

To: Colorado Department of Education Staff: Alyssa Pearson, Deputy Commissioner, Performance and Accountability; Marcia Bohannon, Chief Information Officer; Joyce Zurkowski, Executive Director of Assessment; Jill Stacey, Data Privacy Analyst

From: A+ Colorado Research and Policy Team: Van Schoales, CEO; Lisa Berdie, Policy and Research Director; Laura Valle-Gutierrez, Policy and Research Analyst

Date: December 11, 2018

Subject: Systemic Challenges to Usable Public Data



Table of Contents:

Purpose of Memo

Examples of Current Limitations in Data Utility and Validity:

- Demographic Data
- Assessment Data
- ACCESS Data
- READ Act Data
- Defining Personally Identifiable Information
- Data Lab
- Miscellaneous Challenges

Request for Next Steps

At A+ Colorado, as with many other stakeholders in Colorado's education ecosystem, data is critical to see where students are and how they are performing, as a way to understand what is and is not working for students academically. This is critical for empowering families, educators and communities to best serve students.

Complementary suppression and other data reporting rules and practices are currently implemented to a point where the data that are publicly available have limited utility and validity. Indeed we see many of these practices as severely limiting the reliability and validity of public data, increasing the risk that data is used irresponsibly and ill-informing decisions made at every level of the system.

While we are aware that there are additional issues regarding usability and accessibility of data for families and other stakeholders, our intent for this letter is to focus on the practical issues around our capacity to perform valid and rigorous analyses and draw insights on student performance and progress with the current data.

Examples of Current Limitations in Data Utility and Validity:

- Demographic Data

- Suppression of school-level demographic data makes it impossible to understand where students are. For example, looking at demographic data by Free/Reduced Price lunch status, we don't have information on 10% of schools on just the combined Free/Reduced category. This percentage increases when we distinguish between Free and Reduced price lunch. This suppression includes high poverty districts, such as Center - where we have no school-level information on the percentage of students receiving Free or Reduced price lunch. An additional idiosyncrasy is the fact that the count for two schools is included while the percentage remains undisclosed. Additionally, when school FRL eligible ranges are presented instead of actual proportions, the ranges are far too wide (25 points) to make any real analysis meaningful.
- As is the trend throughout this examples enumerated in this letter, these issues compound when we seek other important information about students, including English language learners, and students receiving special education services.

DISTRICT NAME	SCHOOL NAME	K-12 COUNT	FREE LUNCH	REDUCED LUNCH	NOT ELIGIBLE	FREE AND REDUCED COUNT	% FREE	% REDUCED	% FREE AND REDUCED
CENTER 26 JT	THE ACADEMIC RECOVERY CENTER OF SAN LUIS VA	8	N/A	N/A	N/A	N/A	N/A	N/A	N/A
CENTER 26 JT	CENTER VIRTUAL ACADEMY	8	N/A	N/A	N/A	N/A	N/A	N/A	N/A
CENTER 26 JT	HASKIN ELEMENTARY SCHOOL	297	N/A	N/A	N/A	N/A	N/A	N/A	N/A
CENTER 26 JT	SKOGLUND MIDDLE SCHOOL	121	N/A	N/A	N/A	119	N/A	N/A	N/A
CENTER 26 JT	CENTER HIGH SCHOOL	151	N/A	N/A	N/A	147	N/A	N/A	N/A

- Assessment Data

- Mean Scale Score Suppression
 - There are a number of instances where a mean scale score is suppressed, even with more than 16 students at a school. This example looks at All Grades in a school and isn't disaggregated by any other metric.

School Name	Content	Test/Grade	# of Total Records	# of Valid Scores	# of No Scores	Participation Rate	Mean Scale Score
AVERY/PARSONS E	English L	All Grade	196	188	8	95.9	*
HOLYOKE SENIOR I	English L	All Grade	95	94	1	98.9	*
PUEBLO SCH. FOR	English L	All Grade	73	73	0	100	*
NORTH ELEMENTAI	English L	All Grade	143	141	2	98.6	*

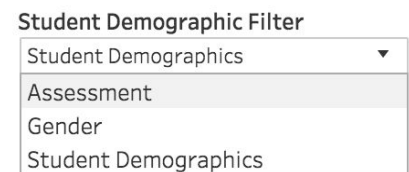
- Suppression of Achievement by Level
 - Along the same vein, there is significant suppression, even without disaggregation by other indicators. Looking at only All Grades for

- ACCESS Data

- A particular case of minimum-n counts
 - ACCESS data is key to understanding how English Language Learners are attaining English language skills. Significant amounts of data are lost from this dataset because of small cohorts from a high minimum n-count. The result is that more than half of the districts, disaggregated by EMH level are missing information. We know that there are still thousands of students that we are missing information on due to suppression. Beyond this ACCESS data doesn't provide information on growth by English language proficiency status nor by English language instruction program type. Finally, similarly to achievement data, information on the percent of students who are growing by a level within the designated time frame for that level is substantially incomplete, as "% on track" must be hand calculated in the underlying data, of which much is suppressed. This metric is an important measure of English language proficiency and more information is missing from this because there is no sense of the overall percentage of students meeting this growth.
- Discrete annual reporting
 - Not connecting ACCESS Data by cohorts is problematic, because the sample is constantly changing. It is very difficult to make sense of progress over time through the publicly available data because the group is particularly dynamic, and is impossible to separate long-term english learners from their peers who are gaining language skills and exiting programming more quickly.

