October 2020 | **ERDC** | Olympia, Washington

Washington State P20W Longitudinal Data System Research Handbook

ERDC
Education Research & Data Center

Foreword

This P20W Longitudinal Data System Research Handbook introduces researchers and analysts to the Washington P20W data warehouse and the use of its products for research and analysis purposes. The Education Research and Data Center (ERDC) ERDC works with partner agencies to conduct powerful analyses of learning that can help inform the decision-making of Washington legislators, parents, and education providers. ERDC's data system is a statewide longitudinal data system that includes de-identified data about people's preschool, educational and workforce experiences.

This handbook is the result of hard work completed by ERDC staff and researchers, in collaboration with partner agencies. The interest and support for building this resource was foundational to the success of its development.

Using the Handbook

The handbook is not a comprehensive look of all the data that goes into P20W or any individual source data set. Rather, this handbook is a body of knowledge that directs researchers to helpful and appropriate resources.

The Table of Contents includes active hyperlinks to each handbook section, for ease of use. Links to ERDC research reports and data dashboards are located at the beginning of each handbook section, where appropriate. Please contact the ERDC directly for detailed information about any of the subjects covered in the handbook.

ERDC ADDRESS	EMAIL	PHONE	FAX
Education Research and Data Center 106 11th Ave SW, Suite 2200 PO Box 43124 Olympia, WA 98504-3113	ERDC@ofm.wa.gov	360-902-0599	360-725-5174

Table of Contents

Introduction

The purpose of the Research Handbook is to familiarize users with the core source data that goes into the P20W system, how to request P20W data, and methods for using it to conduct research. The Research Handbook also contains background information on Washington's Education Research & Data Center (ERDC), a description of ERDC's identity matching process, privacy and security guidelines, and summaries of the source data included in the data warehouse. These summaries include important details about data elements and source file structure, policy and program background information and any limitations or caveats. Also included are methodological insights from ERDC researchers, based on their experience with the Washington P20W data warehouse.[1] This handbook will be updated regularly, to include additional information about new source data and data marts that are produced from the P20W data warehouse.

About the Education Research & Data Center (ERDC)

ERDC was established by legislation (RCW 43.41.400) in 2007 and works with partner agencies to conduct analyses to inform the decision-making of Washington legislators, educational institutions, researchers, families, and students. ERDC's mission is to develop longitudinal information spanning the preschool to workforce system, in order to facilitate analyses, provide meaningful reports, collaborate on education research, and share data in ways that protects the privacy of students. ERDC researchers analyze data and answer critical policy questions that are raised by stakeholders. Data is available to external researchers through ERDC's data request process.

In the beginning, ERDC's major priorities were to:

- Coordinate with other state education agencies to compile and analyze education data
- Collaborate with the LEAP Committee and education and fiscal committees of the Legislature to compile and analyze data, to ensure that legislative interests are served
- Track enrollment and outcomes through the Public Centralized Higher Education Enrollment System (PCHEES), within the Office of Financial Management
- Develop a long-term higher education enrollment plan with other state educational agencies
- Conduct research that focuses on student transitions within and among the early learning, K-12, and higher education sectors of the P-20 system
- Share data from collaborative analyses with education agencies and institutions that contribute data to the ERDC

In 2009, the Washington legislature expanded the mission of ERDC to include:

- Identify key education research and policy questions to address and what data is needed
- Lead the governance of Preschool-to-Grade 20-to-Workforce (P-20W) data
- Serve on the K-12 Data Governance Group and provide them a list of data elements and improvements necessary to ERDC's work

[1] The handbook is not a comprehensive look of all the data that goes into P20W or any individual source data set. Rather, this handbook is a body of knowledge that directs the researcher to helpful and appropriate resources. Please contact the ERDC directly for detailed information about any of the subjects covered in the handbook.

- Monitor and evaluate the versatility and quality of education data collection systems of the organizations and agencies that contribute to ERDC, and ensure that the data they provide is relevant
- Provide recommendations to the legislature that meet the goals and objectives of the comprehensive K-12 data improvement system and K-12 Data Governance Group

The Washington State Longitudinal Data System (SLDS) is a statewide administrative record database referred to as the Preschool-to-Grade 20-to-Workforce (P20W) data warehouse, which connects education data across early learning, K-12, postsecondary, and workforce sectors. The P20W data warehouse is housed within the Washington Office of Financial Management (OFM) and managed by the ERDC. Longitudinal and cross-sector data make it possible to measure long-term progress and differences across cohorts, which informs education policy and practice.

ERDC's Identity Resolution Process

At the core of the P20W data warehouse is the linking of cross-sector data. Through an identity resolution process, ERDC links data files from contributing agencies and institutions to facilitate longitudinal analysis. Identity resolution is the process of identifying records that belong to the same entity (e.g. person or household). The purpose of identity resolution is to create linkages across multiple data sources so that students' early learning records are linked to their K-12, postsecondary, and workforce records. For ERDC's P20W Warehouse, this involves linking individual-level data, such as names and birth dates, across multiple sources to create unique person identifiers. These identifiers are referred to as the "P20ID." P20IDs are assigned to all individual-level data received by ERDC from our data contributors.

First, an identity resolution "token" is created for each record in a dataset. Identity resolution tokens are concatenated sets of identifiers that are guaranteed to be unique to an individual. For example, in workforce data that comes from the Employment Security Department (ESD), more than one individual can ostensibly have the same Social Security Number (SSN). As a result, ERDC cannot rely on SSNs to uniquely identify individuals in this source of data. Consequently, each ESD identity resolution token is composed of the SSN, the employer account number and the employee name. ERDC has found that this set of identifiers is guaranteed to be unique to an individual in ESD data.

Since every source of data has its own set of identifiers and its own set of data challenges, each source's identity resolution token definition is different. For example, different identifiers will be used for K-12 race and ethnicity data that comes from the Office of Superintendent of Public Instruction (OSPI) than what would be used for subsidized childcare data from the Department of Children, Youth, and Families (DCYF). After the identity tokens are created, they are loaded along with their associated set of identifiers into ERDC's Master Data Management (MDM) hub. Since identity tokens may contain personally identifiable information, they cannot be used outside of the MDM hub. Because of this limitation, the MDM hub assigns a unique whole number, the Token ID, to each identity resolution token. The MDM hub then assigns new P20IDs to every new TokenID.

ERDC's identity resolution process involves two steps:

(1) P20IDs are merged using "automerge" match rules. These are conservative rules for merging P20IDs. At this step, inaccurate matches are not much of a concern, as the auto-merge match rules are designed to ensure extremely low false positive rates.
(2) Then, prospective match pairs of P20IDs, along with their underlying tokens and their identifiers, are created using looser match rules than in step one.

These rules are designed so that the false negative rate is low. Unmatched pairs are not a concern at this step, since potential matches are manually reviewed. The resulting prospective match pairs of P20IDs and associated identifiers are brought into a spreadsheet, and the ERDC technical lead evaluates each potential pair of identities for a confirmed match. These P20IDs are then merged, with one P20ID in each confirmed match pair assigned to all the identity tokens of the defunct P20ID. The internal crosswalk of P20IDs to Token IDs is also updated.

No single set of identifiers is common to all data sources, so the identity resolution process and match rules are tailored to the source of data being matched. For example, K-12 data has names and birth dates, whereas ESD wage data has names of mediocre quality and SSNs. As a result, there is no way to directly match these together, but they can be indirectly matched by involving other sources of individual data. Postsecondary education data, for example, can be matched and merged with ESD wage data using their common set of identifiers, names and SSNs. The same postsecondary data can then be matched and merged with K-12 data with names and birthdates, two types of identifiers that are common across the two sources. The Department of Licensing (DOL) data is another important source of data that ERDC uses to bridge between K-12 and ESD wage data. ERDC uses direct and indirect processes to link K-12 students to workforce data, regardless of whether they are enrolled in a public postsecondary education institution.

Figure 1. Flow of Data into the P20W ODS

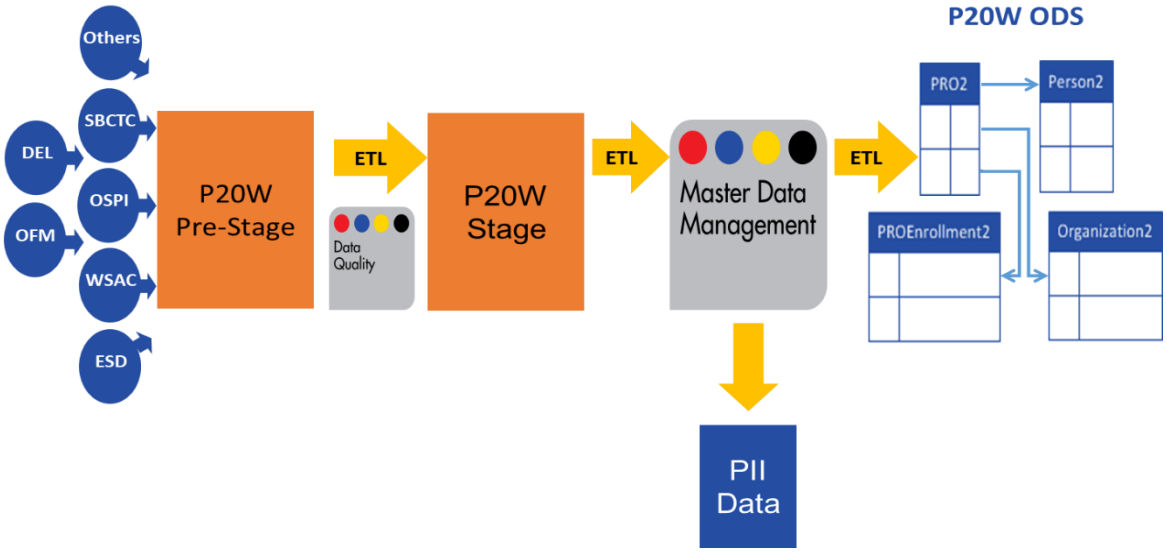

Figure 1 illustrates how ERDC loads source data from contributing agencies. First, the data is loaded to a pre-stage database, then it undergoes a series of quality checks before it is transferred to a stage database. Personally identifiable information (PII) is separated at that point from the rest of

the data and used for identity resolution. Once the identity resolution process is complete, the revised crosswalk of P20IDs and token IDs is incorporated into P20W Education Data Warehouse Operation Data Store (ODS). Once in the ODS, data are de-identified and available for analysis. Identifiers used in the identity resolution process do not advance beyond the MDM hub. The next section of the Handbook describes the core source data files that are loaded into the P20W data warehouse.

Part I. P20W Source Data Summaries

ERDC receives a variety of administrative datasets from agency partners[2] that are incorporated into the P20W data warehouse. These administrative datasets are outlined in Table 1, based on the category of data and data source. These datasets vary in subject matter, from early learning programs to K-20 public school data to workforce data. As such, ERDC's P20W warehouse is the most comprehensive longitudinal education data system in the state.

This section provides a set of data summaries, or quick references to the core data files that feed into the P20W data system. This information is not an exhaustive list of data in the system, nor does it provide the detail needed for a researcher to sufficiently complete an ERDC Data Request Form. Rather, these data summaries are designed to:

- guide researchers toward data that are relevant to their research questions
- provide meta data that will inform research design
- provide examples of how the data is used in research

Table 1. P20W Source Datasets

Data Category	Data Source and Description	Name of Dataset
Early learning and childcare	Department of Children, Youth, and Families (DCYF) DCYF focuses on the well-being of children to ensure that "Washington state's children and youth grow up safe and healthy—thriving physically, emotionally and academically, nurtured by family and community."	Early Support for Infants and Toddlers (ESIT) Program Early Childhood Educational Assistance Program (ECEAP) Child Care Subsidy Programs (CCSP)
K-12	Office of Superintendent of Public Instruction (OSPI) OSPI is the primary agency that oversees public K–12 education in Washington. OSPI allocates funding and provides resources and technical assistance so every student receives a high-quality public education.	Washington Kindergarten Inventory of Developing Skills (WaKIDS) Student Assessment Data Comprehensive Education Data and Research System (CEDARS) Data

[2] For details about partner agencies, please review the ERDC website.

Data Category	Data Source and Description	Name of Dataset
Public community colleges, technical colleges, and GED completion	State Board for Community and Technical Colleges (SBCTC) SBCTC advocates, coordinates and directs Washington state's system of public community and technical colleges. The Board collects student data from each member school and then shares it with ERDC.	State Board for Community and Technical Colleges Data Warehouse (SBCTC) Student Data
Public 4-year university	Office of Financial Management (OFM) OFM provides vital information, fiscal services, and policy support to the Governor, Legislature, and state agencies. OFM works with the state public institutions to track public higher education enrollment and graduation outcomes.	Public Centralized Higher Education Enrollment System (PCHEES) Data
Financial aid	Washington Student Achievement Council (WSAC) WSAC provides strategic planning, oversight, advocacy, and program administration to support increased student success and higher levels of educational attainment in Washington. WSAC collects financial aid award data from Washington Public Colleges and Universities and provides an aggregated financial aid (unit record) dataset to ERDC.	Washington Student Achievement Council (WSAC) Unit Record Data
Apprenticeships	Labor and Industries (L&I) L&I is the state agency dedicated to the safety, health, and security of Washington's 3.3 million workers. One of the agency's roles is to collect applicant information from US Department of Labor apprenticeship programs.	Registered Apprenticeships Data
Workforce	Employment Security Department (ESD) ESD provides services related to unemployment benefits and insurance. This is a valuable resource for data on the employment outcomes of high school and postsecondary graduates.	Unemployment Insurance (UI) Program Data

Early Learning and Childcare

Early Support for Infants and Toddlers (ESIT) Program

The Washington State Early Support for Infants and Toddlers (ESIT) program provides individualized and quality early intervention services to young children (birth to three years old) who have disabilities or developmental delays. The Department of Children, Youth, and Families (DCYF) oversees the ESIT program, in accordance with the federal Individuals with Disabilities Education Act (IDEA), Part C. The ESIT program has two main objectives: 1) find eligible children through screening, tracking, monitoring, and referral services for at-risk children, and 2) provide

early intervention services, including developmental and therapeutic services for children identified as developmentally delayed or who have an established condition for delay. Each year, DCYF provides ERDC with an annual snapshot of the ESIT database (excluding some data tables) for all prior years.

ERDC Reports that use ESIT Data
Identifying Children in Need for Early Intervention Services in Washington State: An Application of Washington State All Payer Claims Database in Education Research
Who Receives Early Intervention Services in Washington State? An Analysis of Early Support for Infants and Toddlers Program Administrative Data

To be ESIT-eligible, a child must have a 25 percent delay or perform at 1.5 standard deviations below their age group in one or more of the five developmental areas[3]. Early intervention services may include but are not limited to specialized instruction, speech therapy, occupational therapy, or physical therapy, which can be provided in a variety of settings including home, childcare, preschool or school programs, and communities. Early intervention services end on the child's third birthday, or upon achieving satisfactory results before the age cutoff[4]. Each ESIT program participant has an Individualized Family Service Plan (IFSP) developed for them to access the intervention services and resources specified in the plan. ESIT also uses a Child Outcomes Summary Form (COSF) to record specific outcomes at program entry and exit[5].

