregated by multiplying the indicator by the weight and summing the multiplied indicators (if indicators are equally weighted, one could sum the indicators and divide by the number of indicators). Options include weighting and adding the ELP Indicator into the other accountability indicators, or adding the ELP Indicator into growth and then weighting the growth indicator as a whole; these two options compare a method that is not sensitive to the proportion of ELs in a school to one that is sensitive to the proportion of ELs in a school[43]. The results will diverge, which may be in line with a state's Theory of Action because the overall result and weight of the ELP Indicator is in line with the representation of ELs in a school.

[43] This example assumes there are only two indicators: content growth and the ELP Indicator (EL progress). Additional indicators would dampen the impact since their (the additional indicators) contribution remains fixed – the smaller the proportion of the composite that content growth and the ELP Indicator are, the smaller the overall impact presented here.

Figure 20 shows the results of comparing the results using fixed weighting versus formula weighting under three different target scenarioes. The results in Figure 20 assume that the minimum N = 20 in the example corresponds to 20% ELs. Lines represented with a (1) are based on a fixed weighting of 0.4 for content growth and 0.1 for the ELP Indicator; also, it is assumed that ELs' content growth is equal to Non-EL content growth, which is approximately true in the sample data (see Figure 11). The lines represented by (2) are based on formula weighting where n meeting EL progress and total EL n are added to n meeting content growth and total n for content growth.

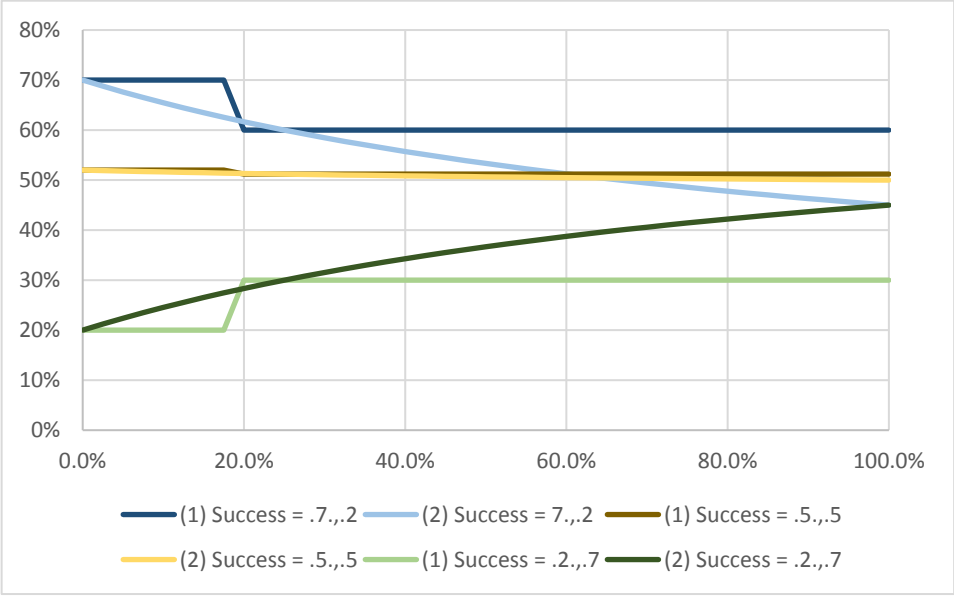


FIGURE 20: THE IMPACT OF FIXED WEIGHTING VS. FORMULA WEIGHTING

The results in Figure 20 demonstrate that if the probability of meeting the content growth target is 0.7, the probability of meeting the EL progress target is 0.2, and if the model is using fixed weighting (dark blue line (1)), then a school's accountability result with fewer than 20 students depends only on content growth; 70% of students meet the growth target and so the school earns 70% of the possible points. At an EL percent of 20, the indicator now also includes EL progress; given that progress is much lower, the school earns 60% of the points. However, because the weight for ELs is fixed, even when the percentage of ELs increases, and progress (in aggregate, content progress and EL language progress) is presumably even slower, the school still earns 60% of its points.

The light blue line (2) uses the same assumptions (probability of meeting content growth = 0.7 and probability of meeting EL progress = 0.2), but uses formula weighting. It is useful to examine the two end points of the light blue line (2) to understand what is

occurring. When there are no ELs in the school, the school earns 70% of the points possible, because 70% of the students (all non-ELs) met their growth target. Schools with 100% ELs earn 45% of possible points. This is due to the fact that even though all the students in the school are ELs (and assuming no RAEL), these students continue to take both the content assessment and the ELPA. Since they have the 0.7 probability of meeting their content growth targets, and a 0.2 probability of meeting their EL progress target, the overall probability of meeting both targets is (0.7 + 0.2)/2, which is 0.45. The light blue line summarizes the relationship between a school's points earned and different percentages of ELs in a school.

At the ELP Indicator weight of 0.1, a school would earn about 60% of possible points under (1) and 62% of possible points under (2). The greater the weight assigned to the ELP Iindicator weight, the larger the divergence. The gold lines show that the more equal the probabilities of meeting the growth and EL progress targets, the more similar numbers of points are earned under each scenario. Finally, the green lines depict a scenario where there is a much greater probability of meeting the EL progress target than the content growth target. The dark green line p(success for: content =.2, EL progress .7) in a school that is 100% EL also results in 45% of points earned because ELs take both the content assessment and the ELPA and so the combined success is (.2+.7)/2) = .45.

There is no a priori reason to believe that an indicator needs to be weighted in direct proportion to the percentage of students contributing to that indicator. This might be especially true if the indicator were highly representative of school "success", a particularly important policy objective, or highly related more distal or less easily measured outcomes. However, if the purpose of the indicator is to provide breadth, then a starting point for determining indicator weight may be the percentage of students represented by that particular indicator in a unique way.

Another option that is used for composite indicators is geometric mean weighting (OECD, 2008). Geometric mean weighting is less transparent than simple composite weights, but results in an indicator that incentivizes improvement on the school's lowest performing indicator. If a school were performing least well on the EL progress indicator, then that is what the school would want to improve under a geometric weighting scheme. Linear composite weights (whatever the weights may be) assume that more of one indicator is an equal substitute for less of another. For example, under a linear weighting scheme where status is weighted 0.4 and EL progress is weighted 0.1, a school could focus on improving status by ¼ of what it needs to improve EL progress in order to obtain the same overall impact. However, one cannot easily substitute in this manner using a geometric mean. The geometric mean is also useful when the indicators are on different

scales. One example of a geometric mean-based composite is the Human Development Index (OECS, 2008).

The geometric mean is:

$$GM = \sqrt[n]{(I1 \; x \; I2 \; x \; \ldots x \; In)} \qquad (5)$$

Where, GM is the Geometric Mean, *n* the number of indicators, and I_1, I_2, I_n are Indicator *1*, Indicator *2*, and Indicator *n*, respectively. For example, if an SEA wanted to apply a quasi-dashboard system to make annual determinations for CSI (Comprehensive School Improvement), their options are:

1) Conjunctive rule not in bottom 5% in any two out of three categories.
2) Composite mean of percentile rankings.
3) GM of percentile rankings.

Option 1 requires the use of the binomial probability formula, ${}_nC_r \times p^r \times (1\text{-}p)^{(n-r)}$[44], where n (in this case) is the number of indicators, r is the number of indicators a school must miss/meet (two or more for example), and p is the probability aligned to the business rule (5% in example). Using this set of rules there is a 99.2% chance that a school would not meet the criteria for CSI.

Table 15 summarizes the results comparing an equally weighted composite (*Equal Weight – Simple Mean*, a not-equally weighted composite (*Weighted – Simple Mean*), and a composite using the geometric mean (*Geometric Mean*).

TABLE 15: COMPARISON OF DIFFERENT COMPOSITE CALCULATIONS AND EFFECTS

Indicator	Equal Weight - Simple Mean			Weighted - Simple Mean				Geometric Mean		
Status	75	82.5	75	0.5	37.5	41.25	37.5	75	82.5	75
Growth	50	50	50	0.4	20	20	20	50	50	50
EL Indicator	25	25	27.5	0.1	2.5	2.5	2.75	25	25	27.5
Composite	50.0	52.5	50.8		60	63.75	60.25	45.4	46.9	46.9

The results in Table 15 assume that a school scores 75, 50, and 25 on Content Status, Content Growth, and EL Progress, respectively. This could be based on percentiles (i.e. aggregate SGPs or some other score). The school's equally weighted mean score is 50. If Content Status were to increase by 10% (second data column), the school's overall Composite score would increase by 5%. However, if the school's EL Progress increased by 10% (third data column), then the school's overall composite would only increase by

[44] Thus with a .05 probability of being in the bottom of each indicator, the probability of being in the bottom of 0 or 1 indicator out of 3 indicators is: $3!/0!(3!) \times .05^0 \times (1\text{-}.05)^3 + 3!/1!(2!) \times .05^1 \times (1\text{-}.05)^2$.

0.4%. This situation is similar when a more common weighting scheme of 0.5/0.4/0.1 for Content Status, Content Growth, and EL Progress, respectively, is applied. In that case, if Content Status improves by 10%, the overall composite increases by 6.25%, while a 10% improvement in EL Progress leads to an overall composite increase of 0.42%.

The last three columns of Table 15 demonstrate the use of the geometric mean. In this case, when any indicator increases by 10%, the overall composite increases the same 3.1%. In this way, the school has an equal incentive to improve each indicator equally. A geometric mean eliminates the compensatory nature of being able to trade improvement across indicators. The tradeoff for this egalitarianism is less transparency because the results are not simple sums or averages.

There are many options for combining indicators to form a composite, and using formula weighting rather than a fixed weight creates an overall score that is more sensitive to the mix of students (and assessments) forming the composite. However, the tradeoff is that the weighting scheme becomes less transparent and may be unnecessary if the proportion of ELs in schools is within a relatively small band. A more transparent version of formula weighting might be to develop business rules that categorizes schools into low, medium, or high, number/percent of ELs and then apply different weights to each category. This system is easier to communicate but the SEA should carefully consider what happens with school scores for schools on the cusp of classifications.

Whichever rules are used to include the ELP Indicator into the overall accountability system, the SEA should examine whether the indicator is neutral (Lyons and Dadey (2017). Schools subject to the ELP Indicator should not be advantaged nor disadvantaged by the way in which the ELP Indicator is part of a school's designation. As noted, is the intent of the indicator to identify schools specifically on performance of this indicator (breadth) or is the indicator designed to add precision to overall results (depth)? It is particularly important to consider the impact on schools of including another indicator in a conjunctive model because overall success (or failure) on a conjunctive model is multiplicative.

Systems using composite weights must still examine the impact of the presence or absence of the ELP Indicator. This includes specifying business rules for weighting when the ELP Indicator is absent. For example, will the points (or weight) assigned to the ELP Indicator be equally distributed among the remaining indicators or to a specific subset (e.g. status and growth, or growth, etc.)? It should be clear that assigning points to indicators that are less "difficult" disincentivizes including the ELP Indicator.

There are several checks the SEA can undertake to examine whether the ELP Indicator is neutral; meaning that, all else equal, including or excluding the ELP Indicator has no impact on a school's overall performance classification. Unfortunately there are no independent criteria with which to test whether the ELP Indicator or the other indicators are unequivocally accurate; the series of checks below provide first approximations of gross inequities. A straight-forward check is to compare overall (composite) scores, classifications, or ranks for schools including the ELP Indicator and schools not including the ELP Indicator. The average difference between these two sets of schools is confounded by the fact that there is no way of knowing whether they may in fact perform equally well except for progress on English language development. A further refinement is to compare these two sets of schools on the aggregate of all indicators excluding the ELP Indicator. This also provides a picture of how schools with the ELP Indicator perform on the other elements of the system and whether simply adding the ELP Indicator impacts that result. Calculating the mean absolute difference in school ranks (Saisana et al., 2005) provides an indication of how much schools' statuses change when the composite is altered. As noted, however, there is no guarantee that a school should perform equally well on all of the indicators.

A simulation provides at least a starting point to help understand the potential outcomes of reassigning ELP Indicator points. Using the assumptions in Figure 20, it is clear that two schools that are equally successful in meeting growth and meeting EL progress will have very different overall scores (using fixed weights) if one has less than 20% EL and the other has 20% or more ELs. This assumes that each component (content growth and the ELP Indicator) contributes points based on the percent of students meeting their targets. Using fixed weights, the SEA might want to use an adjustment for the ELP Indicator. Given the probability of meeting the EL target is 3.5 times more difficult than meeting the content target (this assumes that all the EL points are moving between the ELP Indicator and content growth.) If the ELP Indicator points are distributed among all the indicators then the adjustment becomes more difficult to make. The more divergent the ELP Indicator success rate is from the other indicators, the greater the impact of moving from not including the indicator to including the indicator. To avoid this problem, states can standardize the indicators before applying weights.

Another check is to examine correlations between the ELP Indicator, overall scores, and school input characteristics. This provides a picture of whether there is a relationship between the number of ELs and indicator results (percent of ELs is also informative). A useful plot is overall and ELP Indicator scores against EL N size. The analyst should look for discontinuity around the minimum N (this is discussed in more detail in Part III). Table 16 presents the overall composite score based on three indicators: Status (Percent Proficient); Content Growth; and EL Progress. The results in Table 16 summarize mean

school performance based on various weighting schemes and assume a maximum score of 100. Each weighting scheme consists of two parts: – the weighting of the three indicators if a school meets the minimum N for including the ELP Indicator and the weighting of the two indicators if a school does not meet the minimum N for including the ELP Indicator. In this example means are based on the percent of students meeting targets in each of the indicators; i.e., status – the percent of students proficient, content growth – the percent of students meeting content growth targets, and EL progress – the percent of students meeting EL progress targets. The results indicate that overall, state mean performance is affected by changing the weighting as indicated in Table 16. However, scores do exhibit the expected patterns. Weighting EL progress more raises mean school scores a small amount, as does weighting content growth more heavily. In combination weighting both EL progress and content more heavily results in the highest scores. The overall composite score may not matter if results are relatively similar because decision cut points will be based on scores after weighting decisions are made.

TABLE 16: COMPOSITE SCORES USING DIFFERENT WEIGHTING SCENARIOS.

		Weighting: Status/Content Growth/EL Progress					
Composite	w/EL	.5/.4/.1	.45/.45/.1	.4/.4/.2	.5/.4/.1	.45/.45/.1	.4/.4/.2
	wo/EL	.6/.4	.55/.45	.6/.4	.5/.5	.45/.55	.4/.6
Mean		39.7	40.3	40.2	40.4	41.0	41.5
N		404	404	404	404	404	404
S.D.		12.2	11.2	11.8	11.0	10.2	9.7

It is important to note that in this case, the standard deviation of scores varies and is significantly smaller under the 0.4/0.4/.2 weighting scheme than under the 0.5/0.4/0.1 weighting scheme.

Example 1: Evidence-Based Policy Decision

Ensuring weighs function as intended is an important step in understanding how indicators work together to generate a composite score. There are many possible weighting schemes and several methods to check whether the weighting scheme results in a pattern of outcomes consistent with the SEA ToA. Examining the robustness of the composite can be intensive since this can involve N(N-1)/2 pairs of schools. Lack of robustness is shown when the ordering of any indicator incorporated in the summative score orders schools differently than the summative score (Foster, McGillivray, & Suman, 2013). Kendall's Tau provides a proxy for robustness. As a starting point for the SEA using a composite, they should weight each indicator equally and examine the correlations among the indicators and the summative scores. Indicators may not be

highly correlated with one another but may be moderately-to-highly correlated with the summative score. Given equal weighting, an indicator with a low correlation is a result of the property of the indicator (e.g. lack of variability) which may or may not be a desirable property of the indicator. For example, the correlation (using equal weights) between the status indicator and composite summative score is 0.67, while the correlation between the Content Growth Indicator and the composite summative score is 0.2. This difference is due to the fact that the standard deviation of the Status Indicator is over two times the standard deviation of the Growth Indicator. Table 17 presents the comparisons of the correlations for raw indicators and a raw composite as well as normalized indicators and the corresponding composite based on the normalized indicators. In Table 17 the composite is based on equally weighted indicators.

TABLE 17: KENDALL'S TAU CORRELATIONS

Indicator	Equally Weighted	
Overall	Raw	Normalized
Status	0.66	0.57
Content Growth	0.17	0.16
EL Progress	0.28	0.53
Min N < 20		
Status	0.78	0.57
Content Growth	0.19	0.39
EL Progress	-.010	-.050
Min N >= 20		
Status	0.57	0.57
Content Growth	0.16	0.26
EL Progress	0.53	0.46

The results in Table 17 highlight the impact that the lack of variability of the Content Growth Indicator has on its relationship with the composite score. This is most clearly seen in the correlations corresponding for Min N < 20, because the ELP Indicator is excluded from these schools' composite. The composite for these schools is driven to a large extent by the Status indicator. It is also important to note that for schools that are held accountable for the ELP Indicator (Min N >=20) the ELP Indicator and the Status Indicator demonstrate roughly equal relationships with the Composite – Content Growth, is relatively unrelated to the composite. Normalizing results provides the least ambiguous picture of the relationships among the Indicators and the composite. Status and EL Progress appear to be related as intended while Content Growth continues to demonstrate less influence on the Composite than, perhaps, is desired.

