HERC Longitudinal Database
Frequently Asked Questions

These FAQ's are intended to answer common questions about the HERC longitudinal database. They serve as a supplement to other HERC database documentation, including codebooks and variables index. Please see those documents for more detailed information on these questions, as well as applicable supplemental documents, as noted below.

ONLINE SERVER

1. How can I access the longitudinal database?

   The longitudinal database is stored on a secured, online server. Users must first complete all requirements to be approved by HERC and Rice IRB. Once approved, Rice IT will assist the user with setting up their computer to access the online server and will provide an e-token. The e-token is a USB device that must be plugged into the user's computer every time the user logs into the server. Along with the user's password, the e-token represents "two-factor authentication" that is required for the HERC security system.

2. How is the online server set up?

   Each user will be assigned to a virtual computer on the server. At least two users can log on to a virtual computer at one time. Sometimes more than two users are assigned to a virtual computer (often, the users are part of the same project), but only two of the users can be actively logged on at one time. Users are encouraged to negotiate among themselves to determine when to log in, so that all users on the same virtual computer can have reasonable access to the server. **NOTE: FOR SECURITY PURPOSES, THE ONLINE SERVER IS NOT CONNECTED TO THE INTERNET.**

3. How do I log on and log off of the online server?

   Users will be set up and trained by Rice IT for first-time use. From off-campus, users must first connect to the Rice network through VPN. Once VPN is connected, users can connect directly to the HERC online server through Remote Desktop Connection. Users must connect to their assigned server, and must plug in their e-token before entering their password. Users will connect to a virtual Windows desktop with Stata 13 and Microsoft Office available.

   HERC recommends that users click the Log Off button at the bottom left of the screen to close out their session when leaving the online server. This way, space is not taken up on the virtual computer, and an additional user can log on if necessary. Users intending to leave Stata on for long-running models can click out using the tab at the top of the screen (or use the Disconnect option on the lower-left button), but this leaves the session open.

4. How can I access the longitudinal database on the online server?

   The longitudinal database is actually a set of data files, stored on the S: drive (IT will map this drive on their computers, giving users a dedicated link to the server). Student-level data are stored in one longitudinal data file, based on data from multiple sources and covering six years. In contrast, campus-level data are stored in multiple files, separated by source and year, since students can change campuses from one year to the next. Teacher data are still in process, so cleaned teacher data are not yet available for all HERC users.

   On the server, the student-level data are in the file S:\original data\herc.student.full.08-13.dta. Campus data are placed in the folder S:\original data\campus data, with sub-folders for each data source: campus aggregated data, CCD, magnet program, and TEA school profiles. Other data files are also placed in the S:\original data folder, which may be used or merged with other data as needed. These files have generally been provided to users upon request, but normally have not been cleaned or merged into the student or campus data files. See *HISD Longitudinal Database: Student Codebook Intro* and *HISD Longitudinal Database: Campus Codebook Intro* for more information on the data files.

5. How can I analyze data on the online server?

   Users will use statistical software stored on their C: drive associated with their login to analyze data. HERC recommends that users make an extract file with the variables necessary for their particular research project, and store this extract on their C: drive, along with any .do files and .log files. **USERS ARE PROHIBITED FROM MOVING DATA FILES FROM THE C: OR S: DRIVES TO ANY OTHER COMPUTER OR DATA STORAGE DEVICE OR FACILITY.**

6. What statistical package can we use to analyze the database?

   HERC provides data in Stata format, and all users are provided Stata 13 on the server. SAS has been provided to some users on an as needed basis. If you would like to have SAS installed on your online server, please contact HERC staff at herc@rice.edu. Note: HERC staff cannot provide statistical support for SAS users.

7. How can I store/share data, code, etc. with project team members?

   Users can have shared accounts created in the S:\shares folder. Users can email Rice IT at helpdesk@help.rice.edu. Please include the names of all approved users who should have access to the folder. Feel free to suggest a name for your folder. Users may not store or change files on the S:\original data or S:\documentation folders.

