Education Insights Center
Advancing Research and Policy for K-12 and Postsecondary Education

POLICY BRIEF
JANUARY 2020

Use Drives Quality: Considering Data Quality Issues in California's Pursuit of a Cradle-to-Career Data System

An addendum to the series:
California Education Policy, Student Data, and the Quest to Improve Student Progress

By Colleen Moore

California is Moving Forward on a Data System

California is taking steps toward building a statewide data system to support efforts to improve student progress and outcomes from preschool through higher education and into the workforce (known as a P20W data system), a move recommended by numerous research, policy, and advocacy organizations over many years.[1] The state's Cradle-to-Career Data Workgroup comprises representatives of the public segments of K-12 and higher education, private colleges and universities, workforce development agencies, and agencies administering financial aid, health, and human services programs.[2] The Workgroup will meet over the next 18 months to develop recommendations about the structure and function of a P20W data system, with input from advisory groups of educators, researchers, education advocates, and other stakeholders.

The Education Insights Center produced a series of reports culminating in recommendations for the structure and governance of a P20W data system.[3] This brief follows up on that series, with a focus on data quality; the brief was informed by the author's experience using California's existing education and workforce data systems to conduct research, as well as conversations with 14 researchers and other experts who have significant experience using those data to conduct research, to develop data tools for educators and the public, and for operational purposes.[4]

Data Quality is an Important Consideration

Data quality, as discussed in this brief, refers to the accuracy and completeness of the data, which is an important consideration in the development of a P20W data system. Administrative records, like those maintained by schools, colleges, and state agencies for operational use, offer great opportunities as the foundation for a P20W data system. They allow for observation of individuals over time and across institutions, contain a broad array of information, and are comprehensive rather than representing a sample of individuals. All administrative records have some issues with inaccurate or incomplete information that can

result from data entry errors, miscoding of information, and a number of other factors. However, with sufficient attention to assessing data quality, and appropriate steps to mitigate any issues through good data management, administrative records are uniquely valuable for informing decisions about policy and practice.[5]

Most K-20 and Workforce Data are of Good Quality

The student records maintained by the California Department of Education (CDE), California Community Colleges (CCC), California State University (CSU), and University of California (UC), as well as the employment and earnings records maintained by the Employment Development Department (EDD), are fairly accurate and complete, with a level of quality problems no greater than would be expected in any large administrative data set. Not surprisingly, data elements that are routinely used for a particular purpose, such as the courses students enrolled in, the grades they earned, and the certificates or degrees they were awarded, are of higher quality than elements that serve little operational purpose for the institutions that collect them. Data used to allocate funding, to administer specific programs, or for reporting on accountability metrics are generally of good quality.

The data quality problems that exist usually affect a small share of cases and are of several varieties, including:

- missing data;
- inaccurate data, such as codes in a data element that do not match the data element dictionary (the documentation that defines data elements and their codes) or responses that appear incorrect in the context of other information; and
- inconsistent data, such as different identification numbers or different demographic information for the same student across multiple records.

Problems that affect a larger number of cases are generally well known by the agencies that collect and maintain the data. One example is the assignment by community colleges of appropriate codes to courses and programs (known as Taxonomy of Programs, or TOP, codes), which is known to be inconsistent across (or even within) colleges. Another example is the information in CDE's data system on which high school courses meet the requirements for admission to CSU and UC, which sometimes differs from the official list of approved courses maintained by UC.[6]

Some reasons for poor data quality where it does occur include the following:

- ***Misunderstanding of questions***. Some self-reported information can be inaccurate if the question or the response choices are misunderstood. For example, if a college application asks students about their current educational attainment level, but students misunderstand the question to be asking about their ultimate goal, that can result in responses showing that an 18-year old applicant has a graduate degree.
- ***Newly collected information***. When new data elements are added, it can take time for reporting entities to initiate collection of the information and apply codes correctly.

