Researchers Guide to Using VLDS

Prepared for

Virginia Department of Education

January 31, 2014

Center for Innovative Technology,
Research & Analytic Insights, LLC, and Gibson Consulting

Deborah L. Jonas Author

Will Goldschmidt
Marshall Garland Contributors

CIT CONNECT
CENTER FOR INNOVATIVE TECHNOLOGY
2214 Rock Hill Rd., Suite 600
Herndon, Virginia 20170
703.689.3000

This document was jointly produced by

CIT CONNECT
CENTER FOR INNOVATIVE TECHNOLOGY

Research & Analytic Insights

GIBSON
CONSULTING GROUP

Technical Point of Contact:

Rona Jobe | Senior Consultant, Consulting Services

Rona.Jobe@cit.org | 703.689.3055

Administrative Point of Contact:

Pat Inman | Contract Manager

pat.inman@cit.org | 703.689.3037

INSIGHT POWERED BY VLDS

This Page Intentionally Left Blank

Table of Contents

1 Introduction

Virginia Longitudinal Data System (VLDS) is a pioneering collaboration for Virginia's future, giving the Commonwealth an unprecedented and cost-effective mechanism for extracting, shaping and analyzing educational and workforce development data in an environment that ensures the highest levels of privacy.

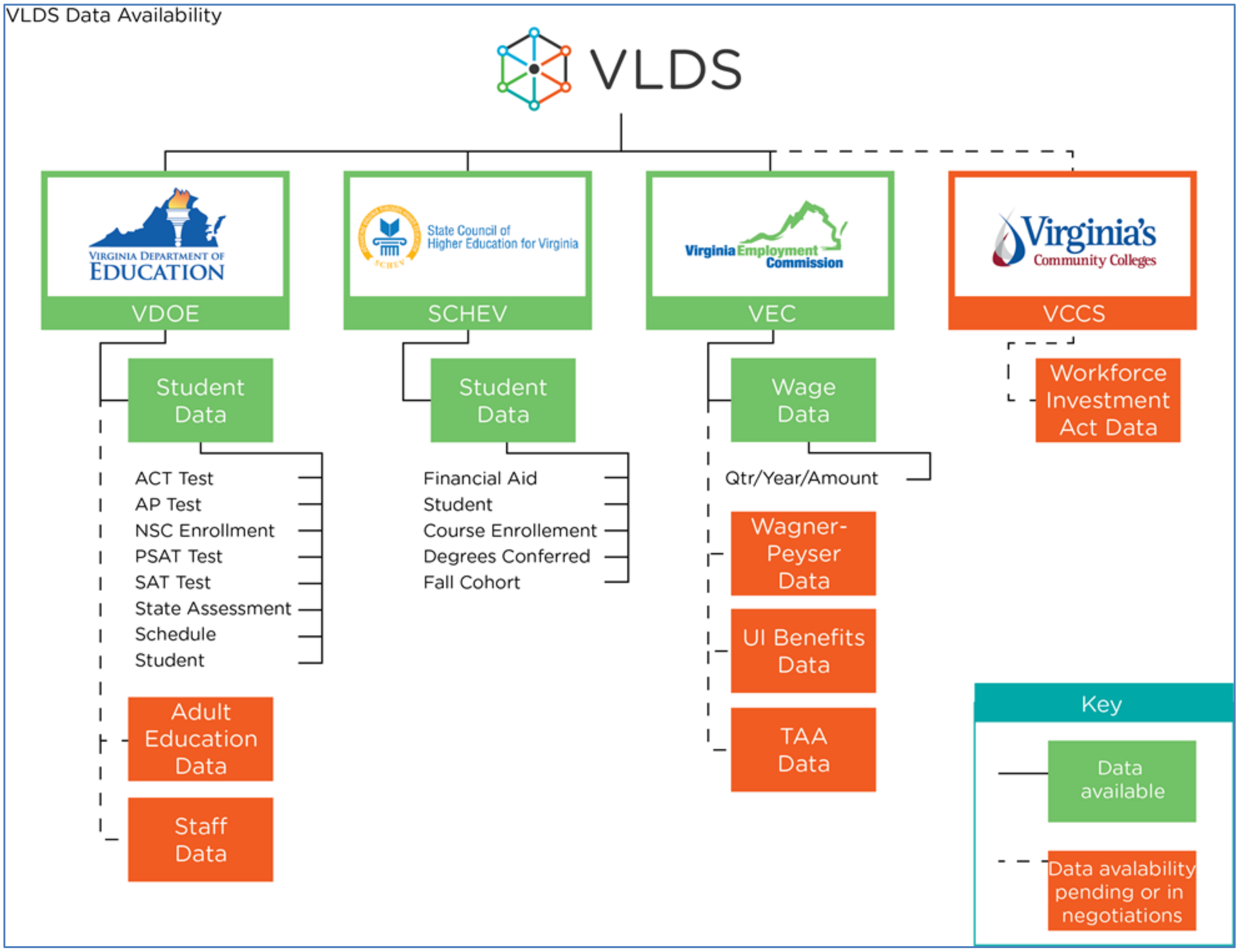
Developed with funds from the 2009 Statewide Longitudinal Data Systems Grant Program of the United States Department of Education, VLDS is comprised of several component technologies that support secure, authorized research addressing today's key educational and workforce training questions. VLDS is the result of a coordinated effort by several Virginia government agencies.

VLDS is built on a "federated" system to merge data across the participating agencies in a complex double, de-identifying hashing process that leaves private data behind the existing firewalls of the participating agencies. This technology was developed, in partnership with VLDS participating agencies, including the Virginia Department of Education (VDOE), the State Council of Higher Education for Virginia (SCHEV), the Virginia Employment Commission (VEC), and the Virginia Community College System (VCCS). Built almost entirely with in-state resources, the agencies partnered with experts from Virginia Tech, Virginia Information Technologies Agency (VITA) and Center for Innovative Technology (CIT) to create VLDS.

VLDS currently leverages data from VDOE, SCHEV, and VEC. VCCS is also a founding partner and is in the process of bringing data to the system. Figure 1 shows agency participation and data sets available as of November 2013.

This paper aims to facilitate researchers' understanding of important details of the VLDS data system, and to inform researchers and authorized users about factors that must be accounted for when proposing and conducting research with VLDS data. We developed this paper to inform the research process based on information available at the time of publication. As the system matures and is updated, and as researchers and agency staff gain more experience, new details may emerge that researchers need to know. As well, some of this information in this original publication may become outdated. Therefore, we recommend VLDS team members revisit and update the document on an as-needed basis.

Figure 1: Agencies and data sets in VLDS as of November 2013



2 VLDS data elements

Researchers from a wide variety of backgrounds and expertise may be interested in using VLDS to conduct research and evaluation that furthers our understanding of the influences of government programs and policies on student and citizen outcomes. In many cases, VLDS can be an appropriate data source. We highly recommend that researchers interested in using VLDS become familiar with the nuances of the system to ensure that the data source can meet project needs. While robust, VLDS and the data contained therein, like all systems, have inherent limitations. In addition, it is important to consider how VLDS structures and processes affect staffing requirements, budgets, and project timeframes. Reviewing this document can be a first step in the process.

2.1 Changes to data elements over time

Researchers interested in using VLDS should carefully review the VLDS Data Dictionary and Selection Tool to understand available data elements. The data dictionary lists the data elements available from participating agencies and provides valid-values. VLDS permits users to download the data dictionary and valid values for offline review.

It is important to understand changes that take place in the data over time. Changes in elements and their associated valid-values take place for different reasons. For example, agencies may add or remove entire data elements, and change or update valid-values. Table 1 lists examples of several types of changes to the data that can impact VLDS users. The table also provides information about where in existing documentation researchers can find information about potential changes in the data of interest.

Table 1: Examples of changes to VLDS data elements, reasons for change, and information to help researchers identify the change in VLDS documentation

Reason for data element change	Example	VDLS documentation of data element change
New data elements	• VDOE added course enrollment and completion data for each student in 2010/11. Course data from prior years are not available, although for some research, state end-of-course tests can provide a reasonable proxy variable for course participation.	• Data Dictionary and Selection Tool, Valid Use Begin Date.