The relationships among indicators allows for the calculation of the effective weight of each indicator; these results corroborate the conclusions based on Table 17. The effective weight of an indicator depends on the nominal (policy) weight as well as the correlation among indicators. The equation is ((Schochet, 2008) :

$$EW_j = WN_j^2 + \sum WN_j WN_k \rho_{jk}. \qquad (6)$$

Where EW_j is the effective weight for indicator j, WN_j is the nominal weight of indicator j, and ρ_{jk} is the correlation between indicator j and indicator k, where $j \neq k$.

TABLE 18: COMPARISON OF NOMINAL AND EFFECTIVE WEIGHTS

	Weights	
Indicator	Nominal	Effective
Status	0.5	0.62
Growth	0.4	0.36
EL Progress	0.1	0.02
Status	0.4	0.47
Growth	0.4	0.42
EL Progress	0.2	0.11

The calculations for Table 18 are presented in Appendix L. Given the correlations and weights among the indicators, status tends to have a greater effective weight on the composite than Growth or EL progress. If the three indicators were equally weighted the effective weights would be .355, .30, and .355 for Status, Growth, and EL Progress respectively – which is consistent with the pattern of normalized weight correlations in Table 17 that indicate status and ELP are equally related to the composite while content growth is least related to the composite.

CHECK

At this stage the SEA should have decided on business rules for including the ELP Indicator. This includes weighting, minimum N, and the distribution of the ELP Indicator in schools that do not meet the minimum N.

This handbook considers four steps[45] in checking the intended functioning of the ELP Indicator.

- Step 1: examine whether the ELP Indicator adequately identifies (meaningfully differentiates) schools in terms of EL progress,
- Step 2: examine whether the ELP Indicator is unbiased in aggregate,[46]
- Step 3: examine sensitivity of the ELP Indicator,
- Step 4: establish how well the ELP Indicator works over time.

Properties of the ELP Indicator at the School Level

Step one revolves around what happens when the ELP Indicator is aggregated. Although the preceding steps focused on individual student progress[47] in building the ELP Indicator, its primary purpose is in aggregate. An important caveat when using school-level means is that there is an implicit assumption that all of the variation in the ELP Indicator is between schools. The proportion of variation in the ELP Indicator (in this case, the individual student score on the indicator) that lies within and between schools can be estimated and will give an indication of the extent to which simply aggregating the ELP Indicator may be problematic—not in terms of the point estimates, but in terms of any statistical tests and inferences regarding statistical significance (this is addressed in more detail in Part III). The intra-cluster correlation (ICC) partitions the variation in the outcome (i.e. EL progress) into relevant components. In this case, the relevant components are the variation in EL progress attributable to students within schools and the variation in EL progress between schools. Previous research on state accountability growth model results indicates that the ICC varies by grade level and state (Goldschmidt, et. al, 2012). The same model applied to different data lead to differential model functioning. The ICC is calculated as:

$$ICC = S^2_b/(S^2_b + S^2_w) \qquad (7)$$

[45] There is significantly more research related to analyzing the quality of a composite set of indicators than a single indicator. If an indicator is based on multiple measures, it is possible to step through Messick's (1995) criteria for establishing validity evidence; however, if the indicator is based on a single measure (e.g. the percentage of students meeting the growth target), then much of the evidence relates to qualities of the underlying growth model (this has received significant treatment as well—see, for example, Goldschmidt, Choi, and Beaudoin, 2012).

[46] The previous discussion notwithstanding.

[47] It is possible that the ELP Indicator was developed to provide aggregate results directly, such as from a model that includes schools as fixed or random effects. These models can address issues with simple aggregates noted above.

Where S^2_b is the between-school variance in growth and S^2_w is the within-school variance in growth. The ICC is readily calculated from the GLM/variance components function or by running a two-level random effects model. Sample SPSS code is provided and the calculation for the ICC is in Appendix G. Based on the data, the ICC is approximately 0.1, which would be considered medium in terms of size (Hox, 2002). This calculation is useful in that the equation (7) can be modified to yield the reliability of each estimate. In this case, reliability refers to how good the sample mean is as an estimate of the population mean for each school. Appendix G demonstrates this calculation. Table 19 presents the reliabilities of the ELP Indicator for different school Ns. When the mean school size is 50, the average reliability of the ELP Indicator is .85 which is high. Reliability is a function of N, so smaller Ns lead to lower reliability while larger Ns lead to higher reliabilities.

$$\text{Reliability}_j = S^2_b/(S^2_b + S^2_w/n_j) \qquad (8)$$

Which is the reliability for school *j* (that consists of N students)[48]. Table 20 summarizes the reliabilities based on different growth or progress models. Current ELP level is included as a basis for comparison. The different models yield relatively similar results, and all improve with larger Ns.

TABLE 19: RELIABILITY OF SCHOOL ELP INDICATOR MEANS

n_j	Reliability
10	0.52
15	0.62
20	0.69
30	0.77
50	0.85
100	0.92

TABLE 20: RELIABILITY OF DIFFERENT PROGRESS MODELS

EL Progress Model	N = 10	N = 20
Current ELP Level	0.61	0.75
Met Growth Target	0.47	0.64
Cumulative Growth	0.49	0.66
VAM*	0.52	0.69
Aggregate SGP	0.42	0.59

[48] Given that the ELP Indicator may be based on a dichotomous outcome (met target/not met target), equations 7 and 8 need additional consideration. One option is to use logistic regression and replace S^2_w with $\pi^2/3$ in both equations, or alternatively, to use a linear probability model with the dichotomous outcome and proceed with equations 7 and 8.

*VAM = Value Added Model

Some researchers argue that reliability, per se, is not as important as the precision of the indicator results (Porter, 1991; Profit, et. al., 2010). The standard deviation of scores represents precision and provides an indication of how much variation there is in the estimate. Precision is an element of reliability but is not the same statistic. For example, El progress can be measured very precisely, but may still lack reliability when reliability is thought of as the ability to distinguish differences in true performance among schools based on observed mean performance. Specifically, we may estimate growth of 50 points per year with a standard deviation of 25 or a standard deviation of 10 (i.e. the within-school standard deviation). The former is more precise than the latter, but both may be unreliable. If every school's growth estimate is 50 points, then it would be impossible to distinguish schools based on growth (because there is no between-school variability in growth)—and reliability would be 0.

Given that under Title III accountability was at the district level, there may be a desire to differentiate timeframes or expectations at the district level. There are three factors to consider: 1) student expectations; 2) business rules at the school level that translate student progress into school points on the ELP Indicator; and 3) the amount of variation in EL progress attributable to districts.

Setting appropriate individual student progress targets or using an appropriate growth model at the student level accounts for much of the reason that districts might want to have differentiated targets. If the progress model, when aggregated to the school level, is related to initial ELD level, as presented in Table 23, then the SEA may want to revise the business rules converting progress into points[49] This approach likely captures the variation in district performance (or expected performance). A useful check is to determine the ICC for districts. The SEA can estimate the extent to which variation in EL progress is attributable to districts; and, the extent to which school ELP Indicator results are attributable to districts. For the former, the ICC uses individual student growth:

$$ICC_{District} = S^2_{District}/(S^2_{District} + S^2_{School} + S^2_{Student}) \quad (9)$$

[49] This can be accomplished by taking the proportion each initial ELP level represents in a school. This implies weighting school aggregate growth model results by initial ELP level or, for example, weighting by the expected proportion of students meeting growth targets given a school's distribution of initial ELP. This would be accomplished by using the values from Table 14. Hence, if 100% a school's current ELs initial ELP level was 1, then it would be expected that 86% of ELs would meet their growth target. One weighting scheme is to divide actual percent meeting the growth target by the expected (86%). In this way a school is not advantaged nor penalized for any specific distribution of initial ELP levels.

This is similar to equation 7. The latter analysis would calculate the ICC as in equation 7 where the $S^2_{District}$ refers to between districts variation, S^2_{School} refers to between schools, within a district, variation, and $S^2_{Student}$ refers to the between student, within school variation in ELP Indicator results.

The standard deviation of ELP Indicator results can be used to set classification cut points. The standard deviation is also useful for examining sensitivity of the indicator (discussed below).

Given that the ELP Indicator has acceptable reliability, the next property of the ELP Indicator that should be examined is whether the indicator is systematically associated with school factors that are beyond the control of the school. Analysts are accustomed to checking whether a school outcome is related to school input characteristics (e.g. the percentage of free/reduced priced lunch-eligible (FRL) students, or a school's percent proficient score).

The ELP Indicator introduces new sources of concern. Does the desired growth model result in different grades and/or school levels (elementary, middle or high) necessarily receiving different scores due to where schools are in the academic cycle as opposed to what language development schools are facilitating? For example, if a state uses some sort of a linear model, but growth is nonlinear (e.g. rapid early then slowing), middle and high schools will earn lower scores because students are less likely to reach expected growth targets.

Table 21 summarizes the relationships among input characteristics and the ELP Indicator. SEAs that create an ELP Indicator should reproduce this table for each measure and the ELP Indicator as a whole. In this example, the ELP Indicator consists of a single measure so only one table is produced. The relationships among student and school characteristics should be examined at both the individual and school levels. Checking for relationships at both levels is important in that there may be different dynamics in occurring at the individual level and in aggregate at the school level. For example,
Table 21 presents the mean values of several growth models for two student characteristics. The results indicate that students with disabilities (SWD) demonstrate consistently lower growth than their non-SWD classmates. The results for FRL students indicate that the models tested are robust to FRL status. The analyst will want to examine the relationship for other characteristics of interest such as RAEL or SIFE.

TABLE 21: COMPARISON OF GROWTH MODEL RESULTS BY SWD AND FRL

Student With Disability		Met Growth Target Table 11	Met Growth Target or Target Level	Domain Sum Score Gain	VAM	Met Growth to Target[50]
No	Mean	0.43	0.45	1.08	0.50	0.34
	N	7394	11138	11059	11043	11264
	SD	0.50	0.50	1.93	0.50	0.47
Yes	Mean	0.33	0.37	0.68	0.44	0.11
	N	1689	3990	4000	3975	4152
	SD	0.47	0.48	1.81	0.50	0.31

Free/Reduced Priced Lunch		Met Growth Target Table 11	Met Growth Target or Target Level	Domain Sum Score Gain	VAM	Met Growth to Target
No	Mean	0.42	0.46	1.06	0.51	0.30
	N	1279	2009	1998	1992	2078
	SD	0.49	0.50	2.01	0.50	0.46
Yes	Mean	0.41	0.42	0.96	0.48	0.28
	N	7804	13119	13061	13026	13338
	SD	0.49	0.49	1.89	0.50	0.45

Growth model results should not be related to initial ELP levels. If a relationship exists between initial ELP levels and the likelihood of reaching growth targets, then schools with students whose initial English language levels vary will be advantaged or disadvantaged based on a factor beyond their control. This may produce disincentives to accept certain students. Table 22 is based on a one-way Analysis of Variance (ANOVA) that uses initial ELD level as the between factor and tests whether ELP Indicator results vary significantly by initial ELD level. A first step when examining ANOVA results is to determine whether there is a significant F-Statistic, which indicates that ELP Indicator results vary significantly by initial ELD level[51]. An important second step is to examine

[50] Growth to Target is defined as Met Growth = yes (1) if current growth >= Target, and no (0) if not. The target for this mode is set as the difference between the exit score and the prior score divided by the number of years needed to attain the exit score. A student past the exit timeframe can meet growth targets if she meets the exit score.

[51] In table 22 all of the results are statistically significant. Using an ANOVA model with a large EL population will likely lead to a significant F-Statistic due to the large degrees of freedom.

the proportion of variation in the ELP Indicator accounted for by initial ELD levels (see Table 24).

TABLE 22: RELATIONSHIP OF INITIAL ELD LEVEL AND ELP INDICATOR BASED ON VARIOUS GROWTH MODELS

Growth Model		Sum of Squares
Met Growth Target Table 10	Between Groups	319.8
	Within Groups	1880.2
	Total	2200.0
Met Growth Target or Target Level	Between Groups	247.3
	Within Groups	3453.4
	Total	3700.7
Domain Sum Score Gain	Between Groups	1000.9
	Within Groups	53627.5
	Total	54628.4
VAM	Between Groups	8.7
	Within Groups	3742.0
	Total	3750.6
Met Growth to Target	Between Groups	124.36
	Within Groups	2978.01
	Total	3102.36

[1]*df* for domain sum and VAM is 4 due to calculating results for all initial ELD levels.

At the individual student level, the analyst will want to examine whether or not the number of years in program is related to the probability of meeting a growth target. This is a relevant check because even if students are on track, it should be equally likely that they can meet the growth expectations of each year. This is also a relevant check because schools that enroll students who have been in the program longer (e.g. middle and high school students), would then be advantaged or disadvantaged because of where these schools fall in the program timeframe. In order to provide all schools an equal opportunity to demonstrate their impact on language development, the relationship between program year and the probability of meeting a target should be minimized to the extent possible.

Figure 21 presents the relationship between year in program and the probability of success on the ELP Indicator (i.e. meeting the various progress targets). It should be noted that the results for "Met Growth Target Table 10" in Figure 21 only captures

students who are On Track and has a maximum of five program years[52]. Also the VAM result for Year in Program = 2 is identical to the result for "Met Growth to Target".

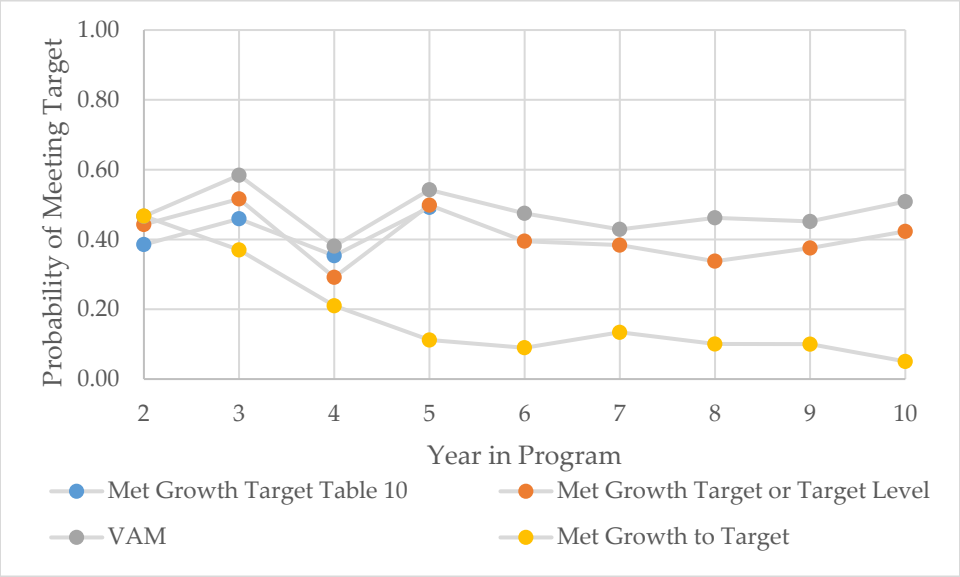


FIGURE 21: RELATIONSHIP BETWEEN YEARS IN PROGRAM AND PROBABILITY OF MEETING TARGET

If growth model results are related to input characteristics such as initial ELP levels or time in program, the growth model can still be used if the business rules apply growth model results to the accountability system, specifically addressing the issues. This is discussed in more detail in the section on incorporating the ELP Indicator into the overall accountability system.

The second step in checking the intended functioning of the in checking the intended functioning of the ELP Indicator is to examine school input characteristics and the ELP Indicator when aggregated to the school level. These are key analyses given the ELP Indicator is primarily intended to meaningfully and coherently provide additional breadth of information about a school's performance, in this instance focusing specifically on ELs' progress in attaining English language proficiency.

Ideally the ELP Indicator is neutral and there is a sense of how hard it is to move on one indicator vs. moving on another indicator (Lyons and Dadey, 2017). Assuming a school is equally "good" at facilitating student mastery of content and content growth as it is at facilitating EL progress, the school's accountability designation should not change with

[52] The value for year five "Met Growth Target Table 10" is virtually identical to the value for "Met Growth Target or Target Level," which is why it is difficult to see in the chart.

the inclusion or exclusion of the ELP Indicator. This assumption can be partly tested by correlating the different indicators, as shown in Table 23 (which presents the squared correlation coefficient) and examining how each of the indicators are related to various school input factors. If indicators are differently related to school inputs characteristics then neutrality is unlikely.

TABLE 23: PROPORTION OF VARIATION IN THE ELP INDICATOR ACCOUNTED FOR BY SCHOOL INPUTS

	Indicators			Composites	
				0.5/0.4/0.1	0.4/0.4/0.2
School Inputs	Growth	EL Progress	Status	0.6/0.4	0.4/0.6
FRL	0.01	0.01	0.05	0.08	0.08
SWD	0	0	0.21	0.19	0.15
Pct Non-White	0.03	0	0	0	0
Total_N	0	0	0.07	0.03	0.02
ELpct	0.01	0	0.02	0.03	0.04
EL_n	0	0	0.01	0.01	0.02
Growth_		0.01	0.01	0.02	0.15
EL Progress_			0	0.01	0.02
Status				0.91	0.7
Composite 0.5/0.4/0.1					0.9
Composite 0.4/0.4/0.2					

The results in Table 23 imply that schools are not necessarily equally "good" at EL progress as they are at other elements of the accountability system[53]. Tables 23 and 24 present the relationship between inputs and indicator/composite scores using R^2 because this relates to the standard error of the estimate (SEE) and a comparison to the SD of the unconditional indicator. That is, a low R^2 indicates that the input characteristic does not contribute a substantively meaningful amount in accounting for the variation of the indicator/composite. A rule of thumb is that if an R^2 less than or equal to .05, it is small enough to consider the input factor's contribution as trivial. Additionally, a small R^2 implies that the standard error of the estimate is not meaningfully reduced; hence, the input factor does not help improve the precision of the estimate (the CI (confidence interval) remains just as wide).

