DATA COLLECTION AND MANAGEMENT IN CHILDREN'S SAVINGS ACCOUNT (CSA): THINGS TO CONSIDER

By Megan O'Brien

CSA Symposium Handout

Center on Assets, Education, and Inclusion (AEDI)

University of Kansas

Acknowledgements

Dr. Willie Elliott, Melinda Lewis, Amanda Jones-Layman, Paul Johnson, and Carol Lei provided assistance on this handout, including consultation regarding data management considerations encountered in CSA administration and editorial assistance in the compilation of these insights. Appreciation is also owed to the CSA program and school district personnel who have generously shared their time and expertise as part of ongoing evaluation efforts.

I. Data: Start with What You Have

Data for management and evaluation of CSA programs can come from a variety of sources. All CSA programs will have some type of *enrollment data*; even those programs that utilize automatic enrollment will do so through reliance on some bank of individual data, such as school enrollment records or vital statistics. As interventions built on financial instruments, CSAs also include some *account data*, such as records of transactions, earnings, and accumulation. CSA *program data* can come from activity records of inputs such as outreach materials distributed, presentations delivered, and engagement efforts conducted. CSAs administered through existing institutions, such as schools, may have ready access to other data sources, including *academic records*. All CSA programs can layer these data with collection and analysis of other information, as well, to contextualize and interpret findings. This may include *census data* on such indicators as neighborhood attributes and household income, and/or data on consumer debt levels and other financial metrics. Depending on the CSA program's design and goals, evaluators may also collect original data through *surveys* or *interviews*. Because few CSA research agendas can be successfully executed without any reliance on external data, it is important for CSA programs to include data-sharing agreements and *Memoranda of Understanding* (MOU) in their designs. [1] These agreements should include not only which data will be shared, with whom, and on what timeline, but also the workload, in terms of which entities will be responsible for accomplishing which parts of the data management. Otherwise, CSA program administrators and practitioners may become quickly overwhelmed with the mechanics of contending with unwieldy datasets, while delays in preparing and, then, analyzing data may slow programs' efforts to make informed decisions.

Enrollment Data

In many cases, opening a CSA requires the completion of at least some paperwork by the account owner, generally a child's parent or guardian. While not required for CSAs that are truly universal and automatic, such as San Francisco's Kindergarten-to-College, this account initiation documentation can provide important information to inform later evaluation, particularly if CSAs

[1] See presentation and materials from Colleen Quint of the Alfond Scholarship Foundation on negotiating data-sharing agreements and Memoranda of Understand (MOU).

design the enrollment process with collection of such data in mind. While these data collection aims must be balanced with the need for an efficient enrollment process, CSA architects should carefully attend to the entire workflow of participant interface in order to identify occasions for data collection.

Planning Enrollment Data

- What data will or can you get at enrollment and what data will you need to get from other sources?
- Important opportunity to collect additional information about parents (account holders)
- For a comparison group, opt-in programs may want to collect the same data from those that do not open an account
- Think about how your enrollment data will be linked to account data and other data sets

Even where accounts are opened automatically, this may be achieved by utilizing an existing dataset, such as records of all births or school enrollment records. In some cases, these datasets will include not only the identities of the account owner and child beneficiary, but also other information valuable for later analysis, such as eligibility for free-and-reduced lunch or subsidized health care. In other cases, enrollment data will need to be merged with other datasets that contain relevant contextual information. In all cases, CSA programs will need to link enrollment data with data from the financial institution holding the accounts themselves.

For opt-in program models, one of the primary limitations of enrollment data as a source of knowledge is that CSA programs seldom collect extensive information from those who do not enroll. This precludes later comparisons that could help to isolate the effects of the CSA on observed outcomes. Whenever possible, then, CSA programs should seek to collect information from those who do not elect to enroll, whether by pulling from available administrative records (as in the case of a school-based CSA), or by surveying a sample that can serve as a control.

Table 1 outlines enrollment process and data considerations for different CSA program designs followed by a real-world, albeit generic, example. CSA designs and processes will vary based on program goals, the will of key stakeholders, and availability of resources.

Minimum Enrollment Data

- Contact information
- Basic legal information to open account
- Child date of birth
- Child race/ethnicity
- Date of enrollment
- Grade at enrollment

Optional Enrollment Data

- Parent date of birth and race/ethnicity
- Parent/household highest education
- Occupation
- Family income
- Grade at enrollment
- Child social/emotional well-being

Table 1. CSA Design Models, Enrollment Processes, and Considerations Related to Enrollment Data