Reason for data element change	Example	VDLS documentation of data element change
Changes in valid values	• VDOE made changes to race/ethnicity codes. • LEP proficiency codes.	• Changes are embedded in the VLDS data structure. VDOE has included different data elements in VLDS to represent the different codes (see also changes in data collection methods).
Removal of data elements from state agency collections	• LEP Proficiency Type was removed from VDOE's data in 2009.	• Data Dictionary and Selection Tool, "Valid use end date."
Changes in data collection policy	• Prior to 2012, it was optional for private institutions of higher education (IHE) to submit course grades to SCHEV. This field became a requirement for all IHE in 2012.	• Information about SCHEV policy change is not currently available in documentation. Researchers using the data may encounter missing grades for entire IHE prior to the change in SCHEV reporting policy.
Changes in data collection methods	• Based on federal requirements, VDOE changed the race type codes and associated collection requirements. • VDOE's homeless flag • VDOE changed the methods by which limited English proficient students' proficiency levels were measured and documented.	• For critical changes, the data dictionary identifies the changes to race/ethnicity codes and data collection methods for the homeless flag. • The data dictionary includes two different proficiency variables for LEP proficiency—one from before and one from after the change in data collection policy.

Reason for data element change	Example	VDLS documentation of data element change
Qualitative changes in the meaning of elements	• Scores on Virginia's SOL tests identify students as being proficient or advanced proficient in content areas. The achievement needed to meet minimum or advanced proficiency changes with each revision of the *Standards of Learning*.	• SOL changes can be identified by requesting SOL test type. This element includes the standards that were measured by the assessment.

Researchers using VLDS can review the critical changes to data elements to learn about changes that affect data and data codes. There can also be qualitative changes in the meaning of elements that may not be obvious by reviewing the data dictionary. One example of a qualitative change to the meaning of data elements is in the K12 *Standards of Learning* (SOL) assessment results. In Virginia, the lowest obtainable scaled score (LOSS) and the highest obtainable scaled score (HOSS) have remained constant since the testing program began in the late 1990s. However, the achievement standards that tests are designed to measure are revised at least every seven years. With changes to the standards come changes to the tests. These changes may not be obvious to researchers who are not familiar with Virginia data. By requesting SOL Test Code, researchers have information about the specific content standards being tested for each assessment. This code includes the year the tested standards were approved. Therefore, it is possible to account for these changes with data elements available within VLDS.

Policy changes can also affect data. SCHEV has collected course enrollment and grades using a common numbering system for more than a decade. SCHEV began requiring course enrollment data from private IHE around 2004; at the time, submitting course grades was optional. In 2012, based on changes to the *Code of Virginia*, SCHEV began requiring course enrollment and grades from both public and private IHE. The result is that some students' course grades at some IHE are systematically missing before 2012.

In general, we recommend that researchers work with the sponsoring agency to ensure that they understand the nuances of the data before making requests and during the research process to ensure accurate interpretation.

2.2 Data elements in multiple agency records

Keeping in mind that VLDS provides records from the same individuals that multiple agencies collect, researchers may not be surprised to learn that some data elements are included in multiple agencies' data. When selecting among these data elements, it is important that researchers carefully consider the costs and benefits associated with choosing one source over another, or incorporating the information from multiple agencies. When selecting data elements, it is important for users to consider the original source and purpose of the data element within an agency's collection and discuss the data quality[1] associated with the element with agency staff. Actual choices about including elements from a particular agency will also depend on the scope of the project (i.e., which agencies are the primary data sources) and the analytic purpose of each data element. Additionally, it is helpful for researchers to collaborate with their project sponsor to understand the nuances of the elements, such as how the data are collected, agency understanding about data quality, and the definition of the elements.

> Both VDOE and SCHEV make race/ethnicity data available from VLDS. Prior analyses show that there is a strong but imperfect correlation between students' race/ethnicity in high school and college. In large part, differences are related to how the element is collected—in high school, schools collect information from parents and send the data to VDOE based on federal reporting requirements. In college, students report their information to colleges directly, which then send the information to SCHEV. Another factor affecting the association is the timing of the collection—the agencies collect race/ethnicity data in multiple collections per year. Parents and students have the ability to change their reported race/ethnicity with each collection. Finally, it is possible that a small (likely trivial) percentage of mismatches is attributable to errors when VLDS matches data probabilistically.

Examples of multiple agencies' records containing similar data elements include demographic characteristics (e.g., race/ethnicity, gender) and achievement or outcomes,[2] such as whether students earned Advanced Studies diplomas in high school and SAT/ACT test scores. Each agency collects these data for its specific purposes. Ultimately, these data are exposed to VLDS for authorized use. When determining whether to choose the element from one or more sources, it is important to consider the research question of interest, data collection method and purpose, and the primary audience for results.

Researchers will need to decide which source to use based primarily on the research question of interest. Research that VDOE conducted as part of its College and Career Readiness Initiative (CCRI; Garland, et al., 2011; Jonas, et al., 2012) was aimed at understanding the high school factors that are associated with college enrollment and success. Because the researchers were focused on high school factors, they used race/ethnicity codes and reporting conventions from VDOE. The K12 community was the primary audience for this work, which also contributed to the decision.

[1] In general, agencies report that data elements associated with funding and that are used to develop public reports have the highest data quality. More information about agency data quality is provided in this paper in Section 4, Data Quality.

[2] As more data sets are added, the types of data with similar meaning are likely to increase.

Researchers who are interested in understanding issues from the higher education perspective would likely choose to use SCHEV race codes.

Another consideration when data elements provide similar information is whether sources provide different levels of detail or quality that matter for the research. For example, SCHEV data can inform researchers about whether Virginia high school graduates earned an Advanced Studies or equivalent diploma in high school.[3] VDOE can provide more precise diploma information for public high school graduates, such as whether students earned International Baccalaureate diplomas as well as details about diploma types for those who did not earn an Advanced Studies diploma. Another example is students' participation in dual enrollment courses. VDOE annually collects categorical data about whether students participated in one or more courses that offered college credit while in high school. Historically, the majority of students have participated in these courses through Virginia's Community Colleges or other in-state IHE. In such cases, more complete information about dual enrollment courses and outcomes may be available by merging high school and college records.

2.3 Data homonyms

Another consideration in choosing data elements that each agency collects is the comparability of data definitions. VLDS users may encounter "data homonyms," or elements with the same name but different meanings. For example, there are multiple examples of "exit codes" and "exit dates" in VLDS that may not have the same meaning, or, because of different state agency coding systems, have different codes that have similar meanings.

Historically, data definitions for Virginia's workforce programs varied. In recent years, significant federal and state efforts have led to a set of common data definitions that the Workforce Investment Act programs now follow. The use of these common definitions is relatively recent and may not apply to all of the data VLDS provides. Further, these definitions do not necessarily align with definitions used by K12 and IHE. To avoid inappropriate data use, it is important for VLDS users to ensure they understand data element definitions when using them for research purposes.

2.4 Deriving variables from multiple data sets

VLDS users will invariably derive new data elements from existing variables. For example, VLDS users might establish a definition for "persistent enrollment" in college, or derive a variable for "multi-program" participation among education or workforce data sets. In some situations, derived variables will use data from multiple data sets or agencies in an effort to obtain a more complete measure than is available within a single data set or agency. Under these circumstances, it is critical to understand and account for limitations in all data elements used to derive the variable, and how these limitations interact. For example, SCHEV captures data about whether college students participated in federal and state work-study programs, and VEC includes records of employment. However, work-study participation may not be included in

[3] Whether a diploma is equivalent to Virginia's Advanced Studies diploma is determined by each IHE.

wage records. Therefore, using the combination of the two elements has the potential to provide a more complete data set of working students. Some research questions would benefit from knowing whether individuals are working, regardless of the type of work they are doing. However, SCHEV data does not include any information about wages. Therefore, the derived variable cannot be used when actual wages earned is a critical variable.