[53] Of course, this might be somewhat tautological because schools can be equally good at EL progress and content status and growth, but the ELP Indicator is simply not adequately capturing school performance on this measure. The SEA can take a step back and correlate the underlying performance using different models to ascertain whether it is the model or the process of facilitating EL progress and content mastery and growth. It is consistent with expectations that EL progress would not be highly correlated with content status.

TABLE 24: PROPORTION OF VARIATION ACCOUNTED FOR BY PERCENT OF INITIAL ELD LEVELS IN A SCHOOL

	Indicators			Composites	
School Initial EL Performance	Growth	EL Progress	Status	.5/.4/.1 .6/.4	.4/.4/.2 .4/.6
Pct ELD Level 1	0.01	0.06	0.18	0.14	0.08
Pct ELD Level 2	0.02	0.00	0.01	0.01	0.00
Pct ELD Level 3	0.01	0.00	0.01	0.00	0.00
Pct ELD Level 4	0.01	0.06	0.08	0.07	0.04
Pct ELD Level 5	0.01	0.00	0.14	0.12	0.08

Step 3 examines how sensitive the ELP Indicator is to changes weighting and N size. This can be accomplished by examining how ranks change among schools using different weighting schemes (Saisana et. al., 2005). An example of this type of analysis is presented in Table 25, which compares the means in absolute changes in ranks as a way to examine the impact of weights. In school accountability, weighting and EL N size form two dimensions on which to assess the impact of weights. Results in Table 25 assume a minimum N of 20. There are three indicators with weights of 0.5, 0.4, and 0.1 for content status, content growth, and EL progress, respectively. Schools that do not meet the minimum N for ELs are scored based on two indicators weighted 0.6 and 0.4 for content status and content growth, respectively (this is the 0.5/0.4/0.1 composite model). Comparison 1 (column 1) examines the impact of doubling the weight of the ELP Indicator from 0.1 to 0.2. Comparison 2 examines the impact of schools with and without the ELP Indicator (even if they are eligible). Comparison 3 examines the impact of more equal weighting (for schools not meeting the minimum N for ELs).

TABLE 25: DIFFERENCES IN RANKS USING DIFFERENT WEIGHTS

		Content Status/Content Gain/ EL Progress		
w/ EL	Wt =	.5/.4/.1	.5/.4/.1	.5/.4/.1
w/o EL	Wt =	.6/.4	.6/.4	.6/.4
w/ EL	Wt =	.4/.4/.2	.6/.4	.55/.45
w/o EL	Wt =	.6/.4	.6/.4	.55/.45
	EL N School	Absolute Change in Ranks		
	No ELs	8	8	10
	1 and 4	6	6	7
	5 and 9	7	7	7
	10 and 14	7	7	7
	15 and 19	5	4	7
	20 and 24	24	19	18
	25 and 29	18	15	17
	30 and 34	12	9	10
	35 and 39	15	13	13
	40 and 44	19	16	16
	45 and 49	27	18	17
	50 and greater	22	18	18
	Total	13	11	12

The results indicate that schools that do not meet the minimum N for the ELP Indicator remain very stable in their ranks (out of about 250 schools). However, schools that do meet the minimum N for the ELP Indicator tend to change ranks more substantially. This is clearly seen by comparing change in ranks for schools that have between 15 and 19 ELs and schools that have between 20 and 24 ELs. Changes in ranking is similar across different weighting schemes, but changes in ranking is very different by eligibility for the ELP Indicator. The SEA should be aware that schools meeting the minimum N for the ELP Indicator will be more impacted by weighting decisions.

Another way of examining the impact of policy is to model whether school summative scores change as a result of being on one side of the minimum N vs. the other. Given there is little evidence to support the notion that schools are equally skilled at facilitating EL progress and content status, a different approach to evaluating the impact is by assuming that schools are not inherently different in their ability to facilitate EL progress and content performance around the minimum N cut score. A Regression Discontinuity approach can be used to examine the impact of placing the minimum N cut score at 20, for example. The process for conducting this analysis is presented in Part III. The results

of examining whether schools are impacted by meeting (or not meeting) the minimum N of 20, indicate that school summative results are not affected overall.
Step 4, which examines stability requires a third year of ELPA results in order to calculate two years of progress with which to compare progress over time. Stability is addressed in the following section, Stability so it can more readily be teased out by the analyst.

Stability

SEAs and stakeholders place much emphasis on stability in school performance year over year. Most stakeholders believe that schools should maintain similar results from one year to the next. It is important to note that it is not desirable to see *no* movement in schools as this might indicate that the accountability system is not sensitive to changes in school performance. One common indicator of stability is the correlation of results from one year to the next - that is, the correlation of school summative scores between year 1 and year 2. This correlation may be attenuated because the same school's scores tend to be within a narrower range than scores among all schools in a given year, which attenuates the correlation.

Another means of examining stability is to see how schools move across classification categories (e.g. A-F, etc.). The number of classification categories is inversely proportional to stability. Also important is how schools move in the tails of the scoring distribution—this impacts how schools move in and out of CSI and TSI (Targeted School Improvemt). It is useful to examine movement across classifications by indicator. Convergent evidence of validity would dictate that the overall changes are corroborated by changes in individual indicators.

Generally, stability is higher when there are fewer classifications into which schools can be placed. Also, examining stability based on the composite vrsus consistency of classifications may lead to different claims about stability. Systems with few categories will be more stable than systems based on a specific score.

Movement across categories is affected by the models chosen and the business rules applied. Figure 21 displays the amount of improvement a school must demonstrate to move from the 5th percentile to the 10th and 25th percentile. Different progress models result in different requirements (or demonstrate more differentiation at the bottom of the performance distribution).

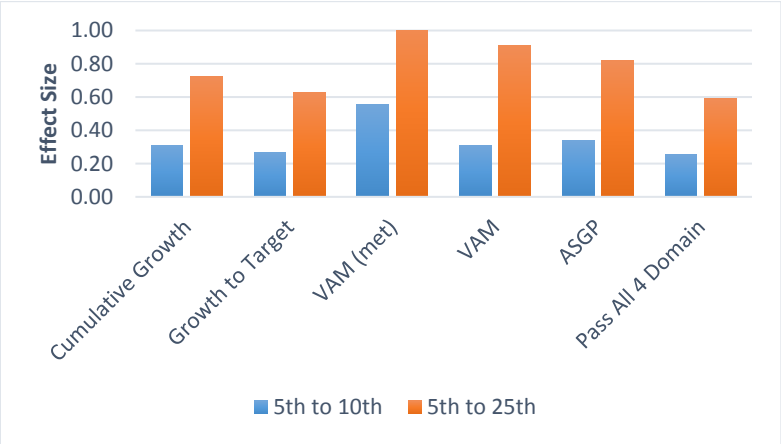


FIGURE 22: ELP INDICATOR PERFORMANCE DIFFERENCE BETWEEN PERCENTILES (IN EFFECT SIZE)

Figure 22 indicates that there are meaningful differences in the amount of improvement a school must demonstrate to move from the 5th to the 10th percentile of performance in the state. For example, a GTT (Growth to Target) model requires that a school improve by about 0.25 standard deviations, while a VAM (met/not met) requires a school to improve by about .55 standard deviations. On the one hand the VAM (met/not met) model provides much greater differentiation between 5th percentile and 10th percentile schools than the GTT model, but on the other hand this differentiation translates into substantially greater improvement required to move levels when using a VAM (met/not met) to monitor progress.

CHECK

At this point the SEA should understand whether ELP Indicator is robust to influence beyond a school's control. The SEA should also understand the impact of the ELP Indicator within the accountability system—whether it contributes as expected and whether claims made about schools align with the SEAs conceptions. For example, does the ELP Indicator result in incentives for schools to remain above or below the minimum N for inclusion of ELP Indicator?

Part III: Evaluating performance of ELs

Part III focuses on continuous improvement through evaluation. Evaluation encompasses a broad spectrum of activities that can broadly be categorized as qualitative and quantitative; consistent with Parts I and II of this handbook, the presentations focus on issues related to quantitative methods. Part I focused on technical issues associated with developing the underlying measures of the EL (English Learner) Progress Indicator, such as exit criteria, time to exit, and how to measure progress; Part II focused on technical issues associated with aggregating the progress measure into a school EL Progress Indicator and various caveats SEAs should consider in determining whether the indicator functions as desired; e.g. the indicator is not unduly influenced by factors beyond schools' control, the indicator is egalitarian with respect to school level, indicator weighting, and the indicator can monitor all ELs until they exit the program. Part III focuses on EL program success – this can encompass using both EL performance in terms of various educational outcomes and the EL Progress (ELP) Indicator. In the broadest sense, evaluating whether incorporating EL progress into Title I accountability has improved EL outcomes may be the first question the SEA is interested in addressing. This question can be addressed either by examining EL performance before and after the ESEA reauthorization, or by comparing schools that are explicitly subject to EL progress accountability (i.e. meet the minimum N) against those that are not. The former requires several years of data, while the latter can be approached through the regression discontinuity design discussed below. Additional specific programmatic evaluation foci are presented below and can be pursued by the SEA or by a district (and often districts have access to additional data that help inform evaluations).

Two important elements the SEA needs to consider in order to develop meaningful systematic evaluations of a program are the Theory of Action (ToA) and an evaluation policy. The ToA solidifies the SEAs conception of the processes that turn inputs into outputs. Importantly, a ToA provides the initial framework for evaluation because the SEA has provided its logic model for success, and it is this model that forms the basis for determining what may be subject to evaluation. The purpose of evaluation is not simply to certify success or failure; rather evaluation may have many goals, from a purely experimental model that compares treatment to control students to a responsive model that is concerned only with the specific time, place, stakeholders, and subjects and not concerned with any form of generalization to other settings. Evaluation can focus on effect size differences between groups to implementation fidelity to staff ownership of processes and outcomes. The varied potential of evaluations implies that the SEA should develop an evaluation policy that guides decisions and actions in developing and conducting evaluations (Trochim, 2009). Developing an evaluation policy is beyond the

scope of this Handbook, but this topic is addressed in detail in *Evaluation Policy and Evaluation Practice* (Trochim, 2009). Evaluation policy includes guidance on when evaluations should be conducted, who conducts them, what criteria should be applied, etc. For example, evaluation policy might dictate that programs are evaluated on a schedule that includes initial process evaluation after year one, fidelity of implementation evaluation during year two, and outcomes evaluation in year three. It should also be noted that cost-effectiveness analysis is least often conducted. Given the reality of fixed resources, this should be considered in some instances, and evaluation policy can provide guidance as to when this element of an evaluation needs to be included (e.g. after year four). An excellent guide to cost-effectiveness analysis is by Levin and McEwain (2001).

It is important to distinguish between accountability and evaluation. While the concepts are related, they are not entirely overlapping. The focus of accountability is on monitoring the performance of schools. An accountability system ought to hold schools accountable for their contribution to student learning and is to a large extent a summative score of performance compared to the previous year, other schools, or a fixed criterion. Accountability system results reflect a measurement process that selects and weights indicators in a manner consistent with the SEA's Theory of Action. Accountability results do not explain why a school meets expectations (aggregation rules not withstanding); rather system results provide a starting point for identifying potential reasons for the observed results[54]. Individual indicators creating the summative score provide some insight for better or poorer performance with respect to an indicator, but they do not provide information about the causes of that performance. For example, an accountability system should be sufficiently transparent that stakeholders can identify that a school's performance is "good" or "bad" due to its success (or failure) in the specific indicators and how the indicators relate to an overall summative determination. A school might receive a low score due to students' lack of progress, but accountability does not address why there is lack of progress. The steps and techniques identified in Part II provide many examples of analyses that SEAs can conduct to examine whether there are any obvious unintended factors unduly influencing school composite results or classifications, and those steps assist in ruling out structural factors[55] contributing to performance and conflating claims about schools.

[54] The initial reason for observed results in terms of the accountability system should be transparent – i.e. the students in school *j* demonstrated *X* proficiency rate, Y growth, *Z* EL progress, etc.; however, why the EL progress rate is *Z*, is not answered by the accountability system – this requires evaluation of services provided.

[55] An example of a structural factor is setting linear growth expectations which will result in middle and high school students less likely to meet progress targets, resulting in lower EL Progress Indicator scores for middle and high schools that reflect their position in the language development timeline and the

Evaluation attempts to understand why scores are what they are and how a program can be improved. Evaluation can encompass many aspects, but the following three are particularly germane to improving student outcomes and low performing schools: understanding why an individual school performs as it does; how systematically performance is influenced by schools; and why specific subgroups (students and/or schools) perform substantively differently from others in the state. In many instances, the proximal outcome of the evaluation is the performance of students on specific measures that are incorporated into the various indicators; the distal outcomes are either the indicators of the accountability system or the overall performance of the school. Importantly, the primary purpose of evaluation is continuous improvement, not simply categorizing schools or programs as successful or unsuccessful; the aim of the evaluation is improved opportunities and services to students. The scope of the evaluation can include identifying school-specific improvements in interpersonal communication skills that are not intended to generalize beyond that specific setting and time. It can also include identifying whether a specific statewide policy has the intended impact, e.g. successfully exiting EL students using a particular set of criteria. Evaluation results can help improve schools and SEA policies and inform refinements of the accountability system.

Part III provides several examples of evaluation and analysis strategies that can be applied to common issues related to ESSA and fostering the SEA's ability for continuous improvement. Again the examples are neither necessary nor sufficient when thinking about evaluation, and, as noted, an evaluation policy provides guidance as to the elements an evaluation should contain. For example, using evidence-based strategies generally means strategies that have been rigorously evaluated using WWC (What Works Clearinghouse) criteria, which are available at: https://ies.ed.gov/ncee/wwc/Docs/referenceresources/wwc_procedures_v3_0_standards_handbook.pdf. The WWC criteria may form the basis of criteria for SEA evaluations in many instances, but the SEA will likely want to develop guidance around when the WWC criteria are applicable and when they are not. Part III does not provide guidance on experimental design – even though carefully thought out designs virtually eliminate the need for sophisticated statistical analyses (Rubin, 2008). The operational reality in most states is that evaluations are either using extant data or interventions and programs that are in response to specific needs and thus do not have flexibility in treatment assignment. Again, having an evaluation policy can provide some pre-implementation steps, minimally in data collection, but potentially in intervention assignment. For example, a systematic policy of maintaining data on all students who are assigned or self-select into

mismatch between expectations and known language development trajectories, rather than reflect on the actual progress of students.

programs irrespective of whether they attend will provide important data for program evaluation.

The remainder of Part III steps through a series of analyses that support evaluation goals the SEA can undertake to examine whether EL progress is being modelled appropriately, whether the ELP Indicator is behaving as expected, and whether the ELP Indicator contributes to the overall accountability results as expected. The Handbook then briefly discusses issues related to making causal claims and three issues related to making causal claims: generalizability, treatment/program assignment, and the importance of thinking not only about statistical significance but also weighing substantive importance. These issues can broadly be conceived as elements to consider when collecting validity evidence. Part III uses as a starting point for considering the impact of interventions/programs/systems or causes of outcomes a process known as Root Cause Analysis (RCA). The handbook does not provide guidance on conducting RCA but rather it aligns RCA with the existing framework of a Theory of Action. Building on this conception, Part III presents four analyses that can be incorporated into evaluations that will address to some extent factors affecting internal validity. Internal validity is made up of a series of elements that represent alternative or rival hypotheses (also known as potential cofounding factors [PCFs]) that might explain the outcomes—thereby making claims related to purported causes untenable. The handbook provides examples related to moderating and mediating variables and explains how to incorporate them into analyses. Adequately including moderating and mediating variables does not address issues related to internal validity; it provides a mechanism through which the specification of relationships is made clearer—aligning analyses with the SEA's Theory of Action and provides a mechanism for formally examining "causes" and contributing factors in RCA.

In order to explicitly address internal validity, the handbook presents an overview of Regression Discontinuity (RD) analysis and the use of propensity scores. RD and propensity scores are presented because the SEA generally does not randomly assign students and/or schools into interventions. For example, at the student level ELs are not randomly exited, rather they meet specific criteria; at the school level, schools are not randomly assigned to CSI status, rather status is based on being in the bottom 5% on a performance metric. Given the inherent purposeful nature of state policy, alternative methods such as RD and propensity score matching provide an avenue to more rigorously understand the impact of SEA indicators, accountability, and policy. Previous research indicates that it is possible to estimate causal effects using non-experimental methods (Shadish, Clark, and Steiner, 2008) but, of course, caution is always warranted (Shadish, 2013). Propensity scores are relevant when the program/treatment assignment

mechanism is not random or not known, while RD is applicable when the program/treatment assignment mechanism is known.

