CADRE

MARCH 2025

Comments on the Colorado 1241 Task Force Accountability Report

Benjamin R. Shear, Elena Diaz-Bilello,
Kaitlin Nath, Erik Whitfield

A report prepared by the Center for Assessment, Design, Research and Evaluation (CADRE) at the CU Boulder School of Education.

School of Education
UNIVERSITY OF COLORADO **BOULDER**

About CADRE

The Center for Assessment, Design, Research and Evaluation (CADRE) is housed in the School of Education at the University of Colorado Boulder. The mission of CADRE is to produce generalizable knowledge that improves the ability to assess student learning and to evaluate programs and methods that may have an effect on this learning. Projects undertaken by CADRE staff represent a collaboration with the ongoing activities in the School of Education, the University, and the broader national and international community of scholars and stakeholders involved in educational assessment and evaluation.

Suggested Citation

Shear, B.R., Diaz-Bilello, E., Nath, K., & Whitfield, E. (2025). *Comments on the Colorado 1241 Task Force Accountability Report*. Boulder, CO: The Center for Assessment, Design, Research and Evaluation (CADRE), University of Colorado Boulder.

Please direct any questions about this report to:

benjamin.shear@colorado.edu

Introduction

In November 2024 the Colorado 1241 Task Force released their final report detailing recommendations and areas for further study to guide improvement of Colorado's Education Accountability System. The report describes 30 recommendations across five domains and four areas for further study. As a consensus document authored by a diverse group of Colorado stakeholders, the Task Force Report provides a valuable resource for the Colorado education community. With 30 recommendations, however, it may not be clear how to prioritize efforts to implement the recommendations. Many of the recommendations are also interconnected; implementation of some recommendations may impact the feasibility and implementation of other recommendations. In addition, and not directly discussed in the report, the scope and cost estimates of the recommendations produced by the Colorado Department of Education (CDE) vary significantly.[1]

More importantly, the Task Force Report does not explicitly consider the theory of action guiding Colorado's accountability system, and does not explicitly state how each recommendation would improve the overall efficacy of the system for achieving the intended aims. Any efforts to revise the Colorado accountability system should be grounded in an explicit theory of action articulating what the accountability system is intended to accomplish and how it is designed to accomplish those aims. Then, each recommended revision should be considered in reference to the theory of action, considering how the revision will make the accountability system more effective at accomplishing the stated aims, which could include reducing the likelihood of unintended side effects or consequences.

To assist in such an effort and to support decisions about implementing recommendations from the Task Force Report, we highlight points for consideration relating to nine of the 30 task force recommendations. We selected these nine recommendations because they relate to our prior experience and expertise; we are not suggesting that these nine recommendations should be prioritized over others. These nine recommendations would benefit from further review before implementation. This brief emphasizes points not necessarily included in the Task Force Report to help decision-makers prioritize recommendations and inform decisions about the recommendations. Below we discuss Task Force recommendations 1, 4, and 8 in the section "District and School Performance Frameworks"; recommendations 9, 12, and 13 in the section "Assessments for Accountability"; and recommendations 11, 16, and 23 in the section "Public Reporting and Engagement: Test Participation."

[1]Preliminary cost estimates and anticipated timelines provided by CDE, for example, range from recommendations that could be implemented in the upcoming academic year (2025-26) with little cost, to recommendations that could require up to $5 million and several years to implement. These cost estimates are preliminary and should not be taken as final. However, the wide range of cost estimates and timelines indicates the widely varying scope of the 30 recommendations. CDE estimates that recommendations 9 and 12, discussed here, would be two of the three most costly recommendations to implement.

District and School Performance Frameworks

Our comments on the School Performance Framework (SPF) and District Performance Framework (DPF) focus on Recommendations 1, 4, and 8. Recommendation 1 proposes lowering student count thresholds (i.e., the "minimum n-size") for accountability calculations and test score reporting. While reducing minimum sample size thresholds could increase transparency for small systems through the reporting of more data points, this strategy would also tend to reduce the reliability of school and district performance metrics. In a forthcoming technical report, we examine the impacts that reducing minimum sample size thresholds would have on the reliability of school and district test score results. Our results highlight two important points. First, while the current thresholds strike a reasonable balance between transparency and reliability, this tradeoff is ultimately a value judgment that cannot be answered statistically. Second, and consistent with the current thresholds, larger sample sizes are needed for growth metrics to reach equal levels of reliability as achievement metrics. These tradeoffs must be weighed carefully before making changes to the current minimum n-size requirements. Recommendation 4 advocates conducting a study to review best practices for minimizing the volatility of framework and other ratings for small systems and to monitor the impact of this volatility. We endorse Recommendation 4 and believe a review of best practices would be beneficial both to Colorado and to stakeholders in other states, as there is little consensus regarding best practices for reducing volatility. Any study carried out to investigate volatility should also consider the issue of students counting in multiple disaggregated groups discussed in the Task Force Report and could consider potential roles for artificial intelligence (AI) to be used to create more holistic ratings based on a wider array of indicators.