8. How are files backed up on the server?

   Files on the S:\ drive are backed up nightly, and the backups are stored for two weeks. Rice IT cannot be responsible for files that are older than two weeks. Users are encouraged to create their own backups of documentation, .do files, or .log files on their own computer systems. **USERS ARE PROHIBITED FROM BACKING UP DATA FILES TO ANY OTHER COMPUTER OR DATA STORAGE DEVICE OR FACILITY.**

LONGITUDINAL DATABASE: GENERAL

1. What is the longitudinal database?

   The HERC database includes data from sources provided by HISD, the Texas Education Agency (TEA), and the Common Core of Data (CCD) from the National Center for Education Statistics (NCES). The dataset spans six years, from the 2007-08 school year through the 2012-13 school year, and it covers the population of students and teachers in HISD schools. The data include student-level background and program participation (PEIMS), testing and achievement data (STAAR, TAKS, Stanford, Aprenda), identification of enrolled school and zoned school, magnet student enrollment, as well as school-level demographics, magnet school identification, financial data, accountability benchmarks, and more (CCD and TEA). Additional data, including teacher data, will be added as they are cleaned and organized.

2. How is the database organized?

   The HERC longitudinal database actually consists of multiple data files. Because the HERC database includes data covering students, teachers, and campuses, the data are separated by unit of analysis. Student data from all sources and years are cleaned and merged into one longitudinal data file. In contrast, the campus-level data are separated by year and by source, since students and teachers can change campuses between years. Teacher data are compiled into one longitudinal data file where possible; other teacher data which may contain multiple observations per individual (such as teaching certifications), may be organized in long form in a separate data file. See codebooks for each data type for more information on how the data are organized.

3. How can users know the source of a particular variable?

   Variables follow a naming convention: the prefix (one or two characters) indicates the data source, and the suffix (two numbers) indicates the year, based on the spring semester of the

academic year (e.g., 08 for the 2007-08 academic year).

ae	Aprenda environment
al	Aprenda language
am	Aprenda math
ar	Aprenda reading
as	Aprenda science
at	Aprenda social studies
c	TEA school profiles
ca	Campus aggregated variables (Based on PEIMS)
ec	CCD school universe
hz	Enrolled/zoned file
ms	Magnet student
mp	Magnet program
p	PEIMS
re	STAAR EOC
rm	STAAR math
rr	STAAR reading
rs	STAAR science
rt	STAAR social studies
rw	STAAR writing
se	Stanford environment
sl	Stanford language
sm	Stanford math
sr	Stanford reading
ss	Stanford science
st	Stanford social studies
tm	TAKS math
tr	TAKS reading
ts	TAKS science
tt	TAKS social studies
tw	TAKS writing
fl	Tracking or analytic sample flag
l	Merged longitudinal data file

4. How do I know which cases have valid data on the variables in the database?

   Whether student, campus, or teacher-level data, HERC has created a number of flag variables to help researchers determine whether they have valid data from a particular data source or set of data sources. HERC recommends using these flags to select variables for data extract files that are specific to their research projects.

   HERC created two types of flag variables. Tracking flags for each source indicate the total number of observations in the data file after deletions were made but with no further restrictions. Sometimes observations were deleted from the cleaning process (duplicate

observations, missing or invalid id, etc.). In addition, for some data sources, we created analytic sample flags with some suggested restrictions that will allow researchers to create the most valid samples that can best be tracked over time. Tracking and analytic flags were created for each individual data source at the time of data cleaning. In addition, for the student data, we created tracking and analytic sample flags for each yearly merged file, before they were merged into one, longitudinal data file. HERC recommends that researchers combine the relevant flags as needed to create longitudinal samples for their specific research projects. For each unit of analysis, the Data Deletions and Analytic Samples section in the *HERC Longitudinal Data Codebook: Introduction* provides the complete list of data flags.