The ESIT database includes all children who were ever referred to the ESIT program from 2009 to 2019. The ESIT database contains 139 tables and 1,203 unique columns. ERDC staff summarized key information into six data views: Child/family information, Referral, Eligibility, Evaluation, IFSP, COSF, and Transition. These six data views are not currently available through the P20W warehouse, but researchers can complete ERDC's data request process to access them.

Key information provided in the data tables are:

1. Child characteristics, family characteristics
2. Name, gender, birth date, contact person, race and ethnicity, language, county, and organization
3. Referral sources, reason, and received date
4. Eligibility determination, evaluation domains and results
5. Functional domain, functional test, diagnosis, eligibility basis, and provider
6. IFSP development and renewals
7. Service type, setting, provider, date, and service plan
8. COSF and reevaluation
9. Outcome type, description, provider, and date
10. Transition reason and transition destination

[3] In addition, children with physical or medical conditions like Down Syndrome, serious hearing or vision problems, Cerebral Palsy, cleft lip/palate etc., are eligible for the ESIT program.
[4] ESIT services can also end voluntarily based on the family's decision, if the family moves out of Washington, or if the family is out of reach from the current service providers.
[5] COSF outcomes are measured in the areas of Positive Social/Emotional Skills, Acquiring and Using Knowledge and Skills, and Use of Appropriate Behaviors. Results indicate if a child has Age-Expected Skills, Decreasing Degree of Age-Expected Skills, or No Age-Expected Skills, as well as Decreasing Degree of Immediate Foundational Skills.

11. Transition age, reason, and provider

Since it excludes children who do not receive ESIT services, the ESIT database represents a limited subgroup of early learners in Washington. The quality of ESIT data is generally consistent across the available years. While some data have a considerable number of missing values, most ESIT data is well-suited for examining participants' characteristics, eligibility, and services received. ESIT data on child characteristics are quite complete, while a substantial amount of data is missing on family characteristics and children's diagnosed medical conditions. ERDC staff found discrepancies within ESIT records indicating transition to ECEAP, when compared to actual ECEAP participation records extracted from ERDC linked data. Missing values do not appear to correlate with specific factors like time or report source.

How race is defined within the ESIT dataset is also unique, compared to definitions in other educational databases like CEDARS. ERDC recommends using the CEDARS demographic fields when combining ESIT with CEDARS tables, since CEDARS data on race and ethnicity is generally more detailed. Furthermore, ESIT column names are not always explicit or do not consistently align with other educational databases. Data columns like Name, Description, and "Dosage" of services may also be ambiguous; while these columns have generic names, they refer to more specific information within the data table.

Some children did not participate in the program after eligibility determination, or out of parents' choice. Children who do not participate due to these circumstances only appear in the referral data. Since the ESIT program provides services to very young children, some participants may exit the Washington State public education system entirely over time. In some instances, a participant's reason for leaving the program is missing or unknown, which could suggest the presence of uncorrected bias within the data. Reporting an unknown reason for leaving may stem from both administrative or participant sources, with no clear way to determine if the missing information is systematic (i.e., reason is not recorded or reason for exit is not represented in the database) or random (i.e., participants move without notifying ESIT, parents' decision to withdraw was not communicated to ESIT, changes to eligibility status, etc.).

ERDC researchers are currently analyzing ESIT data and medical data to examine the impact of health factors on participants' educational outcomes. Exploring the potential connections between childhood health and early learning can help researchers and policymakers establish a solid foundation for children's school-readiness, including at-risk groups. ESIT data on program participants and services is also useful for addressing research questions like those below:

- Explore service recipients by demographic characteristics, reason for eligibility, services received, and outcomes.
- Identify program trends over time based on the critical steps that participants complete, including referral, evaluation, service plan development and review, and program exit.
- Investigate children's pathways post-program participation, when combined with future educational attainment levels.

Early Childhood Educational Assistance Program (ECEAP)

The State of Washington provides preschool to three- and four-year old, low-income children through the Early Childhood Educational Assistance Program (ECEAP). DCYF oversees the program, which served 14,000 children at more than 390 locations during the 2019-20 school year. DCYF provides ERDC with annual data that includes all children who participate in the state funded ECEAP, which ERDC loads into the P20W system. Additionally, the data contains some information about the providers, site curricula, and teachers. The data does not currently include information on Head Start or privately funded early childhood programs.

There are two sources for ECEAP data. ECEAP data prior to the 2011-2012 school year is referred to as "Historical ECEAP" data. DCYF provided Historical ECEAP data as a collection of Excel and text files, which ERDC loaded into the P20W data warehouse. Historical ECEAP data is contained in one table, which is structured with one record per child per enrollment segment. Historical ECEAP data fields contain demographic information (gender, age, race, ethnicity, family income, disability indicator, and primary language) and enrollment information (site, contractor overseeing the site, start date and end date). ECEAP data after the 2011-2012 year comes to ERDC from DCYF's Early Learning Management System (ELMS) database, referred to as "ELMS."

ERDC's ECEAP Reports and Data Dashboards

- Early Childhood Program Participation and K-12 Outcomes
- Early Childhood Program Participation & WaKIDS Outcomes
- Early Learning Feedback Report
- Kindergarten Readiness Among Children Who Participated in the Washington State Early Childhood Education and Assistance Program (ECEAP)
- ECEAP Participation and Kindergarten Readiness Among Hispanic Children in Washington State

While the ELMS database contains over one hundred data tables, ERDC only receives a subset of its data. DCYF's data team prepares an annual extract of the ELMS database and sends the data subset to ERDC at the beginning of each year. ERDC staff created a series of four "data views" that pull key data from the ELMS extract that feed into the P20W data warehouse. These four data views are ELMS Eligibility, ELMS Enrollment, ELMS Organization, and ELMS Site. Each data view has its own unique level of analysis.

The ELMS Eligibility view includes data collected during determination of the child's ECEAP eligibility, so this information is only collected once (though some exceptions apply). The ELMS Enrollment view contains a record for every time a child enrolls in ECEAP. The ELMS Organization view outlines information about ECEAP providers, including contractors, subcontractors, and program sites, for each year. The ELMS Site view includes data on program sites for each year, including location details, facility licensing and operations information, and characteristics about the curriculum used at each site.

- The ***Eligibility*** view contains demographic information of potential value to the researcher: gender, age, race, ethnicity, family income, disability indicator, and language spoken at home.

- The ***Enrollment*** view contains a field for site, class within the site, start and end dates of enrollment, language of instruction, and whether the class is part day, full day, or extended day.
- The ***Organizational*** view contains information about sites and the subcontractor-contractor associations for each year of ELMS data.
- The ***Site*** view is not loaded to the data warehouse at this time but contains key characteristics of sites such as location (including latitude and longitude), whether the site is a licensed childcare facility or is operated by a tribal organization, and the curricula used at the site.

Depending on the nature of the study, the Historical ECEAP data of 2000-2012 is potentially problematic, while the ELMS data that covers the 2013-current years is considered reliable. Missing end dates during the 2011-2012 enrollment years, for example, were identified during the transition from the Historical ECEAP system to the ELMS data management system. This adjustment complicates the process for computing ECEAP "dosage" when measured by the number of days enrolled in ECEAP. ELMS data comes from a relational database, and the quality has improved as a result of this structure and DCYF's continued commitment to data accuracy. ERDC's data views capture all values of key variables. Thus, if a student has multiple race codes, each of the codes is captured and entered into the P20W system. It is up to the researcher to determine how to use all the codes, which are captured in multiple records.

One unique characteristic about the ELMS data is the associations between classes, sites, dosage, and curriculum. Classes are located within sites. Children at a site will receive a set of curricula at that site, but children's dosage is related to a class that could be part-day or full-day. This relationship is different from a college, where students have a relationship with the college that is part-time or full-time, and curricula are associated with classes.

This data is valuable to researchers who want to examine various aspects of state-funded ECEAP, including program-level and individual-level characteristics. Most of the existing research on early childhood programs explore the outcomes of ECEAP graduates as they transition into kindergarten and future grade levels. To address questions like these, researchers must also work with K-12 data. Researchers at the Washington Institute of Public Policy (WSIPP), for example, have conducted research using ECEAP data to explore the progress of ECEAP students within Washington's K-12 system.

Because ECEAP data includes information about participants and programs, it may be especially suited to address the following types of research questions:

- ***Child demographics:*** How does the ECEAP effect on K-12 outcomes such as WaKIDS scores vary among different racial or ethnic groups? Do K-12 outcomes for ECEAP participants vary by gender?
- ***Dosage/class quality:*** Does an increased dosage of ECEAP help children do better on WaKIDS?
- How do WaKIDS scores vary across contractors and/or sites?

ECEAP data does not account for participation in other early childhood education programs, so researchers must interpret their comparison results with caution. If a researcher uses a comparable

low-income group as a comparison group, then this comparison group could include children that participated in other early childhood education programs, like Head Start. This distinction could ultimately dilute the measured effect of the ECEAP program on participant outcomes. If the comparison group is drawn from K-12 students (other kindergartners, for example), then demographics from the K-12 data system must be used to account for that group. However, it is important to note that K-12 demographics may not be at the level of granularity needed for all potential methodologies. Given that ECEAP data reflect annual snapshots in time, researchers may also find the data insufficient for more complex, longitudinal trend analyses. The ECEAP program has steadily expanded over time, so trends in the data could be due to the program effect changing over time and/or potential changes in the population of participating children and families.

Child Care Subsidy Programs (CCSP)

Working Connections Child Care and Seasonal Child Care are collectively referred to as Child Care Subsidy Programs (CCSP). CCSP assists eligible low-income working families[6] by promoting access to childcare and after-school programs that help prepare their children to succeed in school. DCYF is the designated lead for setting CCSP eligibility and payment thresholds, while the Department of Social and Health Services (DSHS) is responsible for delivering CCSP services, determining children's CCSP eligibility, and authorizing payments for CCSP services[7]. Children served by this program range from birth to 13 years old, or up to 19 years old for those with a verified special need or under court supervision.

ERDC is in the preliminary stages of profiling CCSP data for research purposes. Although it has not been loaded to the P20W data warehouse, CCSP data will be available via ERDC's Kindergarten Data Mart in the near future.

DCYF provides ERDC with an annual feed of participant data that contains CCSP eligibility and payment data. The CCSP data extract, sometimes referred to as Subsidy data, includes monthly records from 2009-2019 for each child covered by the program. It comes from administrative payment data records, rather than a survey or questionnaire. This data is structured so that information is covered for each child and each month for eligibility or service. Children served by CCSP can see multiple providers over the course of one month. Providers may be billed for multiple service codes. All providers who receive a payment are required to submit a claim for each invoice. As a result, there is a record for every distinct occurrence of child identifier, month, service provider, and service code. Key variables in this dataset include: Invoice number, authorization number, claimed units, provider number, the month of service, service code, total units paid by the state, and total copay units.

One major limitation of the CCSP data extract is the lack of information on sociodemographic characteristics of the participating children and their families. Researchers can overcome this issue

[6] Families must have incomes at or below 200 percent of the federal poverty level (FPL) when applying, or 220 percent of FPL when reapplying to be eligible to receive the subsidy. The parent must be employed or self-employed in legal, income-generating, taxable activities to qualify.
[7] See WAC Chapter 110-15 for more information.

by merging CCSP data with other available data from the Washington Kindergarten Inventory of Developing Skills (WaKIDS) and/or the Comprehensive Education Data and Research System (CEDARS) databases.

Within the CCSP data extract, there are 44,750 records with a missing claimed unit. Reasons for this missing data include, but are not limited to, case classifications and reviews by program staff, migration, changes in household income, not having access to an eligible provider, or closure of the eligible provider. Many cases in the CCSP data extract have an interruption in their claimed units, which suggests that those participants were considered eligible and authorized to receive the subsidy but did not claim it. This can happen for multiple reasons. Individuals may become eligible for the subsidy, but they may have difficulty finding an eligible provider in their area, and therefore may be unable to receive the subsidy. Providers must also complete a daily report to receive the subsidy, which is a time-consuming process that may lead some providers to not claim the subsidy.

Thus, being CCSP-eligible and having an authorization number does not always mean that individuals received the subsidy. Researchers should be aware of these unmatched records in using the subsidy dataset. Another consideration regarding the subsidy dataset is that the Invoice Provider Year (i.e., the year of authorization) can differ from the year associated with the month of service, or the month of subsidy receipt.

Although ERDC staff have not yet conducted research using the CCSP data extract, there are many opportunities to explore how CCSP may help low-income families improve their financial stability and expand access to childcare resources for their children. In particular, the CCSP data extract may be useful to exploring the following research topics:

- Explore school readiness among children who received childcare subsidy, compared to low-income children who did not receive the subsidy
- Examine the effect of childcare instability on the development outcomes and school readiness of childcare recipients
- Investigate the continuity of childcare and child's development outcomes

K-12

Washington Kindergarten Inventory of Developing Skills (WaKIDS)

Washington Kindergarten Inventory of Developing Skills (WaKIDS) is a legislatively mandated assessment as part of state-funded, full-day kindergarten and is administered by OSPI. It began as a pilot program in the 2010-2011 school year and was implemented statewide in the 2011-2012 school year. WaKIDS is designed to determine whether children exhibit the common characteristics of children entering kindergarten. The WaKIDS evaluation is an observational assessment that happens at the start of the school year. When children enter kindergarten, their teachers conduct the assessment to determine if each child is kindergarten-ready. WaKIDS captures a range of observational evaluations that rate students in six developmental areas: Social-Emotional, Physical, Cognitive, Language, Literacy, and Mathematics. Kindergarten teachers are expected to complete these observations by October 31st of each school year.

At the beginning of each year, ERDC receives a data file from OSPI with kindergarteners' WaKIDS scores, collected by teachers during the fall of the prior school year. ERDC then loads the data into the P20W system and matches students in that data file to CEDARS data. The WaKIDS data contains records for all kindergarteners who have completed the assessment since the 2011-2012 year. Initially, assessments were limited to students who attended high-poverty schools that were required to administer WaKIDS as part of their funding for full-day kindergarten. Over time, the program has expanded so that all incoming kindergartners are now assessed with WaKIDS, unless parents choose to opt their child out. WaKIDS data includes one record per kindergartner per year.

ERDC's WaKIDS Data Dashboards and Reports

Early Childhood Program Participation & WaKIDS Outcomes

Early Learning Feedback Report

Kindergarten readiness among children who participated in the Washington State Early Childhood Education and Assistance Program (ECEAP)

ECEAP Participation and Kindergarten Readiness among Hispanic Children in Washington State

The six domains of the WaKIDS data assessment align with the tool that kindergarten teachers use to evaluate students in the six developmental areas. Within each domain, the assessment data captures the developmental level for each child and if they were considered "kindergarten-ready." Once all observations are made in a domain, the domain score can be determined. The development level of each child is the most fundamental measure in the WaKIDS data. Kindergarten readiness is determined based on whether the child's level of development reaches that of a four-year-old or older. Each level of development is classified by color, as outlined in the table below. Children who reach the blue or purple development level are typically considered kindergarten-ready, though not all students at the blue development level are kindergarten-ready. Readiness flags are populated for all years. Date finalized for each domain is available from 2017-2018 onward. Completion flags were populated for 2011-2012 and 2012-2013 only and are not loaded to the P20W data warehouse. Child birthdates are fully populated from 2012-2013 onward.

