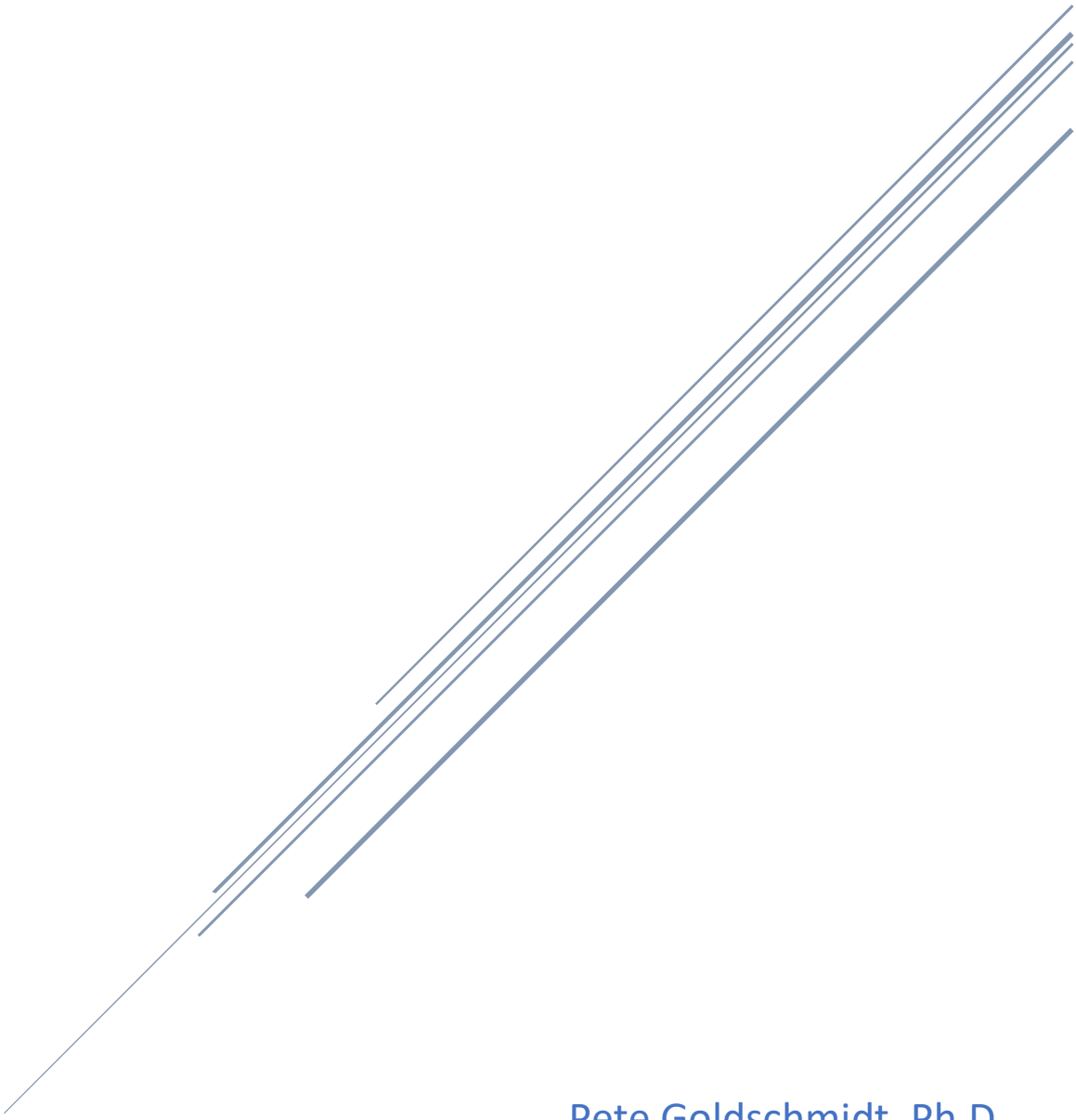
IMPLEMENTING A GROWTH MODEL IN CALIFORNIA

Considerations that support meaningful interpretations for English Learners



Pete Goldschmidt, Ph.D.

