Section 4
How to Set Up the Data Management System
(Sounds Scary, but It's NOT!)

- Dos and Don'ts of Data Entry and Management
- Spreadsheets and Types of Data
- Coding Your Data
- Preparing for Longitudinal Data Analysis
- SUGGESTIONS & EXAMPLES: Sample Spreadsheet and Codebook

A well-structured database is the key to meaningful program evaluation. Setting up the file is an exercise in logic, but it is not difficult. The file becomes a representation of the key elements of the program that you want to evaluate: who the students are, their grade levels, their length of participation, their success in language learning, their academic achievement. It's all contained in that one database. Therefore, that database must be constructed with great care. However, once it is constructed and after you have worked with it, it becomes second nature, and it isn't really all that hard to maintain—if you follow a few data entry and data maintenance tips.

A **sample** spreadsheet and codebook will be available in this section. You can use these as a start to your spreadsheet if you would like.

Dos and Don'ts of Data Entry and Management

A good database[1] is not just an electronic file folder. Data must be maintained in such a way that it can easily be summarized and analyzed. The easiest way to maintain data is in a spreadsheet format such as Excel[2]. Spreadsheets are commonly used, and many people already have some spreadsheet skills. However, to develop the spreadsheet so that the data are easy to analyze, it must be set up according to specific guidelines. Keep in mind the following hints:

[1] We are using the terms database and spreadsheet interchangeably here. There are differences, but we will not get into them here as we are trying to keep this as simple as possible. For those interested in actual databases, there are a number of good database programs (e.g., FileMaker, FoxPro, dBase). However, they are more expensive and take some time to learn.

[2] Excel is part of Microsoft Office. It can be purchased at most computer stores, downloaded online, or possibly gotten for free from the district office. Check these sites for tutorials and use of Excel in the classroom:
Excel tutorial: http://www.usd.edu/trio/tut/excel/

Using Excel in the classroom:
http://www.internet4classrooms.com/on-line_excel.htm

- Make sure there is only one record (line of data) per student to avoid duplicate information. A record can contain an unlimited number of variables, or types of information. The example spreadsheet we will discuss at the end of this section shows what a good relational database looks like as a spreadsheet. Note that each line contains information on just one student, and the student's information appears in only one line.

- Each variable, or column, should contain exactly one and only one kind of information. For example, it might be a last name, a first name, a date of entry, an English proficiency score for a specific year, etc.

- Pay careful attention to the principle of exactly one and only one kind of information. For example, a column might have NCE scores for a reading achievement test in 2005 but not 2004 or any other year. It would contain only NCE scores and not percentiles, grade equivalents or other kinds of scores. If those kinds of scores are of interest, put them in a separate column.

- Do not use zero (0) as a placeholder to indicate missing information. It is possible to specify that zero means "missing information," but it is not necessary. If the person doing the analysis doesn't know to exclude values of zero, those zeros will be entered into calculations, and will significantly distort the results. Enter a zero if and only if it is a real value. Zero does not appear as a valid score in most standardized tests.

- If a certain piece of information is missing on a student, leave it blank. Do not enter explanatory notes such as "absent," "exempted," etc. Just leave a blank. If it is important to know that a student was absent, exempted, etc., create a variable with corresponding codes. For example, you could have a column for a reading test in 2005 that has numerical or alphabetic codes for such information as "absent," "exempted," "couldn't attempt," etc.

- When you decide on a variable (like language background), be consistent in how you enter it, or you will have a nightmare in analyzing the data. For example, you may want to use S to designate Spanish, E for English, and so on. But, what happens if you have a student whose primary language is Swahili or Swedish? Also, be sure that you decide on lowercase or uppercase as E is not analyzed the same as e. Also, be sure that you don't use a lowercase letter l to enter a numeric 1, or an uppercase letter O or lowercase o for a zero 0. The computer reads the letter l and the number 1, as well as the letter O and the number 0 differently. You may think they look similar, and they do, but the computer thinks they are different. If you aren't careful when you enter the data, you will find, to your dismay, that when you analyze the data, you have lots of incorrect information that you have to go back and correct—this is not a fun task!

- Select someone to enter data who is well known to pay attention to details. Sloppy data entry will cause you grief later on! If you don't have anyone who is computer-savvy to enter data at your site, consider asking a parent, or fundraising to hire a good college or high school student, or train a parent or two at your site to do data entry. Once you have the database established, it will not be all that expensive to maintain it.