Assessing EL Progress with Additional Data

Once the SEA receives three years of ELPA results, it can assess whether progress is as anticipated. The first check might be to examine realized growth with the third year's results. The following closely resemble steps identified in Parts I and II:

1) Calculate the gain from the previous year to the current year and create a table (like Table 7) that displays gains by time in program and initial ELD level.
    a) How do the average gains compare to the year 1 to year 2 gains?
    b) What percentages are meeting annual growth targets?
    c) Examine all schools with data, not just those that meet the minimum N-size.
    d) Calculate the growth from year 1 to year 3 by time in program and initial ELD level. Compare whether students meet a cumulative growth target in year three vs. meeting the year 3 gains from the year 2 target. This provides evidence for whether annual growth targets should also include cumulative growth targets, or whether a model adjusting expectations is reasonable.
2) Validity evidence
    a. What is the likelihood of ELs meeting the second growth target if they met the first target vs. if they did not meet the first target? This check can be accomplished with a regression model or a non-parametric table.
3) Aggregate Growth Analyses
    a. Considerations outlined in Goldschmidt & Hakuta, 2016
        i. Stability
            1. N size: are small schools less stable than large ones?
        ii. Reliability
        iii. Correlations
4) Progress and Accountability
    a. Plan against reality-- disaggregate results by and correlations with
        i. Student characteristics
        ii. Other indicators of accountability model
    b. Impact of the ELP Indicator on overall performance
        i. How can the SEA disentangle whether schools are "good" at academics but "bad" (or vice versa) at English language development vs. the ELP Indicator isn't good?
        ii. Realized weight of the ELP Indicator
5) Realized weight of EL progress (factor analysis regression, correlations)
6) Checking adequacy of time to exit

- What percentage of students exit "on time," or "late"
- What are the incremental (yearly growth targets to meet exit time)? Are they linear?
- What percentage of students meets the incremental growth targets?
- Is there significant variation by ELD level, grade, etc.?
- Matriculation over time
- Years to exit
- Years to exit for those who exited
- Years in program for those who have not exited
- Are EL students On Track, Off Track, or Long Term[56]?

Examining in more detail whether ELs are On Track or Off Track is a proactive way of identifying potential Long Term ELs before they reach Long Term standing. Combined with an analysis that provides a sense of whether students who have met or not met targets are likely to meet targets in subsequent years, this can help provide early warnings for potential delays in exiting. Also, these analyses aggregated to the school and district levels can identify systematic strengths and weaknesses.

Table 26 presents one way to begin focusing on continuous improvement by more purposefully examining EL standing in terms of time in program. Time in program is not the only way by which student progress can be monitored—in fact, meeting growth expectations is another method to augment diagnostics presented in Table 26. The results in Table 26 are part of a dashboard that is easy to create and provides many additional variables that can be examined dynamically. A demonstration of how to create the dashboard can be found at:

https://ccsso.webex.com/ccsso/lsr.php?RCID=0ef1e71073cd4aeaab24bbcb3210e3d8

[56] Off and On Track refer to time in program, not to progress (growth/gain). For example, an EL who is initial ELD level 2, may be expected to exit in 4 years – if this student is in year 3, she is On Track; if she is in year 5, she is Off Track.

TABLE 26: STATUS AND STANDING FOR EL STUDENTS WHO HAVE MET AND NOT MET EXIT CRITERIA

Count of Status	Column Labels					
	OffTrack		OffTrack Total	On Track	On Track Total	Grand Total
Row Labels	LT	Not LT		Not LT		
Exit	373	319	692	1252	1252	1944
Not Exit	3431	2066	5497	13845	13845	19342
Grand Total	**3804**	**2385**	**6189**	**15097**	**15097**	**21286**

Average of ELP_Nocc	Column Labels					
	OffTrack		OffTrack Total	On Track	On Track Total	Grand Total
Row Labels	LT	Not LT		Not LT		
Exit	7.34	3.82	5.72	2.16	2.16	3.42
Not Exit	7.58	3.82	6.17	2.21	2.21	3.34
Grand Total	**7.56**	**3.82**	**6.12**	**2.21**	**2.21**	**3.35**

Average of VAM_NL10	Column Labels					
	OffTrack		OffTrack Total	On Track	On Track Total	Grand Total
Row Labels	LT	Not LT		Not LT		
Exit	0.94	0.96	0.95	0.74	0.74	0.84
Not Exit	0.47	0.37	0.43	0.46	0.46	0.45
Grand Total	**0.51**	**0.45**	**0.49**	**0.48**	**0.48**	**0.48**

The key elements of Table 26 are that EL progress and exiting is disaggregated into ELs that are/were On Track and Off Track. Among Off Track, ELs are further divided into Long Term and Not Long Term. The results in Table 26 indicate that students can exit the program either on time (On Track) or late (Off Track), and in some instances ELs were Long Term ELs before exiting. Table 26 also summarizes this On/Off Track information for ELs still in the program. Table 26 identifies growth and years in program by the status (EL Exit/Not Exit) and standing (Off/On Track) variables. Not only do tables like Table 26 potentially identify systematic difference among ELs, but results can be aggregated by initial ELD level and to the school or district level. Moreover, additional factors can be examined such as RAEL status, which would allow for an analysis of whether RAELs' progress differs in meaningful ways. Such information can be used to set expectations or to check whether expectations (annual growth and time to exit) are appropriate for RAELs. To the extent that meaningful differences arise, the SEA may want to evaluate

the impact more formally to determine whether differences are in fact statistically significant and substantively meaningful. Also, the ability to examine individual schools or to aggregate to districts (or regional services areas, for example) allows states to identify how patterns differ among schools/districts and provide guidance as to where the SEA might turn for examples of success.

Moving beyond tables to more formal evaluation models calls for attention to additional details; the Handbook addresses these in the following sections. After a brief discussion of generalizability, treatment assignment and data structure, and statistical and substantive significance, Part III presents several analytic strategies that are appropriate for evaluating EL progress, EL content performance, and policy choices affecting ELs.

Generalizability

Generalizability, from a psychometric conception is the degree to which score properties and interpretations generalize to and across populations, settings, tasks, cues, and raters (Messick, 1995). For the purpose of evaluation, this definition should be amended to be the degree to which interpretations related to causes are consistent with and across populations, settings, tasks, cues, raters, and treatment assignment. For example, it is important that states are required to have statewide exit criteria for ELs, otherwise claims about the impact of exiting in one district would not generalize to other districts and an evaluation of the appropriateness of exit criteria based on a subset of districts would not generalize to the state. Generalizability is one element of validity evidence.

Generalizability is a point of departure between evaluation and research. Evaluation's focus may be narrow and generalizability may, in fact, not be a relevant issue. For example, the impact of exiting EL status in state A may be different than in state B or C and this may be a result of different exit criteria. This does not preclude state A from evaluating its exit criteria; it merely limits the generalizability of results across population, setting, and assignment. The notion that policy (model) results differ across states has been explicitly demonstrated (Goldschmidt et. al., 2012). A specific example is evaluating why ELs in district Z are not meeting progress targets and may be quite nuanced due to district context (e.g. ELs in this district are predominantly newly arrived refugees, whereas in other districts they are highly mobile second generation ELs) – requiring a responsive approach that directly impacts district Z, but would not apply to any other district in the state. The takeaway is that depending on the goals of the evaluation, generalizability and issues with generalizability may not be relevant; however, limitations should be made clear a priori.

Considerations with Respect to Treatment Assignment and the Structure of Data

It is the nature of education that students attend schools[57] and that SEA policy is linked directly to students, teachers, schools, or districts. A full treatment of the impact of clustered data is beyond the scope of this handbook, but two elements worth noting are the effective sample size and understanding the unit of analysis. The effective sample size is important because outcomes clustered within a group (i.e. the progress of ELs attending specific schools) are correlated (i.e. there is a correlation among EL progress scores within a school) and this reduces the effective sample to less than the observed sample size. If this correlation (the ICC, noted above) is not accounted for, standard errors will be too small, potentially causing effects to appear statistically significant when they are not. Consistent with this concept is the unit (or level) of treatment (program) assignment. If schools are assigned a treatment, then analyses of student performance in CSI schools vs. non-CSI schools would be incorrect because students are not assigned to the treatment. Conducting analyses at the student level (i.e. level of aggregation[58]) and including, for example, a CSI indicator variable to account for treatment vs. business as usual would result in biased standard errors that could potentially result in erroneous conclusions about statistical significance. For example, if the analyst wants to compare EL growth on academic English language content in schools in CSI against schools not in CSI then it is incorrect to simply calculate the average growth of both sets of students and use the number of students in each set of schools to calculate standard errors for a t-test. Moreover, this erroneous strategy assumes there is no specific school effect because the effect of the school appears to be the aggregate of its students' performance. It may appear that a simple solution is to conduct analyses that uses school mean performance as an outcome. However, this results in the implicit assumption that there is no heterogeneity in the outcome within schools, and that all of the variation in the outcome is between schools (which is likely not a tenable assumption). Generally this issue is addressed by using a fixed effects model, a mixed effects model, or a correction for clustered standard errors. This is addressed in more detail in Allison (2009), Seltzer (2004), and Raudenbush and Bryk (2002). Generating standard errors that are smaller than they should be has implications for schools that use Cis (Confidence Intervals) for

[57] The increased use of personalized learning and on-line courses does not eliminate the fact that students are in some way clustered, and thus groups are impacted differentially.

[58] Level of aggregation refers to how the data are aggregated *and* how individual students are associated with another unit. For example, a student is usually the most disaggregated unit in the analysis and student performance is often aggregated to the school. Individual students constitute a level and schools constitute another level (and schools can be associated with variables that are not strictly aggregates of student data – e.g. the principal's experience). The association is that a group of students attend a school (they are clustered by school). Associations between and among levels can be quite complex and is covered in more detail in Raudenbush and Bryk (2002).

either exit or entry criteria into CSI or TSI because the Cis will be erroneously small, signaling that schools are more likely to have improved (e.g. for exit) than they really have.

Not only is effective sample size an important consideration, but so too is the interpretation of impact. Evaluations in some instances examine individual student results and in some instances aggregate (school/district/state) results. Hence, evaluation questions focus on different units of analysis (i.e. students, schools, districts, etc.). Analysts need to be cautious in inferences that ignore the nature of the data, even when examining progress where different levels can impact the outcome. One example is the impact of SES (Socio-Economic Status) on EL progress. Individual student SES can impact EL progress because SES is related to opportunities and supports available for the EL at home, while aggregate SES (i.e. school mean SES, or percent FRL, for example) may also impact EL progress, irrespective of the EL's individual SES because school mean SES may be associated with how much summer learning loss occurred in the school that must be systematically addressed[59]. This topic is addressed in Curren and Bauer (2011), who provide guidance on disentangling effects when examining growth (progress).

Statistical and Substantive Significance

In many instances evaluations are concerned with demonstrating statistical significance, but statistical significance should always be accompanied by substantive or practical significance, so the SEA and stakeholders can understand the consequences of the evaluation results. Evaluations do not require formal statistical testing, but it is generally the case that statistical tests provide some level of comfort that the results did not occur by chance (the issues presented above relating to confounding factors and treatment assignment notwithstanding). Statistical significance does not create proof that an intervention or program is effective, it merely provides support for the assertion that evidence supports the claim that the observed impact/effect did not occur (is unlikely to have occurred) by chance.

Of course statistical significance is affected by several factors, including the sample size and the actual magnitude of the impact. Assuming the magnitude of impact remains constant, changing sample sizes affects the inference related to statistical significance. For example, a school may be interested in evaluating whether EL progress of learning 2 new vocabulary words per week is statistically different from no learning (i.e. 0 new vocabulary words). The school then would conduct a *t*-test where $t = (2-0)/se$, where the se is the standard error = SD (standard Deviation)/$\sqrt{n}$. If the SD is equal to 4 and the

[59] This is not to imply is the only mechanism that school mean SES or percent FRL systematically impacts student performance.

school has 9 students, then t = 2/1.33 ~ 1.5, which would not be deemed statistically significant (i.e. learning 2 new vocabulary words is statistically not different from learning no new vocabulary words). If a second school in the same district engaged in the same study and has the same impact and SD, but has 12 ELs, the resulting t = 2/1 = 2, would also not be deemed statistically significant (with 11 degrees of freedom). If the entire district engages in this analysis and finds the same impact and SD, then the resulting t = 2/.87 = 2.3, which would be deemed statistically significant. Even though the impact is the unchanged, by using a larger sample, statistical significance is achieved[60]. A given impact "D" will eventually be statistically significant upon reaching a large enough N, all else being equal. Thus, a small D is eventually statistically significant, and one could reject the null hypothesis that D is 0, but of equal importance is the substantive impact of D. It may be the case that learning two vocabulary words in a week constitutes a significant gain for a large sample of students, but the question is whether learning two vocabulary words is substantively important or practically meaningful. Substantive importance can be evaluated by creating an effect size from D. There are numerous ways to calculate an effect size (Cooper and Hedges, 1994). Generally, effect sizes represent standardized effects and are computed by subtracting the difference in outcomes between the treatment and control groups and dividing by the standard deviation of the control group[61] (Cooper and Hedges, 1994). Effect sizes can be classified as trivial, small, medium, etc., and this classification may be topic-specific. Generally, effects are conceived of as being compared to no effect, but states may compare an effect to a pre-determined value other than 0; i.e. a growth target. The state population of ELs may be learning greater than 0 vocabulary words more than the previous week or year, but this may be insufficient to gain English language proficiency in the required time.

Root Cause Analysis – To Consider Impacts of Interventions/Programs/Systems

A common analytical tool identified in many ESSA state plans to help foster continuous improvement in schools is Root Cause Analysis (RCA). RCA is a process that purports to identify causal factors using a systematic approach by stepping through a series of postulates to arrive at cause(es) (Dew, 1991). RCA is extensively used in the health sciences and particularly by the Veterans Association (VA). Evidence of efficacy is generally available but not conclusive (Percarpio, Watts, and Weeks, 2008). The concept of identifying causal factors systematically has merit, and it appears that no single

[60] Of course the power to detect effects varies with n and small samples would be underpowered. The example uses small numbers to simplify calculations, but the overarching point remains the same: a small effect with a large enough N will be statistically significant.

[61] This is Glass' Δ, but there are many variants in terms of which standard deviation to use, as well as other statistics that can be transformed into an effect size.

method of RCA is more efficient to apply (Doggett, 2004). However, research also indicates that RCA takes between 20 and 90 hours to conduct and is often limited by capacity and access to resources (data, personnel, processes, etc.). The majority of cases lead to conclusions that do not identify a cause with a substantively important impact, and in a majority of instances results do not readily identify solutions (Wu, Lipshutz, and Pronovost, 2008). Importantly, RCA, when effective, tends to be best suited to resolve specific issues (Wu, Lipshutz, and Pronovost, 2008).

RCA depends on creating teams to conduct the analyses, and these teams should consist of various stakeholders who can meaningfully inform the process. For example, an RCA team might consist of the district Title III coordinator, a state Title III coordinator, an EL specialist, a teacher, a principal and a parent. A common method of RCA is to use a Cause and Effect Diagram (CED) (Doggett, 2004). It is important to note that developing a CED results in *potential* causes and does not result in unequivocal identification of a root cause or causes. This can only be accomplished by collecting evidence in support of the claim. A basic CED is displayed in Figure 23.

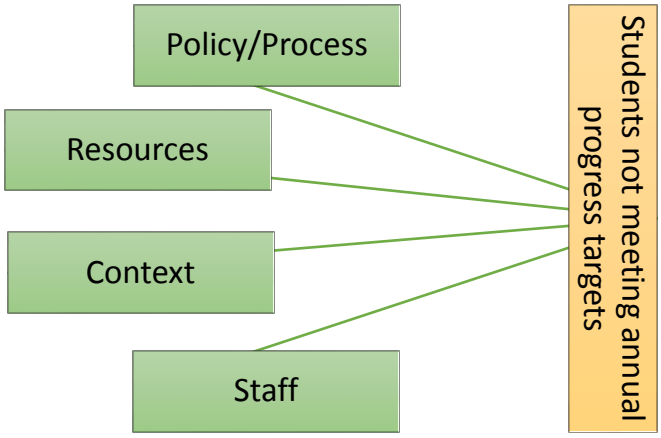


FIGURE 23: CAUSE EFFECT DIAGRAM

The potential causes and contributing factors presented in Figure 23 are a common starting place for RCA, but the RCA team would refine these as to be relevant to the outcome. Once the four basic causes are identified, additional branches are added to refine and isolate causes. For example, staff receive a sub-branch related to knowledge of new EL standards; the policy branch might be expanded to include separate sub-branches for exit criteria.

A Theory of Action is an excellent tool to help develop the CED. For example, the CED in Figure 23 can be informed by a ToA that was previously developed and states:

> If students attend schools that have sufficient resources to provide each student with adequate learning materials (resources), and if teachers and support staff are well trained in updated language standards, as well as appropriate pedagogic techniques that utilize formative assessment practices (staff), and if the assessment adequately reflects EL English language progress (context), then EL students will meet annual progress targets (policy).

Developing a Theory of Action allows states to explicitly consider data collection, measurement of inputs and outputs, and design aspects underlying understanding of the treatment or program impact on the intended audience. SEAs should be cautious, however, in using RCA to back-map potential causal chains in that this results in ex-post facto hypothesizing (developing a hypothesis after the results are known). Ex-post hypotheses are not a valid means of identifying causes since there is no chance the hypothesis can be wrong. Moreover, using RCA to build a ToA after-the fact reflects a process focused on identifying the cause of a problem, rather than developing a ToA that focuses on the appropriate resources and processes that are aligned with success.