CSA Design	Enrollment Process	Potential Data Sources	Limitations	Research Approaches
Universal	Opt-Out	Administrative records (vital statistics, school district)	Will likely have limited information beyond that which is required to open the account (usually, child's name, date of birth, accountholder name)	• Supplementary surveys • Linkage to other datasets through a shared identifier
Universal	Opt-In	Enrollment forms completed by accountholder	May have limited information on those who do not enroll; will have to balance need to streamline enrollment process with need for comprehensive information; if parents complete the forms and information is then entered into a database, there is a host of opportunities for typos and other data entry problems.	• Seek comparison/control group (including information from non-accountholders) • Streamline information needed for account opening to reduce overall paperwork burden • Invest in quality control for data entry and verification
Targeted	Opt-In	Enrollment forms completed by accountholder	May not have information on those who do not enroll; limited information on those who enroll if just get what is needed for the CSAs; will have to balance need to streamline enrollment process with need for comprehensive information; enrolled population may be less representative of the overall population, complicating efforts to find a comparison group; if parents complete the forms and information is then entered into a database, there is a host of opportunities for typos and other data entry problems.	• Seek comparison/control group (including information from non-accountholders) • Streamline information needed for account opening to reduce overall paperwork burden • Invest in quality control for data entry and verification

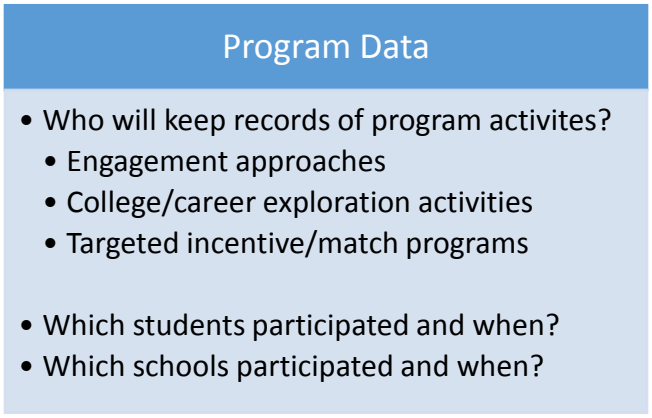
CSA Enrollment Process Example

Below is an example of a CSA program's processes for account opening/enrollment. The steps in the enrollment process for this universal, opt-out CSA illustrate the key lessons learned and important procedures that can help to minimize potential pitfalls at each point.

1. At the beginning of the school year, parents receive notice that they have until September 30 to opt out. This communication occurs through the student handbook, already distributed to each student in the district, so no additional communication is transmitted from the CSA.
   Implications for research: Automating this process facilitates seamless account opening but also reduces the opportunity to collect additional data, particularly about parents/account owners, about whom the school district has relatively little information.
2. The first week in October, the CSA program sends the list of those who have opted out to the school district and requests data on all the kids in the grades that the CSA program covers.
   Implications for research: This information-sharing between the CSA program and the school district should facilitate some comparisons between participating children and those who opt out.
3. The CSA program sends the list of students to the participating financial institution and receives, in return, a list of all the students who currently do not have CSA accounts.
   Implications for research: Back-and-forth data transmission may increase the likelihood of errors in the processing of student information that can, in turn, become embedded in the CSA program's files. If the school district and financial institution have different information on a given student, it may be difficult for the CSA program to reconcile this.
4. The CSA program takes the list of students who have no accounts and 'cleans' it, deleting the students who are not actually eligible (such as children who were not in the school district when the CSA program started or students in private schools who receive only special education services from the district).
   Implications for research: There may be many reasons a student is not eligible for an account, and addressing these discrepancies at this early stage is crucial for research and for effective and efficient program operation, as well. For example, the same child might be in kindergarten two years in a row but only eligible for an account with one of those cohorts.
5. The CSA program takes this cleaned list of students and sends it back to the financial institution in order to open accounts for these students.
   Implications for research: This extends the period between 'baseline' data collection, likely around September, and the initiation of the CSA intervention. Over a period of many years, a few months' delay is not likely consequential, but this lag may be important if attempting to gauge effects within, say, the first year.
6. The CSA program sends a "welcome kit" to the families. The financial institution sends the official bank account number separately.
   Implications for research: This is the first contact between the CSA program and parents, which means that this moment—usually around mid-November, by the time the process is complete—is the first opportunity to gather that contextual information.

Programmatic Data

Enrollment data are not the only records created and maintained by the CSA program itself that will be useful for research. As described in Markoff and Derbigny[2] and in the background sections of some recent CSA research reports, CSA programs utilize a variety of engagement approaches and financial incentives to catalyze saving and cultivate development of college-saver identities. These approaches include outreach materials, which can be mailed to current or prospective accountholders or distributed through other channels; presentations and public information sessions; college and career exploration activities; savings matches; initial seed deposits; benchmark incentives; and other initiatives to build stronger connections between the CSA program and its target audiences. As CSA programs continue to innovate and evolve, these may come to encompass tactics not yet fully conceived, such as social media engagement and the use of apps that allow interface with the CSA. Keeping and monitoring thorough records of the approaches utilized can help the CSA program to track expenditures for outreach and engagement, and linking these records to desired CSA outcomes such as savings performance and, where relevant, account uptake, can help the CSA program to focus on approaches with the greatest return. Here, CSA programs will want to develop mechanisms for tracking not only what was delivered but also to whom, either manually, as in the case of sign-ins for presentations or logs of the accounts that received certain incentives, or, ideally, through automated systems that track the addresses to which outreach materials were mailed or the unique ID numbers associated with earning particular benchmarks. Where possible, these data should be integrated into the CSA program enrollment records; at the least, the datasets should be compatible, to facilitate joint analysis. CSA program administrators and researchers may need to be aware that some of these other datasets may not be disaggregated along the same dividing lines necessary for gauging the effects of the CSA outreach activities, without an added step. For example, school district data may not identify the school students attend, particularly in cases with substantial mobility among buildings or where records are maintained centrally, while CSA programs might tailor outreach efforts to specific schools and, therefore, want to know whether the savings behavior or academic effects they are observing are occurring at a school that received that particular engagement input or not.