- Be **selective**. Remember that each column you add will mean that someone has to collect that information and then enter it into the spreadsheet (and that adds cost to this process). Adding data to the file suggests that it is important data and should be analyzed and interpreted—so be selective about what information is important to add.

- If you decide to do surveys and questionnaires, you may not want to enter that information into the regular database. There will be lots of columns of information that don't need to stay with each student. You might want a different database for that information (or you might want someone to hand tabulate the information), depending on how you want to use the information.

Spreadsheets and Types of Data

If you follow the advice given above, you will have built a relational database. In other words, you will be able to "relate" different elements in the database to each other. By having all the elements on one line refer to the same student, and all the elements in one column in one form (number, code, etc.) and referring to only one kind of information (e.g., English language scores in 2005), you will be ready to conduct analyses that relate all those data elements.

Types of Data

The following types of data will *typically* constitute the database for your evaluation. You may decide on a subset of these or additional data—remember to be selective, as we mentioned above.

- **Student ID –** be very careful about entering this information; if it is wrong, you may misidentify students later on.
- **Student name –** be consistent in the name(s) you use. Don't use nicknames. Over the years, students may have their names changed (through adoption or a parent marriage and name change or by adding on a hyphenated name). If you have to enter survey data or other types of data that does not have the student ID, you will need to use the student name to determine where to enter data. At that point, you will want to be sure that you are entering the right information for the right student.

> **Note**. You need to check with your district administration about maintaining a data file with identifying information on students. Different states and districts may vary in their requirements and procedures for assuring the confidentiality of student information. Policies may be in place limiting access to individual students' information by non-school personnel such as external evaluators or transmitting student data with names over the Internet.

- **Ethnic –** terminology for various ethnic groups differs from district to district, depending on the populations represented in the district. Whether you use Hispanic or Latino or more clearly defined subgroups (e.g., Guatemalan, Mexican) is up to you.

However, having 25 subgroups may be difficult to interpret if you decide to look at group differences according to ethnicity.

- **Native language –** English, Spanish, Korean, Russian, etc.
- **Language group at entry –** determine what terminology you want to use and be consistent (e.g., ELL, IFEP, R-FEP, bilingual, EO). Your terminology should be consistent with your district or state designations so that you can analyze and interpret your data and present it to the district in ways that will make sense to them.
- **Program type** - Determine if there is more than one possible program at your school and if students in other programs will be maintained in the database.
- **Current grade level**
- **Current language proficiency level –** you will want to keep this information separate from the initial level so you can distinguish what students moved from ELL to English proficient.
- **Annual language proficiency scores –** each year, you will want to enter each student's proficiency scores (you will need one column for each year's score)
- **Dates of language proficiency scores –** if it's always the same date (like in the spring), then you don't have to notate it, but if it is sometimes in the fall, sometimes midyear, and sometimes in the spring, you would want the date (of course, interpreting that information would be a nightmare)
- **Academic achievement information –** need to indicate what test, the type of score (e.g., scale score, proficiency category, NCE)
- **Dates of academic achievement information –** this is frequently noted on the scoring forms that the test publisher sends back to the school. Like with language proficiency, if the test is always given at the same time (spring), then it is not necessary to record the date.
- **SES –** you may want to include this in your dataset, using parent education level or whether student participates in the free/reduced price lunch program.
- **Other information about students –** whether the student left the program, was retained in the current grade, was referred for or receiving special education services, was referred for or receiving Gifted & Talented services.

Your district may already have a district-wide database that you can use to develop your own database. If so, you may want to think of variables that are of interest to you (and possibly other schools in the district) and see if the district will add those to their database (then you don't have to go to the expense of collecting that information). However, you may not want all the information that the district has in their database, and you certainly don't want all those students from other schools. See if you can get your school or strand (if you are just a strand in your school and only want your TWI or Developmental Bilingual strand in the database) pulled out into an Excel file. We've done this at different sites and it is a great time-saver (which also means money saver).