The next sections provide examples of analyses that are based on a Theory of Action and address important aspects often included in Theory of Actions. This includes moderating and mediating factors and who/how students/schools are assigned to the intervention. The following descriptions consider elements that are particularly relevant to EL progress (but in no way address every potential issue; further, some of the issues are not only relevant to ELs but to students generally). Three specific issues are considered: at the student level, both the potential impact of being a Recently Arrived EL (RAEL) on meeting the annual progress/growth targets and the impact of mobility on meeting the annual progress/growth targets; and, at the system level, the effect of minimum N on school accountability results. These issues also represent four important analysis techniques: analysis of moderating factors, mediating factors, propensity score matching, and regression discontinuity.

Moderating Factors

A moderating variable is generally exogenous (i.e. the school has no influence on them) and specifies under what conditions a relationship between the input and outcome holds. Moderating variables are often qualitative, such as student characteristics. The Evidence Based Policy Decision (EBPD) analysis below provides an example of informing policy with empirical analysis. The terms "moderating" and "mediating" are often used interchangeably. However, they are not only different in meaning, but also quite

different in the analyses required to determine whether moderation or mediation is taking place. Mediation analysis is presented in the next section.

Testing for moderation means testing whether an interaction or joint-effect of the potential moderating variable is statistically significant and substantively meaningful so that it impacts outcomes. Moderation (and mediation analysis) can be conducted with non-parametric methods such as contingency tables and χ^2 statistics, however it is most common to use OLS (Ordinary Least Squares) regression[62]. Hence:

$$Y_i = C_0 + C_1L + C_2RAEL_p + e_i \quad (10)$$

Where Y_i is the outcome of interest, e.g. meeting the annual growth target or not. Y_i is coded as 1 if the student met the growth target and 0 if they did not[63]. L might be a dichotomous variable that denotes whether a student's initial ELD level is 1 or not (coded: if initial ELD level = 1, then L = 1, otherwise L = 0). In the worked example, L is expanded to specifically represent each initial ELD level. $RAEL_p$[64] is an indicator variable that is coded as $RAEL_p = 1$ if a student is RAEL in the previous year and 0 otherwise (RAEL status from the previous year is required since current RAELs will not have two scores with which to calculate progress). The model coefficients are C_0, the mean probability of a student whose L and $RAEL_p = 0$; i.e., a student whose initial ELD level is greater than 1 and who was not a RAEL the previous year. C_1 is the average effect of having an initial ELD level greater than one on the probability of meeting a growth target. C_2 is the average of having been a RAEL in the previous on the probability of meeting the annual growth target.

Since RAEL status is a moderating variable, the RAEL status will specify the relationship between initial ELD status and meeting the annual growth target; the expectation is that the average effect of both initial ELD level and $RAEL_p$ jointly impact the outcome. This is tested with:

$$Y_i = C_0 + C_1L + C_2RAEL_p + C_3L*RAEL_p + ei \quad (11)$$

[62] As noted above, depending on the analyses the regression models may need to address the data structure by using fixed and or mixed effects models.

[63] Given the outcome is 0/1, the analyst may want to use logistic regression; however, it is also possible to use a linear probability model (an OLS regression with a 0/1) outcomes. The coefficients of such a model are the probabilities of the outcome, Yi = 1 (meeting the growth target).

[64] The analysis uses the previous year's RAEL status because a second data point is required to measure growth for a RAEL.

Including C_3, the joint effect of initial ELD level and $RAEL_p$ status on the probability of meeting the annual growth target changes the interpretation of all the coefficients in the model. This is important to understand because using the results to support policy decisions requires that the evidence is correctly interpreted. Table 27 details what the coefficients represent in equation (11). Table 27 and subsequent descriptions are based on Hardy (1993), which is an excellent source for understanding how to use indicator (dummy) variables in analyses. Table 27 summarizes the interpretations.

TABLE 27: IDENTIFYING EFFECTS

	Not RAEL	**RAEL**
Initial Level =1	C0	C0 + C2
Initial Level > 1	C0+C1	C0+C1+C2+C3

The t-test:
C1 tests whether there is an effect of initial ELD level being greater than 1 among Not-RAEL (it is no longer the average effect of having an initial ELD level greater than 1 on the probability of meeting a growth target).

C2 tests whether there is an effect of being an RAEL among ELD level 1 students (it no longer provides an estimate for the average RAEL effect on the probability of meeting the annual growth target).

C3 tests for the differential effect of ELD level by RAEL, or RAEL by ELD level: the difference in the effect of ELD level among Not-RAEL is C1, and this effect is C1 + C3 for RAEL, so C3 tests whether there is an effect of being an ELD level greater than 1 for RAEL relative to Not-RAEL. The t-test for C3 does not test the significance of net ELD level difference in the probability of meeting the annual growth target among RAELs; rather it tests whether the net differential in meeting the annual growth target between initial ELD level 1 and initial ELD level greater than 1 is the same for Not-RAEL and RAEL. The t-test for testing effect of ELD level for RAELs is:

$$t = (C1 + C3) / [varC1 + varC3 + covC1C3]^{.5} \quad (12)$$

For Not-RAEL it is simply $t = C1/(varC1)^{.5}$ (13)

If, for example, (12) is significant, but (13) is not, then initial ELD level is important among RAEL, but it is not for Not-RAEL students.

The interaction term captures the differences in effects for Not-RAEL and RAEL by initial ELD level. This also accounts for the fact that there may not be a uniform distribution of RAEL and Not-RAEL by initial ELD level (there may be more RAEL in the initial ELD level 1 category than in other initial ELD levels).

Example 2 demonstrates an analysis of a potential moderating variable.

Example 2: Evidence-Based Policy Decision

The SEA might be interested in whether it should adopt a policy to set a different time horizon (and correspondingly appropriate annual growth targets for) Recently Arrived ELs (RAELs). This decision can be informed by examining whether RAEL status moderates the relationship the SEA used in determining annual growth targets for ELs in general. Specifically, accounting for initial ELD level and time in program predicts current ELP level, but the relationship between time in program and current ELP level may be different for RAEL and Not-RAEL students. RAEL potentially moderates the relationship between time in program and the probability of meeting the annual growth target.

Table 28 shows how using equations (10) and (11) focuses on whether RAEL status moderates the relationship between initial ELD level and the probability of meeting the annual growth target in year two[65]. Initial ELD level is recoded such that Initial ELD0 = (Initial ELD level -1). Model 1 is based on (10) and model 2 is based on (11).

[65] This analysis focuses only on meeting the growth target in year two because the dataset had only two years of data; however, there is no loss in generality as an example due to this delimitation of the analysis.

TABLE 28: TESTING FOR MODERATION[66]

Model 1		B	Std. Error	t	Sig.
CO	(Constant)	.796	.011	74.24	.000
C1	Initial_ELD0	-.209	.005	-43.72	.000
C2	RAELp	.186	.015	12.68	.000

a. Dependent Variable: Met_Grth_Target

Model 2		B	Std. Error	t	Sig.
CO	(Constant)	.823	.011	71.90	.000
C1	Initial_ELD0	-.224	.005	-42.30	.000
C2	RAELp	.116	.018	6.33	.000
C3	InitialbyRAELp	.078	.012	6.42	.000

The results (in Table 28) Model 1 indicate, consistent with expectations, initial ELD level is related to the probability of meeting the annual growth target in year two. The probability of meeting the annual growth target is about .796[67] for initial ELD level 1 students (C0). The probability decreases by about .21 for each initial ELD level higher than 1[68] (C1). However, contrary to expectations, RAEL students have a slightly higher probability of meeting the annual growth target, .186 (C2). More precisely, RAEL students who are initial ELD level 1 have a slightly probability of meeting their annual growth target in program year 2 compared to their not-RAEL classmates who are also initial ELD level 1.

The moderating effect of RAEL status is tested in Model 2. The results indicate that all of the variables are significant.

A good way to present the effects is to use the values in Table 28 to fill in Table 27. Table 29 presents these results.

[66] The SPSS syntax for these outcomes is in Appendix I.

[67] The coefficient is interpreted this way because a linear probability model is used.

[68] This model assumes a linear effect of initial ELD level on the probability of meeting the annual growth target – again as a means of simplifying the example.

TABLE 29: PROBABILITY OF MEETING ANNUAL GROWTH TARGETS

	Not-RAELp	RAELp
Initial ELP Level 1	0.82	0.94
Initial ELP Level 2	0.60	0.79
Initial ELP Level 3	0.37	0.65
Initial ELP Level 4	0.15	0.50

The results in Table 29 are consistent with expectations in that lower initial ELD levels are related to a greater probability of meeting growth targets. This pattern is the same for Not-RAEL and RAEL students. However, RAEL status moderates the relationship because the probability of meeting the annual growth target is not only different for RAEL's and Not-RAELs, but this difference varies by amount at each level of initial ELD level. The results demonstrate different probabilities for RAEL and Not-RAEL students in meeting annual growth targets. To help inform policy the C2 and the [C2+C3] coefficients need to be tested for statistical significance. C2 is directly checked using the results in Table 28, model 2. Testing [C2+C3] requires covariance between C2 and C3 which, in this case, is approximately 0. Thus, the significance test for [C2 + C3] is $t = (0.116 + 0.078)/ (0.018^2 + 0.012^2)^{.5}$ which is 0.194/0.02194; t = 8.3, which at a conventional α of 0.05 results in a $p < 0.05$. This rejects the null that C2 and C3 are 0. As noted previously, a policy decision should not simply be informed by statistical significance, rather it should also consider the substantive impact (effect size) of the effect in question. In this case Cohen's *h* would be the appropriate effect size estimate (Cooper and Hedges, 1994).

In the context of this problem, it is likely that analysts would be concerned that RAEL students would be less likely to meet growth targets. The SEA would then alter targets for RAELs which impacts school accountability because it would reduce the relationship between a school's input (RAEL) and indicators of aggregate progress. The analyses demonstrate that in this instance, RAELs are not less likely to meet the annual growth target (and are in fact more likely to do so), so the SEA would have empirical support for a policy that does not differentiate timelines and annual growth targets by RAEL status.

Mediating Factors

A mediator is a variable that is related to both the independent and dependent (input and outcome) variables and tempers or exacerbates the relationship between the independent and dependent variables. Mediating variables are generally malleable. For example, language program might be a mediating variable that provides the mechanism through which the independent variable impacts the dependent variables. Accounting for initial

ELD level and time in program predicts current ELP level, but this relationship might be influenced by the program a student attends; i.e. program type potentially mediates the relationship between time in program and current ELP level.

Testing a simple mediation model requires several steps. It is possible to model a program type and hold schools accountable for differences in success among schools but within program type. For example, a state may use a VAM that includes an indicator that differentiates EL program A from EL program B. Allowing this indicator to vary randomly among schools generates a mean EL program effect and a unique effect of the school on the program. If the random coefficient for the effect of program varies significantly, there is likely also significant variation in program effectiveness among schools.

Testing for mediation is testing whether the relationship between X and Y is significantly enhanced or reduced by introducing an explanatory linkage (M) between X and Ythat explains why or how X is related to Y, using OLS regression[69] analyses. The following steps are based on MacKinnon (2008), which is an excellent resource for statistical mediation analysis.

This is accomplished testing the three models presented in (14-16):

$$Y_i = I_1 + cX_i + e_i \quad (14)$$
$$Y_i = I_2 + c'X_i + bM_i + e_i \quad (15)$$
$$M_i = I_3 + aX_i + e_i \quad (16)$$

In (14), Y_i is the outcome Y for student *i*, I_1 is the intercept, X_i is the input for student *i*, e_i is a random error term, c is the direct effect of X on Y. In (15), Y_i and I_2 are again the outcome and intercept for student *i*, respectively. Also, c' is the relationship between X and Y accounting for M, the mediating variable for student *i*; b is the effect of the M on Y accounting for X, and e_i is the error term. Finally, in (16), M_i is the mediating variable for student *i*, I_3 is the intercept, a is the effect of Xi on Mi, and e_i is the error term.

[69] Here again the structure of the data might require the use of fixed or mixed effects models.

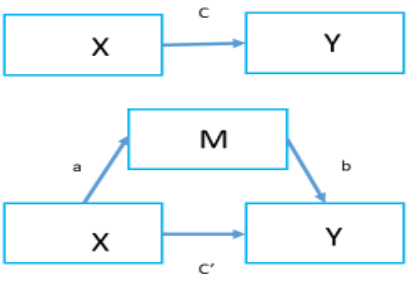


FIGURE 24: PATH DIAGRAM OF POTENTIAL MEDIATING EFFECTS

Figure 24 presents a path diagram that graphically depicts models 14-16. Model (14) tests the strength of the direct relationship between X and Y, c. Model (15) tests whether accounting for the X-Y relationship, there is a relationship between M and Y, b; but also whether, accounting for M, there remains a partial relationship between X and Y, c'. Finally, model (16) tests whether there is a relationship between X and M, a. The mediated effect is equal to the estimated a x b (or also estimated c – c'), while the standard error for the mediated effect is $(a^2s_b^2 + b^2S_a^2)^{.5}$ (MacKinnon, 2008). The total effect is c' + ab.

The purpose of all of these equations is to understand whether the total relationship observed by X and Y actually includes the effect of another factor. It may also be that there is no relationship between X and Y, but that the observed X, Y relationship is due to the fact that X is related to M, and M is related to Y.

Example 3 demonstrates a potential moderating variable.

Example 3: Evidence-Based Policy Decision

The SEA may be concerned that mobility—particularly among ELs—impacts the potential for meeting annual growth targets. This may be related to the SEA's Theory of Action that indicates part of the process of developing English language proficiency as putting students in a stable learning environment. A proxy for stable learning environment could be school-to-school mobility. Given we have only two years of data, mobility is defined as 1 if the student is in a different school from the previous year, and

0 otherwise[70]. Hence, X = initial ELD level (coded 0-4 as in the moderation example), Y=Whether or not the EL met the annual growth target, and M (the potential mediating factor) = mobility, coded as indicated above. Table 30 presents the results of the mediation analysis.

The results in Table 30 summarize the steps used to ascertain whether mobility mediates the relationship between initial ELD level and the probability of meeting the annual growth target in year 2. Model 14 indicates that initial ELD level is, in fact, related to the probability of meeting the annual growth target in year 2. The results indicate that an initial ELD level 1 student has about a 0.88[71] probability of meeting the growth target. And this is reduced by about 0.23 (c)[72] for each initial ELD level above 1. Even though c is not significant this does not preclude a mediating effect. These results are consistent with expectations. The results from Model 15 indicate that the potential mediating variable, mobility, presents suggestive evidence of affecting the probability of meeting the annual growth target ($p < 0.10$). The effect is small – a student who changed schools has a probability of meeting the growth target that is 0.02 (b) lower than a student who did not change schools. It is important to note that the partial effect of initial ELD level (c') is virtually unchanged from the effect in Model 14. Model 16 indicates that there is a significant relationship between initial ELD level and the probability of changing schools. The probability of an initial ELD level 1 student changing schools is about 0.1895, and this is reduced by about 0.03 (a) for each initial ELD level above 1. The total effect C is a combination of the direct and indirect effect.

[70] This simple coding confounds mobility within a grade band and changing schools due to matriculation from elementary to middle, or from middle to high school, but serves as an example of testing for mediation.

[71] These results are essentially the same as in the first step of the moderation example, but differ slightly due to sample difference (associated with missing mobility data).

[72] This refers to the arrows (effects) in Figure 22.

TABLE 30: MEDIATION MODEL RESULTS[73]

Model 14

		B	Std. Error	t	Sig.
I_1	(Constant)	0.8711	0.0091	95.60	0.00
C	Initial_ELD0	-0.2306	0.0046	-50.54	0.00

Dependent Variable: Met_Grth_Target

Model 15

		B	Std. Error	t	Sig.
I_2	(Constant)	0.8777	0.0098	89.35	0.00
c'	Initial_ELD0	-0.2307	0.0046	-50.57	0.00
b	Mobility	-0.0240	0.0132	-1.81	0.07

Dependent Variable: Met_Grth_Target

Model 16

		B	Std. Error	t	Sig.
I_3	(Constant)	0.1895	0.0063	30.18	0.00
a	Initial_ELD0	-0.0301	0.0029	-10.56	0.00

Dependent Variable: Mobility

Total effect =	c' + ab =	-0.2300
Indirect effect =	ab =	0.000723
se of Indirect effect =		0.000404
Z =		1.79
	p <	0.07

Given that the indirect effect is very small (about 0.0007), it appears that mobility, while having some effect on the probability of meeting the annual growth target does not substantively mediate the effect of initial ELD level. The implication is that an SEA would not need to develop business rules to incorporate mobility into time to exit tables or annual growth target tables that are based on initial ELD level because mobility does not meaningfully mediate the relationship of the initial ELD level and the probability of meeting the target; in this case, the marginal impact of mobility is quite small.

[73] The SPSS syntax for these results are in appendix J.

Propensity Score Matching

Propensity Score Matching (PSM) is a useful tool for SEAs to utilize when considering evaluations because it utilizes extant data and can appropriately define the counterfactual, or comparison, group. PSM attempts to account for observable differences among program participants and non-participants, thereby reducing the impact of potential confounding factors (Austin, 2011). Its simplicity and potential complications means that a treatment of PSM is beyond the scope of this Handbook; in fact, this section provides only a brief introduction to PSM.