- READ Act Data

- No full publicly-available data set
 - While the data dashboard is an excellent visualization tool, the fact that the underlying data is not downloadable in a csv format is problematic for any analysis.
- Filtering and cross tabulation
 - Given the differences in assessments and the results in terms of how students are identified for READ Plans (i.e. some assessments setting a lower or higher bar than others), assessment should be the first filter applied to the data, and should likely be a layer (or cross-tabulation) with any other cut of the data.



- READ data past grade 3
 - As schools support students on READ Plans well past grade 3, there should be publicly available data on how students identified with SRDs in K-3 are being served in later elementary and secondary school.

- **Defining Personally Identifiable Information**
 - FERPA's definition of personally identifiable information (PII) is located at two levels: 1) direct indicators of a student's identity such as birth date, home address, social security number and parent's names; and, 2) indirect indicators of identity that results in the direct identification of students. The current application of suppression rules by CDE broadens the interpretation of PII beyond FERPA guidelines and Colorado's data privacy law by assuming that most indirect indicators of identity, even when reported at the aggregated level (e.g., school level), results in the direct identification of individual students. Data disclosure avoidance techniques should match the likelihood that a specific report actually reveals PII. Because the risk of releasing personally identifiable information is fundamentally different in aggregate data sets versus student-level, the Department should better align protections to the risk level. We believe the data request and sharing process could recognize these different risks and facilitate sharing of data sets for both reporting and research purposes.

- **Data Lab**
 - Limited Aggregation Features
 - Aggregation across multiple demographics
 - Significant suppression occurs when multiple demographics are introduced.
 - Aggregation across tests
 - It would be useful to be able to pull in multiple tests into one data lab pull to be able to compare multiple grades in one spreadsheet.
 - New Aggregation Categories
 - There are many levels of aggregation that are collected and coded for students that we are unable to aggregate for on Data Lab (e.g. Migrant Status). It would be useful to have expanded options for aggregation in Data Lab, such as primary home language, or school governance type.

- Output metrics
 - The only metric that is reported in data lab is mean scale score for achievement, and median growth percentiles for growth. Like the former DataLab, it is critical for data lab to also report the percentage of students meeting different achievement levels.
- Cohort Tracking
 - As data is currently reported there is no way to see how cohorts of students are progressing over time. It would be useful to generate a cohort tracking measure that allowed us to see how cohorts of English language learners are doing over time, for example. Or, what proportion of students who exceeded expectations in 3rd grade ELA continue exceeding expectations in 4th and 5th grade ELA.

- **Miscellaneous Challenges**
 - Cross-tabulation
 - Data Lab is an excellent resource for cross-tabulating the assessment results available. Yet that is the only data set where any cross-tabulation is possible. The ability to cross-tabulate demographic, graduation, and other data is critical to understanding which and how students are being served by the education system.
 - Asterisks
 - When asterisks are used to represent a suppressed cell value, it is impossible to use a search and replace function to find those values because an asterisk "*" is the wildcard expression in excel, and will return every cell in the spreadsheet. "Na" or a more traditional null character would be easier to work with.
 - Merged Cells (in Data Lab Exports)
 - Merged cells are difficult to work with because often filtering or summing data for analysis requires the first entry in each row to have a name. If a user is limited to excel remedying this could take a significant amount of time. For example, if we want to see which elementary schools across the state have the highest academic achievement by EMH level, to filter for only Elementary schools you have to unmerge and fill the elementary levels, in order to capture all schools within the field. Clearly this is a rudimentary example, however, the effects compound when we look at multiple levels (e.g. Year > Test > District > EMH > School > English Language Learners).

- CDE Website Navigation
 - Finding specific data sets can be very challenging because the website is not designed in an intuitive way. The READ Act data dashboard, for example, is located in a sidebar that is separate from other tabs in the READ Act page (below the location information) such that the eye doesn't think to go there if you are looking for the data dashboard. Similarly, the growth data is not with other assessment data. While the departments at CDE may be quite separate, for a user, the two sources are often very related and considered in unison.

As stewards of data about Colorado students, schools, and districts, we ask that you take the mantle of empowering the public with critical information about the education system in our state. We hope -- and expect -- the state to continuously iterate and improve upon its reporting practices and systems. To do so effectively and responsively can only occur through ongoing engagement with stakeholders and users of the data.

In particular, we hope to continue conversations about the appropriate n-size, disclosure avoidance techniques, the definition for personally-identifiable information, how data is accessed, what data is collected and reported, and the research and data request process, amongst others.

We look forward to these conversations. Thank you for reviewing our perspective of the current challenges in Colorado's publicly available data.