[2] Shira Markoff and Dominique Derbigny, *Investing in Dreams: A Blueprint for Designing Children's Savings Account Programs* (Washington, DC: CFED, 2015), http://cfed.org/programs/csa/investing_in_dreams.pdf.

Academic Data

Children's Savings Account programs often need to analyze academic data in order to examine CSA effects on children's educational outcomes. This may include standardized tests of reading, math, and other core academic subjects; administrative records of absences and behavioral referrals; and/or measures of social and emotional well-being. As discussed in more detail below, utilizing these data will require not only constructing systems that facilitate linkages, such as unique identifiers, but also understanding the definitions used, the meaning of the measures, and the limitations associated with each of these data sources. Additionally, while much student assessment is standardized across an entire state, there may be some differences by district, including, for example, whether absences are disaggregated by unexcused and excused, or, rather, lumped together. Other items may even vary by building, such as whether student behavioral referrals are consistently recorded. Moreover, school districts vary widely in ability to automatically pull specified data sets and variables, and in some cases may need to compile data for requested variables by hand. It is essential that these process details are understood at the time the data sharing agreement is created, and, while it is not possible to predict or discuss all of these deviations here, CSA programs should plan their research design with an eye to the types of school data they will be able to utilize and the time that will be necessary in order to retrieve, organize, and analyze them.

Planning Academic Data

- MOU/Data Sharing Agreement required
- Schools vary on data they will share - early planning may allow you to collect this information during enrollment/other sources
- How will data be linked to enrollment data?
- Extract data for all students or just those with accounts?

Important Academic Data

- Free/Reduced Lunch Status
- Standardized Math and Reading scores
- Excused and unexcused absences
- Measures of social/emotional well-being

Account Data

While many of the data management challenges CSA programs face are common to most interventions seeking to evaluate their long-term effects, CSAs are relatively unique in their reliance on data from financial institutions as corollaries to their own records. Financial data are essential to a full accounting of the CSA's operations and financial outcomes and, yet, often unfamiliar to CSA administrators, difficult to navigate, and, at times, hard to access, as well. Consideration of CSAs' management of financial data begins with seeing Children's Savings Accounts and the research interests they implicate from the perspective of the financial institutions brokering access to the accounts. While each financial institution's processes will be

distinct, CSA programs can develop practices to work within the constraints of particular datasets by starting with the right questions.

Planning Account Data
• Clear understanding of steps between enrollment how deposits are processed and recorded by the financial institution. Who is responsible? What regulations govern these steps? • Clear understanding of how and when transaction activities are managed, processed, checked for errors, and how this will be communicated to the program • Clear understanding of how the program will communicate match or incentive deposits back to the financial manager • Clear understanding of timing of transmission of financial data to the program to allow avoid delay in program's ability to deposit corresponding matches • Establish process for closing accounts - program and financial institution need to be on the same page about what constitutes an open account and when an account is considered closed versus inactive • Established "point person" for and process for resolution of account issues

Equipped with an understanding of these 'baseline' processes, the CSA program can then work with the financial institution to develop data management protocols specific to the CSA. Crucially, some of these program needs may require some deviation from financial institutions' standard operating procedures, so it is important that CSA programs clearly convey their interest in such data and the purposes these records serve.

CSA program administrators may also find themselves needing to acclimate to different jargon and organizational cultures, in order to navigate the financial institution partner. For example, one CSA program discovered that the financial institution referred to a report that shows just one record for each child with the total as "householded." The first time the CSA administrators heard this term, they did not know what it meant; since then, they have found it helpful to know what the variables they want are called when they approach the financial institution for data. While, again, there are likely differences in these terms and operating procedures, there is also considerable consistency across financial service providers, and learning to traverse this terrain will serve well CSA program administrators seeking financial data.

In an effort to avoid some of these often-complicated requests to retrieve, manipulate, and, then, interpret, financial data, some CSA programs have entered into agreements with data management firms that promise to mechanize many of these functions.[3]

A recent paper by Clancy and colleagues[4] examines financial data definition within the context of the SEED for Oklahoma Kids CSA social experiment. While SEED OK is an outlier in the CSA field in terms of the sophistication of its research design and the extensive capacity of its research team, some of the lessons learned in the process of designing the evaluation and then managing the SEED OK data over the course of nearly a decade hold important lessons for other CSA programs. SEED OK records differentiate between initial deposit, account-opening deposit, savings matches, account owners' net deposits (deposits minus withdrawals), and investment earnings, reported separately and accruing to financial incentives as well as family savings. Different data support examination of different research questions; SEED OK uses total account accumulation to consider asset effects, while transaction data permit consideration of individual account holder interactions with their CSAs.