August 22, 2020

Executive Summary

Implementing a student growth model is an important addition to the California accountability system. A Residual Gain (RG) model is flexible, robust and demonstrates significant potential to help facilitate monitoring of schools and districts. As the SBE weighs evidence and considers implementation options, it is important to take the actual intended use into consideration. These deliberations must be anchored by the context within which the model will be applied: the premise that education is an accumulation of knowledge and skills over time and the SBE's Theory of Action. The following considerations are summarized from *Implementing a Growth Model in California: Considerations that support meaningful inferences for English Learners.*

Monitoring and Reporting of Current ELs and RFEPs Separately

- Results in the analysis in the paper reinforce the need to monitor not only EL student performance and progress but Reclassified (RFEP) student performance and progress as separate subgroups. Monitoring EL progress on English content performance is key to understanding the impact of programs intending to monitor the success of English learners, which includes progress on both English proficiency and academic content. However, as noted, academic content is influenced by English language proficiency level. Similarly, monitoring the continued progress of RFEP students is particularly important as slowing progress, unlike for English only students, can depend on issues associated with language proficiency. Given that one of the stated goals for the use of growth models is to inform and monitor programs, monitoring the progress of ELs and RFEPs as separate subgroups is precisely the type of analyses that should and needs to be conducted to ensure continued success for a substantial population of students.

Inclusion of an English Language Proficiency Level in the Computation of the Growth Model for current Els

- Evidence indicates that the relationship between current academic performance and prior academic performance is different for ELs than for non-ELs. English proficiency levels substantially impact academic content performance. Consequently, RG models are less precise for ELs than for EOs. Including ELPAC scores in the RG model addresses several stated purposes for implementing a growth model and is consistent with the current cross-subject specification. It provides for more appropriate interpretation of residual gains, gaps, and changing gaps. Including ELPAC scores supports, recognizes, and highlights that one cause of the achievement gap is due to insufficient English language knowledge, skills, and abilities. Including ELPAC scores also increases coherence of the accountability system as progress on ELPAC is important not only as it contributes information about English language learning but also as it directly contributes to academic content.

Stabilizing Year to Year Growth

- The SBE should examine various options and tradeoffs for stabilizing year to year growth estimates. It is important to consider the impact of smoothed estimates within an accountability framework and the potential tradeoffs between the proposed EBLP method and other methods specifically within the context of California. Considerations should include overall stability, stability by subgroups (including ELs and RFEPs), relationship between any single year and the smoothed growth scores, transparency (e.g. weighting and changing weights), implementation factors (number of years of smoothing, resetting years included in smoothed estimates, resetting when students change status, etc.), and interpretation for intended uses (school support, program evaluation, gaps closing). It may well be that the EBLP method is the most effective, but it is important to explicitly present and review tradeoffs against other potential solutions. Stability is an important feature of using a model but this aspect needs to be examined with respect to the estimated gains and the inferences afforded by those estimates.

Including a growth model in California's accountability system is an important positive step in improving school and district accountability and providing actionable, policy relevant results. No model or system is perfect so it is important for the SBE and CDE not only to examine a model (and smoothing) but to explicitly place the (preliminary) results into context to understand better how modeling decisions may substantively influence claims about students, schools, and districts.

Implementing a Growth Model in California: Considerations that support meaningful inferences for English Learners

Abstract

A growth model complements an accountability system by providing meaningful additional information about the ability of schools and districts to facilitate academic progress for all students. A residual gain (RG) model is a flexible and robust approach to monitoring student growth and the SBE and CDE should take advantage of this flexibility to ensure that results provide meaningful, policy relevant, and actionable information. This brief presents several recommendations that support valid claims about student progress. The recommendations are to report results separately for English Learners (EL) and Reclassified Fluent English Proficient (RFEP) students, to include additional assessment information in the RG model, and to weigh additional options to improve year-to-year stability in growth results that reflect the SBE theory of action with respect to growth models as part of the accountability system.