Evaluating a program or intervention's impact by simply comparing participants to non-participants is inappropriate and can lead to misleading results. Simply comparing a program to business-as-usual conflates multiple potential causes related to program participation (assuming assignment was neither random, nor based systematically on known criteria). In evaluating the impact of an afterschool program on juvenile crime, for example, a comparison between afterschool program attendees and non-attendees compares students who may be very different on many aspects – including committing a crime. Such an analysis might use PSM to match schools and students within schools to create like groups of schools and students (Huang, Goldschmidt, and La Torre Matrundola, 2014).

Propensity Score Matching is accomplished by estimating the likelihood (probability) that a student receives the treatment. This probability is estimated using as much relevant information that is available about each student. Unlike model building in general, PSM is less concerned with parsimony. Matched students are based on the propensity to have received treatment, but not on any single specific factor. Overall, the impact of creating a PSM sample is displayed in Figures 25 and 26. Figure 25 presents the probabilities of non-participants and participants (AAA_partic = 0 and 1 respectively) participating in the program before matching, while Figure 26 presents the same results after matching. Figure 26 clearly indicates that PSM has produced a comparison group that in aggregate looks similar to the treatment group. However, the analyst needs to examine balance among all the relevant variables before continuing the evaluation.

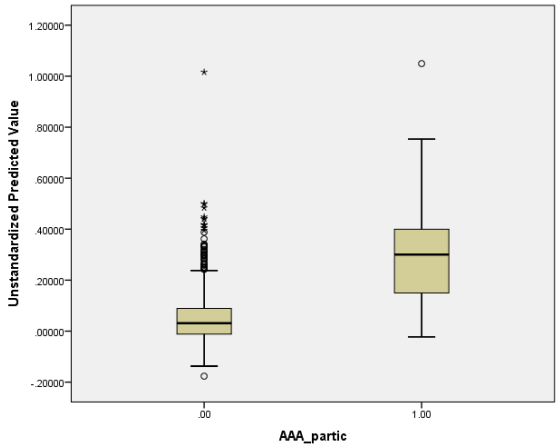


FIGURE 25: COMPARISON OF TREATMENT AND NON-TREATMENT STUDENT BEFORE MATCHING

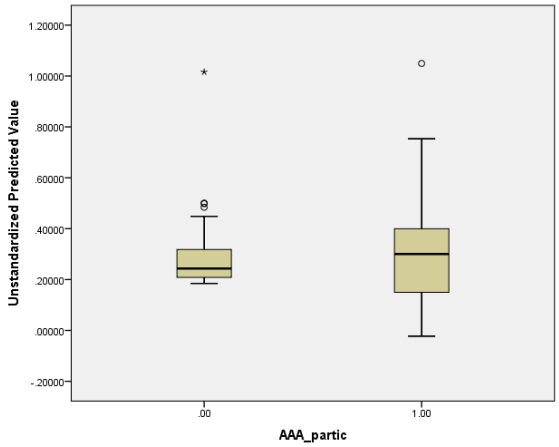


FIGURE 26: COMPARISON OF TREATMENT AND NON-TREATMENT STUDENT AFTER MATCHING

PSM is a potentially useful application for states and districts to refine evaluations because SEAs and districts are generally examining observational data where students/schools were not assigned to the program/treatment through a specific or known assignment mechanism, and because SEAs and district can make use of extensive data archives. However, given the potential misapplication of PSM, examining recent literature such as Shadish (2013) is advisable before using PSM in evaluations.

Regression Discontinuity (RD)

Regression discontinuity is particularly useful, especially as an ex post facto analysis (Thistlethwaite and Campbell, 1960), for evaluating the potential impact of policy decisions on the accountability system, because the analyses are based on the decision point which prompts the identification of a particular status and/or intervention. For example, several studies (Robinson, 2011; Jacob & Lefgran, 2004; Robinson-Cimpian & Thompson, 2016; and Jones, 2016) have used RD analyses to examine the impact of various education-related interventions or programs, with Robinson (2011) and Robinson-Cimpian and Thompson (2016) specifically focusing on the impact of whether EL students benefited from exiting EL status. RD analysis works well because exit status is generally defined by specific criteria that allow data to meet the assumptions underlying RD analysis. RD analysis can also be based on ranks (Keefer, 2016), which is useful given that parts of state accountability systems rely on school ranks (i.e. the bottom 5%) to place schools into various status classifications. On an accountability system level, RD is an excellent tool to help evaluate whether, for example, CSI classification has a meaningful impact on school performance. States vary considerably in how schools are identified for CSI, but one rule is consistent for all states – graduation rates less than 67% - and serves as a general example of where RD is a useful tool. The specifics of the assumptions that need to be met to use RD are discussed below, but, in general, CSI based on graduation rates is an excellent candidate for analysis because there is a known value that causes schools to be classified as CSI and it is quite possible for schools to exhibit graduations rates that are at, or close to, 67% (e.g. 66.6%, or 67.1%, etc.).

There are many potential complexities to RD analysis that can make addressing the assumptions underlying RD analysis difficult. This handbook does not provide a full treatment of RD, rather an overview, and an example. An excellent guide by Jacob, Zhu and Bloom (2012) includes complete guidance on RD, and is available at:
http://www.mdrc.org/publication/practical-guide-regression-discontinuity.
Other resources include:
https://ies.ed.gov/ncee/wwc/Docs/ReferenceResources/wwc_rdd_standards_122315.pdf
https://ies.ed.gov/ncee/wwc/Docs/ReferenceResources/wwc_rd.pdf
http://www.mdrc.org/sites/default/files/full_446.pdf

Given the reality of educational interventions and the use of Target Place-Based Programs (Deng & Freeman, 2011), RD provides a good methodology from which causal estimates, that are as credible as causal estimates based on RCTs (Randomized Control Trials), can be derived. There are several key assumptions that must be met in order for RD to be a viable method of estimating causal effects. If the assumptions are tenable then the actual RD statistical model can be a fairly straight-forward OLS (Ordinary Least Squares)

regression. It is important to note that a general OLS model may look very similar to an OLS model that is part of a RD analysis, but the results can differ in important ways. Point estimates of effects can differ; even if they are similar, statistical significance can differ as well. This is highlighted in Robison (2011) where OLS and RD results (although generally similar in magnitude) would lead to different conclusions regarding the appropriateness of exit criteria. Remember, data structure and analytic methods (i.e. RD) may lead to very similar point estimates of effects as do less carefully constructed OLS models, but contain different standard errors. Incorrect standard errors result in incorrectly rejecting or not rejecting null hypotheses.

The primary assumption underlying RD analysis is that there is a cut score that distinctly identifies two groups (students, schools, etc.) This cut score is based on a "forcing" or "rating" variable and this variable is continuous[74]. The rating variable must be on an interval scale or ranks (Keefer, 2016). One conception of RD is that students very close to the cut score on either side are not generally distinguishable; that is, around the cut score there is a local randomized experiment. For example, if the English language proficiency cut score on an assessment is 5.0, then students scoring at 4.9 or 5.1 are likely not demonstrating substantively different English language skills, rather the variation in scores immediately around the cut is due to imprecision of the assessment. Taking this view, the analyst would use a local linear regression and estimated treatment/program/intervention effects would be limited to students within a particular bandwidth on either side of the cut score. This is the "local average treatment effect". There are methods for developing the optimal bandwidth, but this can also be driven by subjective criteria, as well as testing various bandwidths and checking the robustness of estimates across those bandwidths. Viewing the cut score as a discontinuity along the rating continuum allows for the estimation of an average treatment effect. This latter view is more sensitive to model misspecification, which is discussed below.

There are several assumptions that impact the ability to make causal claims. First, the rating variable is not caused or influenced by the treatment. Second, the rating variable must be measured before the treatment. Third, the cut point must be exogenous. This important aspect implies that the cut score is not manipulated by the subjects (students, schools, etc.) and that it is not set based on who is above or below the cut score. Fourth, there should be a single intervention; if there are multiple interventions, then the causal claim is limited to the combined impact. And finally, the correct functional form must be

[74] The description focuses on a sharp RD design and does not present a fuzzy design which can address some limitations of the sharp design, but which requires significantly more complex modeling (Instrumental Variables) and is beyond the scope of this handbook. Fuzzy designs are described in Jacob, Zhu, and Bloom (2012).

used to estimate the relationship between the rating variable and the outcome (Jacob, et. al., 2012).

The assumptions underlying RD lead to many operational challenges for analysts. These include: (1) manipulation of the rating variable, (2) sample size, (3) distribution of score variable, (4) location of cut score, and (5) specification of the correct functional form (Louie, Rhoads, & Mark, 2016).

(1) Manipulation of the rating variable is particularly relevant for research and evaluations on ELs. For example, in an attempt to evaluate the impact of exiting EL status in a state that use multiple exit criteria and one of these criteria is subjective, states may find that the rating variable is subject to manipulation. One check against this is to plot a histogram of the rating variable and determine whether there is a jump in the distribution at the cut score (Jacob, et. al., 2012).

(2) Sample sizes for RD analysis need to be substantially larger: this is due to the fact that the analysis uses only observations close to the cut score and thus much of the sample is excluded.

(3) It is also important to examine whether the distribution of the forcing variable is normal around the cut score without substantial irregularities.

Another issue is (4) the location of the cut score. If the cut score tends towards a tail of the distribution, it results in a very small N at one side of the cut. Moreover, this places limitations on bandwidth if a local regression approach is desired. In an RD design, the sample size needs to be about 2.5 times as large as in an RCT to achieve the same power (Louie, Rhoads, & Mark, 2016).

(5) The correct functional form of the OLS model is critical. For example if the underlying relationship between X and Y is non-linear than a linear specification may erroneously find a an effect that is, in fact, due to specification error. The analyst is encouraged to read Jacob, Zhu and Bloom (2012) for more details on model specification.

Example 4 demonstrates the use of RD to help provide empirical support for a relevant policy decision.

Example 4: Evidence-Based Policy Decision

One application of RD analysis is to examine the impact of minimum N on school summative scores. The rating variable in this case is the number of ELs (with valid gain scores) attending a school; the outcome is the summative accountability score. This evaluation focuses on the policy decision to set the minimum N at (in this example) 20. We can examine whether schools below the minimum N of 20—meaning they do not include the ELP Indicator in the summative score—score significantly differently than schools above the minimum N, and include a separate ELP Indicator in the summative score. In this case the SEA would want to see no "jump" or discontinuity at the cut score because this would imply that the minimum N is causing schools to have significantly different summative scores simply because they are held accountable for EL progress. Meeting the underlying assumptions is more straight-forward because the results are based on a policy that has not been implemented. The first assumption is tenable because the rating variable is not caused by the treatment. Assumption two is that the rating variable is measured before the treatment and, to the extent that the rating variable is included in the dataset, before the minimum N was set or the summative scores were calculated, this assumption is met. The third assumption is tenable because the rating variable is exogenous. The fourth assumption relates to a single intervention, which is presumed to be the minimum N cut score. The fifth and final assumption relates to the functional form of the model, which is addressed in more detail below.

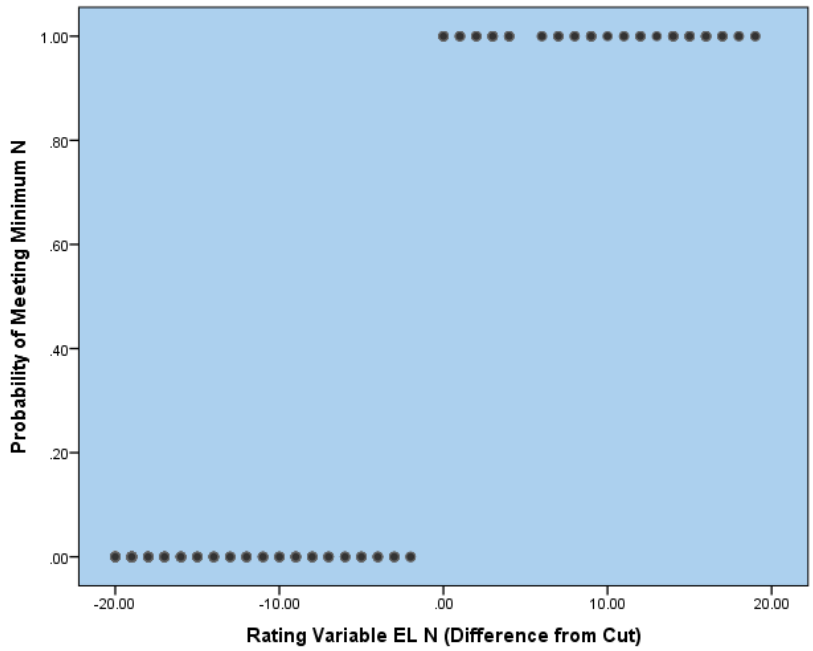


FIGURE 27: PROBABILITY OF MEETING MINIMUM N ALONG THE RATING VARIABLE.

Figure 27 indicates that every school that is below the cut score does not meet the minimum N (20)[75] and therefore does not include a separate ELP Indicator as part of the summative score, while every school above the minimum N does include the ELP Indicator in the summative score.

The SEA would then want to plot the relationship between the outcome and the rating variable. Figure 28 indicates that there appears to be no explicit discontinuity at the cut score. Preliminary analyses indicated[76] that the bandwidth for analyses should fall in the region of 0 to 40 ELs as this encompasses most schools, though it excludes very large schools. This limits the claims regarding the impact of the ELP Indicator to schools within the bandwidth.

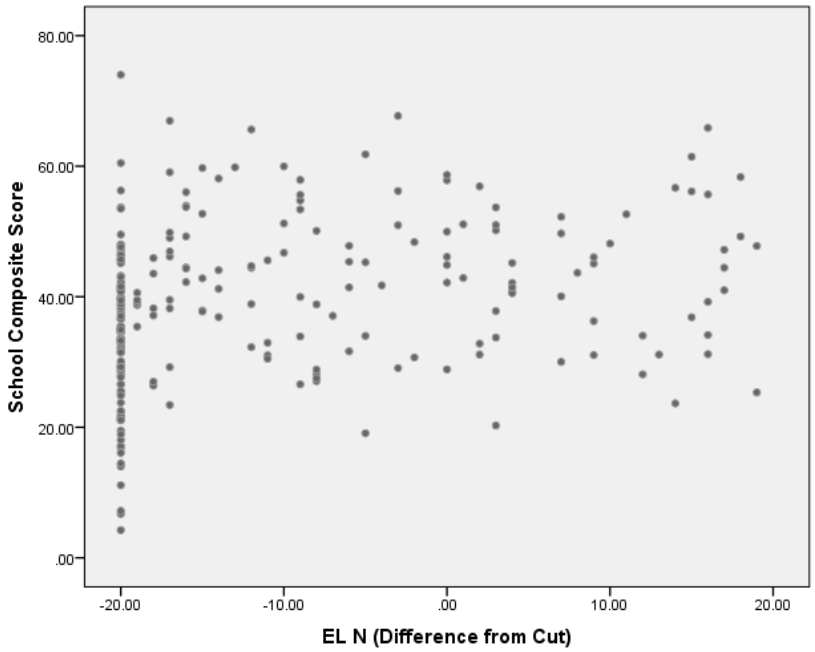


FIGURE 28: SUMMATIVE (COMPOSITE) SCORE AGAINST THE RATING VARIABLE

Another useful check is to examine other variables against the rating variable to determine whether there are any unanticipated discontinuities in variables that should not be impacted by the cut score.

[75] The x axis scale is in EL N—the cut, which is EL N – 20. This way the cut is at 0.

[76] Another way to view this is to consider the entire sample and evaluate the impact of trimming the top and bottom 1% or up to 5%. In this case schools with no ELs were included as an auxiliary analytic interest – the assumption that EL N = 0, is not manipulated may not be tenable.

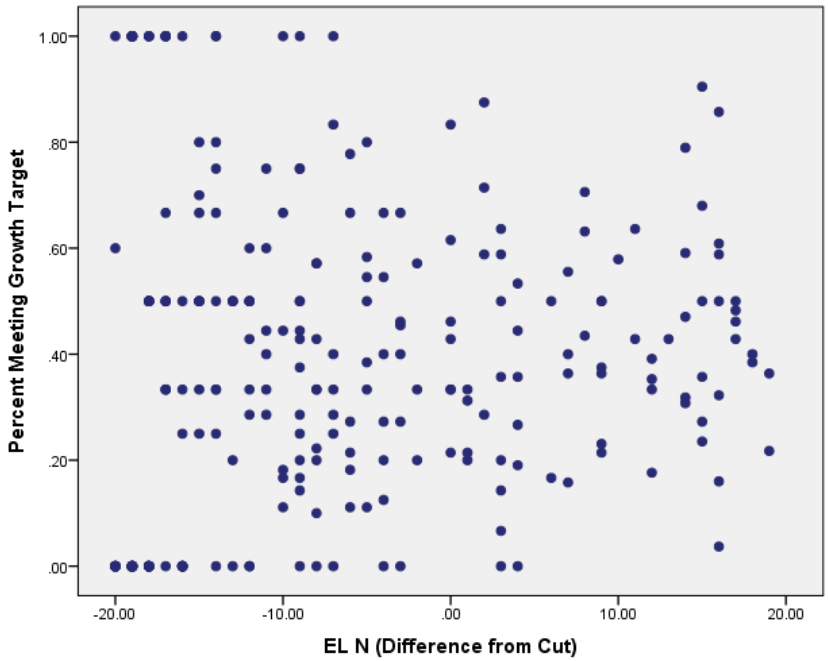


FIGURE 29: EL PROGRESS AGAINST THE RATING VARIABLE

One particularly interesting variable to examine is whether EL progress differs on either side of the cut score. If schools just above the cut score were particularly successful (or particularly unsuccessful), then the fact that a school fell below the cut and didn't receive credit for EL progress would cast doubt on the impact of the cut score as causing schools to have different summative ratings. To check whether there is a discontinuity at the cut score in the percent of students that are proficient or advanced is useful for similar reasoning as for EL progress. Figure 30 displays the relationship.