Important Account Data

- Date of each transaction
  - Can the bank distinguish between multiple transactions made on the same day?
- Ability to distinguish between types of deposits: incentives, match, gifts, family and non-family contributions
- Mechanism of deposit: in-person, online, direct deposit
- Distinguish between bank adjustments and true withdrawals to avoid contamination of future calculations

CSA programs will have to differentiate between summary and transaction-level data and the instances in which they need each type. This will largely depend on the research questions prioritized. For example, if the CSA program is interested in individual and family savings behavior and the outreach and intervention approaches that lead to particular patterns of deposits, only transaction-level data will tell that complete story. In contrast, if the CSA researchers are primarily looking at the accounts' effects on families' asset holdings and, in turn, on children's educational outcomes, these relationships may be obscured in the minutiae of multiple

[3] See presentation and materials from Promise Indiana's Amanda Jones-Layman describing a pilot agreement with VistaShare in order to streamline the delivery of account incentives and more easily monitor account data through an intermediary platform.

[4] Clancy, M. M., Beverly, S. G., Sherraden, M., & Huang, J. (2016). *Testing universal Child Development Accounts: Financial impacts in a large social experiment* (CSD Working Paper No. 16-08). St. Louis, MO: Washington University, Center for Social Development. *Subsequent publication*: Clancy, M. M., Beverly, S. G., Sherraden, M., & Huang, J. (in press). Testing universal Child Development Accounts: Financial impacts in a large social experiment. *Social Service Review*.

transactions. Of course, summary data can be constructed from transaction-level records, but the converse is not true.

Financial institution records are usually agnostic as to the sources and, in some cases, even the types of deposits; at least, financial institutions may not track these dimensions in the same way that CSA programs need. This means that a CSA program may need to stipulate early on that the financial institution records will need to track whether a given deposit was made by the program, in the form of an incentive payment, or by the family; in some cases, programs may even want to know if the family deposit was made via direct deposit, at a school or other third-party location, or in person at the financial institution. This may be particularly difficult when multiple deposits are made on the same day, as some CSA programs may keep track of the dates when incentive payments are transmitted and attempt to distinguish between family and incentive deposits using that criterion. One CSA program struggled to accurately separate types of transactions because a family deposit and incentive payment on the same day were recorded by the financial institution as one "observation" and lumped together. CSA programs need to be able to differentiate among deposits when one is from an earned benchmark deposit (such as for regular school attendance or completion of a financial education session), one from a family net deposit, and one from a savings match. If financial data cannot distinguish between these deposit types, it may be difficult to accurately gauge these items in relation to other program inputs and, in turn, outcomes.

CSA program summary statistics should only include means with caution, since their vulnerability to distortion from a few individuals' large savings makes them not good indicators of typical savings in a CSA program. Therefore, CSAs should report median figures for financial measures, including monthly/quarterly deposits, account balances, and incentives earned. This process should also include searching for outliers, the presence of which may distort findings. In some cases, outliers may be included in descriptions of the individual differences observed in account interaction; in other cases, they may be discarded entirely, with such action accompanied by careful records of the exclusion. Also, when comparing groups, with large standard errors this will limit the likelihood of seeing group differences.

Some financial institutions may close or, at least, 'mothball' accounts that have not seen account activity, which may make it difficult to retrieve these accounts for regular inclusion in data analysis. However, CSA research suggests the potential for significant effects on measures of child well-being even when families and children are not engaging in financial transactions with their accounts. This means that, for the purposes of CSA research, an account that has only ever seen a $25 account-opening incentive, for example, is still very much relevant to analysis, while the financial institution may not see that same account in the same way, at least not without explicit instructions and an understanding of the rationale for inclusion. Unless CSA programs have predicted this potential complication and discussed it with the financial institution partner, a

given savings report—say, for a quarter, or a year—may not include accounts that have been no transaction activity during that period, an omission that may thwart efforts to seamlessly link those accountholders to academic or other datasets, skew summary statistics toward higher-activity accounts, and preclude analysis of the full range of potential CSA effects.

Finally, CSA program administrators and researchers should be vigilant about seeking and identifying financial errors. This quality control process should include both routine spot checks to cross-verify records as well as a scan to look for records that seem incongruous with individual patterns and/or the larger dataset. If these are truly errors, they need to be corrected before analysis, to avoid contamination of the data by erroneously counting the transaction as true account activity on the part of the account holder. If they are not errors, the detection of unusual account activity may point to valuable findings worthy of additional examination.