2.5 General data limitations

State agencies contribute data to VLDS based on existing data collections. Each agency collects data for a specific purpose and designs its data collections primarily around these purposes. Each data set comes with inherent limitations. The following lists the limitations of the current data:

- VDOE's records do not include any data for students who attend private schools or are home schooled, or data from local assessments and programs.
- SCHEV records do not include data from students who attend college out-of-state or from certain technical training programs that Virginia's Community Colleges and private organizations provide.[4]
- VEC wage records are limited to wages for those employed in Virginia by an entity that reports Unemployment Tax to the VEC. Wage records for federal employees, including those in the Department of Defense, are not available. Further, criteria for reporting to the VEC result in some individuals who are employed as consultants and independent contractors (including many psychologists, counselors, barbers, and cosmetologists) being excluded from the records. See *Code of Virginia* § 60.2-219 for more information about VEC reporting requirements.
- While many stakeholders are interested in studying outcomes for students in terms of credentials, VLDS has access to some information about the credentials students earn in public high schools (e.g., in career and technical education programs) and colleges. As workforce agencies bring additional data sets to VLDS, more credentialing data will be available. However, like most other integrated statewide longitudinal data systems, complete data for industry and professional credentials is not available.

These limitations will affect some projects more than others, and, the limitations may have greater impact on some populations than others. It is important for researchers to be aware of and account for these factors when determining whether VLDS is the appropriate data source and to incorporate these limitations in reports and data products used to communicate findings.

[4] Through a subscription to the National Student Clearinghouse, VDOE provides college enrollment and completion data for colleges and universities across the country.

3 VLDS data structures and use

At the time of this writing, VLDS included over 775 data elements.[5] The data elements are organized by partner agency and usually further organized by the source or type of data. Data are available in accordance with each agency's internal data structure, which may differ. For example, VDOE makes data available by school year using a four digit code representing the fall of each school year (e.g., school year 2008 represents the 2008-2009 school year); SCHEV represents school year using a four digit code representing the fall and spring (e.g., 0809 represents the 2008-2009 school year. Similarly, data are stored and therefore delivered to researchers using each agency's internal coding which typically differs. For example, VDOE and SCHEV's codes for students' gender are available with different codes—SCHEV provides data using numeric codes (1, 2, and 4) and VDOE provides data using characters (M, F, and null). The agencies' data codes are available from the Data Dictionary and Selection Tool.

Not all data are available for all years. In general, data from SCHEV is available from 2006 forward and VEC from 2005 forward. VDOE's data system has undergone significant change over the past decade. As a result, the starting year by which authorized users can access VDOE data via VLDS varies by data set and element. For example, state assessment and demographic records are available beginning with the 2005/06 school year; student schedule data in 2011/12. Appendix A provides additional information about each agency's data structure and availability.

In addition to learning about the data available to authorized users, it is important for researchers using VLDS to be aware of the operational processes that are likely to impact level of effort, skills, and time it will take to use data for analytic purposes. These processes were developed to meet agencies policy and legal requirements to maintain individuals' privacy within VLDS.

3.1 Data sets cannot be concatenated

VLDS was developed to meet legal and regulatory requirements that retain individuals' privacy, while enabling researchers to merge records across state agencies. One of the overarching themes during the development of VLDS was the understanding that Virginia's state privacy law (*Code of Virginia 2.2*) prohibits the creation of a single data set or data warehouse that enables the tracking of personal information from cradle to grave.

One feature of VLDS that helps to meet this requirement is the inability to concatenate data sets that are generated at different times. When data from VLDS are returned to authorized users, the system prepares the data using a one-time, one-way hashing algorithm to create a random set of unique identification code for each individual. VLDS then applies the unique identification code to each instance of the individual in all data tables provided from that data request. If a researcher requests data at a different time, even for the same research project, the system will generate a completely new set of unique identifiers for each individual in the data. This feature, developed to comply with state law, ensures that the new data set cannot be linked to the

[5] The number of available elements will change as new agencies join and existing agencies modify data sources that are included in VLDS.

previously requested data set based on unique identifiers.[6] Among the implications of this feature are:

- If a data set within a single request does not contain all necessary elements, the researcher needs to request all new data.
- Researchers conducting longitudinal studies might want to wait until all data are available before making their requests. When possible, researchers might work with the VLDS project sponsor to determine whether they can make a preliminary data request to gain a more nuanced understanding of the data and prepare programming code while waiting for a final year of data.
- When studies are completed or interim reports are prepared before all longitudinal data are available, researchers must re-request all data to conduct analyses from the initial time period of a study. They cannot request data for the additional year and concatenate (i.e., link) the data to previously prepared data sets.

To users unfamiliar with VLDS, this limitation may seem challenging. However, authorized users can include as many agencies as they need into a single request. Another option users have is to parcel a single, large data request into seven or fewer parts that are requested at the same time. This may be easier for complex requests, or, requests for which data elements are added over time. When users create multiple data requests that need to be linked, they must use a single starting agency and submit all of the data requests within a data package. Within VLDS, a data request is a group of selected variables, and can include data from one or more agencies. Multiple data requests make up a data package. Data packages are submitted to VLDS, and, when approved and returned, all records in the package include unique identifiers that can be linked across tables.

3.2 Linking external data sets to VLDS data

Many researchers are interested in connecting data from VLDS to data collected outside of participating state agencies. For example, researchers have inquired about whether they can merge data from student surveys or assessment data that the state does not collect with data from VLDS. Additionally, several non-profit organizations and local government service providers are interested in linking local data with VLDS to evaluate outcomes of locally administered programs. At this time, VLDS does not have the capability to merge individual-level records to external data sources. However, researchers can often link aggregate data, such school- or IHE-level information (e.g., enrollment and completion rates, average test scores); or locality-specific data (e.g., employment data, estimated education levels from U.S. Census Bureau and Bureau of Labor Statistics) to support analyses.

3.3 Agency data are dynamic

Due to the way that agencies collect data, and the reality that original data providers (e.g., school divisions, IHE, businesses) make corrections to data the agencies collect even after they have

[6] This prevents authorized users, including agency personnel, from having a single data base that tracks individuals from cradle to grave.

been locked, data retrieved over time may be slightly different when retrieved year after year. This is a normal and regular part of using state administrative data for research and is well established business practice within state agencies. While these changes are typically relatively minor in relation to all data agencies collect, such updates could affect cross-agency data in unexpected ways. As a result, we recommend that authorized users who retrieve multi-year data sets at different time points analyze the data to determine consistency over time. If significant discrepancies occur, it is important to confer with agency staff to identify any major data changes that would impact results and interpretation.

3.4 Structure of data received from VLDS

Researchers using the data must be prepared to merge records across multiple files to structure an analytic data set that meets project needs. VLDS data users should be comfortable working with large, raw, student-level, unprepared data files. Researchers are likely to receive several data files from VLDS (see Appendix A for more information). Each data file will include records that include a unique, one-time-use linking key that allows researchers to connect records between files. That means that VLDS creates *linkable* data, but researchers are required to do the actual linking and create analytic data sets. These raw, linkable data files use existing agency data structures that may or may not be similar to one another.

Users receive data files organized by View Name, as listed in the data dictionary—one file with all elements selected from each View Name. VLDS users will need to clean, restructure, and merge the files to meet project needs. For example, researchers working on Virginia's College and Career Readiness (CCR) Initiative regularly work with data from the following sources that must be managed and merged prior to analysis:

- VDOE records
    - State assessment records
    - Unique student listing
    - NSC enrollment records
- SCHEV records
    - Course Enrollment table (part 1)
    - Course Enrollment table (part 2)
    - Degrees conferred table

The CCR research team described the process used during a slide presentation at the Institution for Education Sciences-sponsored STATS DC conference in 2013. The presentation is available on the STATS DC conference website.

Based on agency experience, VLDS team members recommend that authorized users who work with VLDS data have strong data management/programming skills. Critical technical skills include identifying and, if necessary, eliminating duplicated records (e.g., students may have multiple administrations of an assessment); managing longitudinal files with multiple records per individual (i.e., panel data), and merging large longitudinal records across tables within different file structures. Some familiarity with scripting would also be advantageous, particularly for large longitudinal files across multiple agencies to facilitate the ability to perform repetitive tasks. Researchers should also ensure that they have sufficient computing capacity to handle large data files resource-intensive data management and statistical procedures can be time-consuming. Current users recommend all teams use a computer with a multi-core processor and at least 8GB of RAM.

Considering the complexity of VLDS data, and other limitations of the system, VLDS team members recommend that:

- Users prepare data management and statistical programs in a way that all data preparation is repeatable/replicable multiple times and over time. Ideally, code should be readily extensible, so that additional years of data could be added without significant effort.
- Research teams have data management/statistical programming expertise and solid documentation of all data steps. For some research projects, the VLDS team may request programming or statistical code from researchers so they can share it with others or replicate some findings in the future.