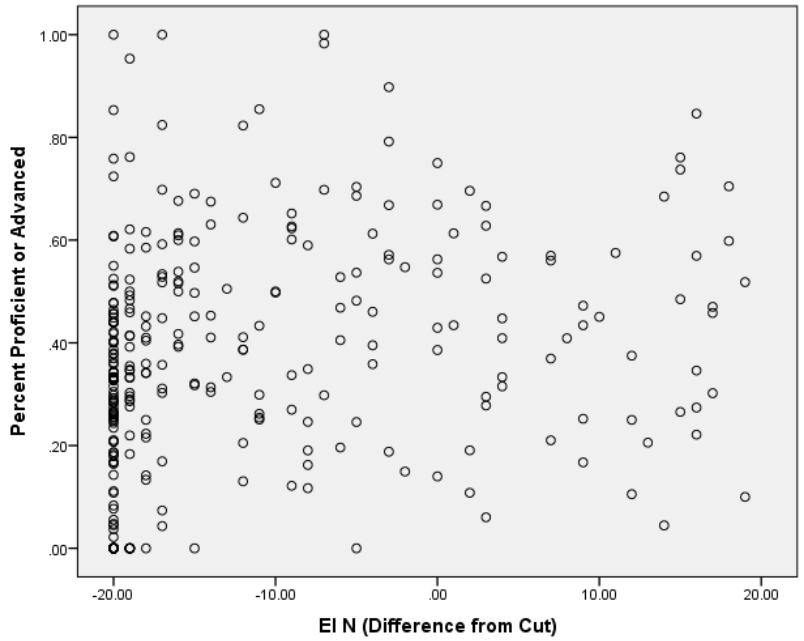


FIGURE 30: PERCENT PROFICIENT OR ADVANCED AGAINST THE RATING VARIABLE

Next, it is useful to plot the distribution of the rating variable to examine whether there appears to be manipulation of the variable; this is shown in Figure 31. Visual inspection is reasonable, but specific tests are available to examine the density around the cut score (Jacob et. al., 2012). Here, there appears to be a slight bump in the number of schools just above the cut score, but overall the general trend is consistent.

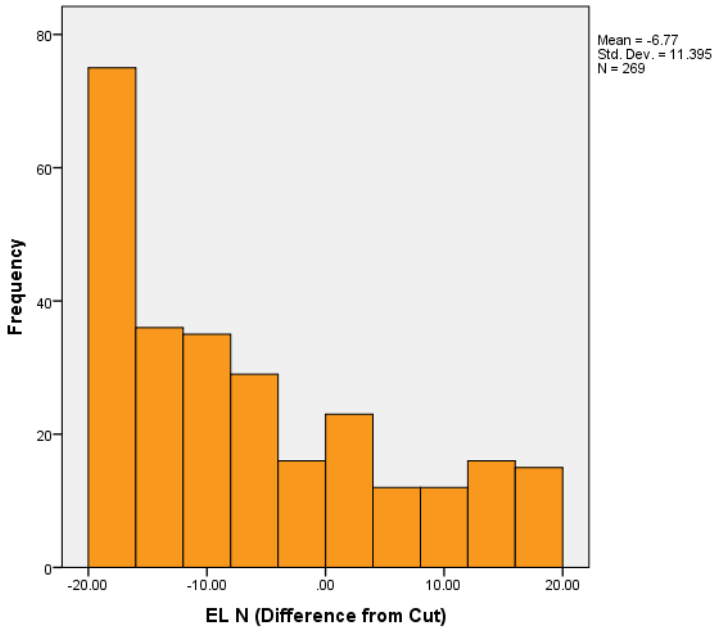


FIGURE 31: DISTRIBUTION OF THE RATING VARIABLE

At this point, the RD model can finally be developed. A linear model is presented but polynomials are generally applied and then eliminated to examine the robustness of the estimates as well as to guard against model misspecification.

This example presents only two models: a linear model and a model with an additional covariate that is clearly warranted to improve model specification.

$$Y_j = b_0 + b_1 Cut + b_2 EL\_N\_DC + b_3 N\_ Interaction + e_j \quad (17)$$

Where Y_j is the composite summative score for school *j*, *Cut* is an indicator variable taking on a value of 0 if a school does not meet the minimum N and 1 if it does, *EL_N_DC* is a variable that is equal to the number of ELs in a school minus 20 (i.e. EL N – the minimum N required to include the ELP Indicator), *N_Interaction* is the joint *Cut* by *EL_N_DC* variable created by multiplying the two variables together, and b0 is the intercept. In Model 17, the key coefficient is b1 which represents whether schools perform differently on Y based on whether they meet the minimum N for including the ELP Indicator or not. The coefficient b3 is also important because it estimates the effect of EL N on both sides of the cut. Model 17 is expanded to:

$$Y_j = b_0 + b_1 Cut + b_2 EL\_N\_DC + b_3 N\_ Interaction + b_4 EL\_0 + e_j \quad (18)$$

Where EL_0 is an indicator that takes on a value of 1 if the school has no ELs and 0 otherwise. The reason for creating a separate indicator for schools with no ELs is discussed below.

Table 31 displays the results for Models 17 and 18. The results of Model 17 imply that there is a significant shift in the composite score of -5.8 at the cut score, meaning there is a relationship between EL N and the composite summative score; this relationship changes at the cut score. This shows that schools that must include the ELP Indicator (near the cut) will be disadvantaged, and that they face a different relationship (i.e. almost no relationship) between EL N and the summative score compared to schools that do not include the ELP Indicator.

However, the Figures 30 and 31 indicate that there are many schools that have a zero value for EL N. One cannot assume the mechanism that results in 0 ELs in a school. One option to address the unknown mechanism would be to exclude these schools from the analyses; another option is to include an indicator variable. It is important not to attach causal reasoning behind any uncovered effects, but, in this instance, it is merely to address the fact that a high proportion of schools have no ELs.

TABLE 31 RD ANALYSIS MODEL RESULTS

Model		B	Std. Error	t	Sig.
17	(Constant)	48.98	2.87	17.04	0.00
	Meet_Min_N	-5.80	3.89	-1.49	0.14
	EL_N_DC	0.67	0.17	4.00	0.00
	N_Interaction	-0.65	0.30	-2.15	0.03
18	(Constant)	42.95	3.28	13.12	0.00
	Meet_Min_N	0.22	4.15	0.05	0.96
	EL_N_DC	-0.02	0.25	-0.07	0.95
	N_Interaction	0.04	0.35	0.12	0.91
	EL_0	-9.45	2.66	-3.56	0.00

a. Dependent Variable: Y = Composite_ALL1a

The results in Table 31 for Model 18 indicate that inferences are significantly different once the model is more appropriately specified. Schools that have no ELs score approximately 9.45 points lower on the summative score than schools that have at least 1 EL. Further, the results now indicate that the cut and the estimated slopes for EL N and the interaction are all statistically not different from 0. This implies that schools required to include the ELP Indicator are not disadvantaged because they have the ELP Indicator. Results reported in Part II are corroborated by the fact that there is no relationship between the number of ELs and the summative score, except here we see there is a substantively meaningful and statistically significant difference between schools with no ELs and schools with one or more ELs. A separate analysis can investigate further why this might be the case.

As noted, RD analysis can become quite complex if all of the technical aspects are strictly adhered to. It is recommended that the analyst use some of the references in this handbook to provide more in-depth guidance. Still, as pointed out in Jacob et al. (2012), graphs that do not demonstrate discontinuities likely do not have discontinuities, nor are they likely to have the statistical models providing supporting empirical evidence.

Summary

This Handbook provides many options for monitoring the ELP Indicator and conducting evaluations on the progress of ELs. It should not be inferred that these are the only options, nor should it be inferred that the various methods (especially in Part III) have received a full and comprehensive treatment. The examples are meant to provide a baseline that analysts can use to begin the process of evaluation. It is understood that SEA staff have limited time, and the intent of the handbook is to provide sufficient information to evaluate programs with existing data. Still analysts are encouraged to expand their methodological toolkit and follow-up with the various references that provide more in-depth presentations of the various strategies – particularly related to more complex situations and the caveats of applying these methods.

Most importantly, SEA senior staff should be encouraged to facilitate consistent monitoring and evaluation of EL progress in both English language development and on content skills and knowledge. These analyses should be undertaken with the knowledge that results may not always align with desired outcomes, and that such results are not a reflection of the motivation, desire, or intent of the facilitators, but an opportunity to improve existing services and programs. SEAs have invested a great deal in developing comprehensive data systems, and in order to realize the benefits of these systems SEAs are encouraged to develop evaluation policies that guide the when, why, and how evaluations of programs need to be initiated.

References

Abedi, J. (fall, 2008). Classification system for English language learners: Issues and recommendations. Educational Measurement: Issues and Practice, 27(3). Retrieved from: http://education.ucdavis.edu/sites/main/files/LEP_Class_EMIP_New.pdf

Aitkin, M., & Longford, N. (1986). Statistical modeling issues in school effectiveness studies. *Journal of the Royal Statistical Society, 149*(1), 1-43.

Austin, P. (2011). An Introduction to Propensity Score Methods for Reducing the Effects of Confounding on Observational Studies, Multivariate Behavioral Research, 46: 399-424.

Baker, E.L. (2003). Multiple Measures: Towards Tiered Systems, *Educational Measurement: Issues and Practice*, 22, p.13-17.

Burstein, L. (1980). The analysis of multi-level data in educational research and evaluation. *Review of Research in Education*, *4*, 158-233.

Castellano, K, & A. Ho (2013). *A Practitioner's Guide to Growth Models*, Washington DC: Council of Chief State School Officers.

Chester, M. (2003). Multiple measures and high stakes decisions: A framework for combining measures, *Educational Measurement: Issues and Practice*, 22, p.32-41.

C. Kilchan, P. Goldschmidt, and K. Yamashiro (2005) Exploring Models of School Performance: From Theory to Practice, *National Society for the Study of Evaluation*, v.104.

Cook, G., Boals, T., & Lundberg, T. (2011, November). Academic achievement for English learners: What can we reasonably expect? Kappan, 93 (3), 66-69. Retrieved from https://www.wida.us/get.aspx?id=485.

Cook, H. G., Linquanti, R., Chinen, M., & Jung, H. (2012). 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; Policy and Program Studies Service. Retrieved from http://www2.ed.gov/rschstat/eval/title-iii/implementation-supplemental-report.pdf.

Cooper H., and L. Hedges (1994). The Handbook of Research Synthesis, Sage, New York.

Deng, L., Freeman, L. (2011). Planning for evaluation: Using regression discontinuity to evaluate target place-based programs, *Journal of Planning and Education Research*, 31(3), p.308-318.

Dew, J. R. (1991). In Search of the Root Cause, *Quality Progress*, 24(3): 97-107.

Foster, J, M. McGillivray, S. Suman (2013). Composite Indices: Rank Robustness, Statistical Association, and Redundancy, *Econometrics Reviews*, 32(1): 35-46.

Goldschmidt, P., K. Hakuta, (2016*). Incorporating English Learner Progress into State Accountability Systems.* Washington DC: Council of Chief State School Officers.

Goldschmidt, P., K. Choi, J.P. Beaudoin (2012). Growth Model Comparison Study: Practical Implications of Alternative Models for Evaluating School Performance, Council of Chief State School Officers, Washington, DC.

Goldschmidt, P., K.C. Choi, F. Martinez, J. Novak (2010). Using growth models to monitor school performance: comparing the effect of the metric and the assessment, *School Effectiveness and School Improvement*, 21(3), 337-357.

Goldschmidt, P., P. Roschewski, K. Choi, W. Auty, S. Hebbler, R. Blank, A. Williams (2005). Policymakers Guide to Growth Models for School Accountability: How do Accountability Models Differ? The Council of Chief State School Officers, Washington, DC.

Hagerty, M., K. Land (2007), Constructing summary indices of quality of life: A model for effects of heterogeneous importance wrights, *Sociological Methods & Research*, 35(4), p.455-496.

Hakuta K., Butler Y., Witt D. (2000). How long does it take English learners to attain proficiency? University of California Linguistic Minority Research Institute Policy Report 2000-1. Santa Barbara: University of California, Linguistic Minority Research Institute.

Hardy, M. (1993). *Regression with Dummy Variables*, Series: Quantitative Applications in the Social Sciences, Sage, Newbury Park.

Huang, D., P. Goldschmidt, and La Torre Matrundola, D. (2014) Examining the Long-Term Effects of Afterschool Programming on Juvenile Crime: A Study of the LA's BEST Afterschool Program, *International Journal for Research on Extended Education*, 2, 113-134.

Henderson-Montero, D., M. Julian, W. Yen (2003). Multiple Measures: Alternative Design and Models, *Educational Measurement: Issues and Practice*, 22, p.7-12.

Herman, J. L., Heritage, M., Goldschmidt, P. (2011). Developing and selecting assessments of student growth for use in teacher evaluation systems (extended version). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Hox, J. (2002). Multilevel Analysis: Techniques and Applications, Lawrence Erlbaum, Mahwah, NJ.

Jacob, B., L. Lefgran (2004). Remedial education and student achievement: A regression discontinuity analysis, review of economics and statistics, 86(1): 226-244.

Jacob, R., P. Zhu, M-A Somers, H. Bloom (2012). *A Practical Guide to Regression Discontinuity*, MDRC.

Jones, E.A. (2016). A multiple cut-off regression discontinuity analysis of the effects of tier 2 reading interventions in a Title I elementary school, *All Theses and Dissertations.* Paper 6097, BYU.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), Educational measurement (4th ed., pp. 17-64). Washington, DC: The National Council on Measurement in Education & the American Council on Education.

Keefer, Q. (2016) Rank-based groupings and decision making: A regression discontinuity analysis of the NFL Draft rounds and rookie compensation, *Journal of Sports Economics*, Vol. 17(7) 748-762 DOI: 10.1177/1527002514541448

Kim, J., Herman, J. L. (2010). When to exit ELL students: Monitoring success and failure in mainstream classrooms after ELLs' reclassification (CRESST Report 779). Los Angeles, CA: UCLA, Graduate School of Education and Information Studies; CRESST. Retrieved from https://www.cse.ucla.edu/products/reports/R779.pdf.

Kolen, M. & R. Brennan (2004). *Test Equating and Scaling: Methods and Practice*, 2nd edition, Springer.

Levin, H. and P. McEwan (2001). *Cost-Effectiveness Analysis: Methods and Applications,* 2d ed. Thousand Oaks, Calif.: Sage Publications, ISBN 0-7619-1934-1

Linn, R. L. (1998). *Standards-based accountability: Ten suggestions* (CRESST Policy Brief No. 2). Los Angeles: University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Linquanti R., K. Hakuta (2012). How next generation standards and assessments can foster success for California's English learners (PACE Policy Brief No. 12-1). Palo Alto: Policy Analysis for California Education.

Louie, J., C. Rhoads, J. Mark (2016). Challenges to using the regression discontinuity design in educational evaluations: Lessons from the transition to algebra study, *American Journal of Evaluation*, 37(3), p.381-407.

Lyons, S., N. Dadey (2017). *Considering English Language Proficiency within Systems of Educational Accountability under Every Student Succeeds Act*, National Center for the Improvement of Educational Assessment.

MacKinnon, D. (2008). *Introduction to Statistical Mediation Analysis*, Psychology Press, Taylor & Francis, New York.

Martinez, J.F., J. Schweig, P. Goldschmidt (2016). Approaches for combining multiple measures of teacher performance: Reliability, validity, and implications for evaluation policy *Evaluation Policy, Educational Evaluation and Policy Analysis*.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*, 741-749.

OECD (2008). Handbook on Constructing Composite Indicators: Methodology and Users Guide, OECD.

Ogawea, R., E. Collom (2002). Using indicators to hold schools accountable: Implicit assumptions and inherent tensions, *Peabody Journal of Education*, 75(4), 200-215.

Percarpio, K., B. Watts, W. Weeks (2008). The Effectiveness of Root Cause Analysis: What Does the Literature Tell Us? *The Joint Commission on Quality and Patient Safety*, 34(7): 391-398.

Porter, A. (1991). Creating a system of school process indicators, *Educational Evaluation and Policy Analysis*, 13(1), 13-29.

Profit, J., K. Typpo, S. Hysong, L. Woodard, M. Kallen, L. Peterson (2010). Improving benchmarking by using an explicit framework for the development of composite indicators: an example using pediatric quality of care. *Implementation Science*, 5(13).

Robinson-Cimpian, J. P., Thompson, K. D. (2016). The effects of changing test-based policies for reclassifying English learners. *Journal of Policy Analysis and Management*, *35*(2), 279-305. DOI: 10.1002/pam.21882.

Rubin, Donald B. (2008). For Objective Causal Inference, Design Trumps Analysis. *Annals of Applied Statistics* 2(3):808–840.