Table 2. Variable Availability by Year

Color	Level	2012	2013	2014	2015	2016	2017	2018	2019	2020
Brown	3rd grade								x	x
Silver	2nd grade								x	x
Pink	1st grade								x	x
Above Purple	Not Applicable	x	x							
Purple	Age 5	x	x	x	x	x	x	x	x	x
Blue	Age 4	x	x	x	x	x	x	x	x	x
Green	Age 3	x	x	x	x	x	x	x	x	x
Yellow	Age 2	x	x	x	x	x	x	x	x	x
Orange	Below Age 2	x	x	x	x	x	x	x	x	x
Red	Not Applicable	x	x	x						

Given ERDC's review and processing of the WaKIDS data, the fields that are used in reporting and in loading to the P20W data warehouse can be fully utilized. The readiness flags and developmental levels are considered to have the most policy-relevant value. This data is valuable to researchers who want to examine how pre-kindergarten programs influence children's kindergarten readiness. Existing research has primarily focused on the ECEAP program's effects on children's educational outcomes. However, the WaKIDS data has also been used to inform other early childhood education programs such as KinderCare or Head Start.

By merging WaKIDS data with OSPI enrollment records, researchers can analyze annual data at the teacher and school level. Analyzing outcomes at the school district level is also possible with this data. WaKIDS source data includes information about the child, the school, and the teacher. However, ERDC considers CEDARS the source of record for K-12 information. In conjunction with other data sources like CEDARS, WaKIDS data can address research questions about kindergarten readiness and other educational outcomes, like those below:

- How might income level, special education status, or type of pre-kindergarten program impact kindergarten readiness?
- How does kindergarten readiness in turn affect performance in later grades?

One challenge to using WaKIDS data is its continued evolution since the program began. WaKIDS started as an assessment tied to full-day kindergarten. Full-day kindergarten began as a program within low-income schools. As a result, the initial years of WaKIDS reflected kindergarten readiness in the poorest school districts, based on the percentage of students who received free or reduced-price lunch. Researchers should account for this distinction when looking at the data over time. Cut-point scores for determining kindergarten readiness also changed prior to the 2016-2017 school year, which may explain some of the visible changes in trend lines at that time. Another issue to note is that observational assessments may be impacted by bias, along with the extent of training and re-training of educators who conduct the assessments.

There are no notable known patterns or reasons for missing values, other than those resulting from structural changes in the program. While it is rare, duplicate student records may exist in the data files, where it appears like the same student took the assessment twice. ERDC considers the most recent assessment to be valid, so duplicate entries are likely an issue with early files when dates of assessments were not available. Other known quality issues are corrected during the data loading process. Students may have more than one record if they changed schools between the start of the school year and October 31st. However, if children transferred schools, then the data collected from the first placement should follow the child to their new placement.

Student Assessment Data

OSPI provides ERDC with an annual file of results of standardized state testing in the K-12 public schools. This data is loaded into the P20W data warehouse and linked to other OSPI data through identity resolution. Statewide assessments measure the progress of students in third grade through 11th grade and within the educational system as a whole. Results are one measure of accountability in the Washington School Improvement Framework—each school is measured on the framework

and OSPI uses results to identify schools for additional support. Assessments are also required for federal accountability purposes.

Included the assessment data are annual student level assessment results for all standardized, statewide tests in Reading, Writing, English Language Arts (ELA), Math and Science, conducted from the 2006 to 2019 school years. In addition to public school students, the file also includes any home school or private school students who took assessments. This file does not include the Washington Kindergarten Inventory of Developing Skills (WaKIDS) or English Language Proficiency Assessments.

The types of assessments and grade levels in which they were administered changed multiple times during the time span included in the data. Between 2006 and 2009, the Washington Assessment of Student Learning (WASL) was used, followed by five years of the High School Proficiency Exam (HSPE) and Measurements of Student Progress (MSP). For Math and Science testing between 2011 and 2014, End of Course (EOC) assessments in Algebra 1, Geometry, Integrated Math and Biology were conducted for high school students. Smarter Balanced assessments (SBA) began in 2015 (optional in 2014) for Math and English Language Arts (ELA) while EOC continued for Science (Biology) through 2017. The Washington Comprehensive Assessment of Science started in 2018. Various alternative tests, for students with significant cognitive challenges, were also administered during this timeframe. See Tables 3 and 4 for more information.

Table 3. Statewide Assessments by Grade Level and School Year

Grade Level	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
3	WASL	WASL	WASL	WASL	MSP	MSP	MSP	MSP	MSP	SBA	SBA	SBA	SBA
4	WASL	WASL	WASL	WASL	MSP	MSP	MSP	MSP	MSP	SBA	SBA	SBA	SBA
5	WASL	WASL	WASL	WASL	MSP	MSP	MSP	MSP	MSP	SBA	SBA	SBA	SBA
6	WASL	WASL	WASL	WASL	MSP	MSP	MSP	MSP	MSP	SBA	SBA	SBA	SBA
7	WASL	WASL	WASL	WASL	MSP	MSP	MSP	MSP	MSP	SBA	SBA	SBA	SBA
8	WASL	WASL	WASL	WASL	MSP	MSP	MSP	MSP	MSP	SBA	SBA	SBA	SBA
10	WASL	WASL	WASL	WASL	HSPE	HSPE/ EOC-Math	HSPE/ EOC-Math/Sci	HSPE/ EOC-Math/Sci	HSPE/ EOC-Math/Sci	EOC-Sci	EOC-Sci	EOC-Sci	SBA
11										SBA	SBA	SBA	WCAS

WASL Washington Assessment of Student Learning
MSP Measurements of Student Progress
HSPE High School Proficiency Exam
EOC End of Course *(Math: Year 1-Algebra 1 or Integrated 1, Year 2-Geometry or Integrated 2; Science: Biology)*
SBA Smarter Balanced Assessment
WCAS Washington Comprehensive Assessment of Science

Table 4. Statewide Alternative Assessments by School Year

2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
DAWL	DAWL	DAWL	DAWL	DAPE	DAPE	DAPE	DAPE	DAPE	DAPE	AIM	AIM	AIM
PORT	PORT	PORT	PORT	PORT	PORT	PORT	PORT	PORT	PORT	SBA - Basic	SBA - Basic	SBA - Basic
	WAMO	WABA	WABA	HSPB	HSPB	HSPB	HSPB	HSPB	AIM			
									SBA - Basic			

DAWL Developmentally Appropriate WASL
PORT Washington Alternate Assessment System Portfolio
WAMO WASL-Modified
WABA WASL-Basic
DAPE Developmentally Appropriate Proficiency Exam
HSPB High School Proficiency Exam – Basic
AIM Access to Instruction and Measurement
SBA Smarter Balanced Assessment - Basic

As outlined in Tables 5 and 6, high school ELA and Math assessments are tied to graduation requirements, which also changed several times from 2013 forward. See the OSPI Graduation Pathways webpage for more information on recent legislative changes to graduation requirements. High school graduation requirement details for the graduating classes of 2015 through 2020 can be found on the Washington State Board of Education (SBE) website.

Table 5. English Language Arts Assessments to Fulfill Graduation Requirements by Class

2008 - 2013	2014 and 2015	2016	2017 Forward
WASL/HSPE Reading/Writing or alternative	HSPE Reading/Writing OR SBA-ELA or alternative	HSPE Reading/Writing OR SBA-ELA or alternative	Smarter Balanced-ELA

Table 6. Math Assessments to Fulfill Graduation Requirements by Class

2008 - 2012	2013 and 2014	2015 - 2018	2019 Forward
No requirement	Math EOC Yr 1 or Yr 2	Math EOC Yr 1 or Yr 2 OR Smarter Balanced Math	Smarter Balanced Math

Data elements include:

- ***Student data*** - Student ID, student name fields, date of birth, school, district and grade level when tested as well as flags indicating if the student attends private school or is home-based;
- ***Assessment information*** - test type, test grade level, test administration time period (e.g. Spring Test Administration), date the student took the assessment and subject name;
- ***Assessment results*** - scale score, performance level, standard met indicator for whether minimum standards were met and attempt code (whether the student did or did not take the test and the reason why).

The unit of analysis is the assessment taken by each student. Assessments are uniquely identified within a year by test administration time period, test type, subject name, test grade, reporting grade, attempt code and score. Students typically have multiple assessment records in a year, usually one for each subject. In a small number of cases, students take the same subject tests multiple times in a year. In such instances, OSPI will take the results from the first time the student tested, using the date the student tested.

ERDC receives one data file from OSPI annually each Spring, covering the previous school year. After a quality review, all records are loaded into the P20W data warehouse, regardless of whether ERDC has an enrollment record for the student. Student identifiers in the file, including name, date of birth and state student ID, are matched to K-12 enrollment data using ERDC's identity resolution process to link enrolled students with their assessment results.

Data quality is generally good, since assessment results are required for federal reporting and are used, along with other metrics, by the state to determine school eligibility for additional support. District and school coverage is comprehensive each year. Our analysis shows that when CEDARS

enrollment data is joined to this file, the proportion of all enrolled students who have assessment records remains stable (around 60%), during the school years 2010 through 2018. This level of stability seems reasonable, considering that not all grades are tested each year.

While student and test information fields are mostly complete for each record and school year, data is completely missing for the Test Grade data element (i.e., grade level of the assessment) for years 2006 and 2007, and the Home Based and Private School data element for years 2006 through 2008.

Assessment results like scores, performance level and standard met indicator, are typically missing for 10 to 20 percent of records. This level of missing data is expected, since they are cases where students did not complete the assessment during that school year due to refusal, being exempt or absent, not being enrolled, or they passed the assessment previously.

The large missing count in Assessment results fields in 2014 is to be expected, because this was the first year of transition to the Smarter Balanced (SBA) test when schools were allowed to choose whether to participate in a pilot of the new Smarter Balanced Assessment (SBA) or take the existing MSP tests. The SBA pilot was only offered in third grade through eighth grade. This is evident in the data, with 91 percent of grade 3 – 8 SBA assessments missing results data, and Measurements of Student Progress (MSP) assessments for those grades 98 percent complete. Researchers are advised to not use data for SBA assessments taken in the 2014 pilot year.

High school grade assessment results show significant missing data for the 2008, 2009, and 2013 school years, with 38, 42, and 58 percent, respectively, of all records in reporting grades 9-12 having no data for the met standards indicator. This is not a concern, especially for 2011 onward. For high school assessments, missing results data are caused primarily by students taking and passing the assessment (i.e., meeting standards) when they were in an enrollment grade level prior to or later than the grade level of the test. In such cases, the results (Scores, Performance Level, and Standard Met Indicator) are usually included in the record for the school year in which the assessment was taken.

An additional record is included for the student in the school year when the assessment would be required, with no results and an indicator in the Attempt field indicating that the student 'previously passed.' This occurred often with Math and Science EOC assessments that were administered between 2011 and 2017. For example, if a student completed a grade 10 Algebra assessment when they were in 8th grade in 2011 and passed, then their scores and other results data would be in a 2011 record with a Test Grade value of 10 and Reporting Grade value of 8. The student would also have a record in the 2013 school year, for the 10th grade math assessment, but with no results data and a code indicating that they had previously passed. Researchers should therefore look across all high school years and middle school years to find assessment score and performance level results for the 10th-11th grade tests for each student, rather than focus on one reporting or test grade level. To determine if students met standards on an assessment, it is important to include cases where the Met Standard Indicator is 'yes,' or the Attempt code indicates that they previously passed.

Of the high school records reported in 2013 with no results data, 90 percent have an attempt code indicating 'previously passed' and, of all test records for the year, 94 percent either include results or have a 'previously passed' attempt code. This latter percentage is lower, at 78 and 75 percent, for 2009 and 2008, indicating that for those years, assessment completeness during high school grades may be a concern.

The entire data set is fully usable for research, with some caveats. Researchers typically want to know if students met standards in a subject area for a particular grade level of the test. Data for years 2006 and 2007 should be used with caution due to entirely missing Test Grade data. Reporting Grade (i.e., student's grade level when they took the test) is well-populated for those years, and OSPI suggests using Reporting Grade as a substitute when Test Grade is missing. However, students may test when they are in either an earlier or later grade than the grade level of the test. This scenario is most common in high school, especially for alternative test types (i.e., students take a test that is at a test grade level lower than their enrollment grade) and End of Course Math and Science tests (i.e., students take a high school-level test when they are enrolled in middle school).

Another limitation is that indicators for Home-school and private school students are missing for years 2006 through 2008. This information is necessary if researchers are using the data as a standalone dataset and want to compare public school students and private or home-school students. Raw scores (Score) are also not useful for comparisons between students who take different test types within a subject and longitudinally. This data element is not standardized across test types, subject or years, and it is no longer provided to ERDC as of the 2016 school year. Scale scores are not much better, because they are not comparable across test types and subjects and, within subjects, they are not comparable across years due to changing standards and scale score values. Therefore, scale scores should be used only for comparisons made within a single school year, test type (e.g. WASL or Smarter Balanced) and subject area. Scale scores are not recommended to directly measure student growth over time. Performance levels, however, can be used to compare growth over time.

Assessment data is typically joined to an enrollment cohort and not used as a standalone file. Assessment results measure student learning and academic performance and, as such, are considered outcomes of K-12 or early childhood education. Researchers may, for example, use 3rd grade assessments as an outcome measure of the impact of early childhood education on a cohort of students who attended early childhood programs. Other research topics examine assessment performance as a predictor of postsecondary experience, relating it to postsecondary course-taking pathways, academic performance or remedial course taking. Assessment results are also often used to make comparisons between student groups based on demographics and/or program participation. Aggregating results to the school or district level can enable researchers to compare groups of schools or districts. OSPI uses assessment data as a key indicator of school and district success, publishing results data in their Report Card.

Comprehensive Education Data and Research System (CEDARS) Data

The Comprehensive Education Data and Research System (CEDARS) contains data reported to the Office of the Superintendent of Public Instruction (OSPI) by each school district in Washington. This information covers the majority of administrative data intake for all school districts, covering topics ranging from enrollment to discipline. Every student who enrolls in any Washington public K-12 school is represented across the files contained within CEDARS and can be tracked over time through their Washington public education career.

ERDC Reports that use CEDARS data
Education Outcomes of Children and Youth Experiencing Homelessness.
The Impact of Transfer in Baccalaureate Completion
The Education and Workforce Outcomes of Youth Who Received a Decline of Jurisdiction

Data Loading

School districts have their own student information systems or vendor-operated systems that are managed at the district level. Districts transfer their local data into the CEDARS system periodically throughout the year, in addition to submitting prior school year data updates. OSPI extracts data from CEDARS and provides it to the ERDC for loading into the data warehouse twice per year. ERDC receives the final data file for the prior school year in the fall and receives the preliminary data file for the current school year in the winter. After receiving a file, the ERDC conducts a data integrity check and consults with OSPI data stewards on inconsistencies or discrepancies found. This work, from CEDARS extraction to final loading, results in a delay of 7-9 months from the end of the school year before final school year data is available for research.

CEDARS Enrollment Data

The K12 enrollment table provided to ERDC is extracted by OSPI from a number of CEDARS Files: Location, School Student, District Student, Student Attributes and Programs, Race and Ethnicity. The core of the file includes student enrollment information by school and student characteristics, as outlined in Table 7. It covers the PreK-12 grades from the 2010 school year forward. Student names and other identifying information also comes to ERDC in this table, and it is loaded into the identity resolution process.