Once data are structured and merged, it is critical to validate the linking process before proceeding with planned statistical analysis.

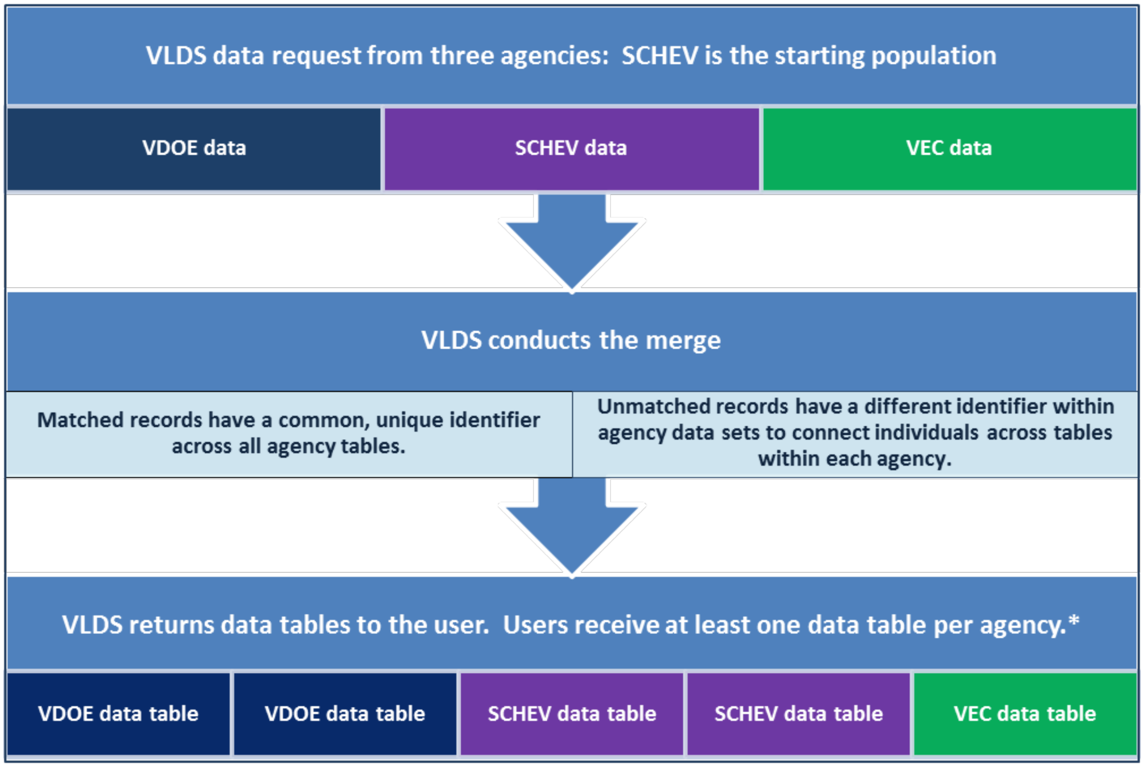
3.5 Matched and unmatched records

When selecting data from VLDS, researchers have the option to include or exclude the unmatched records from relevant agencies. There are times when the unmatched data are less critical, for example, when agency staff use VLDS to prepare regular reports for public disclosure. Unless there were significant shifts in data (e.g., due to population or enrollment changes), the unmatched data are not likely to provide new or critical information.

Researchers using VLDS will typically want to receive unmatched records to assess whether there are systematic differences between those individuals who are matched and those who are not matched. For example, prior research has shown that Virginia high school graduates who enroll in an IHE differ on a variety of achievement measures compared to those who do not enroll in an IHE (Garland, et al., 2011; Jonas, et al., 2012; Lichtenburg, et al., 2010). In an ongoing VLDS project, researchers are examining high school graduates who enrolled in college and are working. In the project, VLDS will return unmatched records for high school graduates to ensure that the research team can identify systematic differences between matched and unmatched populations. For example, the project team will assess differences in college-going and non-college-going high school graduates *and* characteristics of students working in college (matched between VDOE, SCHEV, and VEC wage records) and college-going students not found in the wage records (VDOE matched to SCHEV and NOT matched to wage records). When selecting data for research, it is important for VLDS users to think carefully about appropriate filters to apply to the data to minimize unnecessary receipt of unmatched records—at least one filter is required for each agency included in the request. For example, using the Data Selection Tool, users can filter the data request to include only relevant years; graduates with certain credentials; or individual characteristics that align with project needs.

When VLDS users require matched and unmatched records, it is important to structure the data request to ensure that the requisite data are available. To structure the request, it is important to understand that the VLDS matching process is a one-to-one process, and, that data packages can only be structured to match records from the starting population to each of the other agencies separately. Users who choose to receive unmatched records will receive data tables from each agency included in the query. Matched records between agencies will have a common identifier in every data table that VLDS returns. In addition, all records will have a unique identifier that allows the user to link records (matched and unmatched) across files within each agency. We show this process in Figure 2, which is representative of a request used in a project assessing college and workforce outcomes for a high school graduating class.

Figure 2: Matching process that results in matched and unmatched records

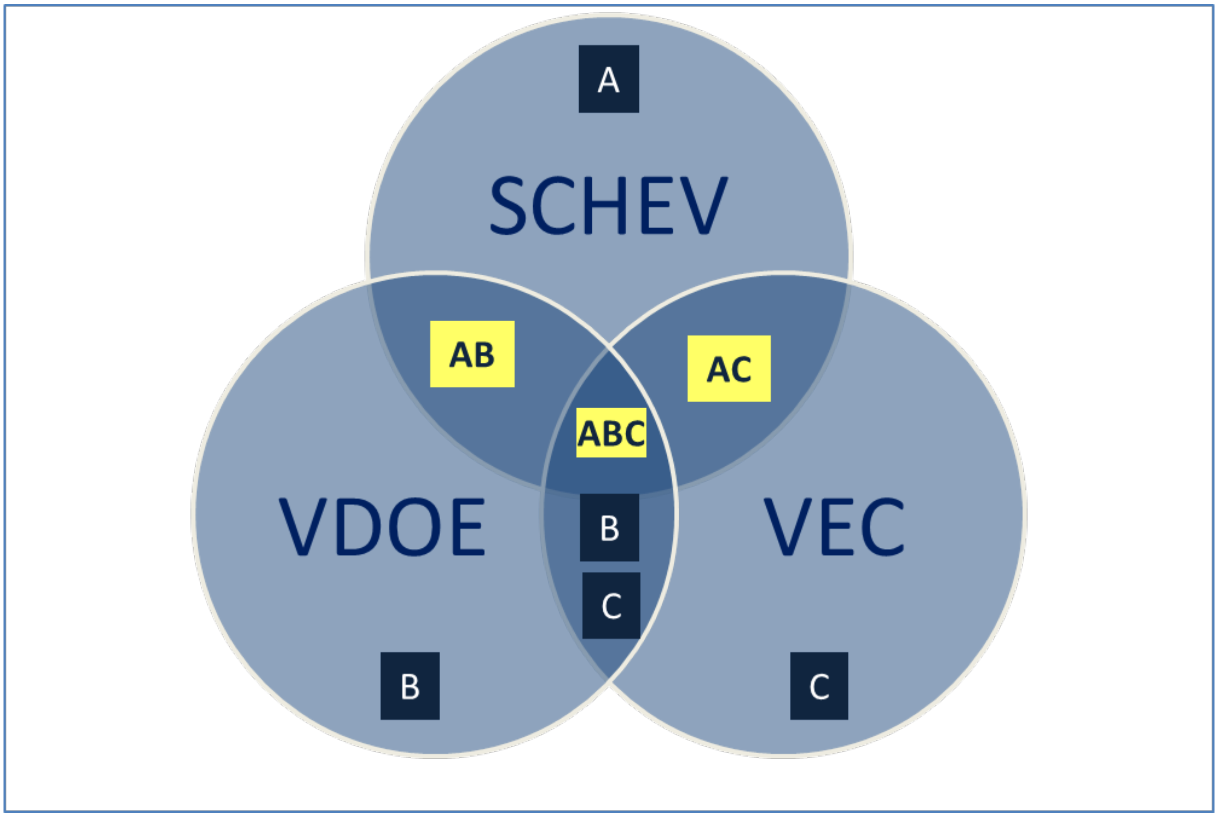

VLDS data request from three agencies: SCHEV is the starting population
VDOE data
SCHEV data
VEC data
VLDS conducts the merge
Matched records have a common, unique identifier across all agency tables.
Unmatched records have a different identifier within agency data sets to connect individuals across tables within each agency.
VLDS returns data tables to the user. Users receive at least one data table per agency.*
VDOE data table
VDOE data table
SCHEV data table
SCHEV data table
VEC data table

*In the example shown here, VLDS returned two data tables from SCHEV and VDOE, and one from VEC.