Saisana M, A. Saltelli, S. Tarantola (2005). Uncertainty and sensitivity analysis techniques as tools for the quality assessment of composite indicators. *Journal of the Royal Statistical Society Series A*. 168:307–323.

Seastrom, M. (2017). Best practices for determining subgroup size in accountability systems while protecting identifiable student information. (IES 2017-147). U.S. Department of Education, Institute of Education Sciences. Washington DC.

Shadish, W. (2013). Propensity Score Analysis' Promise, Reality, and Irrational Exuberance, Journal of Experimental Criminology, 9:129-144.

Shadish, W., M. Clark, and P. Steiner (2008). Can nonrandom experiments yield accurate answers?, Journal of the American Statistical Association, 103: 1334-1343.

Thistlethwaite D., D.T. Campbell (1960). Regression discontinuity analysis: An alternative to the ex-post-facto experiment, *Journal of Educational Psychology*, 51(6):309-317.

Schochet, P. (2008). *Technical Methods Report: Guidelines for Multiple Testing in Impact Evaluations* (NCEE 2008-4018). Washington DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, US Department of Education.

Willms, D. Raudenbush, S. (1989). A longitudinal hierarchical linear model for estimating school effects and their stability, *Journal of Educational Measurement*, 26(3), 209-232.

Wu, A., A. Lipshutz, P. Pronovost (2008). Effectiveness and efficiency of root cause Analysis in Medicine, *Journal of the American Medical Association*, 299(6): 685-687.

Appendix A: Data description

District, school, and student IDs can be fictitious but need to be constant over time in the dataset so that all are linked for four years (if four years are available).

Variable	**Description**
District_Code	
School_Code	
STID	(Student ID)
Year	(School Year, e.g., 13-14 = 14, 14-15 =15, etc.)
School level	(Elementary, Middle, High)
EL_n	(number of ELs in the school)
REL_n	(number of RELs in the school)
Grade	(Student's current grade K =0, 1 = 1, etc.)
English Proficiency Status	(e.g., EO, EL, REL, IFEP, etc.)
• EL	(0= Not EL, 1 = EL)
• REL	(0 = Not Reclassified EL, 1 = Reclassified EL)
• EO	(0 = Not English Only, 1 = English Only)
Ethnicity	
• WHITE	(0 = not White, 1 = White)
• BLACK	as above
• HISPANIC	""
• ASIAN	""
• NATIVE	""
Gender	(0= male, 1 = female)
FRL (Free/Reduced Lunch)	(0=Not FRL, 1 = FRL)
SWD (Students with Disabilities)	(0= Not SWD, 1 = SWD)
PL_M	(content performance level – Math)
PL_R	(content performance level – English Language Art [ELA])
PnA_M	(content Proficient or Above– Math)
PnA_R	(content Proficient or Above – ELA)
SS_M	(content scale score – Math)
SS_R	(content scale score – ELA)
SS_Mp	(content scale score prior year– Math)
SS_Rp	(content scale score prior year– ELA)
SS_Mp2	(content scale score 2 years prior– Math)
SS_Rp2	(content scale score 2 years prior – ELA)
SS_Mp3	(content scale score 3 years prior– Math)
SS_Rp3	(content scale score 3 years prior – ELA)
ELP_SS	(English Language Proficiency Assessment [ELPA] scale score)
ELP_SSp	(ELPA scale score prior year)

ELP_SSp2 (ELPA scale score 2 years prior)
ELP_SSp3 (ELPA scale score 3 years prior)
ELP_SSp4 (ELPA scale score 4 years prior)
ELPss_init (Initial ELPA scale score)
ELP_Nocc (ELPA test occasion number of test)
ELD_lvl (ELD level current year)
ELD_lvl_int (Initial ELD level)
ELD_lvlp (ELD level prior year**)**
ELD_Prof (0= Not English Proficient, 1 = English Proficient)
FAYnot (Full Academic Year) (0= FAY, 1 = Not FAY)
REL_Yrcnt (Year as REL)
RAEL (0= not recently arrived EL, 1 = recently arrived)
RAELp (0= not recently arrived prior year EL, 1 = recently arrived prior year)
RAEL_Yr (Year as Recently Arrived EL)
SIFE (Student with Interrupted Formal Education)
(0= Not SIFE, 1 = SIFE)

Note: Other state specific relevant indicators that warrant additional attention can be added to the list

Variables to Calculate:

ELPss_Gain = ELPss – ELPssp.
ELPss_Gain2 = (ELPss – ELPssp2)/2.
ELP_GainBase = (ELPss – ELPss_init)/ELP_Nocc.
ELDlvl_Chg = ELD_lvl – ELD_lvlp.
ELP_SSm = ELPA mean by Grade.
ELPsspsq = ELPssp^2.
ELPsspcb = ELPssp^3.
Grade2 = Grade^2.

Note: If there is a desire to adjust gain scores by the reliability of the assessment, then the ELPA reliability by grade is needed.

Dataset Structure

The structure is long and wide. Below shows only a few variables.

School_Code	STID	Year	Grade	EL	REL	ELPss	ELPssp	ELPss_Gain
001	100	13	3	1	0	250	225	25
001	100	14	4	1	0	300	250	50
001	100	15	5	1	0	315	300	14
002	100	16	6	0	1	350	315	35

Appendix B: Initial ELD Level by Initial Grade

Initial Grade	Initial EL Level Level 1	Level 2	Level 3	Level 4	Level 5	Total
K	3293	1832	1774	3923	323	11145
1	842	369	245	201	12	1669
2	626	237	125	135	17	1140
3	674	202	84	82	9	1051
4	632	180	78	89	17	996
5	558	149	109	85	12	913
6	645	123	75	74	10	927
7	500	165	74	69	9	817
8	377	109	59	76	10	631
9	493	244	143	131	31	1042
10	143	136	101	119	16	515
11	73	74	79	101	28	355
12	32	8	8	20	8	76
Total	8888	3828	2954	5105	502	21277

Appendix C: Distribution of Time in Program by Grade

Grade	Number of Years EL				Total Number of ELs
	1 - 3 Yrs	4 - 6 Yrs	7 - 9 Yrs	10 - 12 Yrs	
K	2406	0	0	0	2,406
1	2721	1	0	0	2,722
2	2135	204	0	0	2,339
3	713	1276	1	0	1,990
4	662	1150	5	0	1,817
5	593	767	119	0	1,479
6	678	371	614	1	1,664
7	615	269	452	9	1,345
8	644	289	324	100	1,357
9	740	275	160	153	1,328
10	772	257	135	132	1,296
11	534	173	104	86	897
12	241	221	122	62	646
	13454	5253	2036	543	21286

Appendix D: Data and Calculations for Decision Consistency (DC)

ELP Level	Proficient: No	Proficient: Yes	Total
1.00	7	0	7
1.25	13	0	13
1.50	3	0	3
1.75	17	0	17
2.00	68	0	68
2.25	161	0	161
2.50	214	0	214
2.75	252	0	252
3.00	372	1	373
3.25	456	0	456
3.50	574	1	575
3.75	516	3	519
4.00	861	5	866
4.25	878	11	889
4.50	872	31	903
4.75	856	32	888
5.00	212	23	235
5.25	192	24	216
5.50	181	27	208
5.75	107	38	145
	6812	196	7008

A Sum No	B Sum Yes	C =A+B Total	C/Total N DC
2137	194	2331	33.3%
2653	191	2844	40.6%
3514	186	3700	52.8%
4392	175	4567	65.2%
5264	144	5408	77.2%
6120	112	6232	88.9%
6332	89	6421	91.6%
6524	65	6589	94.0%
6705	38	6743	96.2%

Appendix E: Percentiles of ELA Performance by ELD Level and EO

	Percentiles				
ELD Level	5	25	50	75	95
Level 1	181	216	226	242	260
Level 2	185	220	236	249	269
Level 3	201	234	252	268	291
Level 4	236	260	275	289	310
Level 5	260	283	295	309	330
English Only	244	285	308	329	357

Appendix F: Decision Consistency between the Listening Domain and Domain Sum Score

	Listening Levels						
Doman Sum	1	2	3	4	Total	Listening Level 3	Listening Level 4
4	309	0	0	0	309		
5	311	204	0	0	515		
6	258	986	2	0	1246		
7	149	1487	30	4	1670		
8	67	2066	84	1	2218	58.8%	21.6%
9	21	3050	245	2	3318	68.3%	31.9%
10	2	2105	1041	43	3191	81.5%	47.4%
11	1	1080	1819	22	2922	86.3%	61.9%
12	0	367	2361	72	2800	82.7%	75.3%
13	0	39	1931	183	2153	73.1%	87.7%
14	0	0	476	297	773	63.4%	96.0%
15	0	0	50	163	213	59.8%	96.9%
16	0	0	0	102	102	58.8%	96.3%
Total	1118	11384	8039	889	21430		

Appendix G: Calculating the ICC

```
VARCOMP Outcome BY School
  /RANDOM=School
  /METHOD=MINQUE(1)
  /DESIGN=School
  /INTERCEPT=INCLUDE.
```

Using the Handbook data the results are:

Variance Estimates

Component	Estimate
Var(School_Code)	.025
Var(Error)	.229

ICC = ICC = $S^2_b/(S^2_b + S^2_w)$ = .025/(.025 + .229) = .098.

Unit Reliability, ρ, can be estimated from the variance components as well:

$\rho = S^2_b/(S^2_b + S^2_w/n_j)$, where n_j is the number of students in school j.

For example with n_j = to 50 (the mean of the sample):

ρ = .025/(.025 + .229/50) = .025/(.025 + .0046) = .025/.02958 = .845.

Appendix H: Probability of Meeting Growth or Level Targets –All Years

Initial_ELD		P(Meet)	N	S.D.
1.00	2	0.86	1574.00	0.35
	3	0.62	1216.00	0.49
	4	0.33	891.00	0.47
	5	0.34	1159.00	0.47
	6	0.26	837.00	0.44
	7	0.27	702.00	0.44
	8	0.25	473.00	0.43
	9	0.21	304.00	0.41
	10	0.27	228.00	0.45
	11	0.19	150.00	0.39
	Total	0.46	7534.00	0.50
2.00	2	0.66	797.00	0.47
	3	0.33	511.00	0.47
	4	0.18	423.00	0.39
	5	0.32	237.00	0.47
	6	0.22	142.00	0.41
	7	0.27	136.00	0.45
	8	0.17	63.00	0.38
	9	0.26	43.00	0.44
	10	0.21	34.00	0.41
	11	0.27	37.00	0.45
	Total	0.39	2423.00	0.49
3.00	2	0.48	600.00	0.50
	3	0.33	446.00	0.47
	4	0.13	359.00	0.33
	5	0.27	188.00	0.45
	6	0.14	139.00	0.34
	7	0.26	76.00	0.44
	8	0.19	42.00	0.40
	9	0.29	42.00	0.46
	10	0.20	25.00	0.41
	11	0.19	27.00	0.40
	Total	0.31	1944.00	0.46
4.00	2	0.16	1726.00	0.37
	3	0.31	930.00	0.46
	4	0.11	539.00	0.31
	5	0.33	76.00	0.47
	6	0.23	60.00	0.43

	7	0.38	24.00	0.49
	8	0.36	14.00	0.50
	9	0.44	16.00	0.51
	Con't			
	10	0.00	6.00	0.00
	11	0.15	13.00	0.38
	Total	0.20	3404.00	0.40
5.00	1	1.00	499.00	0.00
	5	1.00	1.00	
	Total	1.00	500.00	0.00
Total	1	1.00	499.00	0.00
	2	0.52	4697.00	0.50
	3	0.44	3103.00	0.50
	4	0.21	2212.00	0.41
	5	0.33	1661.00	0.47
	6	0.24	1178.00	0.43
	7	0.27	938.00	0.44
	8	0.24	592.00	0.43
	9	0.23	405.00	0.42
	10	0.25	293.00	0.44
	11	0.20	227.00	0.40
	Total	0.39	15805.00	0.49

Appendix I: Testing for Moderation

```
compute Initial_ELD0 = INitial_ELD -1.
compute InitialbyRAELp = Initial_ELD0 * RAELp.

TEMPORARY.
select if (ELP_Nocc eq 2).
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT  Met_Grth_Target
  /METHOD=ENTER Initial_ELD0 RAELp .

TEMPORARY.
select if (ELP_Nocc eq 2).
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS BCOV R ANOVA change
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT Met_Grth_Target
  /METHOD=ENTER Initial_ELD0 RAELp InitialbyRAELp.
```

Appendix J: Testing for Mediation

```
*** Model 9***.
TEMPORARY.
select if (ELP_Nocc eq 2).
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS BCOV R ANOVA change
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT Met_Grth_Target
  /METHOD=ENTER Initial_ELD0 .

**Model 10****.
TEMPORARY.
select if (ELP_Nocc eq 2).
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS BCOV R ANOVA change
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT Met_Grth_Target
  /METHOD=ENTER Initial_ELD0 Mobility.


**Model 11****.
TEMPORARY.
select if (ELP_Nocc eq 2).
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS BCOV R ANOVA change
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT Mobility
  /METHOD=ENTER Initial_ELD0.
```

Appendix K: RD analysis

```
COMPUTE Meet_Min_N =0.
if (EL_N ge 20) Meet_Min_N =1.

compute EL_N_DC = EL_n - 20.
compute N_Interaction = EL_N_DC*Meet_Min_N.

if (EL_N eq 0) EL_0 =1.
if (EL_N gt 0) EL_0 =0.

REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA change
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT Composite_ALL1a
  /METHOD=ENTER Meet_Min_N EL_N_DC N_Interaction
  /method=enter EL_0.
```

Appendix L: Calculating the effective weight of an Indicator in Excel

$$WE_j = WN_j^2 + \sum WN_j WN_k \rho_{jk}$$

Where WE = the effective weight, WN = the nominal, or policy, weight, and rjk equals the correlation between indicator pairs.

Assuming three indicators, then.

The pairwise correlations among indicators is:

	Status	Growth
Status		
Growth	-0.11 (B)	
ELP	0.06 (C)	-0.11 (D)

Each correlation is denoted with a letter for identification in the next step.

The next step is to multiple and then sum all the Indicator weights and correlations as presented below.

	(A)			(M)		(E)		(F)	(G)		(H)	(I)	(J)	(K)	(L)
				Effective											
	Nominal			L1/Sum(L)		A1^2		A1*A2	*B		A1*A3	*C	F*G	H*I	E + J + K
1	0.4	Status		0.46		0.16	+	0.16	-0.11	+	0.08	0.06	-0.0176	0.0048	0.1472
				L2/Sum(L)		A2^2		A1*A2	*B		A2*A3	*D	F*G	H*I	
2	0.4	Growth		0.42		0.16	+	0.16	-0.11	+	0.08	-0.11	-0.0176	-0.0088	0.1336
				L2/Sum(L)		A3^2		A1*A3	*C		A2*A3	*D	F*G	H*I	
3	0.2	ELP		0.11		0.04	+	0.08	0.06	+	0.08	-0.11	0.0048	-0.0088	0.036
														SUM =	0.3168

Column E - Squares the nominal weight.
Column F – multiples the nominal weight of Indicator j and Indicator k1.
Column G – is the corresponding correlation between Indicators.
Column H - multiples the nominal weight of Indicator j and Indicator k2.
Column I - is the corresponding correlation between Indicators.
Column J – multiples columns F and G.
Column K – multiplies columns H and I.
Column L – Sums columns E, J and K. The sum of these sums is at the bottom of this column.
Column M – divides the row sum in column L by the sum of all rows in column L.

Appendix M: Mean Gains by Year and Occasion

Initial ELD Level	Years in Program	Mean Gain	N	S.D.	Cumulative Gain
Level 1	2	1.52	1459	0.95	1.5
	3	0.91	1193	0.96	2.4
	4	0.64	882	0.90	3.1
	5	0.52	1155	0.83	3.6
	6	0.42	831	0.82	4.0
	7	0.36	701	0.81	4.4
	8	0.32	472	0.81	4.7
Level 2	2	1.05	769	0.87	1.0
	3	0.34	508	0.88	1.4
	4	0.15	422	0.85	1.5
	5	0.40	234	0.76	1.9
	6	0.18	142	0.75	2.1
	7	0.28	136	0.64	2.4
	8	0.22	63	0.75	2.6
Level 3	2	0.36	575	0.77	0.4
	3	0.23	442	0.88	0.6
	4	-0.08	356	0.77	0.5
	5	0.43	187	0.67	0.9
	6	0.17	136	0.63	1.1
	7	0.27	76	0.71	1.4
	8	0.19	42	0.67	1.6
Level 4	2	-0.24	1678	0.76	-0.2
	3	0.22	927	0.71	0.0
	4	-0.08	539	0.66	-0.1
	5	0.61	72	0.84	0.5
	6	0.18	58	0.79	0.7
	7	0.46	24	0.61	1.2
	8	0.42	14	0.64	1.6

Appendix N: Logistic Model SPSS Code

```
LOGISTIC REGRESSION VARIABLES PnA_R
  /METHOD=ENTER ELP_SS ELP_SSsquare
  /SAVE=PRED
  /CRITERIA=PIN(.05) POUT(.10) ITERATE(20) CUT(.5).

GRAPH
  /SCATTERPLOT(BIVAR)=ELP_SS WITH PRE_17
  /MISSING=LISTWISE.
```