The adoption of a growth model by California is a step in the right direction and undoubtedly improves the monitoring of schools and districts as they facilitate student progress towards college and career readiness. The California State Board of Education (SBE) and the Department of Education (CDE) appears to have settled on a Residual Gain (RG) model, which provides flexibility, advantages and disadvantages (Goldschmidt & Hakuta, 2016; Castellano and Ho, 2013; Goldschmidt, et al, 2012). The purpose of this brief is to highlight additional consideration for operationalization of a growth model for California. The brief presents recommendations and considerations related to utilizing additional assessment information for EL students, the related importance of examining modeling options, and reporting results for EL and RFEP students separately as subgroups. This is followed by considerations related to addressing stability. The brief concludes with a short summary of recommendations.

Alternative Specifications of the RG Model

No growth model can address every aspect of student progress; there are tradeoffs, as has been noted by previous reports and analyses presented to the SBE. These tradeoffs must be weighed in context of the Theory of Action; i.e. the logical schema that links school and district monitoring of student progress through a growth model to actual student performance and progress and the mechanisms that facilitate the improvement of student academic outcomes. The purpose of implementing a growth model is to provide districts and schools a tool that supports the development of local goals, allows for the evaluation of programs, provides information on gaps and the closing of gaps. This is accomplished by implementing a growth model that is technically sound, can be part of an accountability system, and provides stable results from which claims about schools and districts can be made. Based on comments from SBE board members at the previous (July 2020) meeting there was still interest in having coherence in results from students to districts.

That is, that growth results from individual students could be seamlessly interpreted and aggregated to subgroups, schools, and districts.

Although the SBE has settled on a Residual Gain (RG) model details of how to operationalize this model are still under consideration and this important step warrants the same level of examination as the initial model choice. Importantly, examining model functioning as it is intended and with actual state data is important as models do not function exactly the same across contexts (Goldschmidt, et. al, 2012). Options and trade-offs among options are important to explicitly consider. Alignment between the SBE Theory of Action and how the model is implemented is arguably the most critical step in implementing an accountability model that intends to help improve outcomes for all students in California.

A key consideration in using a RG model is noting the interpretation of results. The RG model provides an estimate of relative growth and compares a student's current performance relative to students who had the same prior performance (ETS, 2018). This assumes that all students with the same prior score(s) should have the same growth trajectory, i.e. relationship between prior and current performance (and importantly that schools should be held accountable for equal progress). This assumption may not be tenable for a substantial portion of students.

Academic content performance and measurement of progress using multiple assessments is influenced by access to that content as well as the ability to adequately demonstrate knowledge skills and abilities on each assessment. Evidence suggests that English language proficiency plays a key role in academic content performance and, importantly, in progress (Bailey & Carrol, 2015; Bailey & Huang, 2011; Wolf & Leon, 2009; Butler & Stevens, 2001). English Learner (EL) students and a non-EL students with the same prior score may be on different trajectories because the EL student is not only progressing on content but on an antecedent skill – English language. Recent evidence suggests that ELs benefit from additional instructional practices designed to address language that help facilitate learning academic content (Vaughn, Martinez, Wanzek, Roberts, Swanson, & Fall, 2017) and further corroborates that langue proficiency influences academic content growth. These language effects can impact math assessment results as well as ELA (Martiniello, 2009).

The SBE has previously indicated that it does not intend to use student background characteristics as it creates unintended consequences with respect to expectations. RG (type) models in different states have taken various approaches to student background and given that student background is not allowable under ESSA, states do not utilize student background for federal for school accountability. Although there are some calls for its use in California (Polikoff, 2019). Generally, RG (type) models with and without student background result in substantially similar results and models without student background afford somewhat more straight-forward interpretations[1].

[1] However, some argue that the interpretations are more straight-forward when including student background as they can account for unequal distributions of student subgroups in schools, for example.