Recommendation 8 advocates that the state re-evaluate the weighting of indicators used in the DPF and SPF accountability ratings. The state repeatedly emphasizes student growth as the most valuable part of the accountability system and the growth indicator is currently weighted most heavily in SPF calculations to reflect this. Our analyses of Colorado SPF data, however, show the achievement indicator often has the largest impact on SPF scores because achievement varies more across schools than growth (Shear & Nath, 2024). We agree that the recommended evaluation should be carried out and would be relatively straightforward to conduct. In addition to considering volatility and the correlation of ratings with student demographics, an underlying theory of action outlining the goal of the accountability system should inform the assignment of weights. The more challenging aspect of this recommendation is determining which weights are most aligned with the intended aims of the accountability system, as this is a policy decision rather than a technical choice.

Assessments for Accountability

Our comments on Assessments for Accountability focus on recommendations 9, 12, and 13. Recommendation 9 would require the state to make assessments at all grade levels available in additional languages beyond English and Spanish. Any changes to the state assessments should be made with the aim of improving the validity of the resulting scores for their intended uses. Before creating additional test translations, clarity is needed regarding which languages

the tests will be translated into and how these translations would enhance the validity of test score inferences and uses for multilingual learners. Developing and evaluating the validity of translated tests is an extremely costly and time-consuming process. Moreover, test translation processes can result in changes to the construct being assessed by a test and may not always produce more valid scores for multilingual learners, particularly when a student's language of instruction differs from the language of the test. A recent report by two language and assessment experts provides detailed consideration of these and other issues related to test translation for state accountability testing (Solano-Flores & Hakuta, 2017). If the primary goal is to make the state assessment results more accessible to multilingual learners and their families, translating score reports would be a much less costly and more viable option to consider.

Recommendation 12 advocates two changes to the CMAS testing program: create a computer-adaptive version of CMAS and create a vertical score scale that would allow scores to be directly compared across grades. These changes are independent but not mutually exclusive, meaning it would be possible to implement either one separately or both together. There are three motivating goals stated for Recommendation 12: 1) reduce testing time, 2) allow for cross-grade level inferences, and 3) add back a writing subscore. While adaptive testing designs can usually produce more reliable test scores with less testing time, this often comes at the cost of reduced construct representation as adaptive tests rely on a larger proportion of multiple-choice items or other items that can be scored quickly by a computer while the assessment is being completed. Creating an adaptive version of CMAS would also require creating many additional test items at each grade level, because adaptive tests require larger item banks than fixed-form tests. As noted in the Task Force Report, AI has the potential to address some of these issues through its use to rapidly score constructed-response items or to develop larger item banks.

Regarding cross-grade inferences, changes in scale scores from the current CMAS design cannot be calculated across grade levels because CMAS does not employ a vertical scaling design. Moving to an adaptive test on its own would not allow scale scores to be compared across grade levels; a separate process is required to create a vertical score scale, which would substantially increase the cost of re-designing CMAS and maintaining the vertical scale over time. Reporting writing subscores would likely require adding additional test modules to an adaptive version of CMAS, potentially counteracting reductions in testing time gained by the switch to an adaptive test. Ultimately, greater clarity is needed about the goals of creating adaptive or vertically scaled versions of CMAS, and a feasibility study would be needed to determine whether an adaptive version of CMAS could meet those goals. In addition, we strongly recommend that CDE play a leading role in investigating current and potential uses of AI to improve state assessments – regardless of the decision to develop an adaptive version of CMAS.

Recommendation 13, to improve the timeliness of reporting assessment results, is the most feasible of these three recommendations. Yet we recommend further study before implementing this recommendation to better understand the current reporting process and to clarify the intended goals of faster reporting. Clearly identifying the most time-consuming steps in the current reporting process is critical to evaluating the scope and feasibility of this recommendation. Moreover, the factors contributing to delays in reporting could change significantly if either of the above recommendations (translating tests or creating adaptive versions of CMAS) were implemented. Clarifying the goals of faster reporting is critical to ensuring the revised method of reporting results will support the intended aims. Depending upon how assessment results are used, simply reporting the same results more quickly may not

accomplish the intended aims. If the primary goal is to return results more quickly so that results can be used to inform instruction, for example, this should be viewed skeptically as there is little evidence that state assessment results alone can provide instructionally useful information given, among other issues identified over two decades, their broad content coverage (Blazar & Pollard, 2017; Supovitz, 2009). A systematic investigation of the ways state assessments are currently used by different stakeholders would be essential to addressing these questions and could provide valuable information for improving the design, reporting, and use of Colorado's state assessments more generally.

