California Cradle-to-Career Definitions Subcommittee Meeting Summary

July 21, 2020

The California Cradle-to-Career Data System Definitions Subcommittee met on July 21. This document provides a summary of the key points that emerged from the meeting, which focused on developing a framework for how information could be aligned for the Cradle to Career data system, including developing policies related to specific data alignment dimensions. The suggestions from this group will be used to craft specific definitions for elements in the P20W data set.

More information about the meeting, including support materials, a recording of the meeting, and the PowerPoint, are available at https://cadatasystem.wested.org/meeting-information/definitions-subcommittee (click on "Meeting Materials").

The following Definitions subcommittee representatives attended the meeting:

Randy Tarnowski, Association of Independent California Colleges and Universities; Joanna Murray, Bureau for Private Postsecondary Education; Todd Hoig, California Community College Chancellor's Office; Jerry Winkler & Channa Hewawickrama, California Department of Education; David Sanabria & Jennifer Schwartz, California Health and Human Services Agency; Patrick Delaney, California Department of Social Services; Monica Malhotra, California State University; Rima Mendez & Amy Tong, California School Information Services; Adrian Felix, California Student Aid Commission; Erin Skubal, California Teacher Credentialing Commission; Patrick Getz & Daniel Rounds, Labor and Workforce Development Agency; Chris Handy, University of California Office of the President; and Todd Britton, University of LaVerne

Agenda Review and P20W Data Elements

The facilitator provided an overview of the P20W data set elements, dashboard, and query builder. The facilitator affirmed that several issues of concern were being addressed in other subcommittees, including:

- A deidentification policy to ensure that small cell sizes are appropriately treated, so that individual's identities cannot be discerned
- Deidentification will be implemented by the managing entity, in accordance with this policy
- Timeframes for data uploads have been determined (once per year for the data used to construct the dashboards, query builder, and partner research data set)

Polices and Principles for Aligning Data Definitions

The group discussed a draft set of policies and principles for aligning data definitions.

David Sanabria of CHHS noted that versioning should be used to track when there are changes in the data, particularly so that earlier versions of figures are not erased. Monica Malhotra of CSU suggested that rules be established about when prior data can be updated. For example, for IPEDS, you may only change the numbers in a one-year timeframe.

David Sanabria of CHHS suggested that metadata (information about the data shown) should also be available to the public. For example, each data element should be time-stamped to clarify when it was first recorded, when it was uploaded, and if it was changed. As part of the lineage documentation, information on any recoding of the source data from individual K-12 or postsecondary institutions that is implemented by the state agencies should be documented. Jennifer Schwartz of CHHS added that it would be helpful to clarify what types of metadata the partner entities need to provide and how to best share this information with the public. For example, there could be a webpage that explains the data in less technical terms. Many subcommittee members agreed that defining metadata and making it available should be a priority.

Monica Malhotra of CSU noted that it may be useful to allow for different definitions of the same element to be available. For example, if there is a difference between the federal and California definitions, as is the case for sex/gender. She also cautioned that education institutions may not capture student characteristic data each year (such as disability or homeless status), so it will be difficult to determine when a specific item was valid.

Note: An updated copy of the policy document is on page 4 to be used for reference in future meetings.

Small Group Work: Data Alignment Dimensions

Subcommittee members broke into small groups to discuss several data dimensions—factors that need to be determined when crafting data definitions. The notes below reflect the full discussion after each small group reported out.

Timeframes: The IPEDS definition for academic years should be used, which means that leading summers should be included (sometimes referred to as a college year).

The group agreed that there should be a clear notion of an event so data can be tracked for those events. For example, "Effective Start" and "Effective End" timestamps for program and service involvement could be used. However, given that education institutions collect data only on a term or annual basis, these concepts may not translate easily into the Cradle-to-Career data system. Even enrollment in an education institution can be open-ended. For example, in community college, students often stop taking courses for a period of time, and then enroll again. Therefore, it will be important to establish a definition for an "exiter" that is more expansive than whether a student earned an award.