- ***Changes to data elements***. Data codes sometimes evolve, and it takes time for reporting entities to learn about and correctly apply the changes, such as recent efforts to expand race/ethnicity codes to allow for more disaggregation.
- ***Misunderstanding about what should be reported or how to report it***. Different institutions, departments, or individuals can have different interpretations of data reporting instructions, resulting in variation in the information collected.
- ***Limited resources and capacity***. Some reporting institutions have limited staff capacity; provide inadequate training to their staff; have less effective processes for data entry, management, and reporting; and lack the sophisticated technology that can make it easier to identify and mitigate data quality problems. The central offices collecting and aggregating the data from institutions may have limited resources to run data quality checks, to make timely updates to reporting instructions, and to respond to questions from institutions about data discrepancies or other issues.

While the quality of individual records at the public education segments and the EDD is generally good, some other data sets need improvement in order to be included in the P20W system. Student-level data on early learning, adult education, and workforce training programs should be incorporated, but such data are currently not collected and reported in a systematic way with appropriate quality controls. The quality of student-level data at private colleges and universities may vary, an issue the Association of Independent California Colleges and Universities (AICCU) is currently exploring as part of efforts to develop a longitudinal student data system on behalf of its member institutions.[7] In addition, while current procedures for matching records across data sets with different identifiers—based on individuals' full name, birth date, and other characteristics—yield good match rates, the matching process could be facilitated with the development of a single, unique identifier.

Perception of Data Quality Problems is Often Larger than the Reality

Several issues can lead data users to think that data quality issues are a more significant problem than is actually the case, resulting in a lack of trust in the data and in the agencies putting them out. For example, there has been a proliferation of dashboards and other data tools created by the educational institutions, their systemwide offices, and other entities, which often present similar metrics with slightly varying definitions. The measures of graduates' earnings calculated by the CSU and UC offer one example, as each system uses a different definition of the time elapsed since graduation. Such differences can be confusing and make it difficult for users to interpret the information or understand how and why it varies. The problem here is one of definition and interpretation rather than the accuracy and completeness of the underlying records used to create the metrics, but it contributes to the perception of poor data quality. Timing of data reporting, and updates made to the data over time, can also contribute to this perception. Data reflected in statewide reports or data tools may not match an institution's own records because the submitted data do not reflect recent updates made to the information, which can also undermine trust in the data.

Significance of Quality Issues Depends on the Purpose for the Data

The intended uses of data in a P20W data system affect the significance of any data quality issues. If the data are intended for use in providing direct services to individual students, such as determining whether a student has met prerequisites for enrollment in a course or completed the requirements to earn a degree, then, as one interviewee put it, the data "have to be bullet proof." If, however, the data are to be used for reporting, calculating metrics, creating data dashboards, and conducting research to improve policy and practice, then having a small share of records with quality issues—records that cannot be correctly matched across institutions, or that contain some missing or incorrect information—does not pose a significant problem. Good data management and research practices can mitigate such issues.

Researchers interviewed for this report could not recall any occasion when a data quality problem seriously hindered their analyses. They emphasized the importance of taking substantial time to assess the data for any potential quality problems by, for example, running descriptive statistics (such as a frequency tabulation) on every variable to look for responses that are not included in the data element dictionary, values that fall outside specified ranges, or conflicting information across different variables or in the same variable across different data files. Researchers described actions they take to deal with issues identified during quality checks, such as focusing their analyses on higher quality data elements, excluding the small share of cases with incorrect data, and "cleaning up" the data based on a set of decision rules. Documenting these actions and any limitations posed by data quality issues can aid interpretation and application of the results.

California has Some Experience and Success Using Linked Data

California's K-20 education and workforce data have been used for important reporting and research purposes. Each of the public education systems uses its student information system to populate data tools that respond to state accountability reporting requirements and help educators and the public understand the progress and success of its students.[8] Some tools bring together data from multiple sources based on limited data sharing agreements, including metrics provided by the higher education systems showing students' employment and earnings outcomes, determined by matching student records to EDD data. A P20W data system will rely on the same underlying data, expanding the options for data tools that track student progress and success across the various education systems. This will fill the gaps in our understanding about what happens to students at the transitions, and about the impact of prior educational experiences on students' current progress and ultimate outcomes.