Public Reporting and Engagement: Test Participation

Our comments on public reporting and engagement focus on three interconnected recommendations: 11, 16, and 23. While recommendations 11 and 16 are discussed earlier in the Task Force Report, they are closely tied to recommendation 23 and collectively relate to test participation reporting and engagement. The primary premise of recommendation 16 is that the confusion stemming from a dual accountability system could be alleviated by clarifying participation rules under federal and state accountability frameworks. Since the development of the SPF, participation rules and the determination of "who counts" have been annually updated and communicated to schools and districts. The accountability division also maintains a frequently asked questions page with a section on participation and parent excusal on their website to explain these rules and policies. Still, conflicting policies regarding participation persist, and these unresolved discrepancies are likely to continue to generate confusion. For instance, some high-performing schools, exempt from state-level oversight, are flagged for targeted support and interventions under the federal system due to the underperformance of specific student groups.

This conflicting policy environment also complicates the state's ability to effectively implement recommendation 11. On the surface, having the state provide guidance to schools about encouraging, or at least not discouraging, test participation seems straightforward. However, this task is complicated by a tangle of federal laws, state statutes, board policies, and district-level guidelines. For example, C.R.S. 22-7-1013(8)(c) prohibits districts and schools from encouraging parents to excuse their children from state assessments or from discouraging students from taking these assessments. At the same time, C.R.S. 22-7-1013(8)(a) mandates that districts establish written policies and procedures allowing parents to opt their children out of state assessments. Adding to this complexity, the Colorado Association of School Boards in 2017 noted that the State Board upheld a dual accountability system to prevent student opt-outs from negatively affecting school and district ratings within the state accountability system. Given these contradictions, it is unclear what additional materials or guidance CDE could develop to further support schools, beyond reiterating the importance of state assessments, while adhering to existing laws and policies supporting parental opt-outs.

Finally, recommendation 23 suggests implementing corrective action plans for districts and schools with low test participation. While this approach could motivate efforts to address low participation rates, we propose reframing this recommendation to focus on understanding the issue more deeply. Instead of adopting a punitive stance, we suggest conducting a comprehensive study to examine the root causes of low participation. During the height of the

opt-out movement, a 2017 national survey by Teachers College (Pizmony-Levy & Cosman, 2017) identified various motivations for opting out, many of which were influenced by information disseminated on social media. Exploring the factors driving low participation across schools and districts could yield valuable insights. These findings could help identify misinformation about testing that needs to be addressed and provide a foundation for integrating stakeholder feedback into efforts to enhance testing and accountability practices statewide.

Conclusion

The 1241 Task Force Report does not provide a thorough consideration of the goals for an effective school accountability system or leverage the existing theory of action to articulate how Colorado's accountability system is intended to achieve these goals. Yet implementation of any task force recommendations should be directly tied to a clearly stated theory of action articulating the goals for Colorado's school accountability system and explaining how implementation of the proposed recommendations would make the accountability system more likely to achieve these goals. Evading hard conversations about the theory of action makes it difficult to prioritize which aspects of the accountability system need to be reimagined, which areas should be finetuned, and which studies should be pursued. In this brief we discuss nine of the report's 30 recommendations, highlighting issues and considerations that will help to inform the consideration and potential implementation of these recommendations within a broader theory of action. Ultimately, the Task Force Report's recommendations are highly interdependent and can neither be considered nor implemented in isolation. Before implementing any of the recommended revisions to Colorado's school accountability system we urge extensive further review and study to clarify the purpose and scope of each recommended revision, both on its own and for how it would interact with other revisions.

References

Blazar, D., & Pollard, C. (2017). Does test preparation mean low-quality instruction? *Educational Researcher, 46*(8), 420-433. https://doi.org/10.3102/0013189X17732753

Pizmony-Levy, O., & Cosman, B. (2017). How Americans view the Opt Out movement [Research Report]. Teachers College, Columbia University. https://www.tc.columbia.edu/media/news/docs/How-Americans-View-the-Opt-Out-movement---v8-COMBINED.pdf

Shear, B. R., & Nath, K. (2024). Nominal and effective weights of composite accountability ratings: A demonstration using Colorado's School Performance Framework. Boulder, CO: The Center for Assessment, Design, Research and Evaluation (CADRE), University of Colorado Boulder. https://www.colorado.edu/cadre/media/140

Solano-Flores, & Hakuta, K. (2017). Assessing students in their home language. Stanford University, Understanding Language Website. https://capri.utsa.edu/wp-content/uploads/2017/09/assessing-students-in-their-home-language-Hakuta.pdf

Supovitz, J. (2019) Can high stakes testing leverage educational improvement? Prospects from the last decade of testing and accountability reform. *J Educ Change 10*, 211–227 (2009). https://doi.org/10.1007/s10833-009-9105-2