A limitation in VLDS exists when two agencies cannot directly link records. Currently, this applies to VDOE linking to VEC—there is currently no reliable way to directly link data between these two agencies. However, using SCHEV records and a process like the one shown in Figure 2, users can connect VDOE's data to wage records for a part of the population—those high school students who at some point were enrolled in a Virginia's IHE. An approved research team recently requested these data and successfully demonstrated the ability to obtain matched and unmatched records from a specific high school graduating class. In the data request, the starting population was SCHEV records. Data requested included matched and unmatched VDOE records from two high school graduating classes, and matched records from VEC within specified years. The resulting data tables included all records from the starting population (SCHEV records within specific years); all records from the requested high school graduating classes, including records that matched SCHEV records and those that did not; and records from VEC that matched to SCHEV based on selected filters (years).

Figure 3 shows a representation of the resulting matched and unmatched records.[7] In the figure, agency data are represented by a letter, A, B, or C. The overlapping areas between agencies represent those records that occur in multiple agency records. In the figure, areas of overlap that can be matched are marked with yellow boxes and multiple letters (e.g., ABC represents matched records between three agencies). The returned data sets include unique identifiers that allow researchers to link VDOE, SCHEV, and VEC records in these areas of overlap. Note that one area of overlap—the VDOE/VEC overlap that excludes SCHEV—cannot be linked, and therefore, is indistinguishable from other unmatched records.

Figure 3: Matched and unmatched data from a data request starting with the SCHEV population and including VEC and VDOE data.



Note 1: VDOE and VEC overlap cannot be determined except in the case where it also overlaps with SCHEV data.
Note 2: For illustrative purposes only. VLDS users can request data from each agency with filters that limit the returned data to that which is needed for research projects.

[7] Note that VLDS will not permit users to submit a request without limiting filters.

4 Data quality

VLDS relies on both deterministic and probabilistic matching methods to connect records between agencies. When VLDS returns data to users, the tables include a "match type" data element that specifies the matching method used to connect records across agencies. Data matched between SCHEV, VEC, and other workforce agencies are typically matched using a deterministic process based on a common and unique identifier available in those agencies.[8] All matches with VDOE are based on a probabilistic matching algorithm. Under some circumstances, SCHEV obtains state testing identifiers (STIs) using a probabilistic matching process before data enter VLDS. VLDS then uses these identifiers to match data to VDOE records. As a result, VLDS technology may show some matches between VDOE and SCHEV as deterministic.

Researchers working with VLDS data will need to determine the quality of the matched data. In some situations, researchers may be working with data that have never been matched before, and in other situations, data may have been matched in one or more prior research projects. The effort it takes to determine data and matching quality for first-time matching projects can be significant. It is good practice for all researchers to verify that the data they have adequately represent the population of interest, and are of sufficiently high quality to use for the analysis of interest. This section of the report describes approaches to conducting both internal and external data validation for VLDS data. The process may begin at selection, for example, if two agencies collect similar data elements (see Section 2.2 of this report). However, a significant effort will also be required after the data are merged, cleaned, and structured for analyses.

4.1 Consider source data quality

Individual agency staff typically understands data quality for their own agency's data. It is important for VLDS users to work with their project sponsor to learn about each agency's perspective on data quality for the elements needed for research. A first step is to learn about whether key data needed for research are most likely to be of the highest quality. These include data that are:

- Used for fiscal purposes, such as elements that determine eligibility for benefits, and those that determine funding levels and decisions.
- Subject to regular auditing, or other types of monitoring and validation, such as data from the Workforce Investment Act .
- Used in public reporting, such as results of public school (K12) assessment results and graduation rates and IHE retention and graduation rates.

Researchers might also want to consider the data collection method. For example, are the data collected via:

[8] As described previously, these identifiers are hashed before leaving each agency and again before data are returned to authorized users, and therefore, no personal information is shared between agencies or with authorized users.

- Individual self-report (e.g., race/ethnicity in college; prior credentials earned for workforce participants)
- Original source records, often subject to local audit (e.g., school-, IHE- , workforce program credentials awarded; program administrators enrollment or participation records)
- Program administrator report (e.g., VDOE's graduate plan ID)

In general, self-reported data are considered the least reliable, although some types of self-report data, such as customer satisfaction data, cannot be appropriately collected using other methods. Data from original sources, particularly those that are audited or otherwise subject to monitoring and validation at the local level are considered high quality, perhaps as high as data that are specifically used to make funding decisions. Finally, program administrator reports can be highly variable. Agency staff can typically provide guidance about data quality for specific elements from the agencies.

Another consideration is whether the data represent official or unofficial records. For example, VDOE's state assessment data are based on official records collected through the state testing program. However, VDOE and SCHEV course participation and grades are not official student transcripts. These records may be missing important information that is included on transcripts, such as previously earned credits (e.g., through Advanced Placement or out-of-state dual credit courses), and information about courses that were, by local or institutional policy, excluded from transcripts.

Finally, there are some data sets included in VLDS for which agency staff have limited information about data quality, particularly when merged with other data. Use of VLDS, therefore, provides an opportunity for state agency staff to learn from researchers who use the data, and where appropriate, implement approaches to improve data quality within the agency.

4.1.1 Internal consistency

Once researchers have merged data sets, it is important to invest in assessing the internal consistency (reliability) of the data. While each agency data set may be highly accurate and complete for its own purposes, the merged data may not be comprehensive. VLDS users might consider using data elements that are similar across agencies to learn about the match quality. For example, researchers might assess correlations between students age at certain times, gender, locality, race/ethnicity. In some cases, these data elements are part of the probabilistic matching algorithm (e.g., birth month and year) and therefore, should be near perfect. Perfect correlations even among variables used in the matching process are rare. This is because the matching process is conducted based on a point-in-time matching, where individual records are collected over multiple time points where data may have changed.

In some data sets, researchers can assess internal consistency by directly comparing match rates for similar data collected using different methods. Appendix B shows the results of matching records from Virginia high school graduates to two different sources: SCHEV records and records from the National Student Clearinghouse. Both methods use probabilistic matching algorithms. The match rates for enrolled students are similar, although overall, NSC's method is matching a larger number of students as enrolled and earned credentials than VLDS. As well, the information shows that each method matches some unique students relative to the other (i.e., there are matches with SCHEV not matched by NSC, and matches with NSC not matched with SCHEV). Some of the differences are likely attributable to the matching algorithms. However, NSC also suppresses some student-level data returned to VDOE at the request of IHE or students whose data they collect. This may account for a sizable percentage of the students matched with SCHEV but not NSC. Other reasons for discrepancies are briefly discussed in Appendix B.

> One researcher using VLDS realized during the data validation process that one important data element was missing. The researcher re-requested data only from the agency that housed the missing data element. These internal validation checks quickly reminded the researcher of the VLDS process that requires you to re-request all data—because data sets cannot be concatenated (see section 3.1).

4.1.2 External validity

In addition to conducting internal consistency checks, it is important for researchers to validate VLDS data against other data sources. A good starting point is to compare VLDS data to agency-published reports and those from relevant external sources. For example, via their web sites, all state agencies report a substantial amount of data; VDOE and SCHEV report enrollment and credential data for Virginia's public schools and IHE, and VEC reports on a variety of economic and labor indicators. Researchers can review published data (from agencies and external sources that provide data to agencies) to determine whether data sets are complete and that analyses are consistent with agency data use.