This is a student-level data set that includes every student enrollment segment in Washington K-12 public schools for each school year. The enrollments are the basis for loading other CEDARS-sourced data into the P20W system, in that only the records with a corresponding school enrollment are loaded from these tables (e.g., absence, program, discipline). Data can be analyzed at the student, school, or district level for most of the data. Reporting restrictions may apply for smaller groupings or cell sizes to protect student privacy.

Table 7. Availability of CEDARS Enrollment Data Elements by School Year

	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
School Year	x	x	x	x	x	x	x	x	x	x
District	x	x	x	x	x	x	x	x	x	x
School	x	x	x	x	x	x	x	x	x	x
Is this the School that is Primarily Responsible for the Student?	x	x	x	x	x	x	x	x	x	x
Gender	x	x	x	x	x	x	x	x	x	x
Grade Level	x	x	x	x	x	x	x	x	x	x
Date Enrolled in District	x	x	x	x	x	x	x	x	x	x
Date Exited from District	x	x	x	x	x	x	x	x	x	x
Date Student Enrolled in School	x	x	x	x	x	x	x	x	x	x
Date Student Exited from School	x	x	x	x	x	x	x	x	x	x
School Withdrawal Code	x	x	x	x	x	x	x	x	x	x
School Choice Code	x	x	x	x	x	x	x	x	x	x
Federal Race/Ethnicity Rollup (calculated)	x	x	x	x	x	x	x	x	x	x
Student Primary Language Code	x	x	x	x	x	x	x	x	x	x
Student Language Spoken at Home	x	x	x	x	x	x	x	x	x	x
Graduation Requirements Year	x	x	x	x	x	x	x	x	x	x
Student Expected Year of Graduation	x	x	x	x	x	x	x	x	x	x
Cumulative Grade Point Average	x	x	x	x	x	x	x	x	x	x
Credits Attempted	x	x	x	x	x	x	x	x	x	x
Credits Earned	x	x	x	x	x	x	x	x	x	x
Initial USA School Enrollment	x	x	x	x	x	x	x	x	x	x
Number of Months of Non-US Attendance in School	x	x	x	x	x	x	x	x	x	x
Cumulative Days Present this Enrollment Period	x	x	x	x	x	x	x	x	x	x
Disability Code	x	x	x	x	x	x	x	x	x	x
Disability Flag (calculated)	x	x	x	x	x	x	x	x	x	x
Disability Description	x	x	x	x	x	x	x	x	x	x
Is Student and Approved Private-School Student Attending Class Part Time?	x	x	x	x	x	x	x	x	x	x
Is Student a Home-Schooled Student Attending Class Part time?	x	x	x	x	x	x	x	x	x	x
Is Student from a Foreign Country with an F-1 Visa? (Student Exchange Status)	x	x	x	x	x	x	x	x	x	x
Is Student Homeless?	x	x	x	x	x	x	x	x	x	x
Military Parent or Guardian								x	x	x
Confirmed Transfer In										x

While there are missing cases in some of the columns, a significant number of the columns show no missing cases. For example, there are no missing records in the columns that cover race/ethnicity, date enrolled, grade level, and days present. Other columns have missing data as expected, based on OSPI business rules or data collection changes. Not all students have a disability, so most records do not have data in that column. In addition, withdrawal code only applies if a student leaves a school. As state above, certain columns only apply to high school enrollments. Missing cases for military

parent/guardian status are expected, because the collection of this data did not start until the 2017 school year.

An exception to expected missing cases is with language spoken at home, which has missing cases ranging from 6 to 14 percent for the 2010, 2011 and 2012 school years, but no missing cases after 2012. Other columns with small counts of unexpected missing cases, in 2010 only, are gender, primary school, primary language, and homelessness. This gap as well as the missing language spoken at home cases are likely because 2010 was the first year of the CEDARS system and reporting in the early years was generally inconsistent and incomplete. While there are no significant limitations beyond these missing values, it is important to note that this data often must be merged with other data to conduct research. Researchers should be aware of the missing data noted above for some columns as they consider their study design.

CEDARS Program Data

The CEDARS program data provided to ERDC contains administrative records from school districts on students who participated in or received services from specific PreK-12 programs, eligibility for free/reduced meals and selected student attributes, from the 2010 school year forward. This data is collected in the CEDARS Student Programs and Attributes file. There are 49 unique programs included in this data, 31 of which that are active as of the 2018 school year (the most recent year analyzed). There have been changes over time. Some programs have begun and/or ended, while some programs have been moved to or from other collections. Therefore, not all OSPI programs are captured in this table in all years. For example, Title III language instruction for English learners and immigrant students was moved to the English Learners collection in 2012. Further, Special Education and Career and Technical Education programs are not collected in this table. See Table 8 below for details.

All variables in this table are reported in all school years, with the Program ID variable being the key component. Columns contain information at the student level on:

- name of the program that the student participated in
- school and district in which they participated in the program
- date that the student began receiving services in the program
- the reason why the student qualified for the program
- date that the student exited the program and the reason for exit

Table 8. Program Availability by School Year

	2010	2011	2012	2013	2014	2015	2016	2017	2018
21st Century Community Learning	x	x	x	x	x	x	x	x	x
College Bound Scholarship Application			x	x	x	x	x	x	
Disability status						x	x	x	x
Early Education	x	x	x						
Free Reduced-Price Meals	x	x	x	x	x	x	x	x	x
Gifted acceleration					x	x	x	x	x

	2010	2011	2012	2013	2014	2015	2016	2017	2018
Gifted funded by combination of state district and/or local funds				x					
Gifted general education classroom					x	x	x	x	x
Gifted outside the traditional school setting					x	x	x	x	x
Gifted services or program funded by district or other local highly capable fund				x					
Gifted state Highly Capable Program funds	x	x	x	x					
Gifted unique highly capable program					x	x	x	x	x
Graduation Reality Dual Role Skills (GRADS)					x	x	x	x	x
Learning Assistance Program Behavior							x	x	x
Learning Assistance Program English Language Arts							x	x	x
Learning Assistance Program Graduation Assistance		x	x	x	x	x	x	x	x
Learning Assistance Program Language Arts	x	x	x	x	x	x			
Learning Assistance Program Math	x	x	x	x	x	x	x	x	x
Learning Assistance Program Readiness		x	x	x	x	x	x	x	x
Learning Assistance Program Reading	x	x	x	x	x	x			
Migrant Education Program	x	x	x	x	x	x	x	x	x
NCLB Supplemental Services	x	x	x	x					
Plan 504	x	x	x	x	x	x	x	x	x
Readiness To Learn (RTL)							x	x	x
Reading Corps								x	x
Recruiting Washington Teachers									x
Reengagement Program							x	x	x
Title I Neglected Delinquent Supplemental Services		x	x	x					
Title I Part A Services Local Neglected Students					x	x	x	x	x
Title I Schoolwide Additional Assistance Language Arts			x	x	x	x	x	x	
Title I Schoolwide Additional Assistance Math		x	x	x	x	x	x	x	
Title I Schoolwide Additional Assistance Reading		x	x	x	x	x	x	x	
Title I Schoolwide Additional Assistance Science			x	x	x	x	x	x	
Title I Targeted Assistance Career Technical Education									x
Title I Targeted Assistance Language Arts	x	x	x	x	x	x	x	x	x
Title I Targeted Assistance Math	x	x	x	x	x	x	x	x	x
Title I Targeted Assistance Other									x
Title I Targeted Assistance Reading	x	x	x	x	x	x	x	x	
Title I Targeted Assistance Science	x	x	x	x	x	x	x	x	x
Title I Targeted Assistance Social Sciences									x
Title I Targeted Assistance Social Studies	x								
Title III Immigrant	x	x							
Title III Native American English Language Development	x	x							
Title VII Indian Education Supplemental Services		x	x	x	x	x	x	x	x
Truancy Action								x	x

	2010	2011	2012	2013	2014	2015	2016	2017	2018
Unaccompanied Youth			x	x	x	x	x	x	x
Washington State Seal of Biliteracy Earned							x	x	x
Washington State Seal of Biliteracy Proficient								x	x

Every record has a Program ID, district code, school year and start date. All records have a school code, except for the Migrant Education Program. The Migrant Education Program lacks a school code because it is a district-level program. Some school codes have very few records because they are either for schools with low enrollment, or they are codes for non-school entities like community centers, group homes, skills centers, and homeschool centers. The Program Table contains records from 87% of all schools.

Since funding streams are often tied to program participation, this data is regularly used for state and federal reporting, which makes it subject to audit and review. Except for the Migrant Education Program, the programs most likely to be used in research have consistent data, with few missing values. The Migrant Education Program has consistent data starting with the 2015 school year, when a new data collection system was implemented. Low record counts for migrant education in 2013 and 2014 school years are a result of the change in data system, rather than a reflection of actual program enrollment during those years.

Data in this table are also relatively consistent regarding exit dates. A record contains an exit date if the student exited the program. If in any given school year, a student no longer receives services associated with the gifted program but continues to receive 21st Century Community Learning Program services, then an exit date would appear in the gifted program record and no exit date would be included in the 21st Century Community Learning Program record. Nearly one-third of records (32%) have an exit date, and of those records with exit dates, 82% have an exit code. Records that have an exit code but no exit date are rare. A small number of records that have exit dates in the future are not actual exits. The percentage of future school year exit dates are low during the 2010-2013 school years, while around 3% of the records with exit dates during the 2014-2018 school years. Future school year exit dates are also used differently among districts, due to differences in district practices for rolling over to a new CEDARS year. For example, in the 2018 school year, 17% of districts have at least 10% of their exit dates occurring in future school years. Exit dates are also used differently across programs. Of the most frequently researched school programs, the Free and Reduced-Price Lunch (FRPL) program and Section 504 program have high percentages of exit dates occurring in future school years, especially during 2015 to 2018. The Migrant Education Program has the highest occurrence of future exit dates, with 100% of exit dates in a future school year during 2013-2018.

The Qualification code is only required for certain programs: the Seal of Biliteracy, Reengagement, Disability, and GRADS. While the qualification codes are collected for FRPL, they are not provided to ERDC. For the Seal of Biliteracy, the qualification code is the language code, and students can earn the seal in multiple languages. The Reengagement program qualification code describes where the student is receiving services. Both the Seal of Biliteracy and GRADS qualification codes contain no

missing values, and the reengagement program has a very low qualification code missing rate. For disability status, the qualification code is the disability code describing the type of disability.

This dataset could be used to identify student attributes or to study program participation and effectiveness. The most commonly used data elements from the program table found in ERDC reports are FRPL eligibility as an indicator for economic disadvantage (Chen et al., 2019; Chen, 2019), a flag for students who have a 504 plan (Hough, 2019), and a flag for students in the migrant education program (Pyle & Chen, 2019). There is a report that uses the Section 504 and low-income variables in addition to disability status (Gertseva & McCurley, 2019), and a report dedicated to re-engagement program participation (Hough, 2019).

Beyond the ERDC studies conducted with this data, researchers may create specific "flags" in the data to examine the effectiveness of service delivery methods and/or explore differences in outcomes among students. For example, students can be flagged as gifted if they participate in the gifted program during a selected period of study, and this flag can be used to compare student outcomes by student group. Changes in coding categories over time make it harder to identify differences in program outcomes, but the current categories would enable researchers to study the effectiveness of different district choices in program delivery. Unaccompanied youth could be used as a category as well, to study how Washington is serving this population and to understand the characteristics of this student group and how they change over time.

Researchers must be cautious with using Section 504, disability status, and/or FRPL eligibility information. FRPL eligibility extends from the previous school year to the first 30 serving days of the new school year, or until eligibility is determined. For improving accuracy of this measure, researchers may consider excluding students whose eligibility carries over from the previous school year but discontinues after the first 30 days. Students who are eligible for free meals and students who are eligible for reduced priced meals cannot be delineated using this data.

When utilizing disability status and Section 504 information, it is important to recognize that very few students with a 504 plan have a disability, and few students with disabilities have a 504 plan. Eligibility for protection under Section 504 is a physical or mental impairment that limits at least one major life activity. Impairments can be permanent or temporary, so it is possible for a student to have a 504 plan for only a few months, only a few school years, or for their entire K-12 career.

CEDARS Special Education Data

The CEDARS Special Education data table provided to ERDC has information on students who receive special education services from Washington K-12 public school districts. This data table, collected from school districts in the CEDARS Student Special Education Programs File, covers information on the level of support and services students receive, how long students receive services during their K-12 career, and why they leave the special education program from the 2010 school year forward.

There is one record for each enrollment of a student into the special education program at a location for each school year. Typically, there are not multiple entries for a student within both a location and a school year. The key components of this dataset are Least Restrictive Environment

(LRE) where a student receives their education, program exit reason, start date, and exit date. LRE offers information on the level of support and services being provided to students. Program exit reason answers whether a student leaves the program because they no longer have a need for special education services. The start date and exit date indicate how long a student receives services. These are useful pieces of information for understanding student progress and program effectiveness related to special education. A disability code is not included in the special education table but can be found in the CEDARS program table. Note that not all students with a disability will have records in the Special Education table.

Every record has a school year, an LRE code and description, a start date, and a district code. Since records only contain an exit date when a student leaves the special education program, only 13% of records have exit dates. Consistent with the business rules in the CEDARS manual, all exit dates occur after start dates. A small number of records have an exit code with no exit date, or they have an exit date with no exit code. These records appear in only a few districts during the 2010 to 2013 school years. In the years that last evaluation date is collected, an annual average of 82% of records have a date listed. Due to some variables being optional and situational, missing values are more frequent for last Individualized Education Program (IEP) review date, initial eligibility date, and initial referral date, with respectively 43%, 23%, and 16% of records containing a date during the years that these variables are collected. Not all variables in this table are collected during all years. See Table 4 below for specific years when variables are available.

The LRE codes in CEDARS changed over time, with some codes discontinued and some new codes introduced. They can be grouped by age ranges 0-2, 3-5, and 6-21 years old. Exit reason codes have also gone through similar changes. Coinciding with the start and exit date definition changes, "graduated" and "transferred" exit codes ended in 2014, and the "no longer enrolled in district" code was added in 2015.

Start date and exit date had definition changes in the 2014 school year. From 2010-2013, start date is defined as the first day (or the earliest known day, if the first day is unknown) that the student attends or receives services from a special education program anywhere in the state. From 2014 onward, the start date field is defined as the date that the student began receiving services each school year in the Special Education program in the reporting district, or the date the student had a change in the LRE. Additionally, initial WA service date was collected alongside the new start date definition in 2014. This variable reflects the old start date definition during the 2014 transition year. Like the start date changes from 2010-2013, exit date is defined as the last day a student attends or receives services from the Special Education Program in Washington. From 2014 onward, the definition for exit date is the last day a student attends or receives services from the Special Education Program at the reporting district or the date that the student had a change to their LRE. Despite the definition changes to include LRE start and exit, there are typically not multiple entries per student per year at a location.