Excluding student background characteristics in the model does not obviate the need to examine whether claims about students within a subgroup are equally accurate, unbiased, and equally valid claims about students, schools, and districts are tenable. Figures One and Two are based on data from a CA district[2] and demonstrate that the relationship between prior performance and current performance is not the same for all students. The results in Figures one and two highlight that: current EL students are predominantly low performing compared to their Reclassified, Fluent English, and English Only classmates; the relationship between prior and current scores (the dashed lines) do not represent the same slope (relationship between the scores). Lower performance may bias results overall Lockwood & McCaffrey, 2014), while differential slopes support the contention that students with the same prior scores may not progress in a like fashion.

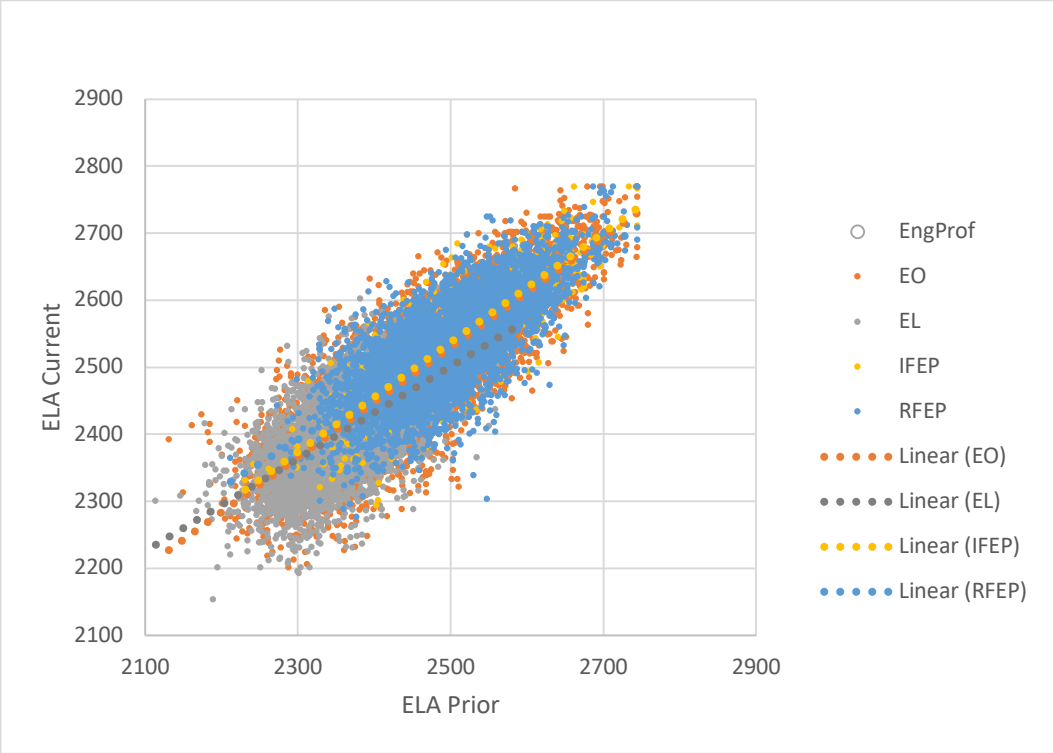


Figure 1: Relationship between current and Prior ELA Performance by Language Subgroup

The dotted lines in Figures One and Two summarize the relationship between current and prior results by English language proficiency (EngProf). The results in Figure One indicate that both ELs and RFEPs have different slops than initial English speakers (EO and IFEP) and are close to parallel to one another. This implies that although RFEPs have higher absolute performance than ELs (and on par with EOs) the link from prior to current scores is more similar to ELs than EOs.

[2] This example is not intended to imply that these are the state-wide relationships, but that these relationships should be examined and considered explicitly.