In addition to the data reports and tools, researchers have used the data that will feed into a P20W system to address important questions related to education policy and practice. For example, studies have demonstrated the importance of reforming assessment, placement, and remediation policies in the CCC, and have assessed the early impact of recent reforms on student progress.[9] Studies on the labor market outcomes of career education in the CCC have demonstrated positive returns to a wide variety of certificates and associate degrees, results that can guide colleges in their efforts to offer programs most likely to lead to positive

returns, and help them direct students to appropriate courses of study.[10] Data quality issues have not been an impediment to such studies, but the work is severely limited by the disconnected nature of the state's current data systems and the need to negotiate data sharing agreements with multiple agencies. A P20W data system with established procedures for assessing and addressing quality issues and providing access to the data for researchers will facilitate this work and lead to improved education policies and practices.

Using the Data Will Improve Their Quality

> *"When data are used, data quality issues can be revealed as people really see the data for the first time. Using the data is a part of data quality. Data quality is a process, not an end."*—Paige Kowalski, Data Quality Campaign

Creating and using a P20W data system could help identify and fix data quality issues that are not currently recognized. Linking data sets together provides more opportunities to cross-check information and identify problems and inconsistencies. The data collection and management processes of reporting institutions will likely improve as a "data for compliance" mindset gives way to an understanding that the data are widely shared and are critical for continuous improvement. As an example, one interviewee pointed to the state's recent inclusion of dual enrollment participation in the College and Career Readiness Indicator in the California School Dashboard, noting that the record keeping on dual enrollment has since improved.

Ensuring data quality must be an ongoing focus as California establishes and maintains a P20W system, with opportunities at the institutional, system, and state levels to improve quality. Some examples of such actions include

- ***Reporting institutions,*** such as schools and colleges, should work to improve their capacity, perhaps with support from the state as needed, to train staff responsible for collecting and coding data, to ensure both a good understanding of reporting instructions and an appreciation for the importance of accuracy given how the data are used. Institutions should also consider having data reviewed by the offices responsible for the relevant activity prior to submission, as those offices may be in the best position to recognize errors in the data.
- ***Systemwide offices/participating agencies*** that receive and aggregate institutional data should ensure that their data element dictionaries and reporting instructions are clear and up to date, and provide adequate notice of changes. They should review and, where needed, improve their quality control processes (by, for example, checking data against prior terms/years to identify significant deviations that could indicate incomplete or inaccurate data). They should assist institutions in their efforts to provide staff training, by providing manuals, webinars, or other supports. They should consider evaluating the offices responsible for submitting data (such as college institutional research offices) to ensure they meet a standard for effectiveness and to identify common challenges that might require system action.

- ***The managing entity*** assigned to develop and administer the P20W data system should work closely with participating agencies to develop data matching processes, data definitions, and reporting schedules, with quality control procedures and a certification period built in that allows agencies to review and confirm their data. As part of data matching processes, the entity should develop and maintain a unique individual identifier. The entity should ensure that processes for sharing data with researchers include commitments to conduct quality checks and respond appropriately to any issues. Transparency is critically important to engendering trust in the data, so the entity should provide comprehensive and accessible documentation of data definitions, data submission and quality control processes, known data quality issues, and contact information for staff who can answer questions.
- ***State policymakers*** are responsible for establishing an effective structure for administering the P20W data system, one that ensures some independent authority for collecting the data and setting standards for data submission. They must ensure the managing entity has sufficient funding and staff capacity to administer the data system and maintain good data quality.

Finally, ensuring that participating agencies and institutions get value for themselves out of having a statewide data system will help to improve and maintain data quality. The data system should be used to construct reports and dashboards that are useful to the agencies in their own work, as well as to address the interests of a wide array of stakeholders related to education planning, labor market analysis, safety net planning, and other broad concerns. As one interviewee said, “when it comes to data, use drives quality.”

Acknowledgments

A number of researchers and other experts on California’s education and workforce data generously shared their time and expertise on data quality issues, including Kramer Cohen, Laura Coleman, Kevin Cook, Marisol Cuellar, Anthony Dalton, Betsey Friedmann, Jacob Jackson, Hans Johnson, Paige Kowalski, Michal Kurlaender, Alyssa Nguyen, Patrick Perry, Sherrie Reed, and Jesse Rothstein. I am grateful for the review and comments provided by Jacob Jackson, Paige Kowalski, Sherrie Reed, and Andrea Venezia, which helped improve an earlier draft of this report.