5. How can I combine data files in Stata?

   The student-level data are already joined in one longitudinal data file with data from multiple sources, while the campus-level data are in separate files by year and data source (TEA, CCD, etc.). A *merge* command joins additional variables from the using data set to the same observations in the master data set, matching on one or more key id variables. In contrast to *merge*, the *append* command joins additional observations from the using dataset to the end of the master dataset.

   Merge. Users may want to combine student-level data with one of the many campus-level files. In Stata, users will first want to read in the student data, either using the full student file or a pre-made data extract:

   use "S:\original data\herc.student.full.08-13.dta"

   The *merge* command specifies the type of data merge, in this case, a *many-to-one* merge that joins multiple individual students to a single campus in which they were enrolled in the given year. Students are joined to campuses based on the campus id campusYR that will be in both the student file and the specific campus file. Thus the *merge* command to merge CCD data from 2011-12 to the student file would be as follows:

   merge m:1 campus12 using "S:\original data\campus data\ccd\ccd.scun.12.dta"

   If users are starting with campus-level data to which they are merging student data, the merge command would be a *one-to-many* merge:

   merge 1:m campus12 using "S:\original data\herc.student.full.08-13.dta"

   Combining campus-level variables across data source, but in the same year, would be a *one-to-one* merge. For example, starting with the CCD 2011-12 data:

   merge 1:1 using "S:\original data\campus data\magnet program\hisd.magp.12.dta"

   Append. Users who wish to combine campus-level data files across years will want to use the *append* command, since these files will not have a common id variable from the same year

(e.g., campus12). Essentially, users will be combining observations from different years. It is expected that the data files will have the same variables, but this is not required; missing values will be generated for all observations from the master data set that do not have matching variables in the using data set, and vice versa. For example, the command for adding CCD 2010-11 data to CCD 2011-12 data would be:

append using "S:\original data\campus data\ccd\ccd.scun.11.dta"

Researchers should use the *help merge* and *help append* commands in Stata for more information and examples of these Stata commands.

STUDENT DATA

1. What is PEIMS? Why is this data source so important?

   The Public Education Information Management System (PEIMS) is a database of student background, demographics, and program participation, collected each academic year and reported to the state and federal government. Thus, the PEIMS is a reliable indicator of all students enrolled in HISD, as recorded on a particular day in October of each year. See *PEIMS Variables: Table of Contents* for more information.

   Since the PEIMS includes basic demographic and background information on all students enrolled at the PEIMS date, and the data are verified and reported the federal government, HERC uses PEIMS as a master data file to which all other data sources were merged when creating the longitudinal database. HERC recommends using PEIMS as an indicator of a minimum level of valid information available for each student.

2. How can I track students in the database?

   The student data are organized in wide form; that is, each observation corresponds to one student who has multiple variables repeated for each year, with a suffix indicating the year (08 for 2007-08, for example). Each student has a unique student ID number (id). All student-level data are merged to the PEIMS based on student ID. There are also year-specific copies of the student ID variable (id08, id09,…,id13), which can indicate whether the student has valid data from at least one data source in the specific year; otherwise, the variable should have a blank missing value. In addition, there are year-specific student ID variables (pid08, pid09,…,pid13) from the PEIMS file, which can indicate the availability of PEIMS student background information in each year. See the **Data Structure** section in the *HERC Longitudinal Data Codebook: Student Data Introduction* for more information.

3. The number of students with valid id variable for a particular year (id08) is greater than the number of students with PEIMS data in that year (pid08). How can this be?

The PEIMS data include all students enrolled in HISD on a particular day in October of each year. Additional students could be in the HERC database because they enrolled in HISD after the PEIMS date, they left HISD before the PEIMS date, or other reasons. Therefore, a student could have data from other sources (and thus could have a valid idYR variable), while having missing data from PEIMS (and thus a missing pidYR variable). Because PEIMS includes important background information for each student, HERC recommends that researchers limit analyses to students who have valid PEIMS data, if possible.

4. How can I determine the number of years that a student is in the longitudinal database? How can I track students over time?