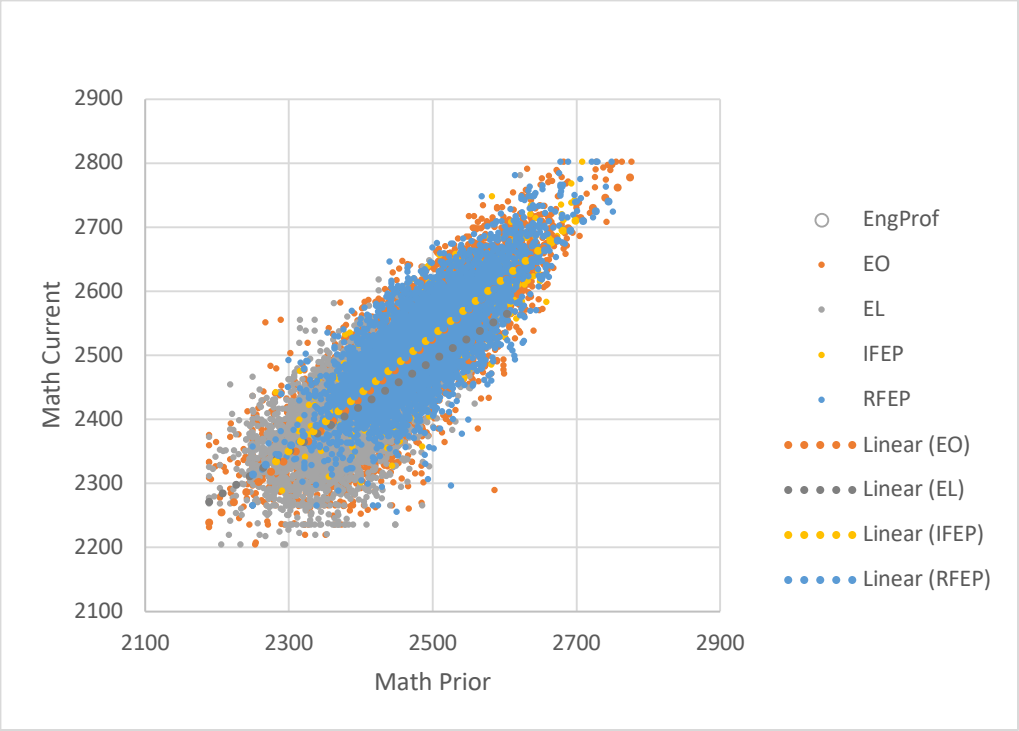


Figure 2: Relationship between current and Prior Math Performance by Language Subgroup

The results in Figure Two highlight that ELs exhibit the same pattern in math as in ELA and that RFEP students are more similar to their EO classmates in the link between current and prior year performance than on the ELA assessment. These results are consistent with expectations.

The results in Table 1 further highlight that the relationship between prior and current scores is different for current ELs and non-ELs by indicating the amount of variability in the current score accounted for by the prior score (R^2). A lower R^2 reflects greater error, which not only contributed to uncertainty in the current year, but also to instability across years.

Table 1:
Variation in Current Year Performance Accounted for by Prior Year[1]

	ELA	Math
EO	0.71	0.70
EL	0.42	0.41
IFEP	0.70	0.69
RFEP	0.54	0.59

[1] R^2

These results reinforce the need to monitor not only EL student performance and progress but Reclassified (RFEP) student performance and progress as separate subgroups. Monitoring EL progress on English content performance is key to understanding the impact of programs intending to monitor the success of English Learners, which includes progress on both English proficiency

and academic content. However, as noted, academic content is influenced by English language proficiency level. Similarly, monitoring the continued progress of RFEP students is particularly important as slowing progress, unlike for English only students, can depend on issues associated with language proficiency. Figures One and Two highlight the role of language proficiency as the prior-current link in ELA is generally similar to ELs while that same link in math is less similar to ELs[3]. Given that one of the stated goals for the use of growth models is to inform and monitor programs, monitoring the progress of ELs and RFEPs as separate subgroups is precisely the type of analyses that should and needs to be conducted to ensure continued success for a substantial population of students. Monitoring RFEP performance and progress separately may be challenging at schools with small N-sizes, but in those instances district aggregates still provide insight into RFEP academic progress.