Table 9. Variable Availability by School Year

	2010	2011	2012	2013	2014	2015	2016	2017	2018
School Year	x	x	x	x	x	x	x	x	x
District Code	x	x	x	x	x	x	x	x	x
Start Date	x	x	x	x	x	x	x	x	x
Exit Date	x	x	x	x	x	x	x	x	x
Exit Reason Code	x	x	x	x	x	x	x	x	x
Exit Reason	x	x	x	x	x	x	x	x	x
LRE Code	x	x	x	x	x	x	x	x	x
LRE	x	x	x	x	x	x	x	x	x
Initial Referral Date	x	x	x	x	x	x	x	x	
Initial Eligibility Date	x	x	x	x	x	x	x	x	
Initial WA Service Date					x				
Last IEP Review Date	x	x	x	x					
Last Evaluation Date	x	x	x	x					

Since funding is tied to special education program participation, this data is regularly used for state and federal reporting, which makes it subject to audit and review. For this reason, the overall quality of the CEDARS special education table is considered reliable. The data components of interest for use in research are highly complete and allow for longitudinal analysis. Data are consistent with business rules, in that there are no missing values for any of the required fields. Most districts have consistent enrollment trends from year to year. Note that not all entities with a disability code participate in special education. In each of the date variables, there is a small percentage of unrealistic dates that are likely to be errors. Due to the definition change for start and exit dates in 2014, dosage can only be studied from 2014 onward.

Typically, this dataset is used to identify students in special education. Reports by the ERDC use a special education flag to indicate if students receive services in special education during the period of study and to include students in special education in breakdowns comparing outcomes by student group (Chen et al., 2019; Chen, 2019). This dataset could also be used to study LRE trends to determine if they tend to become more or less restrictive over time, possibly even merging it with other data to examine differences by race, gender, economic disadvantage and disability. Another potential use for this data is to explore gender, race, and economic differences in the grade level that students enter special education.

The limitations of this dataset are minor and can be addressed with simple strategies. To measure the length of time that students receive special education services, the definition change for start dates may present a challenge. For analysis purposes, researchers should expect to treat pre-2014 data differently than data from 2014 onward. If researchers are interested in studying school-to-school differences related to special education, then the data's district-level nature is a limitation. However, linking this table with enrollment data would address that issue.

CEDARS English Language Learners Data

The CEDARS English Language Learners (ELL) data provided to ERDC includes information on students who receive services from Washington K-12 public school districts in the State Transitional Bilingual Instructional Program (STBIP) and, starting in 2017, Native American Title III English Language Development Services. The data is collected in the CEDARS English Learners File. The ELL table contains administrative records from school districts on English language learners receiving PreK-12 services from 2010 to present, representing 73% of schools.

STBIP is a federal Title III program. As OSPI describes on their Bilingual Education Program webpage, "both programs share the same goal: develop language proficiency that enables meaningful access to grade level curricula and instruction." To determine program eligibility, students' English language proficiency is tested. Students who are tested and do not qualify for services are not included in the files provided to ERDC. Information about students' primary language is not covered by this table, but those details can be found in the CEDARS enrollment table.

There is one record for each enrollment of a student into an English Language Learning program at a location for each school year. The key components of this dataset are program designation, program exit reason, start and exit dates, school year, instructional model, placement test, placement status, and grade level at placement. Not all columns of this table are included during all years. Initial US Placement Date, Number of Months of US Attendance, and Number of Months of Non-US Education are only collected in the 2010 and 2011 school years. Additionally, Placement Test fields and Grade Level at Placement are collected from the 2013 school year and on. All other variables are collected during all years. Every record has a start date, district code, school code, school year, and program code.

Overall, the ELL table contains high quality data. At the district level, year-to-year program enrollments are generally stable. Most columns are highly complete, but missing rates by column are higher in the 2010 and 2011 years. Among the date variables, outlier dates are found in 0.03% or fewer records. Some fields should be used with caution due to high missing rates or non-conforming logic. Initial US Placement Date, Number of Months of US Attendance, and Number of Months of Non-US Education have high missing rates for the two years they were collected. All other columns are considered relatively complete. A small percentage of records violate the business logic, with exit dates occurring before start dates, having an exit code with no exit date, or having an exit date with no exit code. This issue occurs most frequently in the early years of CEDARS and improves after the 2012 school year.

Researchers should be aware that, starting in 2018, the meaning of the value for test level status varies based on the placement test taken by a student. Before the 2018 school year, the test level status code corresponded to a single test level status description. For details, refer to appendix N in the CEDARS manual.

This dataset is typically used to explore the characteristics and participation of students who receive English Language Learner services. Reports by the ERDC use such a flag to compare

outcomes by student group (Chen, 2019; Hough, 2019). It could also be used to study the effectiveness of particular instructional models, examine postsecondary outcomes for Native American students who receive Title III English Language Development Services or study the relationship between postsecondary completion and the timing of bilingual program exit.

CEDARS Absence Data

The CEDARS Absence file provided to ERDC is an event history file representing the occurrences of student absences in Washington public K-12 schools. The information, collected from districts in the CEDARS Student Absence File, describes the type of absence (full day or partial day, excused or unexcused) and date of absence for the school years between 2013 and 2019. According to WAC 392-401-015, a student is absent when they are: (a) not physically present on school grounds; and (b) not participating in the following activities at an approved location - (i) instruction; (ii) instruction-related activity; or (iii) any other district or school approved activity that is regulated by an instructional academic accountability system, such as participation in district-sponsored sports.

The Absence table includes one record for each student served in the district during the current school year, for each absence associated with the student for each school the student is enrolled. Even when a student leaves the school related to these absence records, these records continue to be reported through the remainder of the school year. Absences must be reported for students in grades K–12. If attendance is tracked for preschool students, then those absences may also be reported in this file. Available data covers school years from 2013 forward for every school district in Washington. Only students with a recorded absence at a particular school are included in this dataset.

The base analytic unit in the data is at the student level, but additional aggregations at the date, school, and district level are possible. Aggregating data by absence type is strongly discouraged, because the reported values are potentially subjective. Types of absence and absence date are the primary analysis components of the dataset. The type of absence captured in the data includes "excused full-day absence," "excused part-day absence, "unexcused full-day absence," and "unexcused part-day absence."

Although there are no missing cases across the columns, it not clear if any data are missing or unreported. It is possible to have absences for the same student at multiple schools. Also, a student could have both excused and unexcused full- or part-day absences on the same day. ERDC has not determined whether such patterns appear consistently throughout all data collection years or just in earlier years.

CEDARS Race and Ethnicity Data

The CEDARS Race and Ethnicity data tables provided to ERDC are extracted from the CEDARS Race File and Ethnicity File. The data include parent/guardian-reported (self-report) or observation-

reported data on the race and ethnicity of each student in each district, using a two-part question, starting with the 2010 school year.

The Office of Superintendent of Public Instruction (OSPI) implemented this two-part race and ethnicity question as part of the CEDARS data collection process. Schools were first required to submit detailed race and ethnicity information for each student during the 2011 school year, using 49 race codes and ten ethnicity codes, though schools could submit it in the new format during the 2010 school year.[8]

For students that did not self-report, federal guidelines specify that for Washington state,

> (B)y law, a student (or the parent/guardian on behalf of the student) is not required to identify their race and/or ethnicity on school forms. However, if a student (or parent/guardian on behalf of the student) does not complete the two-part question on race and ethnicity, by law, school personnel must use 'observer identification' to select the race and ethnicity of the student. Source.

In 2019, expanded ethnic and racial categories were added to the CEDARS collection, with full implementation required by the 2022 school year. See the OSPI publication, "Race and Ethnicity Student Data Task Force Guidance for the Washington State Public Education System" for further details. The gradual implementation of the expanded categories is evident in the data. For the 2018-19 school year, about 10% of students had reported racial and ethnic information based on the expanded categories. The remaining 90% had not been resurveyed, as their data was carried forward from earlier years.

These data tables include information for all students enrolled in Washington public schools between the 2010-11 and 2018-19 school years. They consist of at least one record for each student served in the district during the school year, along with ethnicity and race data unique to each the student. Each student in a school district is represented by an annual record for each racial and ethnicity category they identified with in their responses. Every record includes District Code, and school year as identifiers and contains Ethnicity or Race Codes and Descriptions, and Collection Method (beginning in 2019). The data is the source for the Federal Rollup Race/Ethnic field included in ERDC's CEDARS Enrollment Data Table. For this field, student race and ethnicity responses are aggregated into a single field that includes high level race codes. If the student identified with any type of Hispanic code, then they are assigned to the 'Hispanic' category.

This data is considered complete and sufficient for open analysis at the state level, as race and ethnicity data is required to be submitted in conjunction with student enrollment data. In any given year, less than ten students lack data in the Federal Rollup Race/Ethnic categories in the Enrollment file. Each of the ethnic and racial groups are large enough to be analyzed at the state level. At the school district level, numerous heterogeneous communities should have data available for many of the racial/ethnic subgroups. This data provides a richness not currently available in the

[8] Prior to 2011, school districts were only required to report race and ethnicity by seven aggregate groups: American Indian/Alaskan Native, Asian, Black/African-American, Hispanic, Native Hawaiian/Pacific Islander, White, or Two or more races.

federal rollup categories, as the multiracial data in this dataset is more detailed than the federal multiracial category.

Numerous reports from ERDC and research partners have used earlier versions of this data to investigate subgroup differences in Racial/Ethnic categories. By creating a dataset with binary flags for each subgroup category ERDC and external researchers can construct various meaningful aggregations relevant to their research goals (see here). For example, a multi-racial student can be used for calculations for all the subgroup categories selected. As long as researchers are careful to construct their numerators and denominators to include the proper subset of students, then the rates will be calculated properly. As these ethnicity and race data are new to the ERDC data warehouse, there aren't any reports that are currently using the current data. Reports prior to this publication utilize race and ethnicity data found in the enrollment file which lacks subgroups.

Ethnicity and race data can provide meaningful context when they are linked with other OSPI files. Once data is linked, however, researchers should exercise caution with small subsets of ethnic/racial groups. Certain data combinations may lead to totals of less than 10 students, which violates the threshold outlined in the Family Educational Rights and Privacy Act (FERPA) guidelines and ERDC's requirements for publication. Data suppression may also occur at the school district level. These data exist starting in the 2011 school year and continuing years.

Despite the gradual implementation of new race and ethnicity questions starting in 2019 and some bias in the data due to over-responses from parents (i.e., parents selecting all possible options for both the ethnicity and racial categories), these enhanced data elements will enrich the research conducted by ERDC and affiliated researchers. This data enables more in-depth, detailed analysis of student outcomes for racial and ethnic subgroups. Parental response bias may be enough to limit analysis of small subgroups at the school district level, but the minimal noise introduced at the state level should not significantly alter any results for groups or subgroups.

CEDARS Discipline and Exclusionary Discipline Data

The discipline data provided to ERDC and collected in the CEDARS Student Discipline File (2013 through 2018) and the Student Exclusionary Discipline File (2019 forward), captures information regarding behavior and discipline actions for students involved in incidents during school or school-related activities. The data collection standards were developed by a special student discipline taskforce, including "elements of education services, petitions for readmission, credit retrieval, and school dropout as a result of disciplinary action", were incorporated into the CEDARS collection (Source).

This data includes information on behaviors and disciplinary actions for students enrolled in all grades of Washington public schools who were involved in incidents during school that resulted in removal from their regular education setting for 2013 forward. The Discipline data are a collection of administrative records reported to OSPI based upon the discretion of each school district to determine when a behavior is an incident for which disciplinary action was taken. For the 2013 through 2018 school years, only the final or most extreme disciplinary action (intervention) is

included in a single record. For all years, the most extreme behavior is entered into the Behavior Code field, and any other behaviors go in the Other Behaviors field.

This reporting changed in 2019, when OSPI changed the reporting process and created the new CEDARS Exclusionary Discipline file. According to the 2019 CEDARS Manual, reporting now contains

> "… a record for each exclusionary action taken for each student involved in an incident during the current school year. If a student has multiple exclusionary actions for a single incident, each exclusionary action must be reported in a separate record. If multiple students are associated with the same incident, then one record must be submitted for each exclusionary action for each student being disciplined. If a student is involved in an incident that results in more than one exclusionary action or if an exclusionary action is modified and/or converted to another exclusionary action, each exclusionary action must be reported as its own record."

Therefore, as of 2019, if more than one intervention is applied per incident, then each one is captured in a separate record. In loading the Discipline data to the data warehouse, ERDC combines the data from both the old and the new file formats into one file structure. Every record includes District Code, School Code, Incident ID, Incident Date, and School Year. Key fields for research refer to Behavior Type, Type of exclusion/intervention, Date of exclusion/intervention and number of days or total time of exclusion/intervention. See the OSPI CEDARS Manuals for a full list of data elements. Incident IDs connect all students associated with a particular incident and could serve as a unit of analysis. Data could also be aggregated by behavior or type of intervention.

ERDC is uncertain if any data are missing. However, different variables have different starting years. See Table 10 for more information. Several new fields were added in 2016. Known data issues are mostly due to the changes that began in the 2019 school year. To correct known issues, OSPI is currently building validations for: 1) Exclusion days are missing, but exclusionary time is more than one day; 2) The total in the Exclusionary Time field is greater than the number of days reported in the Exclusion Days field; and 3) If one or the other is greater than zero, then the other must also be greater than zero.

Table 10. Variable Availability by School Year

	2013	2014	2015	2016	2017	2018
Academic Services				x	x	x
Appeal				x	x	x
Behavior*	x	x	x	x	x	x
Behavior Services				x	x	x
Intervention**	x	x	x	x	x	x
Is in IAES (Interim Alternative Education Settings)		x	x	x	x	x
Other Behavior IDs			x	x	x	x
Petition for readmission				x	x	x

	2013	2014	2015	2016	2017	2018
Reengagement Plan				x	x	x
Weapon	x	x	x	x	x	x

* Revisions to categories 2014-2016

** Emergency Expulsion (EE) added in 2015, Interim Alternative Education Settings (IAES was moved to its own field in 2014. In School Suspension added 2014.

Researchers should be aware of limitations with this data. School years 2013-2018 can be used fully with consistent data across all years, while the 2019 and future school year files have the added feature of capturing all exclusionary actions applied to the student per incident (as described above). Analyses using all available years will require special handling. Between 2013 and 2015, several changes were also made to improve the data collection process and better address reporting requirements. Fields impacted by these changes include intervention, emergency expulsion and behavior fields.

Most of the research centered on discipline data aims to address equity issues in education. Additional data must be joined with these data to examine inequities in student discipline, e.g., race and ethnicity. OSPI released the "Equity in Discipline" report, which explores the concept of disparate discipline, or when the rate of discipline is greater for one group compared to another group. This report also focused on severity of punishments where students in one group display the same behaviors as another group, but punishments for one group may not be as severe as for another group (i.e., exclusion days).

The most common elements displayed and analyzed by OSPI researchers using discipline data address the following two measures:

> "Discipline rate is a measure used to monitor the use of out-of-school exclusionary discipline actions in schools. Discipline Rate is calculated by counting the number of distinct students who have received an out-of-school exclusionary action divided by the number of distinct students enrolled. Exclusion Days Rate is a measure used to monitor the length of time students are excluded for out-of-school exclusionary discipline actions in schools. Exclusion Days Rate is calculated by counting the number of distinct students who have received an out-of-school exclusionary action for a given exclusion length timespan and dividing that count by the number of distinct students who were excluded." (Source)

Currently, only the out-of-school exclusionary data are reported for the OSPI Report Card Data for Schools and Districts. Due to small numbers of exclusions for many schools and districts, only non-redacted data or top/bottom coded data are reported (see Suppression Rules for Public Reporting). Due to these small numbers, some districts may not be able to report data for one school (elementary, middle, or high) if redacted data would reveal identifiable student information. This issue is further exacerbated if data on race/ethnicity, gender, program participation (e.g., Free or Reduced Meal Service, English Language Learner) are linked, which further reduces of the number of cases at the school or district level. Caution is exercised in the dissemination of these data so as

not to violate FERPA requirements. Most of the analyses and group comparisons may only be performed at the state level, or for the largest and most diverse districts.