   Although the longitudinal database includes data for six years, most students are not in the database for all six years. Each year students leave due to graduation, moves, or other reasons, and each year new students are added. HERC created several variables that researchers can use to count the number of valid observations for each student in the database. The variable lctid is simply a count of the number of years with a valid student id idYR; a nonmissing student id means that there is data in a particular year from at least one data source. The variable lctpeim similarly counts up the number of years with valid PEIMS data. As stated above, HERC recommends using PEIMS as the master data file, so lctpeim should be a count of the number of years with a minimum of background data on each student in the database. We also created several variables counting up the number of years with valid analytic sample flags from the yearly merged files, depending on the testing outcome variable, with the prefix lct*. For example, lctflsm is a count of the number of years with a valid analytic sample flag for Stanford math (flmansmYR).

5. How can I use the database to analyze student achievement over time?

   The longitudinal database allows for analyses of cohorts of students across several years. However, the possible cohorts differ by test, grade level, and subject, due to the administration schedules of the TAKS, Stanford, and Aprenda tests. For example, the TAKS was given to students in grades 3-11 from 2007-08 through 2010-11; the TAKS began to be phased out in 2011-12, so it was only given to grades 10-11 in that year. Thus, researchers need to understand the test subjects, grade levels, and years for which valid data are available for a particular cohort of students. See the Data Structure, Longitudinal Data Structure section in the *HERC Longitudinal Data Codebook: Student Data Introduction,* and *Longitudinal Data Files: Intersection of Year and Cohort* for more information.

6. What test score variable should I use from the TAKS data?

   HERC recommends using the TAKS scale score variable, as opposed to the raw score variable, as the appropriate indicator of student achievement. However, there are actually two different types of TAKS scale scores. The vertical scale score can be used to measure growth

in student achievement over time, while the "regular" or horizontal scale score can be used for direct comparisons of performance within grade and subject, and for indicators of whether the student "met standard" or achieved "commended performance" based on state standards. Students in grades 9-12 only have horizontal scale scores; vertical scale scores are only available for students in grades 3-8 in each academic year, although students in these grades also have horizontal scale scores in the 2007-08 and 2008-09 academic years only.

For each academic year, the HERC longitudinal database includes a combined scale score variable (trvrscYR and tmvrscYR); the value for students in grades 3-8 is a vertical scale score, and the value for students in grades 9-12 is a horizontal scale score. Vertical scale scores are considered by some to be the best option for some growth or score gain models (see Jorgensen 2004). For 2007-08 and 2008-09, the database also includes separate variables of horizontal scale scores (tmrsc08,tmrsc09, trrsc08,trrsc09), and vertical scale scores (tmvs08,tmvs09, trvs08,trvs09), in addition to the combined score variables. While these variables are included in the database for completeness, HERC recommends using the combined scale score (trvrscYR and tmvrscYR) for longitudinal analyses examining achievement growth, and restricting models to grades 3-8 so that the combined score used is actually a vertical scale score. See the Data Sources, TAKS section in the *HERC Longitudinal Data Codebook: Student Data Introduction* for more information.

7. What are the variables beginning with l?

   These are variables that were created in the merged longitudinal file, after all yearly files were merged together. These variables are labeled with an l prefix to indicate that they did not come directly from a source such as Stanford or PEIMS, but instead were constructed by HERC based on longitudinal data. These variables include counts of valid years of observations, recodes or combinations of variables from multiple sources, and variables that were cleaned to deal with inconsistencies across years. See the Cleaning the Longitudinal Database section in the *HERC Longitudinal Data Codebook: Student Data Introduction* for more information.

8. What are the lfemaleYR, lethnicYR, and lgradeYR variables? How are they different from similar variables from other sources, such as PEIMS?