The most significant challenge but also a critical step in validating the data may be to determine the validity of the merged data sets. VLDS team members recognize that if data sets were readily available to study priority issues, agency personnel would already be using them. As such, it is difficult and perhaps impossible to get completely accurate information about the result of cross-agency data merges. Nonetheless, in many situations, researchers, policymakers, and others have already developed estimates of key outcomes of interest that researchers can use to compare merged outcomes. For example, a variety of publications exist to estimate high school graduates' college enrollment at the state level (e.g., NACHEMS) that can be used to verify VDOE to SCHEV matches. In addition, researchers have concluded that certain demographic groups are often under-represented in probabilistic matches between high school and college (Dynarski, et al., 2013; Holian & Mokher, 2011). The Bureau of Labor Statistics reports estimates of annual employment and unemployment numbers by locality, which may be helpful when validating employment data.

As of this writing, researchers have explored relatively few data sets after merging across Virginia's agencies. The more the data are used, the more we all can learn about the match quality, and in general, strengths and limitations of merged data sets.

5 Summary and recommendations

This report provides the research community with important information that will help them determine whether VLDS is an appropriate data source for their projects as well as information about the data to help inform project staffing, budgets, and timelines. In preparing this report, we identified a number of limitations associated with VLDS use, and, important ways that researchers and VLDS partner agencies can learn from each other's expertise. In preparing this report, our team developed the following recommendations for VLDS consideration:

- For each project, particularly those in which data sets are matched for the first time, consider requiring researchers to provide agency staff with information about match quality, specify limitations, and make recommendations for improving existing data quality (e.g., via strengthening internal validation checks during the data collection process). VLDS agencies can then use this information to guide agency-specific improvements, working on those areas that are of highest priority to the agencies as resources permit.
- For individual research projects, consider whether researchers should provide data management or statistical code as part of project deliverables. This information can help the agencies and other researchers replicate (or identify concerns with) project results.
- As VLDS expands, and more data are added, update this document—or specific sections—to ensure that researchers have a meaningful document to review before using VLDS.
- Consider developing a more comprehensive list of data changes that are not obvious in the data sets or critical change lists (see section 2.1).
- It might also be possible to create a social media site or other sharing mechanisms for VLDS researchers to share information gained from agency staff about the data. This could minimize staff burden while maximizing information available to authorized users.

6 References

Dynarski, S.M., Hemelt, S.W., & Human, J.M. (2013). *The missing manual: Using National Student Clearinghouse data to track postsecondary outcomes, working paper 19552.* Cambridget, MA: Natonial Bureau of Economica Research. http://www.nber.org/papers/w19552.

Garland, M., LaTurner, J., Herrera, A.W., Ware, A., Jonas, D., and Dougherty, C. (2011). *High School Predictors of College Readiness: Determinants of Developmental Course Enrollment and Second-Year Postsecondary Persistence in Virginia.* Richmond, VA: Virginia Department of Education. http://www.doe.virginia.gov/instruction/college_career_readiness/research/high_school_predictors_of_cr_in_va_2011.pdf.

Holian, L., and Mokher, C. (2011). *Estimating College Enrollment Rates for Virginia Public High School Graduates.* Issues & Answers Report, REL 2011-No.104. Washington, DC: U.S. Department of Education, Institute for Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Appalachia. http://ies.ed.gov/ncee/edlabs

Jonas, D., Dougherty, C., Herrera, A., LaTurner, J., Garland, M., and Ware, A. (2012). *High School Predictors of College Readiness: Determinants of High School Graduates' Enrollment and Successful Completion of First-Year Mathematics and English College Courses in Virginia.* Richmond, VA: Virginia Department of Education. http://www.doe.virginia.gov/instruction/college_career_readiness/research/determinants_of_enrollment_and_completion_of_english_and_mathemathcs.pdf.

Lichtenberger, E., Dietrich, C., Kamulladeen, R., and O'Reilly, P. (2010). *Postsecondary Enrollment: Summary Phase I.* Richmond, VA: Virginia Department of Education. http://www.doe.virginia.gov/instruction/college_career_readiness/research/ps_enrollment_phase1_summary_.pdf.

Appendix A: Agencies data structures in VLDS

At this time of this writing, VLDS Data included over 750 data elements.[9] These data elements are organized by Participating Agency and usually further organized by the source or type of data. Not all data are available for all years. In general, data from SCHEV are available from 2006 forward and VEC from 2005 forward. VDOE's data system has undergone significant change in the past decade. VDOE has provided data that vary in their start year. For example, state assessment and demographic records are available beginning with the 2005/06 school year; student schedule data in 2011/12.

The information below is an overview of the data each partner agency has made available within VLDS. It is not all-inclusive and not all data are available for every year. As well, VEC and the Virginia Community College System (VCCS) are in the process of exposing additional data to VLDS. These agencies will add data with their own data structures to the system. The VLDS data dictionary contains specific details about available data elements.

A.1. SCHEV data

SCHEV organized VLDS data into four principle areas:

- Course enrollment;
- Degrees conferred;
- Fall Cohort; and
- Financial Aid.

Within these principle areas, VLDS provides authorized users with access to data such as:

- Student characteristics such as gender, location of domicile, race/ethnicity, tuition status;
- Student status such as visa-status and college majors;
- Student financial aid status and awards;
- IHE enrollment;
- Courses enrollment by semester/quarter, course credits available, and grades;
- Credentials earned and;
- The data also includes is a significant amount of institution information, course information, aid information, family information, in state status and tuition status information which a researcher may receive for each student being studied.

A.2. VDOE data

VDOE organized the VLDS data into four principle areas:

[9] The number of available elements will change as new agencies join and existing agencies modify data sources that are included in VLDS.

- Test Results (including ACT Test Results, AP Test Results, PSAT Test Results, SAT Test Results, State Assessment Results);
- National Student Clearing House Records;
- Student Records; and
- Student Schedule.

Within these principle areas researchers can receive student data by school year such as:

- School and school division;
- Demographic information (e.g., gender, ethnicity, birth year);
- Student characteristics and program participation (e.g., limited English proficient; student disability status; gifted program participation; GED program participation; career and technical education program information);
- Specific assessment information (e.g., content-specific scores; proficiency levels; test type; test location);
- State-developed variables (e.g., graduation year); and
- Student outcomes status (e.g., diploma type).

At the time of this writing, VDOE was working on adding data from adult education programs to VLDS, which may result in additional principle areas. As well, VDOE is in the process of adding Phonological Awareness Literacy Screening (PALS) data to VLDS. VLDS team members expect these data to be available in 2014.

A.3. VEC data

VEC initially exposed a limited amount of wage records data from the current unemployment insurance data collection. Specifically, VEC exposed:

- total wages earned per quarter;
- quarter wages were earned; and
- calendar year.

VEC does not provide any other information about the wage earner. For example, VEC does not provide information about the employer's industry or the wage earners' employment location. VEC is currently working to expose data from three programs to VLDS: Wagner-Pyeser, Unemployment Insurance Benefits, and Trade Act Assistance. Data from these programs are expected to be available in 2014.

Appendix B: Comparison of match rate for VLDS and National Student Clearinghouse data

As part of Virginia's College and Career Readiness Initiative, researchers are using VLDS to follow high school graduates into college for four years. The project will document enrollment and persistence in college, and assess the association between high school outcomes and college enrollment, persistence, and credentials earned for Virginia's public high school graduates.

The research team is using two primary data sources to track students' enrollment, persistence, and credentials earned in college: SCHEV data and data from the National Student Clearinghouse. As part of the project, the research team assessed the match rates using these two different data sources. This appendix shows the results of the team's initial analysis aimed at understanding match quality.