CEDARS Grade History

The CEDARS Grade History File (CEDARS-GH), as provided to ERDC, includes student course-level records from Washington K-12 public schools. CEDARS-GH was originally developed as a transcript-like collection that would provide information on all courses taken for high school rigor credit, including transfer courses obtained outside the reporting district. The data collection process requires districts to report each student's high school rigor course-taking history each year. With this framework, the data can be used to answer questions about credit accumulation, course-taking trajectories, or courses taken within a single school year.

High school rigor course data has been collected annually since 2010. However, ERDC does not recommend using the data prior to 2013 due to uncertainty in data completeness and quality. Students who enrolled in Washington public schools who ever attempted to take high school rigor courses would be included in the data. High school-level students are the major cohort for this dataset. However, there are few course records indicating students of earlier grade level (i.e. 8th grade) who completed high school rigor courses prior to progressing to high school. All course records are reported each year, so duplicates are often found from source data.

Currently, OSPI uses CEDARS-GH data for both public reporting and internal analysis. OSPI's dual credit and Community and Technical Education (CTE) reporting both rely on this data. The Washington School Improvement Framework (WSIF) also includes a metric called "9th Grade on Track," which uses this data to identify the proportion of first-time 9th graders who passed all credits attempted. Internally, OSPI uses this information to explore course-taking trajectories and examine course information focused on specific content areas like math or art.

Researchers should be aware of several potential pitfalls to using this data. CEDARS-GH provides course records for all students of Washington public K-12 schools. However, inconsistent data completeness across districts and school years has been identified from early years of data collection. Duplicate records are common, which may be due to incomplete reporting during entering and loading procedures from schools to districts, school districts to OSPI, then OSPI to ERDC. Researchers are strongly encouraged to examine missing patterns and conduct record deduplication before beginning any data analysis to address these inconsistencies.

Another potential issue is that even though the variable "state course code" is used to categorize complex coursework structures, the definition of each category is generally at a high level. There is no information about level of course rigor, sequence, or specialty. However, some (but not all) course codes and titles recorded by each school or district provide a proximate guess about course content. There is currently no crosswalk between state course code and actual courses provided by each school and district. The lack of crosswalk makes it challenging to pursue a statewide analysis or even a between-district analysis. In addition, credits attempted and earned are not always reliable, as there are several records showing an abnormally high number of credits earned in a

single school year. It is unclear whether some students actually earned a large amount of credits, or if the elevated amount are the result of data entry or loading errors.

There is also an issue with using course names/codes as proxies for rigor. The variable "course designation code" is used to identify type of courses (e.g. AP, IB, Honors, etc.). Those codes are often automatically set in the district database and may not get updated systematically when changes occur. The lack of systematic updates may result in inaccurate counts and ratios of specialized class types and should be compared to official counts before beginning any analysis.

CEDARS-GH data may also not be the most accurate representation of course-taking for the entire Washington student body. CEDARS-GH only collects high school rigor courses. It does not provide information for studies attempting to track course-taking from earlier grade levels, nor does it include summer school courses. Transfer courses, including Running Start course taking, are available in original OSPI raw data, but ERDC extracts excluded those records until 2018.

CEDARS-GH also has no direct measures of course level or sequence. The CEDARS-GH requires districts to submit the State Course Code data field, based on the five-digit NCES-SCED course codes. This coding scheme provides only information about subject area (i.e. English language and literature, mathematics, etc.) and course identifiers within each subject area. The other three crucial elements are course level, sequence and Carnegie unit, but that information is not included in CEDARS-GH. Without these three elements, it is challenging to analyze statewide course-taking patterns and outcomes across schools, districts, grades, and/or years.

The current CEDARS-GH file was not designed to observe whether policy changes align with the implementation of more rigorous state assessments. ERDC does not recommend using CEDARS-GH data for policy evaluation related to coursework, such as high school graduation requirement and Common Core State Standard. CEDARS-GH is also not ideal to use in longitudinal studies. The current CEDARS student grade history data includes inconsistent data elements and definitions. For example, state course codes are assigned to courses by districts or individual high schools, with no centralized system, so course code assignments are not consistent across the state. Further, districts may have recorded the same courses differently across time (i.e. a math course was coded as art). Comparisons across cohorts over time should be taken cautiously.

Researchers looking to examine associations between high school coursework and student outcomes should use this data with caution. Measures like course level and sequence are not directly collected. Proxy measures may be constructed with careful articulation of the data collection process and the definitions of each course coding logic. It is important to test the validity and reliability of each construct by considering the consistency and completeness across districts and years. This dataset also provides information about coursework for students in dual credit programs (CTE Dual Credit, Advanced Placement, International Baccalaureate, Cambridge International and College in the High School) and flags that identify Career and Technical Education (CTE) completers.

Public Community and Technical Colleges and GED Completion

State Board for Community and Technical Colleges Data Warehouse (SBCTC) Student Data

The State Board for Community and Technical Colleges (SBCTC) Data Warehouse collects includes enrollment and awards data (degrees and certificates) from Washington public community and technical colleges. The Data Warehouse supports research and reporting conducted by the SBCTC and by colleges in the state system. It is the primary source of information supporting policymaking and allocation of funds for the community and technical colleges. The warehouse is built to enable each of the 34 colleges in the system to submit data for each quarter.

ERDC Reports that use SBCTC data
The Impact of Transfer on Baccalaureate Completion
Institutional Impact of Upward Transfer on Baccalaureate Degree Attainment
The Characteristics and Experiences of Students Who Transfer to Four-Year Institutions
Postsecondary Education Assessment in Washington State: Earnings Premium Estimates for Associate Degrees

The SBCTC academic year begins with summer quarter and ends with spring quarter, in line with the Department of Education's definition. Full-time equivalent enrollment is 15 credit hours. The federal financial aid definition is 12 hours for undergraduate students. All colleges operate on a quarterly system.

SBCTC enrollment and course completion information is reported in several files each term. SBCTC extracts data from these files and provides them to ERDC on a quarterly basis for loading in the P20W data warehouse. Students are linked to all other P20W data through the identity resolution process. These files capture information about students, students' coursework, credits, and degree completion, as outlined below:

- The ***Student*** file contains demographic information and some summary items for the reporting term, including credits and FTE enrolled by funding source, total FTE and credits earned to date, GPA, residency information, dual credit standing, veteran's benefits status, and whether the student took any eLearning course in the quarter. A variety of descriptive flags are also in this file to indicate aspects of student intention and planning or to describe their purpose for enrollment.
- The ***Student-class*** file relates each student to the specific courses taken in a term. Designations that pertain to the type of course the student is enrolled in are included here as well. Clock hour conversion information is included in this file and is important for research using data prior to 2008.
- The ***Class*** file provides detailed information about the courses such as Classification of Instructional Programs (CIP) designations, clock hour / credit weights, funding source, day, time, and location information, instructor identification and Full-time Equivalent (FTE), student enrollment FTE, 10-day enrollment counts, clinical activity information and workforce indicators.
- The ***Transcript*** file provides information on credits earned and reported letter grades.

- The ***Completions*** file contains one record for each student for each degree or certificate earned in a term.
- The ***GED Snapshot*** file provides information on GED® awards.

The data extracts that ERDC receives from the SBCTC data warehouse follow the standard quality review process as other data source files, ultimately loaded to the P20W system for research use. The lag time is generally 3-4 months from the end of the quarter to the time data is available for research. ERDC does not load any data elements from SBCTC data warehouse files that can be calculated using other data elements or were determined by SBCTC to be very unreliable.

SBCTC data is available from academic year 2004-05 forward; however, prior to summer quarter 2008, many programs were reported in terms of clock hours rather than credits. The transition to credit hours across the system began in summer of 2008, with programs continuing with clock hours until the current cohorts matriculated from the system. The conversion process took over two years to fully accomplish. When using data from this timeframe, programs with courses that are recorded as clock hours must be converted to credit hour courses. This approach is essential for appropriately calculating GPA and FTE.

When analyzing SBCTC data, users should be concerned with type of student included. Data not only includes students who take courses for credentials, but also those with other objectives like basic skills training or personal enrichment. Data also includes students who are incarcerated. For more information about business rules, please contact ERDC.

Public Four-Year Colleges and Universities

Public Centralized Higher Education Enrollment System (PCHEES) Data

ERDC's PCHEES Data Dashboards
Earnings for Graduates
Time to Degree Visualization
Public Enrollment Four-year Dashboard

The Public Centralized Higher Education Enrollment System (PCHEES) collects data for enrollment and awards (degrees and certificates) from Washington public baccalaureate institutions. PCHEES was originally developed for state budget and accountability purposes and has evolved into a comprehensive database supporting a variety of research and reporting needs. PCHEES data is a critical resource in the production of tools and reports widely used by stakeholders, legislative staffers, and the general public. PCHEES was created under RCW 43.62.050 and RCW 28B.10.784, which require the Office of Financial Management to collect and report higher education enrollment data. PCHEES data is regularly extracted from the PCHEES database and loaded into ERDC's P20W data warehouse, where it is linked to all other data through the identity resolution process.

PCHEES enrollment information is collected in three major files each term: Admissions, Student, and Registration.

- The ***Admissions*** file includes information about the student at the time of initial enrollment, including type of student, term of admission, county of origin, high school

attended/graduated, high school GPA, dual credit affiliation, standardized test scores, previous institution attended, previous degrees, and previous credits earned.
- The ***Registration*** file contains a summary of each student's enrollment in a single term, relating each student to their specific courses taken. This file includes institution, campus, specific course and section information, funding source, credit hours attempted and earned, and letter grade received.
- The ***Student*** file contains identifying information that may change from term to term including name, birthdate, gender, race/ethnicity, class standing, field of study, term GPA, cumulative GPA, fee-paying status, and financial aid indicators.

Five additional files are submitted by the institutions or created by ERDC: Institution, Campus, Program, Term Dates, and Course. These files provide supporting details related to the data elements in the Admissions, Student, and Registration files. PCHEES data is collected at the student level, but analysis at the school or student level is appropriate for these data.

Data are collected on all students enrolled at any of the six public baccalaureate institutions for two different points in time: Enrollment Day 10 and the Final Enrollment Day. Data representing a snapshot of enrollment around the 10th day of each term are submitted by the institutions, depending on their term structure and whether they enroll students during that term (e.g., semester institutions typically do not enroll students for a winter term). A data snapshot that reflects the final enrollment for summer term is submitted in the February after the end of the term, and the remaining final term files are submitted in the following fall.

Day 10 enrollment, sometimes referred to as "census day," is also required for Integrated Postsecondary Education Data System (IPEDS) reporting, the mandatory federal data collection system for all schools approved for Title IV student financial aid funding. Within state government, the Day 10 data are used to calculate budget projections.

Data are available beginning with academic year 2007-08 forward. Day 10 data are available shortly after the files are submitted, depending on assessment of data quality. For final data files, there is typically a one-year lag from the end of the academic year because final files are not due until the fall of the next academic year.

When data files are submitted by an institution, they are uploaded to a restricted area of the ERDC server where automated checks are done, and any error messages are returned to the institution for action. Institutions can view reports about errors in the submitted data, correct the errors, and resubmit the files. After the submitted files pass all automated checks, the data is loaded into the PCHEES database. Once institutions review the loaded data, they release the data to OFM. OFM conducts a manual inspection of the data for quality issues. If any concerns are noted, OFM notifies the institution about the concern. OFM and the institution work together to address the concern and, if needed the institution adjusts and reloads their data. After OFM confirms the data quality to be sufficient for research purposes, users with authorized access to the PCHEES application can download reports and predefined datasets.

An institution can reload their data at any time, even after a submitted file has been released for all users. The active submission is the only version of the data that is released to all users and available

for research. There is no simple way for a researcher to know if a more recent version of the data may become available for use in the near future. If having the very latest data is critical to a project, then prospective users should contact a PCHEES data steward to determine if the data will be updated during the preferred timeframe.

The Day 10 data for 2007-08 through 2009-10 was migrated from an earlier version of PCHEES and does not contain the complete set of data elements and enrollment types in the current system. The Final collection data for the same academic years was submitted through the current system, so it contains all the additional data fields and completion fields. The fully featured set of Day 10 data for research is from the 2010-11 year forward, though ERDC does not have this data for the University of Washington's 2011-12 summer term. The following data elements are only available in the Final collection data for the 2007-08 year forward: credit hours earned, PELL grant and Washington College Grant, term GPA, cumulative institutional GPA, cumulative overall GPA, and the degrees awarded fields in the completion file.

Overall, the PCHEES data is substantially complete for most fields. Each institution establishes its own policies about what to report and what not to report, within reason. As a result, some fields are mostly complete for some institutions, and mostly incomplete for others. Across all six institutions, fields that reflect student K-12 information like K-12 ID, last high school attended, and high school GPA, are mostly incomplete. SAT and ACT score records are also mostly incomplete across all institutions. GPA scores are not available in the Day 10 files. Within the Final files, overall cumulative GPA (which includes transfer credits) is largely incomplete, while institutional cumulative GPA is generally complete. The veteran's military affiliation and benefit type fields are generally missing, along with the field for teacher certification ID. The Evergreen State College is an in-state institution that uses a Pass/Fail system instead of awarding grades, so GPA is not computed for their students.

Some characteristics of PCHEES reflect its original objective: to support the state budget. The PCHEES academic year begins in the summer term and runs through spring. Within PCHEES, undergraduate state-funded full-time-equivalent enrollment is 15 credit hours, which typically leads to degree completion in four years (excluding summer enrollment). The federal financial aid definition is 12 hours for undergraduate students. For graduate state-funded enrollment, full-time-equivalent enrollment is 10 credit hours.

A variety of detailed documentation is available for researchers to better understand the PCHEES fields, what data they represent, and how the fields can be used. A detailed submission guide is maintained by the data stewards; updates are reflected in the release notes document provided in the guide. Additional valid values tables are also available for some fields that are not included in the submission manual.

Financial Aid

Washington Student Achievement Council (WSAC) Unit Record Data

ERDC Reports that use related Unit Record data
Analysis of Alternative Financial Aid Interventions
Outcomes of Need-based Financial Aid: Choice of Major and After-Graduation Earnings
Determinants of Need-based Financial Aid
Impact of Need-based Financial Aid on College Completion: An Event History Analysis
Unmet Need Among Financially Needy College Students in the State of Washington

Financial aid data is provided to ERDC by the Washington Student Achievement Council (WSAC) on behalf of Washington's public colleges and universities. This information reflects the administrative records of students who matriculated at one or more of Washington's forty public postsecondary institutions (34 community and technical colleges and six four-year institutions) during each academic year (Fall through Summer). Institutions collect these administrative records when financial aid is awarded to students. The Unit Record dataset is generally delayed six months from the completion of the collection time period. Once the data is received by ERDC, data integrity checks are conducted, in order to identify the presence of missing information or business rule inconsistencies. After consulting with WSAC on any concerns, ERDC links the data to existing records through the identity resolution process and loads it into the P20W system. Processing by ERDC can take several weeks to a few months, depending on the quality of the submitted data.