Coding Your Data

Coding allows you to put students and their data in categories that can be separately analyzed or compared. The most obvious category of interest to TWI programs or other programs for

English language learners is the student's language group: Are the students English learners, native English speakers, or bilingual (entering a program already speaking both languages)? The best way to code data is to use a single column, which may represent the variable of interest (e.g., type of program, language group, SES). A code will be entered for each student, e.g., TWI for students enrolled in the two-way immersion program, DB for students who may be enrolled in a Developmental Bilingual program within the school or district, SEI for ELL students who are enrolled in a Structured English Immersion program. If some students are not enrolled in a language assistance program (regular English mainstream), they might be coded None. You can use whatever terminology you want; just be consistent! An example of this coding would look like the following:

Student ID	Program
11111	TWI
11112	TWI
11113	DB
11114	DB
11115	DB
11116	TWI
11117	SEI
11118	DB
11119	SEI
11120	NONE
11121	TWI

Now imagine if you had a person entering the data who was not very careful. You could have data entry that looked like this:

Student ID	Program
11111	TWI
11112	Twi
11113	DB
11114	db
11115	Db
11116	twi
11117	SEI
11118	D.B.
11119	s.e.i.
11120	NONE
11121	T.W.I.

Then, when you tried to run analyses and see if students in the TWI or Developmental Bilingual program scored differently from students in SEI or English mainstream (None), you could not answer your question. You would find that you had not 4 different programs (TWI, DB, SEI, None) like you thought, but you actually had 4 TWIs (TWI, Twi, twi, T.W.I.), 4 DBs (DB, db, Db, D.B.), 2 SEIs (SEI, s.e.i.), and one None – and we didn't even deal with typos. This is why consistency is so important!

USING NUMERIC CODES

It's very helpful to use numeric codes even for such variables as language group, ethnic group, etc. In fact, we strongly recommend using numeric codes. It is easier for three reasons:

1. You don't have the problem we just saw with inconsistent use of labels (TWI, T.W.I.)
2. It makes it easier to select certain groups for data analysis after you acquire some familiarity with a statistical program.
3. It also makes data entry much easier, since you can use the number pad on the computer keyboard and simply type in the numeric codes.

The trick is, of course, to keep a record of what the numeric codes mean!

One trick is to use numbers to represent sequential levels. For example, if you record three levels of SES, you can use 1 for the lowest, 2 for the middle, 3 for the highest. The logic of it serves as a memory aid. Another trick is to use 0 for no and 1 for yes, and that can apply to membership as well. For example, let's say you have some students who are in the two-way program and some who aren't. You can have a column "2-way", and for each student, 1 means yes, in the program, and 0 means no, not in the program. Anything that makes sense to you can work.

Using the example about coding Program categories above, we could define program codes as:

1=TWI (Two-Way Immersion)
2=DB (Developmental Bilingual)
3=SEI (Structured English Immersion)
4=None/English mainstream

The number or order is not as important as being consistent with the definition and use of the codes. So, the above spreadsheet would look like the following if we used numeric codes instead of names.

Student ID	Program
11111	1
11112	1
11113	2
11114	2
11115	2
11116	1
11117	3
11118	2
11119	3
11120	4
11121	1

Make sure to have a Codebook (see the sample codebook at the end of this section) on hand for the next person who may not share your sense of logic. That means, write out your coding system and store it in the computer, print it out, and keep it on hand as a reference tool. That way, if you lose the copy or you have a new person work on it, or you haven't worked on it for a while and forgot the codes, you will be able to remember and keep your data coding consistent.

There are other ways to code data or organize a spreadsheet, but in the interest of keeping this Toolkit as simple and user-friendly as possible, we are not discussing all the possible ways of coding data.

Preparing for Longitudinal Data Analysis

There's practically no limit to how large a spreadsheet can be. It could contain all the students' data from the time they start the program in kindergarten until they leave it at 5th grade or higher. However, it is much easier to manage data files that reflect one year's data at a time. Some information can simply be copied from the original file to a new one, such as students' names, ID numbers, language group, initial language proficiency, etc.—any information that will not change over time. Make sure the student ID number is kept in each year's file. That will allow you to merge files, bringing together data for longitudinal analysis, such as progress over time in language proficiency.

If you are using Excel, you can save last year's file with the name of the file for the current year. Be SURE to name the new file something different (Like Outcomes_2005-06). If you save the old file with the new name, then you can delete the information that you will need to add for the current year. A bunch of information will stay the same – student ID, name, language background at entry, date of entry, etc. Current information will need to be updated, so just highlight all the cells that will need to be changed for this year, then delete the information. Then rename your variable names so that they have the current years. And, in just minutes, you are ready to go with your new dataset for the current year. Of course, you'll need to add new students and all their background information, and delete students who left (or graduated).