Attending Multiple Institutions: Students should be included in the display for all institutions they attended, if they attend multiple institutions in the same year. Then, these figures should be deduplicated when a higher-level view is selected. For example, if a student attended two colleges in the same community college district, at the college level disaggregation, the student should be included at both institutions, but at the district level or state level views, the student should only be shown once. It may be helpful to include a flag that indicates students attended more than one institution in the query builder.

Milestone Attainment Years: For the purpose of the query builder, a student should be included in a cohort the first time they attained that milestone (such as the first time they enrolled in third grade). However, it would be important for researchers to be able to determine whether a student had repeated third grade when examining test scores.

Multiple Values: The subcommittee discussed how to develop policies to assign student characteristics when there are multiple values, such as if the same individual is reported as having two different races in the same time period by different institutions. Participants recommended that an entity should report all values they have on the same individual.

Data Gaps: Different flags are needed to distinguish when data are not captured by partner entities, when a value is missing, when different values are collected by different partner entities, and if missing information has been provided by a different source. Notes should be included for metrics that are likely to have reduced representation, such as employment and earning data that is limited to those with valid social security numbers and working for specific types of employers in the state of California.

Rates: Due to variability in whether each value is available for all individuals, the subcommittee recommended that, for rates, the number of individuals in the denominator should be displayed. Displays could also use alternative approaches for showing rates, such as using the number per thousand.

Given that different dimensions may be applicable for different elements or calculations, and for different data providers, the group members decided that all dimensions should be examined for each data point individually. These dimensions include:

- For all definitions: geographic domain (such as within the same school or any school), which value to use if there are repeats (first time, most recent time, if ever), time stamps (start and end dates, data collected, data uploaded), lineage (recoding by state agency)
- For notes: language regarding missing information, lagged information, low quality information, divergence from national definitions, information only available from a smaller subset of individuals
- For enrollment and exit definitions: how a cohort would be constructed (including dual enrollment status and enrollment thresholds)

As part of this examination, the subcommittee should consider recommendations about whether or how data could be displayed in the public-facing tools to reduce the likelihood of misinterpretation.

Next Steps

The group agreed to a process for developing data definitions for the public-facing tools and documentation for the P20W data set where each partner entity that is contributing data provides documentation on the relevant data that they collect and WestEd constructs a proposal that for what would appear on the public tools and flag issues about missing or lower-quality data. This proposal will include possible ways to handle the topics identified in the data dimensions document, such as when there are multiple values for the same individual in the same timeframe. At each subcommittee meeting, subcommittee members will discuss the proposal, make edits, and confirm a final product.

Updated Policies and Principles for Aligning Data Definitions

When presenting information in public facing tools such as dashboards and query builders, definitions for the Cradle-to-Career Data System will be determined according to the following policies and principles:

- **Intent**: Articulate the intent for aligning information
- **Granularity**: Provide information at the most granular level possible, while ensuring that data suppression keeps information on individuals secure
- **Expansion**: Identify opportunities for expanding categories in future phases, but begin with elements that most partner entities collect
- **California Focus**: Evaluate when it is preferable to use California definitions over federal definitions, clarify when non-federal definitions are being used in public-facing tools, and ensure both definitions are available to researchers
- **Lineage**: Clarify the origin of each data element and describe how information gets adjusted as it moves from local institutions to state agencies to the Cradle-to-Career Data System
- **Missing Information and Gaps**: Create a consistent process for handling missing information, which distinguishes between partner entities that do not collect that variables, instances where the data are not present, and individuals who declined to state the information
- **Translation Tables**: In cases where partner entity data are being grouped into a broader category, create a translation table, clarify when one status trumps another, and indicate when specific categories are being assigned to an "Other" category
- **Changes**: Document changes to definitions and note when data definitions or values have changed in a manner that is easy for the public to understand
- **Transparency**: Provide information on metadata to the public such as when it was recorded and uploaded