The full Unit Record dataset covers the financial aid information of matriculated students in Washington public postsecondary institutions who were awarded and accepted need-based financial aid. Not all students who attend postsecondary institutions use financial aid, so this dataset is not a complete record of all enrollments. However, it does include students who did not complete the academic year and a small number of students who did not actually receive aid. See Appendix B for additional information. This data includes demographics, student need, and sources of financial aid. Additional columns in this dataset also summarize aid programs by type, denoted by "total" in the column name.

Data represents students enrolled during the academic years (AY) 2004-2005 through 2017-2018, though reporting requirements vary by academic year. Program requirements, financial award caps, and calculation procedures used by schools may have changed over time. Researchers should be wary of utilizing data representing both multiple years and multiple institutions concurrently, given these variations (see Appendix B for details). Information for AY2004-2005 through AY2007-2008 contains mainly summative data. AY2008-2009 through AY2012-2013 represent data with inconsistent submissions between schools and years, but it is also a highly complete dataset. AY2013-2014 through AY2017-2018 is the most consistent submission period with high completeness. Enrollment for each academic year is indicated by five time blocks denoted by semester (fall, winter, spring, summer, summer 2)[9]. Since the dataset includes information across

[9] Enrollment information for Washington State University is generally missing because the school's quarter-based system does not conform to the data.

multiple semesters/quarters, portions of the data are aggregated by the institutions into a single student/year report, which is represented as a single row for each institution the student attended. Students who attend multiple institutions will have multiple rows per academic year.

This dataset is generally considered to be reliable; however, systematic missing data are known to occur for a variety of reasons. The Unit Record dataset has distinct patterns, which may impact research quality. Data is separated into five distinct time blocks, as noted above, which should be considered when selecting research samples.[10] Most years of data have incomplete rows because not all schools award every financial aid option. These missing data points are generally consistent at the school/award level over a given set of years.

Students who receive aid from merit-based awards, students who do not qualify for federal aid (such as international students or students without documentation), or students who pay for school out-of-pocket are generally not captured by this data. These students are considered "unknown missing," since their information cannot be directly observed within this dataset. This data should not be considered a reliable proxy for ability to pay or student income research. The Free Application for Federal Student Aid (FAFSA) is not required by many institutions, particularly two-year programs; therefore, some institutions may have more frequent missing financial aid data.

Some data columns that researchers may find important are completely missing for AY2008-2009 to present. This "known missing" data occurs because the program was not available, or an institution did not award funds from a program during a given time frame. Further, award rates of particular programs may differ across colleges, because some schools may not offer certain financial aid options, or they may formulate their aid packages differently. Researchers should carefully consider the institutional influences on awards before utilizing this data. For cross-school comparisons, ERDC recommends relying on "total" columns rather than component columns, where possible.[11]

Known missing data for AY2004-2005 through AY2007-2008 are most frequently caused by administrative data not being collected by schools, or programs not being offered by schools. In later periods, known missing data is almost exclusively the result of a program not being available. Unknown missing may occur when data is not submitted to WSAC.

Research questions that could be answered with this data, or linked with other data, include:

- What is the impact of increasing the award amount for a grant on student graduation?
- Is there an equity issue in financial aid awards between racial groups?
- Is there a difference in student outcomes for unsubsidized loan recipients between two-year and four-year schools?

[10] No limitations were identified at the institution level, except for Seattle Vocational Institute (WSAC Institutional Code 5750), which is missing from AY17/18 data. Data for that school are now reported as part of Seattle Central Community College (WSAC Institutional Code 4450), which is the parent institution.

[11] During the 2010-11 fiscal year, the legislature rescinded some of the funding for the State Need Grant program. Public institutions were required to maintain grant aid to students by using their own funds to compensate for the rescinded funds, thus eliminating any reduction in funding to students. See Appendix B for more details.

Apprenticeships

Registered Apprenticeships Data

The Registered Apprenticeship dataset is provided annually to ERDC by the Department of Labor and Industries (L&I) and includes cumulative participation records for each apprenticeship program participant, by program/occupation enrollment segment. Data includes participants in registered apprenticeship programs sanctioned by the US Department of Labor and administered by L&I. The main components of typical apprenticeship programs include Business Involvement, Structured On-the-Job Training, Rewards for Skill Gains, and a Nationally Recognized Credential.

Prospective participants must first contact the apprenticeship program of interest to determine if applications are being accepted. Each program has their own set of admission requirements. In general, apprentices must meet the four criteria below:

- Be at least 16 years or older.
- Be able to perform the work, with or without reasonable accommodation.
- Have the knowledge, skills, and abilities needed to learn the job.
- Provide proof of age, high school diploma or equivalency (GED), and/or honorable military discharge.

L&I's apprenticeship dataset contains over 93,000 records of Washington registered apprenticeship participants from 2000 to 2019. Ninety-five percent (95%) of all apprentices are Washington State residents. Data is captured at the record or individual level, reflecting participation in 365 programs across 372 different occupations. L&I's apprenticeship data includes 21 variables that reflect key characteristics about the programs and participants. While they could start their apprenticeship before the year 2000, as denoted within the WorkStartDate variable, only apprentices who were active as of the year 2000 are included in this dataset.

The key data elements are county of residence, apprenticeship program, registration date (when apprentice entered a program), status date (the date the apprentice experienced a change in status), status (completed, suspended, active, canceled) and apprentice occupation. Other key components include gender, race, ethnicity and prior education level.

Unlike most education data, L&I's apprenticeship data are not structured by participants' school graduation/completion year or enrollment academic year, which makes it challenging to use in conjunction with education data. The time to complete an apprenticeship can vary greatly depending on the type of work, company policies, and the apprentice's ability to master the skills required for program completion. Since apprenticeships typically last one to four years and include on-the-job training, they more closely resemble the internship format than the traditional postsecondary degree trajectory. Although a prior education variable exists within the data set, ERDC advises caution in how it is used given the overwhelming number of missing values.

ERDC is in the early stages of profiling the Registered Apprenticeship dataset for research purposes. These data can be effectively utilized to track and evaluate apprenticeship program activity, including participants' initial program enrollment through completion or non-completion,

including any potential breaks, transfers, or suspensions among participants. It can answer questions regarding apprentice demographic characteristics, historical participation trends by program or occupation, rates of completion, and/or breaks in program continuity. Since demographic variables like Race and County have a considerable number of missing values, researchers should account for this systemic data limitation in their analysis.

Workforce

Unemployment Insurance (UI) Program Data

UI Wage Data Dashboards and Reports
Outcomes of Need-based Financial Aid: Choice of Major & After-graduation Earnings
Education and Employment Characteristics of Incarcerated Young Adults
Earnings for Graduates Data Dashboard
High School Graduates Outcomes Data Dashboard

The Unemployment Insurance (UI) Program is a federal-state program financed by payroll taxes paid by employers, and in a few states paid by the employee. The U.S. Department of Labor sets broad criteria for program eligibility and coverage, while states determine the specifics of program implementation. In Washington State, the Employment Security Department (ESD) is responsible for the administration of the UI Program. Nearly all employers are required to participate if they pay wages to employees, regardless of dollar amount. Employers must register with the state, submit quarterly reports, and either pay unemployment taxes or reimburse ESD for benefits paid to all their part-time or full-time employees.

The UI Wage data, from which an extract is provided to ERDC by ESD, includes records for all individuals employed in wage-paying positions for employers based in Washington State, though some exclusions apply. While exclusions are subject to change, individuals in the following paid positions are not included in the UI data collected by ESD:

- Spouse, children under 18 and student workers of small farm operators – those with payroll less than $20,000 and fewer than 10 employees.
- Employees performing domestic services in a private home, college club, fraternity or sorority, if the total wages paid are less than $1,000 per quarter. If payroll exceeds $1,000 in any quarter, wages must be reported for the entire year and the following year.
- Non-profit preschool staff, if fewer than four staff.
- Business owners. Sole proprietors do not report themselves, their spouses or unmarried children under 18.
- Corporate officers are required to cover themselves for UI unless they opt out by January 15th each year.

Depending on the circumstances, employers may not be required to report the following additional types of employees:

- Self-employed workers
- Church employees

- Work-study students, if the employer is a non-profit 501(c)(3), state government or local government
- Licensed insurance agents
- Railroad employees
- Licensed real estate agents, brokers, and investment company agents
- Federal employees, such as U.S. Postal Service (USPS), federal civilian employees, and active duty and retired military

Employers must submit two files quarterly to ESD: one wage file at the employee level, and one summary file at the employer level. The detailed wage file data includes the amount of wages paid quarterly to each employee and other employment information, as outlined in the table below. The quarterly wage detail report filed by employers includes the following elements: Year, Quarter, Employer account number, Employee identifiers, Wages paid during quarter, and Hours worked during quarter.

ERDC receives quarterly wage record files from ESD. Data is loaded into the P20W Data Warehouse, and the identity resolution process links employees to other data. During the loading process, the data is cross-checked against previous data to ensure completeness and data quality. ERDC's UI Wage dataset includes wage data for the first quarter of 2000 through the most recent quarter received. There is typically a four-quarter lag in the UI Wage data that ERDC receives, based on the current quarter. Once received, this information is considered complete, given the exceptions above and contains no unintentionally missing values.

As with many administrative data sources, there are known missing values and unknown missing values in the UI Wage dataset. Known missing values are the result of specific groups of employees that are excluded from submitting data, as described above. ESD estimates the rate of coverage at 92%, but the rate may vary by industry. Researchers must account for the exclusion of some paid employees, such as work-study students, especially when analyzing education and wage data concurrently. Similarly, individuals who receive non-wage income (i.e., an ownership stake in a corporation or partnership, receipt of royalty income, etc.) will not have their total income reported because certain types of income are excluded from this data.

Beyond the known missing, there is also an issue of unknown missing. While not specifically excluded, gig workers have operated as both self-employed and wage-paid workers, depending on the employer. As rules change, gig workers may show artificially deflated wages and, in some cases, appear in the data intermittently. Similarly, not all employers report employees based in Washington as Washington employees. The structure of some organizations can limit the ability to determine an employee's industry or location. Comparing and analyzing wages by region can also be difficult because employers may only report by their corporate office address and not the location(s) where employees physically earn wages. This is particularly important in areas that border other states, because where an employee works determines where the UI tax is paid. For example, if a company is based in Washington, and their employee works in California, then UI taxes are paid (and data is reported) only in California instead of Washington. See ERDC's Technical Report 2012-01: Employment Data Handbook, July 2012 for additional information.

P20W Administrative Data Limitations

While all the datasets above are processed to the highest quality standards by the source agencies, it is important to recognize that inaccuracies may exist within administrative data. Unlike other data, where both cross- and within-subject controls are possible, such measures are often unfeasible and impossible to incorporate in administrative data. Administrative data is collected as both transactional and summative datasets by local administrators and submitted to an agency authority, making variance among data collectors a potential source of bias in each dataset. Quality control processes may be imposed after data is submitted to agency authorities, which could impact data quality in ways that are difficult to detect within the final dataset.

These limitations as described in this Handbook, are not meant to suggest that the administrative data loaded into the P20W data warehouse is unreliable. ERDC advises researchers to keep these potential concerns in mind as they request data and conduct their research. Given the large datasets that are created from administrative data, it is tempting to apply complex statistical models to generate results that are hidden on more simple analyses. Administrative data must always be thought of as the combination of both the collected data and the process used to collect the data. The data summaries in this handbook delve into these processes, but they are not the researcher's only resource. Researchers who use ERDC data for analysis purposes should review all the available data documentation and adjust their models according to the research question and the administrative data collection procedures.

Part II. P20W Research Methods

Administrative data often suffers from inherent systematic and structural bias (both known and unknown), because the data is primarily collected for reporting and administrative purposes. Researchers should consider how potential data issues may affect the chosen methodology and interpretability of study findings. ERDC does not attempt to suggest what methodologies are appropriate for conducting education research. Rather, this section explores some statistical methods approaches that have been used or are relevant to answer education research questions with Washington's P20W data.

Descriptive Statistics

Descriptive statistics can be a useful starting point for analyzing administrative data. They can help researchers understand the availability, completeness, and appropriateness of the data, and aid in establishing baselines and hypotheses for further analysis. Descriptive statistics are also generally the most straightforward way to highlight program outcomes to policymakers. Most of ERDC's dashboards and data visualizations rely entirely on descriptive statistics. Even when inferential statistics are used in a program evaluation, data visualizations will typically present descriptive statistics that provide meaningful context for those results.

Although descriptive statistics are useful for establishing baselines and generating further questions, they are usually insufficient on their own for research purposes. ERDC research reports

that rely entirely on descriptive statistics carefully acknowledge the relevant biases and identify key questions for future study. In addition, descriptive cohort analysis can be used to determine which demographic variables are essential for inferential analysis.

Inferential Statistical Methods (Quantitative methods)

Researchers use inferential statistics to determine how each independent variable impacts the dependent variable with a statistical model that best fits the data being analyzed.

Regression modeling with cross-sectional data

A regression model is the most popular inferential statistical method used to answer research questions. While there are several types of regression models, identifying the most appropriate regression model for the data has enormous implications on the quality of the research. The selection of any regression model primarily depends on whether the outcome variable(s) in question are continuous or discrete.

Linear Ordinary Least Square (OLS) regression model. The ordinary least square (OLS) regression model is likely the most widely used inferential statistical method in education research, in which dependent variable is continuous. While it does require researchers to make certain assumptions about the data (i.e., the normal distribution of the standard error), OLS can generally provide an unbiased estimate of the relationship between independent and dependent variables.

Generalized Linear Models (GLM). Unlike OLS, Generalized Linear Models (GLM) allow for flexible distributions of error terms, and allow the dependent variable to have a different relationship with the independent variables. The link function allows for modelling counts or categorical dependent variables.

Probit and Logit regression models. Univariate probit and logit regression models can be useful when dependent variables are *categorical or dichotomous* (e.g., enrolled vs. not enrolled). Logit and probit models make different assumptions about the distribution of the error terms. Probit models assume a normal distribution, while logit models assume a logistic distribution that is logarithmic or left-skewed. Logit is better at distinguishing outcomes where there is a clear difference between options (only A or B), while probit is more appropriate to use when exploring probable outcomes (closest to A or B).

Ordered and multinomial regression models. Ordered and multinomial models are two types of categorical regression models. The difference between these ordered and multinomial models, however, is the structure of the dependent variable. Dependent variable of an ordered regression models takes several finite and discrete values that contain ordinal categorical data. In contrast, the dependent variable of a multinomial model takes finite and discrete values that have no set order.

Multilevel Models

Multilevel models are extensions of regression, in which data are structured in groups or clusters. In education research, multilevel models are frequently used to distinguish the individual-level and

group-level effects, in which individuals (i.e. student) are nested under group (i.e. classroom or school). For instances, to estimate the effect of special education on student test scores and whether the effect differ by school, fitting a multilevel model to identify the effect at student level and school level is more appropriate than classic regression. Another application is to distinguish individual effect from time effect using panel data, which is a critical feature of P20W data. While cross-sectional data is a set of values that illustrate a single time period, panel data is a blend of individual-specific effects and time-specific effects.