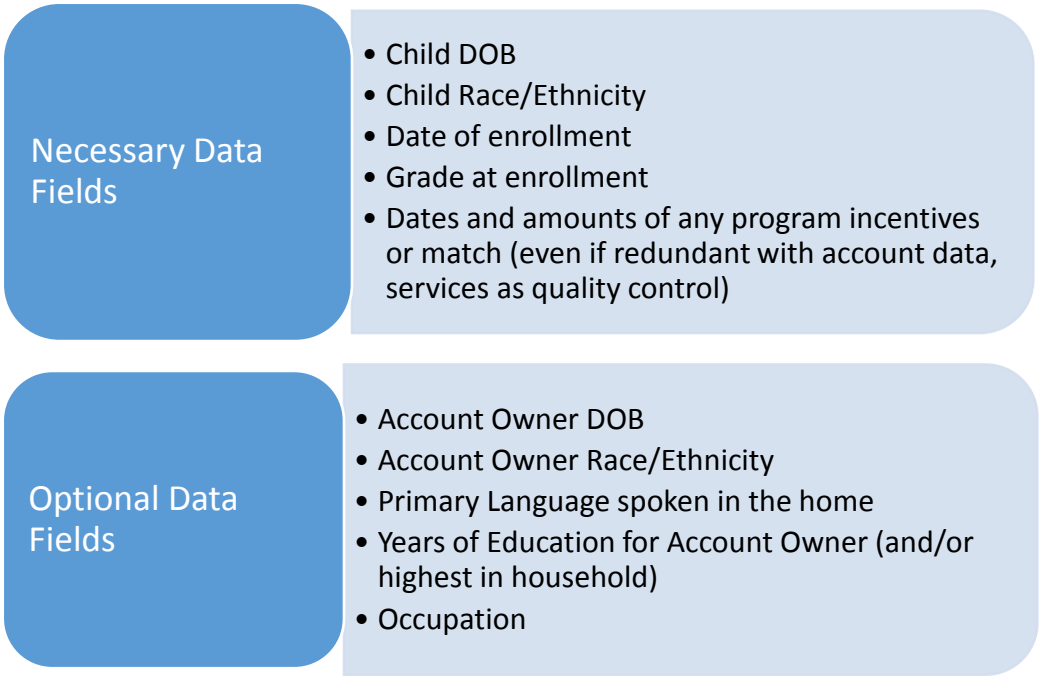
II. Making the Data Work for You

As with planning what data to receive, early attention to setting up these data sets will ensure that program staff and evaluators can extract the information they need. Here, we provide suggestions for settings up datasets, merging them together, and common analyses. It is important to note that these are merely suggestions based on our experience, and are not meant to encompass the totality of CSA data management or preclude other approaches.

Setting up Your Data Sets

Enrollment Data. Generally, all data should be maintained at the account (child) level. In your enrollment data set, this would be reflected as one record (row) for each child. If, for example, a parent signs up for accounts for siblings, each child is still recorded and tracked separately as a unique case. Each variable is recorded in its own field (column).

At minimum, the following variables should be collected in the enrollment database:



Any data that CSA programs collect could be useful for research. For example, CSA enrollment materials might ask how families heard about the program or why they decided to enroll. Programs using automatic enrollment might ask accountholders about trusted sources of

information about the CSA, higher education, or financial matters. This can all be recorded in a database.

In addition to the layout of your dataset, one of the most important aspects of setting up datasets that will provide useable information is to define and label each variable and categories within each variable. Even if no outside evaluation is anticipated, this information is best documented rather than assumed self-explanatory, or held in the memory of one or two individuals. Defining variables has two main parts: determining what the variable means and, then, standardizing how this meaning is to be communicated (or coded).

Example: Academic data from the school contains a variable called SPED.

What does this mean?