Given the potential substantive differences in the nature of progress for ELs, additional considerations are warranted. Not including student background in the RG model is a reasonable choice but should not preclude the SBE from considering the use of additional assessment information that can improve model functioning and account for the interplay of antecedent English language development and content performance and progress. It is possible to use a RG model and include additional prior assessment results. One form of this was presented in Goldschmidt & Hakuta (2016). Including language proficiency levels[4] explicitly addresses the importance of language progress on content performance. Including EL proficiency assessment results is a method used by at least one other state as part of their growth model. There are known benefits to including additional prior scores to reduce bias in RG models (Wright, 2008) and the added precision will likely improve stability. The R^2 results in Table One for ELs are improved by approximately 14% when ELPAC scores are included in the model.

Including EL ELPAC scores requires some reconsideration of the RG model specification compared to how it is currently designed by ETS (2018), but this does not change the overall conceptualization of the RG model, it does not introduce student background, and utilizes additional assessment results to support more appropriate claims with respect to EL student progress. Including ELPAC results for EL students induces additional coherence among elements of the accountability system. By including ELPAC scores in the RG model it strengthens the link between the progress ELs make towards English language proficiency and the progress that should be made on academic content. Program effectiveness for ELs cannot be based solely on EL progress on language but must include all facets relevant to a student's educational opportunities and success.

[3] Math requires English language skills, but clearly less than ELA and thus we would expect language effects (if any) to be greater on ELA results.
[4] The model could use proficiency levels, scale scores, or other conversion of ELPAC results.

Additional Considerations for Addressing Year-to-Year Stability

One tenant of the growth model is that the results are stable across years, which is an important aspect because excessive year to year fluctuations not only make results less actionable but also less credible. SBE has rightly sought out solutions to improve stability. It is interesting to note that the RG estimates of stability provided by ETS (2018) are somewhat lower than expected compared with previous analyses of state-wide datasets (Goldschmidt, et. al., 2012). As noted, this may be due to specific state context but is also likely due to the nature of the RG model. There are several means by which stability can be improved and careful examination and comparison of the options is certainly warranted. As with the selection of a growth model, each potential remedy has trade-offs and there exists no perfect solution. SBE is currently evaluating the Empirical Best Linear Predictor (EBLP). The EBLP is based on dynamically weighting multiple years of growth estimates (results from the RG model) and smoothing growth over time. It appears, based on presentations thus far that the EBLP estimates smooth growth substantively more for smaller samples than for larger samples. The EBLP method is quite sophisticated but open-source software produces the estimates which a positive feature of this application. Also, given the larger impact on smaller N-sizes, sub-groups would demonstrate more stable growth estimates than without the adjustment. Considering the impact on inferences about schools based on smoothed results is an important step before adopting an adjustment strategy. The most recent results (ETS, 2020) support the need to explicitly consider context in which results will be utilized before adopting a strategy to improve stability. The results (ETS, 2020) generally confirm that EBLP method substantively improves accuracy and stability, but that the impact is not uniform. The impact for subgroups by school and LEA differ thus substantiating the need to examine results by relevant subgroups; e.g.. ELs and RFEPs. The dynamic nature of EL and RFEP status implies that N-size change is continual and meaning of smoothed gains become more difficult to interpret.

The most recent ETS (2020) results are promising, but inferences still need to be considered with respect to accountability, program evaluation, and analyses of gaps – given these are three important intended uses of the growth model. There are several issues that can be examined:

whether post-hoc smoothing is the best option;
how the proposed EBLP approach compares to less complex approaches;
the sensitivity of results overall and the sensitivity of results for all subgroups;
the disconnect between any single year's growth and adjusted results;
the impact of differential (dynamic) weighing of estimates; and,
operational and reporting considerations.

Stability of results based on growth has received attention as an important issue for some time (Goldschmidt & Swigert, 2002; Wilms and Raudenbush, 1989) and there are many potential options. Staying within the RG framework, however, it is possible to structure the model in such a way as to produce direct estimates of trends over time, rather than post-hoc adjustments. Such models produce both an RG and a trend component over time. Whether this approach is better

than the EBLP is unknown, but it does provide a comparable alternative that the ETS staff involved with this project have ample capacity to develop and evaluate.