Endnotes

[1] For example, see Vernez, G., Krop, C., Vuollo, M., & Hansen, J. S. (2008). *Toward a K-20 student unit record data system for California*. Santa Monica, CA: RAND Corporation; The Education Trust–West (2010). *No time to delay: Delivering the statewide data systems California's students deserve*. Oakland, CA: Author; Taylor, M. (2013). *Improving workforce education and training data in California*. Sacramento, CA: Legislative Analyst's Office; Jackson, J. & Cook, K. (2018). *Modernizing California's education data system*. San Francisco, CA: Public Policy Institute of California.

[2] For information on the Workgroup and the state's planned process to design a California Cradle-to-Career Data System, see https://cadatasystem.wested.org/.

[3] The reports in the series can be found on the EdInsights website at http://edinsightscenter.org/Publications/Research-Reports-and-Briefs/ctl/ArticleView/mid/421/articleId/2198/California-Education-Policy-Student-Data-and-the-Quest-to-Improve-Student-Progress.

[4] See the Acknowledgments box on page 6 for a list of the people interviewed for this brief. The author has considerable experience using student records from the California Community Colleges and the California State University for research, as well as earnings records from the Employment Development Department's base wage file.

[5] For discussions of the value and limitations of administrative data, including data quality issues, see Boruch, R. F. (2011). *Administrative record quality and integrated data systems*. Philadelphia, PA: Actionable Intelligence for Social Policy, University of Pennsylvania; Rothbard, A. (2013). *Quality issues in the use of administrative data records*. Actionable Intelligence for Social Policy, University of Pennsylvania; Connelly, R., Playford, C. J., Gayle, V. & Dibben, C. (2016). The role of administrative data in the big data revolution in social science research. *Social Science Research*, *59*, 1-12.

[6] UC's Course Management Portal is the authorized database of all approved "a-g" courses. Information on courses that meet the requirements can be found on UC's website at https://hs-articulation.ucop.edu/agcourselist.

[7] As described in *Report on Phase One (December 2018-June 2019)* of the Intersegmental Coordinating Committee's A Planning Grant on California Intersegmental Data and Postsecondary Educational Metrics on Student Outcomes, AICCU is interested in participating in the state's P20W data system on behalf of its member institutions. The Bureau for Private Postsecondary Education (BPPE) collects aggregate institution-level data about student outcomes from private postsecondary institutions in the state, including institutions that are not members of AICCU. The AICCU and BPPE are included as representatives on the state's Workgroup.

[8] The California Department of Education's website includes the California School Dashboard at https://www.caschooldashboard.org/, and provides other summary data through DataQuest at https://dq.cde.ca.gov/dataquest/. The CCC's Student Success Metrics and other tools can be found at https://www.cccco.edu/College-Professionals/Data. Also see the California State University's Student Information Dashboards at https://www2.calstate.edu/data-center/institutional-research-analyses, and the University of California's Information Center at https://www.universityofcalifornia.edu/infocenter.

[9] For example, see Willet, T. (2013). *Student-Transcript-Enhanced Placement Study (STEPS) Technical Report*. The Research and Planning Group for California Community Colleges; Rodriguez, O., Johnson, H., Mejia, M. C., & Brooks, B. (2017). *Reforming math pathways at California's community colleges*. San Francisco, CA: Public Policy Institute of California.

[10] Stevens, A., Kurlaender, M., & Grosz, M. (2019). Career technical education and labor market outcomes: Evidence from California community colleges. *Journal of Human Resources*, *54*(4), 986-1036; Bohn, S., Jackson, J., & McConville, S. (2019). *Career pathways and economic mobility at California's community colleges*. San Francisco, CA: Public Policy Institute of California.

Education Insights Center
Advancing Research and Policy for K-12 and Postsecondary Education

SACRAMENTO STATE