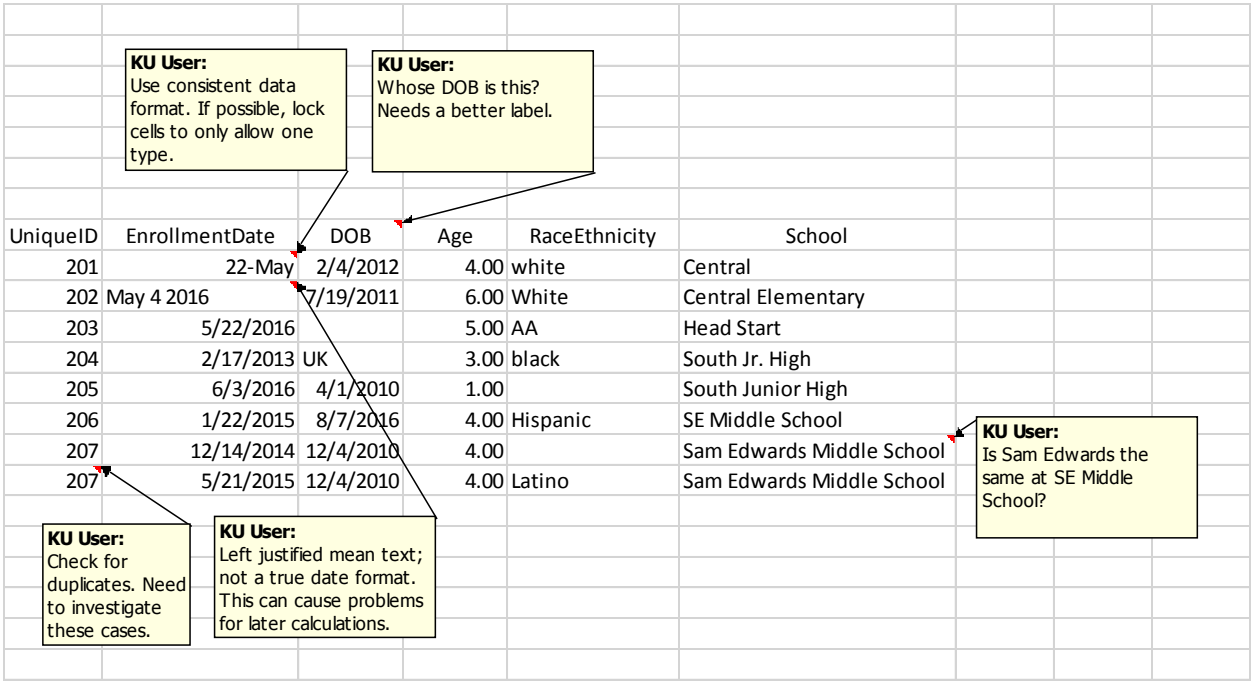
Yes, Special Education, but some SPED definitions include any student with an IEP, such as students identified as intellectually gifted and/or those with physical disabilities, while other SPED definitions are limited to only those students who receive remedial special education services. Either definition can be used, and either may help to provide important context for understanding the effects observed from CSA participation. What is critical is to ensure that the definition is applied uniformly and with a clear understanding of its true intent. It is often important for CSA program administrators and researchers to work closely with experts in order to gain assistance in interpreting unfamiliar variables. Here, a Yes/No dichotomy may mask a lot of underlying complexity in the underlying definition of SPED.

How is it to be coded?

You receive the data, one record for each child. You notice that some cells have 'Y' and some have 'N' while others are blank. Do not assume that blank cells indicate missing. It is not uncommon for blank cells to actually indicate 'No'. If this is the case here, all of the blank cells would need to be recorded as 'N' for No.

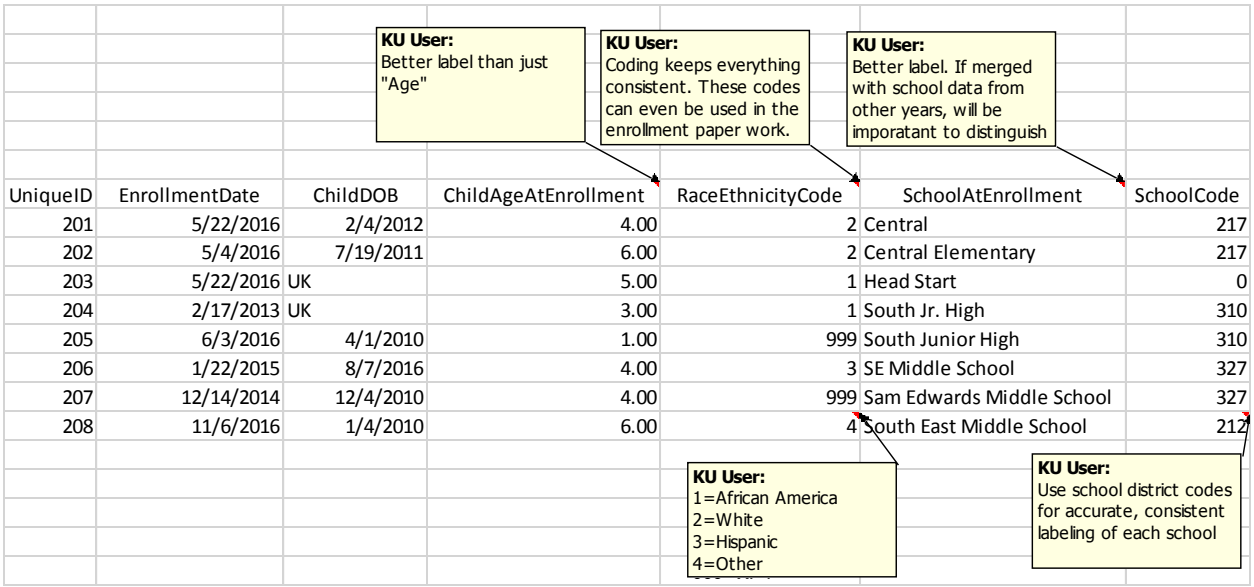
Below are examples of two enrollment databases. The first represents common approaches to recording data and highlights drawbacks. The second example, represents the same data but cleaned up with more precise labeling and coding. CSA programs need protocols to minimize the likelihood of data errors, including setting up databases to 'lock' fields, avoiding string data entry, utilizing unique identification numbers rather than names, and, whenever possible, triangulating with other datasets. Individuals doing data entry should have training and oversight and, whenever possible, work from an established code list.