There are also some other data structures appropriate for multilevel models. For instances, in settings where overall time trends are important, multilevel data with repeated measurements sometimes called time-series cross-sectional data. It is also possible that multilevel models could be applied to non-nested data that individuals are not completely nested under specific group (i.e. students vs. neighborhoods and schools). Multilevel models could be fitted in different ways, depending on whether the focus is on individual-level or group-level effects, or both. The model specification on intercept, slope, and constant terms, determines which model to apply. Commonly used methods may include linear mixed effects, random coefficient effects, random slope, hierarchical linear models, and growth-cure models.

Quasi-experimental methods

In some situations, it is useful to approach estimation from a causal standpoint. While true causal estimation requires random assignment, there are ways to utilize longitudinal data to generate causal inferences. These methods tend to rely on special features of the data like discontinuities to differentiate between groups of similar students. Discontinuities are not always appropriate for determining causal relationships, so quasi-experimental methods should be used with great caution to avoid inappropriately strong conclusions.

Difference in Difference (DiD) model. DiD is one of the most common quasi-experimental methods used in education research. For this method to be successful, data must have two key features: 1) A discontinuity in a continuous variable that extends beyond the change point and 2) two statistically similar samples that differ by a treatment or intervention before and after the discontinuity. In effect, this method requires a treatment and control group that experience a change at the same time, leading to the creation of four subgroups: 1) Pre-Treatment, 2) Pre-Control, 3) Post-Treatment, and 4) Post Control. These subgroups are then compared to each other to determine the impact of the discontinuity and the treatment.

Regression discontinuity (RD) model. RD is useful in situations where a significant change occurs, often over time, while data is being collected for a continuous dataset. The RD model compares the dataset from before and after the change to the dataset, in order to identify potential differences that occurred. In an ideal dataset, these differences can be attributed entirely to the impact of the cutoff. However, with administrative data it is often difficult to account for all possible influences on the outcome variable. Thus, this method is most useful for analyzing sudden policy changes or external shocks (such as a recession) on a well-defined program.

Matching methods. Matching methods may be useful when comparing similar individuals within groups, especially when there are known differences between the demographics and experiences of each groups (for example, those who pursue a higher education degree and those who do not). This can be done using a variety of matching techniques. Some applications of matching methods are briefly introduced below.

- ***Propensity score matching (PSM).*** PSM relies on the estimated *propensity* score to match similar individuals in the treatment and control groups, or match individuals with the same propensity score. The propensity score is calculated based on observed pre-treatment characteristics that are associated with the selection into treatment(s). Then, a control group is formed by matching individuals with propensity scores like those of the treatment group. If two students fall within the same propensity score range but are in different treatment groups, then the assignment is assumed to be random.
- ***Coarsened exact matching (CEM).*** CEM is another approach to compensating for selection bias by matching and comparing similar individuals in the dataset. Unlike PSM, the CEM does not start estimating the propensity score. Instead, the researcher identifies clusters of individuals using coarsened data (e.g., instead of *age*, they might use *age groups*) and matches individuals within those clusters for the final analysis. This approach maximizes the number of individuals within a dataset to include in the analysis, instead of focusing only on individuals with exact matches. For this reason, CEM can be useful when PSM might otherwise reduce the size of the dataset too much, or when exact matching is not possible.
- ***Generalized propensity score (GPS).*** While the propensity score is developed for use with binary treatment, GPS is used for quantitative or continuous exposures. Examples of continuous treatment include income or years of education. In the context of education research, a good example is using the Quality Rating and Improvement System (QRIS) scores to assess the impact of early childcare center quality on child outcomes.
- ***Marginal mean weighting through stratification (MMWS)***. The matching methods introduced above do not account for the weighting scheme of treatments. When a study sample is disproportionally distributed into treatment groups, MMWS provides a weighting method that makes use of propensity scores. It stratifies a sample based on the propensity score for each treatment, and then computes the weight according to the proportion of individual units within a stratum assigned to the corresponding treatment. As a result, this method enables pairwise comparisons between the treatment groups.

Structural Equation Model (SEM). SEM is a comprehensive and flexible statistical technique analyzing the structural relationship between measured variables and latent constructs with multiple pathways. It combines factor analysis and multivariate regression analysis to estimate causal relationships. A common practice in applying SEMs is to construct a diagram that specifies the model, where each latent variable is defined with its observed indicators variables, and the relationship between variables.

Interrupted time series (ITS) design. ITS design works similarly to DiD design, with some caveats. While DiD evaluates a program's impact by whether the treatment group deviates from its baseline mean at a greater rate than the comparison group, ITS controls for differences in the

baseline mean and trends between the treatment and comparison groups. Compared to DiD design, ITS has more stringent data requirements, typically requiring a sufficiently longer time series. ITS can estimate the effect of an intervention on outcome variables for a single treatment group or when compared with one or more control groups. With a single treatment group and no control group, the intervention trend is projected into the treatment period as counterfactual.

For more information on how these methods have been applied in educational research, please see Appendix A.

Requesting Data from ERDC's P20W Warehouse

ERDC has an established process for researchers to request and explore data from the P20W Warehouse for research and analysis purposes. Because the P20W Warehouse includes data from multiple educational institutions and workforce agencies, ERDC must comply with a variety of federal and state policies to protect the privacy of individuals. These policies are important to determining what level(s) of data are made available to researchers. This section summarizes the important privacy policies by which the ERDC abides, the three levels of data produced by ERDC, and the steps to request data from ERDC.

Data Privacy

One of the ERDC's key roles is to maintain the privacy of individual data that is managed within the Washington P20W data warehouse. The Federal Educational Rights and Privacy Act of 1974 (FERPA) requires educational institutions and state agencies to safeguard the confidentiality and privacy of personally identifiable information in student records. Workforce-related data are also protected and secured by federal laws like Section 303 of the Social Security Act, for which the U.S. Department of Labor has promulgated Title 20, Chapter Five, Part 603 of the Code of Federal Regulations.

To comply with these laws, ERDC only publishes aggregate information and never publishes information that can be used to identify individuals. All researchers who use P20W longitudinal data from ERDC are required to abide by these regulations, which are included in the formal data sharing agreement with ERDC. ERDC's data request and product review process is designed to ensure that these privacy protections are upheld, and that the data is used appropriately.

ERDC compiles and analyzes three different types of data: Highly Restricted-Use Data (Level 1), Restricted-Use Data (Level 2), and Public-Use Data (Level 3). ERDC has different approaches to handling each type of data for research purposes, as outlined below.

- ***Highly restricted-use data (Level 1).*** Level 1 data is highly restricted because it includes information about the identity of individuals or employers. This data is strictly confidential and requires specific procedures to protect confidentiality per FERPA regulations and other state and federal requirements. ERDC uses this information only for record matching purposes as described in the identity resolution section of this handbook. Level 1 data is

always kept secure. It is very rarely shared, and when it is, it is only shared under strict protocols.

- ***Restricted-use data (Level 2)****.* Level 2 data is unit record data such as individual-level data but with no direct identifiers. Though stripped of direct identifiers, this data is still considered potentially identifiable, since it may be possible for someone with direct knowledge of a student's specific characteristics (i.e., a person of known race, gender, age, college enrollment, and high school experience) to infer their identity. Level 2 files are sometimes shared with partnering institutions and researchers for research purposes under strict precautions to ensure privacy and security. Partnering researchers and institutions must have the technical proficiency to store data in a secure environment, confirm their understanding of relevant data privacy laws and regulations, and agree to strict protocols regarding how the data is used. Before data can be shared, ERDC staff first establish that the researcher's stated use of the data is both *legal* under relevant laws and expressly *authorized* by our data contributors.
- ***Public-use data (Level 3).*** Level 3 data is aggregate data for public use and can be published. Examples of Level 3 data include college attendance rates for high school graduates by school district, student transfers between two-year institutions and four-year institutions, and the kindergarten readiness of early learning participants. Precautions are taken with Level 3 data to protect individual identities. ERDC may provide more information about some groups of students than others.[12] It may be possible to compare the graduation rates of White students with Hispanic or African American students, but if there were only a few Native American students in the cohort, then we cannot be as detailed about them as a group. For more information, see Suppression Rules for Public Reporting.

Requesting Data from ERDC

Data requesters that are asking for Level 3 aggregate data should complete ERDC's Data Request Form B. Form B is for requests that involve no redisclosure and requests that do not require substantial statistical analysis. Data requesters that are asking for Level 2 unit record data should complete ERDC's Data Request Form A, especially when the request is for detailed datasets that fall under the "redisclosure" category, or non-redisclosure requests that involve substantial statistical analysis. If the request requires redisclosure, requesters must also complete ERDC's Data Security Form. Requestors should submit their completed form(s) to the ERDC Inbox. Researchers should review ERDC's Data Request Process webpage for more detailed information. Before submitting the request, requesters are encouraged to contact agency partners to discuss their project and how it aligns with their interests.

ERDC must only approve requests for projects that are an audit or evaluation of an education program. For ERDC to legally share education data with third party requesters, the study must

[12] All researchers using P20W data from ERDC should consult Technical Brief #3 published by the Privacy Technical Assistance Center from the U.S. Department of Education. This document explores the redaction and suppression logic that must be followed when publishing any research using P20W longitudinal data. Revealing even aggregate information (such as averages) of small groups could still allow people to infer information about individual students.

constitute the audit or evaluation of a state- or federally-supported education program. For the purposes of FERPA, the Department of Education defines "education program" as:

> Any program that is principally engaged in the provision of education, including, but not limited to, early childhood education, elementary and secondary education, postsecondary education, special education, job training, career and technical education, and adult education, and any program that is administered by an educational agency or institution.[13]

ERDC's approval of a request depends on the framework of the research project. For example, a project examining the education outcomes of foster care program participants might not meet this requirement. However, this project may qualify if a researcher is looking into how the state's education programs serve a particular student group (foster care participants). Researching the education outcomes of justice-involved participants does not meet this requirement, since justice participation is not an education program. However, it is possible to investigate the justice outcomes of education program participants. Differences are more than merely rhetorical, because the framework impacts the definitions of the cohort.

[13] As defined in the U.S. Department of Education's "Integrated Data Systems and Student Privacy" Technical Brief, which also explains FERPA as it relates to integrated data systems like ERDC's P20W warehouse.

Appendix A. Washington's P20W Research Method Bibliography

Descriptive Statistics:

- Education Outcomes of Children and Youth Experiencing Homelessness, 2019. ERDC researchers use bar charts to demonstrate how student performance (by subjects and ethnicity) varies between homeless and non-homeless students.
- State Need Grant Recipients' Educational Progress and Degree Completion, 2017. ERDC researchers looked at the outcomes of State Need Grant recipients, comparing across student demographics, institution type (two-year vs. four-year) for eight years.

Regression Modeling:

- Determinates of Need-based Financial Aid. This ERDC research uses *ordinary least square regression* to study the association between several financial-aid determinants and the amount of four financial aids received.
- Cowan, J. and Goldhaber, D. (2015). How Much of A "Running Start" Do Dual Enrollment Programs Provide Students? *The Review of Higher Education*, 38(3): 425-460. This study utilizes linear probability model with school fixed-effect model to study the impact of a dual enrollment program- Running Start, on college enrollment and completion.

Quasi-experimental Methods:

- The Earnings Premium of Washington Higher Education: Gender Deficit in Earnings Among Washington College Graduates. This ERDC report uses propensity score matching to examine and compare earnings among students who obtained a college degree with the earnings of peers who were likely to obtain the degree but only received a high school diploma.
- Fumia, D., Bitney, K., & Hirsch, M. (2018). The Effectiveness of Washington's College Bound Scholarship program. Washington State Institute of Public Policy researchers use propensity score matching to study the effect of College Bound Scholarship on educational outcomes.
- Goldhaber et al. (2019). Pledging to Do "Good": An Early Commitment Pledge Program, College Scholarships, and High School Outcomes in Washington State. This study applies advanced *difference-in-difference* analysis to study the effect of College Bound college scholarship on low-income middle school students' high school outcomes.
- Bania, N., Burley, M., & Pennucci, A. (2013). The Effectiveness of the State Need Grant program: Final Evaluation. Washington State Institute for Public Policy (WSIPP) applies *regression discontinuity* to study the effect of the Washington state college financial aid. Change in grant eligibility rules is adopted as the cutoff to process the RD analysis.

Appendix B. State Need Grant Technical Notes

State Need Grant Local Funds for Public Institutions, Fiscal Year 2010-11

During the 2010-11 fiscal year, the legislature rescinded some of the funding for the State Need Grant program. Public institutions were required to maintain grant aid to students by using their own funds to compensate for the rescinded funds, thus eliminating any reduction in funding to students. The total value of institutional funds substituted for rescinded State Need Grants (SNG) funds was $25,363,500. Researchers must use caution in evaluating SNG for this year.

Institutions were given two options in substituting local funds for rescinded state funds. The first option was to repackage student financial aid awards and replace state need grants with institutional grants. All universities except Central Washington University (CWU) chose this option. The second option was to return the rescinded funds in a lump sum and keep in place the state need grant awards that were in students' financial aid packages. This option was chosen by CWU and all SBCTC institutions. Both options required distinct methods for recording institutional substitutions for rescinded funds. Institutions that repackaged their financial aid awards reported the local funds used to replace the rescinded funds in the appropriate institutional aid categories of the Unit Record Report. For the institutions that returned a lump sum, the Higher Education Coordinating Board made proportional reductions in the SNG amounts submitted on the Unit Record Report. Prorated reductions began with spring term awards and worked backwards until the lump sum total for the institution was reached. Thus, for lump-sum institutions, the SNG amount currently on the Unit Record Report represents the initial SNG amounts packaged by the institutions minus the prorated values of rescinded funds.

The institutional funds substituted for rescinded funds are recorded in the variable [InstSubstitutionStateNeedGrant].[14] These funds are not counted in the sums for the variable [StateAidTotal]. Since the institutions that repackaged financial aid awards included the local funds in existing Unit Record Report aid fields, the funds were included as usual in the sums for the aid bucket variables [GrantAidTotal], [InstAidTotal] (institutional aid total), and [AllAidTotal]. For lump-sum institutions, WSAC added the prorated local-fund values it calculated into the sums for [GrantAidTotal], [InstAidTotal], and [AllAidTotal]. Adding values for [InstSubstitutionStateNeedGrant] to any of these aid buckets will result in double-counting local funds.

The variable [InstSubstitutionStateNeedGrant] must be used carefully. If the purpose is to conduct a longitudinal study of SNG program effects, then researchers may want to add the local funds to the SNG funds to get a value that represents all funds (state and institutional) that students received by virtue of the SNG program. The resulting variable would represent the SNG funds that would "normally" have been expected for the year 2010-11, had the mid-year rescission not occurred. If this course is pursued and [InstAidTotal] is used as a control variable, then local funds must be subtracted from [InstAidTotal] to avoid double-counting.[15]

[14] WSAC datamart variable SNGLocalFunds.
[15] There are a small number of mis-reported records for which the value of local funds exceeds the institutional aid total. It is not clear how to reconcile the data reported in these cases.