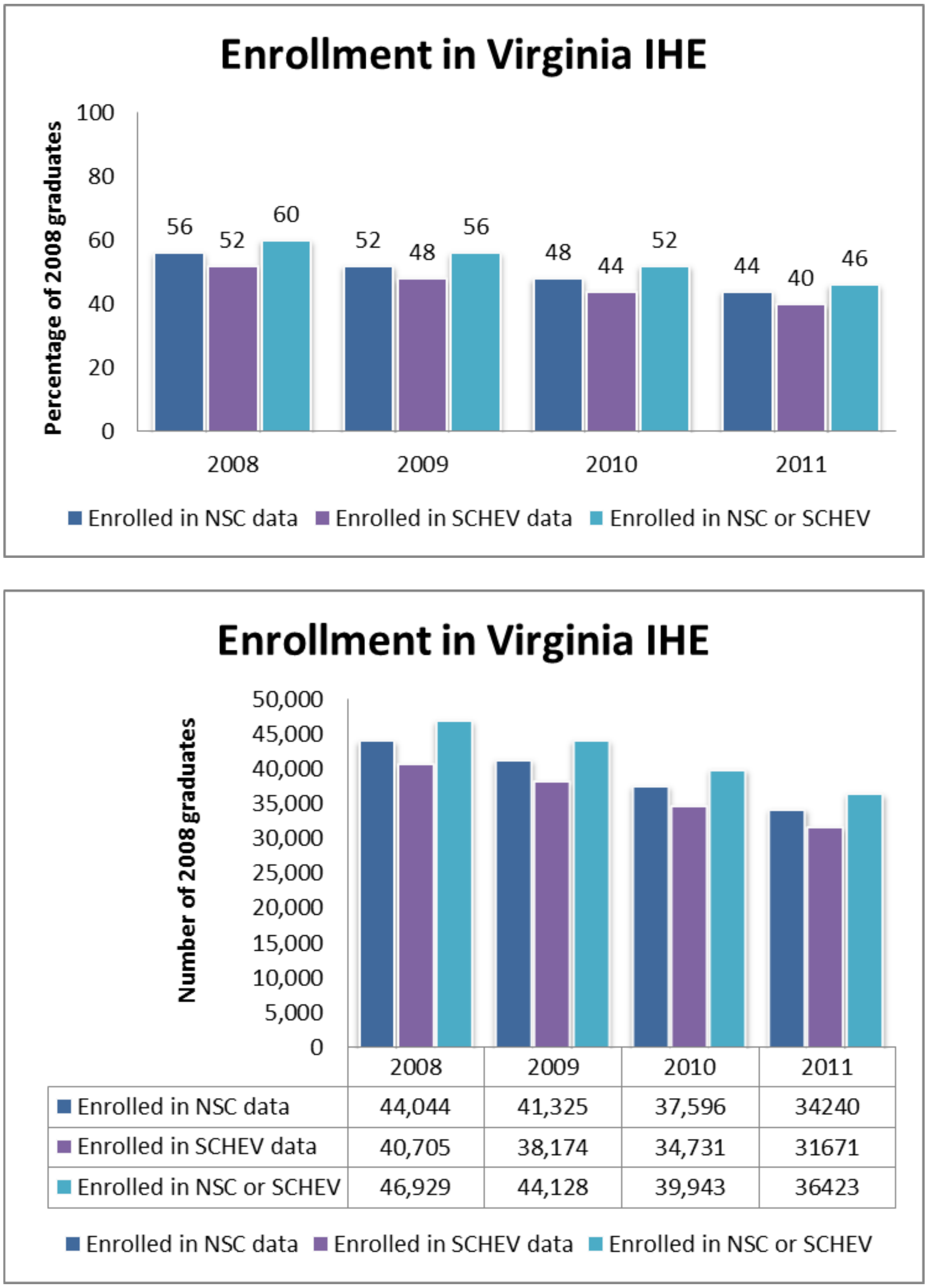
The following graphs show the number and percentage of the 2008 cohort of high school graduates who were matched as enrolled or earned credentials from a Virginia IHE. The college enrollment data provide a snapshot, meaning that the research team counted all students who were enrolled in a given year regardless of when they first enrolled or whether they persisted. The data show the results of students found in the NSC matching method; VLDS matching method; and, students who were enrolled in at least one of the two data sets (mathematical union of SCHEV and NSC matches).

In all cases, the union of SCHEV and NSC matches yields a higher number of students than using either data set alone, even when the records were limited to students who were enrolled in Virginia IHE. After identifying this discrepancy, the team conducted analyses to determine whether there were systematic differences in the students who were matched in the different data sets, and to better understand the reason for the discrepancy.

B.1 Postsecondary enrollment in Virginia IHE

Results from postsecondary enrollment matches in Virginia IHE, across all IHE types (i.e., two-year, four-year, less than two-year), show that NSC matches four percent more students each year relative to SCHEV enrollment data. NSC matched 3,339 more students in 2008 compared to SCHEV; the numbers drop some each year as overall enrollment decreases, such that NSC matched 2,569 more students in the 2011/12 school year. When student enrollment based on matches from NSC and SCHEV are combined, such that students are included in analyses if they are found in at least one data set, we end up with the most students matched. These results, by year, are shown in Figure B-1.

Figure B-1. High school cohort, 2008, postsecondary enrollment in Virginia IHE, 2008 through spring 2012

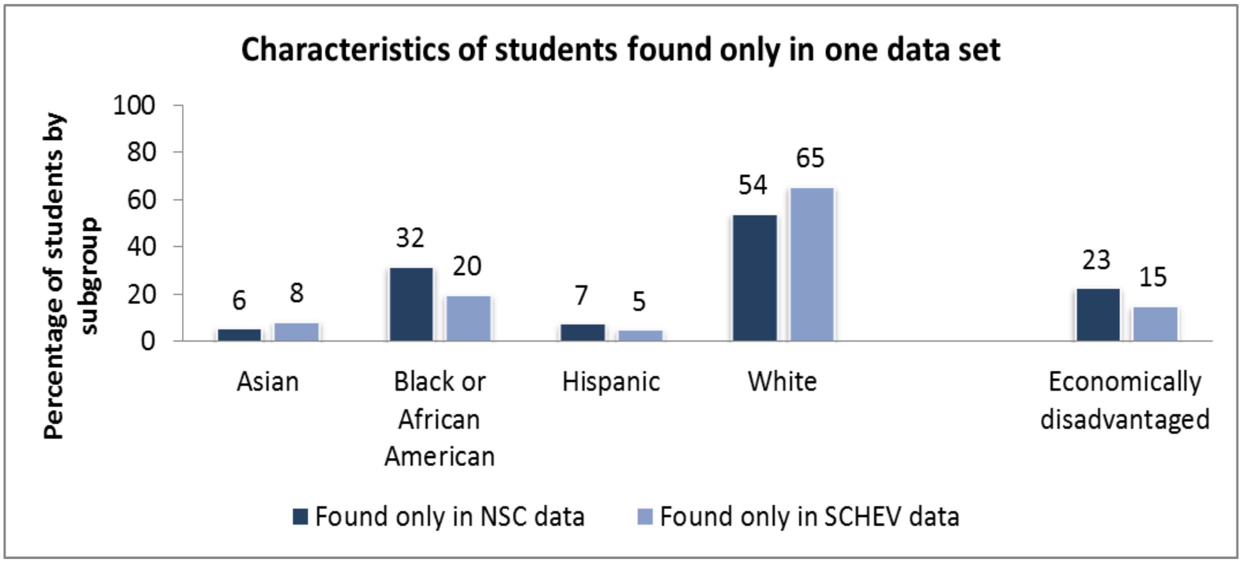


	2008	2009	2010	2011
■ Enrolled in NSC data	44,044	41,325	37,596	34240
■ Enrolled in SCHEV data	40,705	38,174	34,731	31671
■ Enrolled in NSC or SCHEV	46,929	44,128	39,943	36423

There are several potential reasons for the differences in the enrollment data when matching with two different sources. First, NSC includes private IHE in Virginia that are not part of the SCHEV data collection. This includes technical schools such as ECPI and ITT Technical, as well as other institutions. Second, SCHEV also includes some IHE that are not included in the NSC data set. Finally, we hypothesize that the two matching algorithms have some differences that impact results. The NSC algorithm is proprietary, and therefore, there is no way to learn whether the differences are related to the matching process itself.

Knowing there are differences, it is important for researchers to examine ways that limiting a data set to one population (e.g., SCHEV only) systematically differs from more complete enrollment data. For example, Figure B-2 shows that students found in NSC and SCHEV data sets differed by race/ethnicity. The SCHEV-matched data resulted in proportionally more white and fewer economically disadvantaged students relative to the NSC data. Researchers continue to investigate the cause of these differences, estimate the impact of the students not matched in a data set used, and appropriately qualify conclusions based on these types of data limitations.