Less complex post-hoc approaches than the EBLP methodology may provide similar results but be more transparent to stakeholders. For example, a three-year moving average[5]. Using the district data presented in Figures One and Two, a moving average produces a correlation of approximately 0.88 for consecutive estimates. A moving average may not be as successful as the EBLP in proportionally improving estimates inversely to N-size, but it minimally should be presented as a baseline comparison. Also, a moving average tends to be moderately to highly correlated with any single year growth – which is important in establishing credibility – The single year to three-year correlations range from .74 to .91 in the district used for this presentation. The disadvantage of a moving average compared to the EBLP is that each year has a fixed weight[6] while the EBLP takes advantage of dynamic weighting.

Within an accountability framework, a tradeoff is that while dynamic weighting produces more precise estimates than a simple average, dynamic weighting can lead to confusion and credibility issues with year to year estimation because some stakeholders will take issue with "arbitrary" weights that are estimated post-hoc each year. Choosing the EBLP method requires training and trust-building among stakeholders.

Another issue related to weighing is the fact that any method that combines years to create a current year estimate will be less sensitive to changes in growth. This is likely exacerbated for subgroups, like ELs or RFEPs as they generally have smaller N sizes and, as noted, the continual changes to status that is unique to these subgroups. This means that changes in school progress will be smoothed out and potentially delays recognition of improvement or challenges.

Hence, any method used to smooth results requires attention to how it will be operationalized. A moving average is straight-forward and the EBLP seems to be quite flexible; this simply requires running these options under desired accountability contexts, for example, the extent to which smoothed results are used to inform the status and or change in status of a school, program, or achievement gap. How does smoothing affect claims about schools or districts? Also, if the RG model is used for identification of support categories, does the smoothing impact the ability of a school to exit status? This may be particularly important for subgroups such as ELs due to small N-sizes (since small Ns are smoothed more than large Ns). With respect to subgroups, an operational question might be whether growth is smoothed or reset when students are reclassified from EL to RFEP status. It is important to monitor this transition and smoothing across it may obfuscate important signals about the continued success of former ELs. Additional operational rules would need to be considered; for example, if a school enters support status should growth be

[5] One potential benefit of the EBLP method is that it appears as effective using two years instead of three to smooth estimates (ETS, 2020).

[6] In the example presented here, each year was weighted equally; however, policy weights can be used to weight more recent year more heavily, for example. This may be possible with the EBLP method as well.

considered only from the year of entry into status? Another example might be whether a multiyear smoothed estimate should reset with a new principal? This is not an exhaustive list of considerations and the SBE and CDE should place model results into context and how they will be used and whether the tradeoffs are warranted.

Summary and Recommendations

Implementing a student growth model is a useful addition to the CA accountability system. A RG model is flexible, robust and demonstrates significant potential to help facilitate monitoring of schools and districts. As the SBE weighs evidence and considers implementation options it is important to take actual intended use into consideration. These considerations and recommendations include:

Examining evidence supporting the use of a model (and potentially smoothing technique) by the EL and RFEP subgroups. EL students are currently considered separately, and the reasons are well understood. Considering the impact on RFEPs separately is equally important as these students have demonstrated substantial improvement in English language skills while also progressing in academic content. The effectiveness of EL programs does not stop once the EL is reclassified, rather continued monitoring (which is required under ESSA) should include reporting separate academic results for this group. This is particularly germane to growth as changes in RFEP growth takes on additional meaning with respect to language support and potential systematic school and/or district effectiveness in providing RFEPs appropriate continued opportunities to learn.

Including ELPAC scores in the RG model as this potentially addresses several stated objectives and purposes for implementing a growth model. It provides for more appropriate interpretation of residual performance, gaps, and changing gaps. Including ELPAC scores supports, recognizes, and highlights that one cause of the gap is due to insufficient English language knowledge, skills, and abilities. Including ELPAC scores increases coherence of the accountability system as progress on ELPAC is important not only as it contributes information of English language learning but also as it directly contributes to academic content. This connectedness is consistent with the intent of ESEA reauthorization that brought EL progress into Title I accountability.