Example Enrollment Database – A Good Start



UniqueID	EnrollmentDate	DOB	Age	RaceEthnicity	School
201	22-May	2/4/2012	4.00	white	Central
202	May 4 2016	7/19/2011	6.00	White	Central Elementary
203	5/22/2016		5.00	AA	Head Start
204	2/17/2013	UK	3.00	black	South Jr. High
205	6/3/2016	4/1/2010	1.00		South Junior High
206	1/22/2015	8/7/2016	4.00	Hispanic	SE Middle School
207	12/14/2014	12/4/2010	4.00		Sam Edwards Middle School
207	5/21/2015	12/4/2010	4.00	Latino	Sam Edwards Middle School

Example Enrollment Database – Much Better



UniqueID	EnrollmentDate	ChildDOB	ChildAgeAtEnrollment	RaceEthnicityCode	SchoolAtEnrollment	SchoolCode
201	5/22/2016	2/4/2012	4.00	2	Central	217
202	5/4/2016	7/19/2011	6.00	2	Central Elementary	217
203	5/22/2016	UK	5.00	1	Head Start	0
204	2/17/2013	UK	3.00	1	South Jr. High	310
205	6/3/2016	4/1/2010	1.00	999	South Junior High	310
206	1/22/2015	8/7/2016	4.00	3	SE Middle School	327
207	12/14/2014	12/4/2010	4.00	999	Sam Edwards Middle School	327
208	11/6/2016	1/4/2010	6.00	4	South East Middle School	212

Special Definitions for CSAs

- *What do you mean by savings?* Variable definition is even more important if the program decides to collect its own data. While a comprehensive discussion of survey design is beyond the scope of this guide, instruments that will collect original data should be constructed with a careful consideration of how key components will be defined and explained. For example, if a CSA program is going to ask about 'savings', the administrators and researchers will need to decide if they are interested in savings specifically in the CSA account, in any savings held in the child's name, in school-designated savings held in any vehicle by any accountholder in the household, or in savings behavior, for any purpose. These definitions could be further delineated, if, for example, the CSA program's research includes considering how the presence of other children in the family shape savings behavior for the target child and/or how the presence of other liquid assets for non-educational purposes might affect children's savings. "Savings" could mean the habit or practice of depositing or the accumulation of assets in an account or elsewhere, so those distinctions would need to be addressed in the formulation of survey questions, as well. There is a valid argument to be made for attending to these seemingly minor distinctions. In the CSA evidence base, for example, "savings designated for school-related purposes may be associated with improved children's math scores, even among children from households of similar income level," and effects of college-specific savings seem to be greater than for savings in general.[5]

Special Coding Issues for CSA Programs

In particular, CSA programs often struggle with:

- Inconsistent recording of dates, which can complicate efforts to automate calculation of measures such as children's ages and tenure of account ownership
- Inconsistent application of race and ethnicity codes, which can make it difficult to paint an accurate picture of CSA participants' demographics and also complicate alignment with other datasets
- Errors in school names, since schools may be referenced in different ways by different stakeholders, and since even minor differences will cause misalignment between fields
- Errors in individual student or parent/account owner names

[5] William Elliott and Kelly Harrington. *Identifying Short Term Outcome Metrics for Evaluating Whether Children's Savings Accounts Programs Are on Track*. Community Development Issue Brief 1, April 2016. Boston: Federal Reserve Bank of Boston.

    - One CSA program shared experiences with errors in their enrollment data, where many parents struggled to correctly enter children's names online, when unaccustomed to computers, which prompted the CSA to switch to paper forms, problematic themselves when CSA data managers cannot read parents' handwriting.
- Blanks in datasets, particularly where 'blank' means something other than missing data. For example, when a variable is yes/no, some datasets may have a 'yes' inputted, where applicable, but leave the space for the 'no' blank. This makes it impossible to know, later, if the blank means 'no' or if the information is missing.
- Similarly, if the value is truly zero, the field should not be left blank. If recording unexcused absences, zero absences should always be indicated with an actual zero. Blank cells should indicate missing; however, ideally, CSA programs will avoid blank cells entirely, replacing them instead with a value representing BLANK such as '999'. Leaving a cell empty makes it difficult for anyone to know what really should be there. Also, when transferring to other programs, blanks are read differently, which can cause serious problems.
- All datasets will likely need to be cleaned before they can be relied upon for data analysis. This includes checking for duplicates and errors and using crosstabs to look for mislabels, such as a 7th grader in an elementary school or a child whose birthdate does not align with age at enrollment. The process of data cleaning should consider the source of the data, whether hand entry or automatic, in order to anticipate the most likely types of errors, while maintaining vigilance in order to identify outliers warranting additional attention.