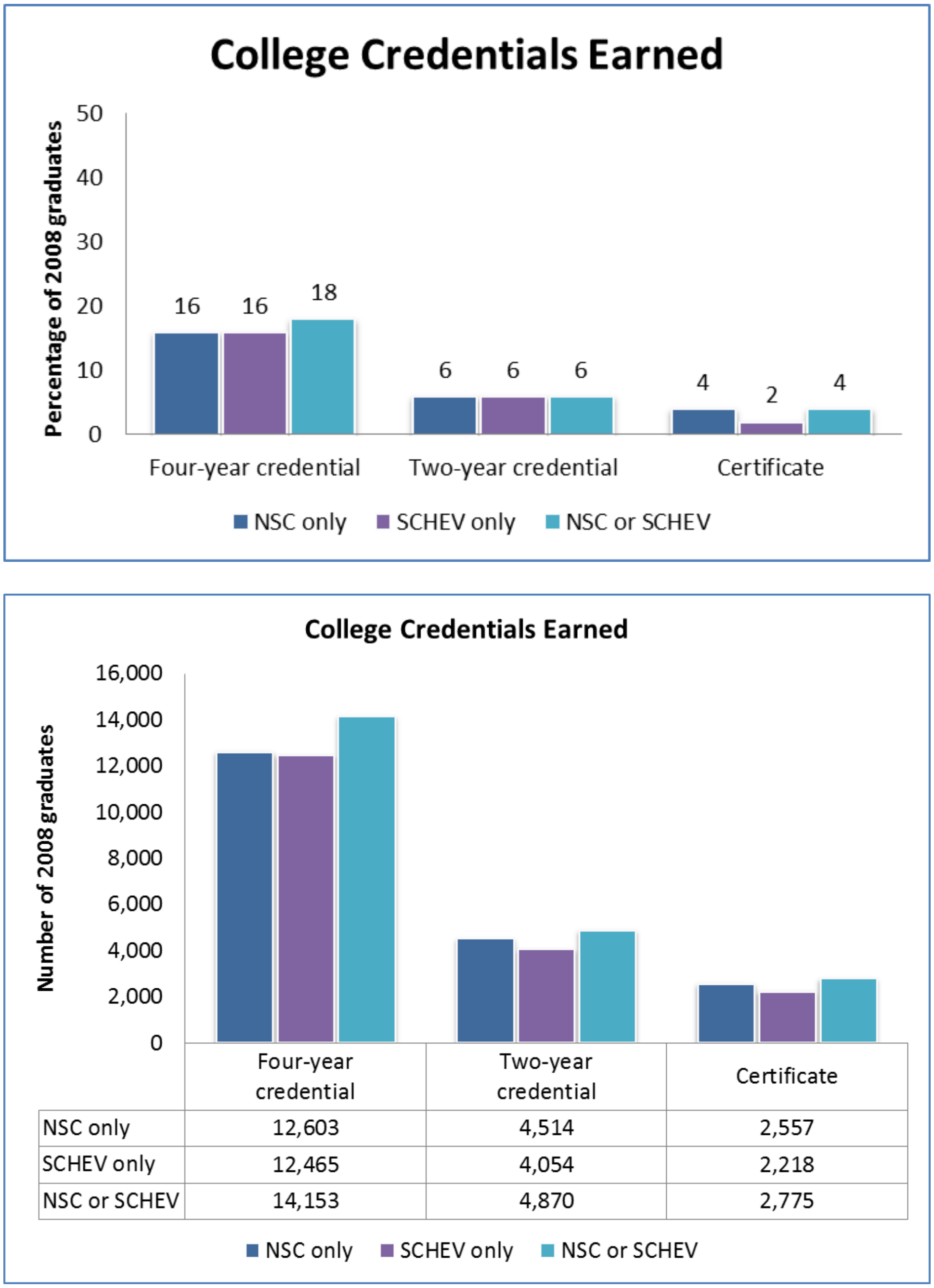
Figure B-2. Percentage of students found enrolled in NSC or SCHEV data by race/ethnicity



B.2 Postsecondary credentials students earned from Virginia IHE

The research team is also examining differences in credentials earned. As shown in Figure B-3, NSC data results in a larger number of students earning credentials compared to SCHEV data; researchers obtain the highest yield from using both NSC and SCHEV data. These differences may result from the same factors identified as contributing to differences in enrollment numbers. However, the research team determined that the majority of the discrepancy in the number of credentials earned resulted from the timing of credential interacting with data availability. The NSC data for this project were available through the summer of 2012, and SCHEV data were only available through the spring 2012. The majority of credentials that were only documented in the NSC data set were issued in the summer of 2012—a time-period not available from SCHEV for the project. Researchers who encounter similar differences in the data can choose to limit the sample (based on the data VLDS provides) so that SCHEV and NSC data are from the same time period.

Figure B-3. Number and percentage of college credentials earned in Virginia IHE



	Four-year credential	Two-year credential	Certificate
NSC only	12,603	4,514	2,557
SCHEV only	12,465	4,054	2,218
NSC or SCHEV	14,153	4,870	2,775

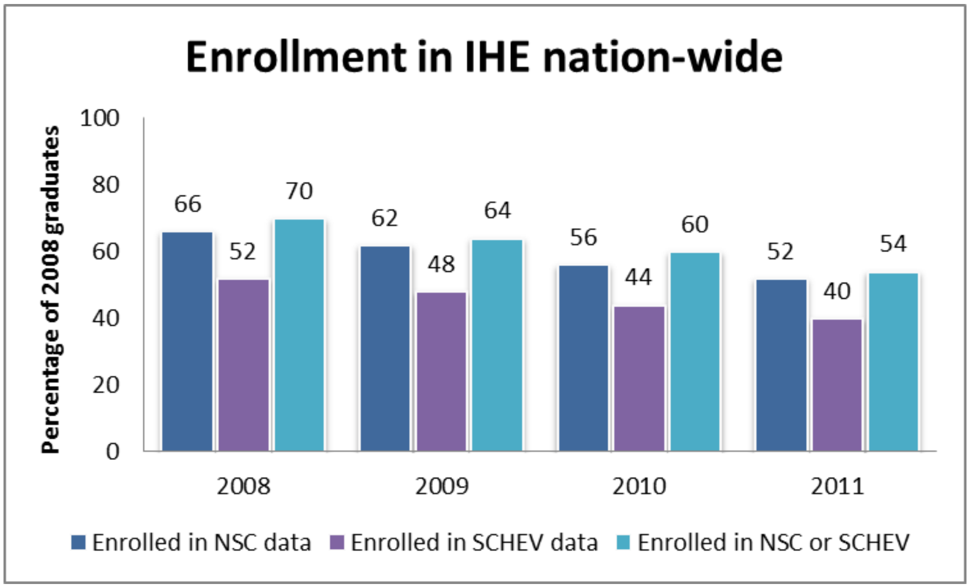
B.3 Additional analysis needed to understand the match quality

The CCR research team has determined that the combination of NSC and SCHEV data[10] yields the most comprehensive data set to use for enrollment, persistence, and credentialing analysis. In

[10] The combination refers to the mathematical union.

addition to the data shown here, the research project will include data available in the NSC data set from high school graduates who enrolled in IHE outside of Virginia. As shown in Figure B-4, this results in a larger percentage of students identified as enrolled in college—70 percent of high school graduates in the 2008 cohort.

Figure B-4. High school cohort, 2008, postsecondary enrollment in IHE in and outside of Virginia, 2008 through spring 2012



NOTE: SCHEV only collects data from Virginia IHE; NSC data include in-state and out of state IHE.

Understanding whether there are systematic differences in the students who are matched differ using the NSC and SCHEV matching methods is important for understanding the impact of these differences on statistical estimates and the resulting inferences that researchers can make using these data. The research team is continuing to conduct analyses to better understand the differences. Some possible reasons for discrepancies:

- Students identified only in the SCHEV data were enrolled in IHE that do not participate in NSC services, and therefore could not be included in the data set.
- Students enrolled in Virginia IHE and identified only in the NSC data were enrolled in IHE that do not report to SCHEV, which are private for-profit institutions.
- The NSC and VLDS matching algorithms produce different results for some students.
- Errors in probabilistic matching.

The research team can also learn more about the match quality by:

- Further investigating systematic differences between the uniquely matched students in each data set (e.g., high school from which they graduated; college attended; high school achievement).
- Comparing the results to known enrollment statistics in Virginia, such as those available on the SCHEV website. These data will include high school graduates from outside the

Virginia public school system, and thus, will not provide perfect information. However, they can help determine whether the biases that researchers have determined exist in the NSC data also exist in the high school data matched to SCHEV data.

- Determining how the above factors influence inferences made from statistical models applied to these data.