Even if each year's data is kept in a separate file, make sure the headings indicate which year they refer to. Otherwise, when you merge the files, you won't know which data are for 2003, which are for 2004, etc. You may think you won't forget this information, but once you've been away from it for a while, it's easy to forget.

Anticipate the kinds of information that need to be kept on a long-range basis, how the information will be used, and what that implies for recording information in the database. For example, in a TWI program, you will want to conduct separate analyses for the language groups represented. Anticipate the kinds of questions you will want to ask of your test data, and make sure you are maintaining the right kinds of test scores. (See Section 6, Analyzing and Interpreting Assessment Data, for guidance on test scores.)

Note that a longitudinal analysis means the data represent the exact same students over several years' time. In a true longitudinal analysis, students who had scores some years and not others, due to dropping out, late enrollment, etc., would be excluded from the analysis.

ALWAYS, ALWAYS, ALWAYS KEEP BACK-UPS OF YOUR DATASETS—AND NOT ALL ON THE SAME COMPUTER!

SUGGESTIONS & EXAMPLES:

Sample Spreadsheet and Codebook

If you have not begun to develop a spreadsheet for your data, you can download the Excel spreadsheet from the Toolkit. You can download the Excel spreadsheet to enter data directly on the computer OR you can print out the spreadsheet to use as a hard copy to enter information by hand.

- Below you can see an Excel spreadsheet for California schools
  (filename is: samplespreadsheet_ca.xls)

- Below you can see an Excel spreadsheet for non-California schools
  (filename is: samplespreadsheet_other.xls)

You can also view an Excel spreadsheet with data filled in—this is an example spreadsheet for third and fourth graders at one school site. We will use it for some data analyses we will complete in Section 6 and in Section 9 for the Step-by-Step guide to data analysis and presentation.

- This is an actual completed Excel spreadsheet
  (filename is: example03-04gr34.xls)

The codebook can be downloaded in Word or PDF. Of course, downloading in Word means that you don't have to re-enter all the information, but can start from our example to make changes appropriate for your spreadsheet and site needs. Remember, you can develop your own codes, you can delete some of the variables, add new ones and so on.

- See the EXAMPLE Codebook

Codebook for Dual Language Spreadsheets

Use the Excel spreadsheet to record the information.
BE CONSISTENT IN YOUR DEFINITION OF CODES
Variable names appropriate for SPSS are listed in parentheses next to regular variable name[3]

THESE ARE RELATIVELY FIXED – not likely to fluctuate from year to year

Student ID (ID) -- enter the district or state ID number, whichever is appropriate (Double check your entry). In California and some other states, each student has an individual identifier number.

Student NAME (NAME) -- Last name, first name

ETHNIC (ETHNIC) -- use the following categories **ONLY** to code the student's ethnic background, OR use the codes for California[4] or your state:

1 = Hispanic/Latino/Mexican
2 = European American, White, Other than Hispanic
3 = African American, Black, Other than Hispanic
4 = Asian American
5 = Native American/Eskimo/Aleut
9 = Other

LANGUAGE (Language) -- Primary language or language spoken at home

1 = English

2 = Spanish (if your 2nd language is something other than Spanish, you would want that here)

3 = …

LANGUAGE GROUP AT ENTRY – ELL, EO, IFEP (LANGROUP) -- indicate whether this student ***__began__*** the program as LEP/ELL or EO (not what the student is now)

1 = ELL
2 = IFEP/EO

You need to determine whether you have a significant BILINGUAL population that you want to code for 3=BIL or IFEP

THE VALUES FOR THE REMAINING VARIABLES MAY CHANGE FROM YEAR TO YEAR and should be noted by the appropriate academic year.

[3] SPSS requires that variables names be no longer than 8 characters in length.

[4] In California, you should distinguish the following ethnic groups: American Indian or Alaska Native, Asian, Pacific Islander, Filipino, Hispanic or Latino, African American-not Hispanic, White-not Hispanic, Multiple or no response.