Examine various options and tradeoffs for stabilizing year to year growth estimates. It is important to consider the impact of smoothed estimates within an accountability framework and the potential tradeoffs between the proposed EBLP method and other methods. Considerations should include stability overall, stability by subgroups (including ELs and RFEPs), relationship between any single year and the smoothed growth scores, transparency (e.g. weighting and changing weights), implementation factors (number of years of smoothing, resetting years included in smoothed estimates, resetting when students change status, etc.), and interpretation for intended uses (school support, program

evaluation, gaps closing). The SBE should be presented with results that simulate the use of the results which goes beyond simply reporting stability estimates but includes the actual estimated gains, EBLP adjusted gains, other potential adjustments. These results should be reported for all relevant subgroups for schools and districts. The CDE should work with these results and provide some potential inferences and claims associated with these results for a small sample of schools and districts so the SBE can determine whether the inferences and claims are consistent with their intended purposes and goals.

Given the complexity of RG models as well as any second step stability procedure, it is useful to place school or district growth into context. Similar to previous systems (e.g. the United Kingdom, New Mexico) multiple growth estimates should be presented. This might include presenting both the current year, all prior years (with N counts) used for smoothing, and the smoothed estimates. In this way transparency is improved and stakeholders will develop trust for the growth scores.

Including a growth model in California's accountability system is an important positive step in improving school and district accountability and providing actionable, policy relevant results. No model or system is perfect and so it is important for the SBE and CDE to not only examine a model (and smoothing) but to explicitly place the (preliminary) results into context to better understand how modeling decisions may substantively influence claims about students, schools, and districts.

References

Avenia-Tapper, B. & L. Llosa (2015). Construct Relevant or Irrelevant? The Role of Linguistic Complexity in the Assessment of English Language Learners' Science Knowledge. Educational Assessment, 20(2).

Bailey, A. & B. Huang (2011). Do current English language development/proficiency standards reflect the English needed for success in school? *Language Testing*, 28(3).

Bailey, A. & P. Carrol (2015). Assessment of English Language Learners in the Era of New Academic Content Standards. Review of Research in Education, 39(1).

Butler, F & R. Stevens (2001). Standardized assessment of the content knowledge of English language learners K-12: current trends and old dilemmas. *Language Testing*, 18 (4).

Castellano, K. E., & Ho, A.D. (2013). A practitioner's guide to growth models. Washington, DC: Council of Chief State School Officers.

ETS (2020). Exploring Empirical Best Linear Prediction for Aggregate Growth Measures, California State Board of Education September 2020 Agenda Item #02.

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

Goldschmidt, P. & S. Swigert (2002). Oxymoronic *Program Evaluation: The Short-Term Longitudinal Analysis Dilemma, Paper presentation, AERA, New Orleans, LA.*

Lockwood, J. R., & McCaffrey, D. F. (2014). Correcting for test score measurement error in ANCOVA models estimating treatment effects Journal of Educational and Behavioral Statistics, 29(1), 25, 52.

Martiniello, M. (2009). Linguistic complexity, schematic representations, and differential item functioning for English language learners in math tests. Educational Assessment, 14, 160–179. doi:1080/1062719090342290.

Vaughn, S., Martinez, L. R., Wanzek, J., Roberts, G., Swanson, E., & Fall, A.-M. (2017). Improving content knowledge and comprehension for English language learners: Findings from a randomized control trial. *Journal of Educational Psychology, 109*(1), 22–34. https://doi.org/10.1037/edu0000069

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.

Wolf, M., & Leon, S. (2009). An investigation of the language demands in content assessments for English language learners. *Educational Assessment*, 14, 139–159. doi:10.1080/10627190903425883.

Wright, S. P. (2008). Estimating educational effects using analysis of covariance with measurement error. Paper presented at CREATE/NEI Conference, Wilmington, NC, October 2008.