Putting It All Together: Linking Data Sources

The most meaningful evaluation of Children's Savings Account programs comes from linking data from a variety of sources to ascertain relationships among different variables and patterns in outcomes. Sometimes, this linking must happen across entirely separate datasets. For example, understanding the relationship between the level of savings and academic outcomes requires knowing which savings accounts go with which test scores. Determining how account owners' economic statuses affect their observed performance on such indicators as frequency and size of deposits often requires linking financial records with information gleaned from enrollment data, sometimes also with added layers from original surveys. In other cases, the linkage needs to happen temporally, as when math and reading scores from a previous year need to be linked to CSA account activity data this year, and, in turn, to math and reading scores for future years, as well, or even when CSA account transaction data from previous periods needs to be considered within the context of current accumulation. Similarly, social and emotional well-being might be measured when a child is as young as age 3 or 4 and then linked back in time to CSA account initiation as well as forward, to measures of later academic achievement. In the long term, children's outcomes in postsecondary education will need to be considered in light of their history as CSA participants, understanding of which will require juxtaposition of records from

the K-12 system, higher education/financial aid, and the CSA program itself. In short, CSA programs will not answer the pressing questions facing the field until we have reliable individual-level data that are linked with other relevant data such as surveys, academic records, and account information. While the discussion below describes some of the technical considerations involved in linking datasets, this is not a purely logistical enterprise. Instead, linking datasets may raise additional confidentiality considerations, if, for example, sensitive information about a student's disability or behavioral record is contained in one data file and identifying information such as name and parents' name in another, in which case the file that links the two may need to be maintained separately. Additionally, the proprietary nature of many datasets may necessitate extensive negotiation with the parties that possess each piece of information.

The Importance of Unique Identifiers

While linking these datasets is an often-daunting administrative challenge, it is not beyond the realm of possibility, even for relatively small CSA programs. Unique identifiers, capable of bringing data systems together, are essential in this task. At enrollment, each account should be assigned a unique identification number. This number should always stay with the account and never be reused. Even if the account is closed, that identification number should remain associated with that closed account and not assigned to a new account. If the person reenrolls later, he/she will resume account holding under that same, unique, number. This may seem extreme, but it is the same protocol used by the Social Security Administration in the issuance of Social Security Numbers and in many other institutions that have to track people over long periods. Substituting or transferring identification numbers leaves too much room for later error, particularly in CSA programs enrolling large numbers of accountholders, where not only the sheer number of accounts but also the likely duplication on other dimensions multiplies the chance of mistakes.

Important Issues for Unique Identifiers

- Always check all data sets for duplicate IDs and names
- Never reuse a unique ID
- If using names to link data sets, use the full name and date of birth if possible
- Linking by name can be complicated if names are recorded differently in each data set
- Consider creating an ID for the person who opened the account to allow for examination of deposits by family

CSA programs have options for the creation of these unique identification numbers. The process they select may depend on the size of the program and the institutional context. Some just use a 4-digit number starting at 1000. Some create numbers that are alphanumeric with the program initials. Some use the student school ID number or other, previously-generated ID. This can be useful because they are pre-existing and likely will not change, but there may be resistance to exporting those numbers to CSA programs operated outside of the school system, even if the target populations overlap. It is technically feasible to use the first, middle, and last name – and

maybe date of birth – to uniquely identify a child for matching to other datasets such as school test records. However, even in small CSA datasets, duplicate names can be found. Unique identification numbers at enrollment—whether parent-initiated or automatic—help avoid this altogether.

Even if the program assigns unique IDs, it can still take some work to link with school records and account records. If the account ID is the same as the unique student ID, the technical process of linking should be seamless, although securing the data-sharing agreements that facilitate such alignment may be more complicated. If these numbers are not the same, CSA program administrators will need to complete the intermediate step of linking the account ID to the unique student ID by using the student's name. This is not difficult to do; however, as with all data merges, it requires attention to detail. Names (first, middle, last) and date of birth should be set up in the same exact format in each file. It is very common for school administration files to differ on how a student's name is recorded compared to how a parent may fill it out at enrollment, and how it might have been transferred from paper during data entry (for example, ineligible handwriting or ethnic variations on spelling and punctuation). Encouragingly, however, more schools' transitions to fully electronic student registration may reduce the frequency of these errors.

It is also helpful to have identification numbers at the account owner level as well. While many of the metrics of interest to CSA programs unfold on the level of the individual child, others relate to the household. CSA programs will be hindered in their ability to analyze these dimensions without a way to connect different children's accounts to the same guardian. For example, if a CSA program is interested in looking at how much money a family is saving or how much they now hold in total educational assets, the program would need to know which accounts are held by the same owner. Considering the effects of the CSA on the household's material hardship, would necessitate assessing not only one student's account, but all those to which the family has access. For these reasons, the dataset might have an ID for the account owner as well as IDs for each unique account. Without a family ID, it is nearly impossible to tell which accounts go together. One cannot depend on similar last names or addresses or, conversely, assume that differences on those variables separate different households. Even if cases have the same last name, without identification numbers and variables to link the two cases, there is no way of knowing whether these children are siblings, if the account owners share a household, or whether or not these accounts should be considered part of the same family's overall financial picture. All of these are potentially important inquiries for CSAs aimed at strengthening household financial well-being and/or bringing parents into the financial mainstream, but the dataset was not configured to allow this examination. This is another reminder of the core truth of data management for research: knowing what questions you are interested in answering ahead of time will be important for deciding what types of systems (in this case, identification numbers, for the individual child and for the account owners) are needed and how to obtain them.