   Generally, HERC recommends using the PEIMS variables (prefix p) in instances where there are multiple measures of the same variables from different sources. However, in some instances, data from PEIMS are not consistent in ways that we might expect. For example, reports of student gender and ethnicity are not consistent across years. These inconsistencies could reflect shifting gender or ethnic identities, or they could reflect measurement error. In addition, reports of grade level sometimes included discrepancies from one year to the next. For example, some students were reported with the same grade from one year to the next (pgrade08=8 and pgrade09=8), moving up two grades in one year (pgrade10=7 and pgrade11=9), or other discrepancies (sometimes moving up 5 grade levels in one year, or moving backwards from grade 11 to grade 10). Certainly, some of these could be legitimate

discrepancies if students were retained or accelerated up a grade. But without clear confirmatory data at the time, HERC was challenged to determine whether these grade discrepancies were valid and, if not, to determine what was the "true" grade level for a particular student in a particular year.

The HERC Data Manager undertook to clean these three variables for a particular research project, as reconciling data across years is particularly important for research using longitudinal growth models. As such, HERC created several variables to indicate cleaned, or at least stable, measures of gender, ethnicity, and grade level. For gender and ethnicity, HERC sought to create consistent indicators for all years in the database (based on the first or most common report), and also to include indicators of change or inconsistency for cases that did change their gender or ethnicity reports over time. In the case of grade level, HERC used a strategy of starting with pgradeYR as the variable most likely to be valid, and using grade level reports from testing data files (TAKS and Stanford) to confirm or edit the pgradeYR value, as necessary. See the Cleaning the Longitudinal Database section in the *HERC Longitudinal Data Codebook: Student Data Introduction* for more information.

Constructing these longitudinal indicators required some judgment calls in situations without perfect information; HERC cannot know what students or parents intended to report, only what was actually reported in the administrative data. Users are not required to use these longitudinal, "cleaned" variables, but they are available for use if desired.

Note: These longitudinal variables were constructed for the 2007-08 through 2011-12 data only (N=383,300), as this work was done before the 2012-13 data were available. Thus, researchers should use these cleaned variables only when working with data up through 2011-12. If using 2012-13 data, researchers should use original PEIMS measures. Due to the considerable time investment that went into cleaning these variables, as well as concerns about altering measures that are already in use by researchers, HERC has no plans to "re-clean" the gender, ethnicity, or grade level variables by incorporating any data after 2011-12.

CAMPUS DATA

1. Some of the campus data measure the same construct but from different data sources. How should researchers use the campus data?

   There may be some content overlap between three sources of campus-level data: TEA school profiles, CCD school universe surveys, and campus aggregated variables (based on PEIMS) constructed by HERC. However, the TEA school profiles data have much more detail and many more variables than CCD or campus aggregated data, as it includes accountability indicators across a wide variety of demographic breakdowns. The CCD data are derived from the NCES, rather than from the state of Texas, although they are based on reports made by the state to the U.S. Department of Education. CCD data are mainly useful for counts of students by demographic categories, as well as federal classifications of schools, and for tracking school openings and closures over time.

However, both the TEA and CCD data are only available through 2011-12 at this time. Researchers needing campus-level data in 2012-13 may need to use campus aggregated data as well. Since the campus aggregated data are based on the PEIMS student-level data, which are reported to the federal government, they should align well with the CCD data. In addition, the campus aggregated data may include variables (such as counts of economic disadvantage categories by grade) that may not be available in other data sources. HERC makes no recommendations about which data source to use for specific research projects.

2. What does magnet program description mean? Do some schools have more than one program?

   HISD schools all have only one magnet program; however, the program descriptions are created by each school. When cleaning these data, HERC endeavored to collapse each unique program description into categories, to help researchers group programs that shared common characteristics. We followed the magnet program descriptions in the HISD School Choice Options: 2013-2014 Magnet and Specialty Schools catalog. Some schools had program descriptions that seemed to fit more than one category; for example, Spanish and Technology. In those cases, we counted the school in more than one category; for example, Spanish and Technology would count in both mpprgstemYR and mpprglangYR. The variable mpprgcountYR is a count of the number of categories in which a school's program fits; most schools fit only one, but some fit two categories. These categories are constructed by HERC, not by HISD, and are not intended to suggest that a school has more than one program, simply that a school's program fits more than one descriptive category.