CURRENT DATA (Represent current by last 2 digits of academic year. Academic year 2006/07 = 06 in examples below

PROGRAM TYPE (PRGTYP06) -- Enter the program type in which the student is participating – if all students are in the two-way/dual language program, no need to enter this information.
1 = TWI
2 = DB
3 = SEI
4 = English mainstream

GRADE (GRADE06) – current grade level of student
0 = kinder
1 = first, etc

CURRENT LANGUAGE GROUP -- ELL/R-FEP/EO (LANGRP06) -- indicate whether this student ***is currently*** ELL, R-FEP (redesignated as FEP) or EO
1 = ELL
2 = R-FEP
3 = IFEP/EO

CURRENT ENGLISH LANGUAGE PROFICIENCY SCORE – enter the student's language proficiency score in ENGLISH. *(COMMENT: If you have a state or district designated test, you may want to have one or more columns for that test and one column for a rubric, like SOLOM or FLOSEM, if you are using these. You will need one column for each score). You should stay current on the requirements of your state for testing instruments, scores, and interpretation of scores.*

EXAMPLE: California requires the CELDT test (Form F).

CURRENT CELDT Score (CEL_T06) – Total:
CURRENT CELDT Score (CEL_L06) – Listening
CURRENT CELDT Score (CEL_S06) – Speaking
CURRENT CELDT Score (CEL_R06) – Reading
CURRENT CELDT Score (CEL_W06) – Writing

1 = Beginning
2 = Early Intermediate
3 = Intermediate
4 = Early Advanced
5 = Advanced

EXAMPLE:

CURRENT FLOSEM Score (FLOSE06) – this would be a score from 5-30

CURRENT SPANISH LANGUAGE PROFICIENCY SCORE – enter the student's language proficiency score in SPANISH.
EXAMPLE:

CURRENT FLOSEM Score (FLOSS06) – this would be a score from 5-30

(Note that the SPSS names for the FLOSEM are FLOSE06 with the E for English, and FLOSS with the S for Spanish – both end in 00 designating 2006/07 as the academic year)

CURRENT ENGLISH ACADEMIC ACHIEVEMENT SCORE:

READING/LANGUAGE ARTS – enter the student's academic achievement score in ENGLISH – in ____ (*COMMENT: Select appropriate type of score)*

EXAMPLE:

CURRENT CST[5] Scale Score (CELA_S06)
CURRENT CST Proficiency Level Score (CELA_P06)

1 = Far Below Basic
2 = Below Basic
3 = Basic
4 = Proficient
5 = Advanced

MATHEMATICS – enter the student's academic achievement score in ENGLISH – in ____ (*COMMENT: Select appropriate type of score)*

EXAMPLE:

CURRENT CST[6] Scale Score (CMth_S06)
CURRENT CST Proficiency Level Score (CMth_P06)

1 = Far Below Basic
2 = Below Basic
3 = Basic
4 = Proficient
5 = Advanced

[5] CST = California Standards Test, see glossary. You would enter your appropriate state's test name (and variable name for SPSS if you're using SPSS)

[6] CST = California Standards Test, see glossary. You would enter your appropriate state's test name (and variable name for SPSS if you're using SPSS)

CURRENT SPANISH ACADEMIC ACHIEVEMENT SCORE: READING/LANGUAGE ARTS – enter the student's academic achievement score in SPANISH – in NCE (*COMMENT: You can select NCE or scale score)*

EXAMPLE:

> **CURRENT APRENDA3 NCE Score (SRT06N – S for Spanish, R for Reading, T for total as opposed to subscores, 06 for academic year 06/07, and N for NCE)** – score would be between 1-99

MATH – enter the student's academic achievement score in SPANISH – in NCE (*COMMENT: You can select NCE or scale score)*

EXAMPLE:

> **CURRENT APRENDA3 NCE Score (SMT06N – S for Spanish, M for Math, T for total as opposed to subscores, 06 for academic year 06/07, and N for NCE)** – score would be between 1-99

Answer 1 for YES only where applicable (a blank will assume a NO answer) -- to the following:
SES (SES06) -- Eligible for free/reduced price lunch or parent education (check whether there are state or local guidelines you should use)
RETAIN (RETAIN06) -- Retained in Grade in past year
SPEC ED (SPEDED06) -- Special education in past year
GIFTED (GIFTED06) -- Gifted education in past year

Remember that these are all examples and suggestions -- not requirements. You may collect whatever data you choose and name it whatever you wish in your spreadsheet.

The Example Excel Spreadsheet examp03-04gr34.pdf should follow this.