(2017) In (Eds.) A. Weinberg, C. Rasmussen, J. Rabin, M. Wawro, and S. Brown (Eds.), Proceedings of the 20th Annual Conference on Research in Undergraduate Mathematics Education (pp. 129-140). San Diego, CA.

Data Cleaning in Mathematics Education Research: The Overlooked Methodological Step

Aleata Hubbard
WestEd

The results of educational research studies are only as accurate as the data used to produce them. Drawing on experiences conducting large-scale efficacy studies of classroom-based algebra interventions for community college and middle school students, I am developing practice-based data cleaning procedures to support scholars in conducting rigorous research. The poster identifies common sources of data errors in mathematics education research and offers a framework and related data cleaning process designed to address these errors.

Key words: Research methodology, Efficacy studies, Algebra

The results of educational research studies are only as accurate as the data used to produce them. Screening data for potential errors and ensuring anomalies do not influence analyses is an essential step of the research cycle (Wilkinson, 1999). Odom and Henson (2002) demonstrated how regression models of high school seniors' mathematics achievement varied depending on the level of screening applied to the publicly available High School and Beyond National Survey data set. As another example, Whalley (2011) analyzed the Panel Study of Income Dynamics data set to show how the choice of trimming procedures (i.e., methods for removing outliers) effected estimates of the relationship between education and labor income volatility. Educators and policy makers rely on study results to make decisions that influence the lives of many. It is important that scholars understand and apply appropriate methods for ensuring high quality data.

The process of identifying, resolving, and documenting data inconsistencies is called *data cleaning* (Rahm & Do, 2000). Despite the importance of data cleaning in rigorous research practice, most methodology courses only give cursory attention to the topic (Osborne, 2013). Therefore, scholars often acquire cleaning strategies heuristically, making it difficult for others to accurately judge or replicate studies (Leahey, Entwisle, & Einaudi, 2003). Furthermore, my conversations with scholars new to large-scale research suggest many underestimate the amount of time and resources required to properly clean data. Data collection and data preparation each can take about 20% of project time (Munson, 2012). Well-established standards for data cleaning could facilitate the integration of this topic into researcher training and support educational researchers in accurately planning their studies.

Drawing on my experiences conducting statewide and nationwide efficacy studies of mathematics interventions, I am developing practice-based data cleaning processes to support research in classroom settings. While data cleaning can be applied to many forms of data (e.g., interviews, observations, documents), I focus on quantitative data gathered from surveys, questionnaires, assessments, and demographic records. Specifically, I ask: (1) What are the sources of data errors and challenges in educational research studies conducted in authentic mathematics learning environments? (2) How can a data cleaning process be designed to consistently produce accurate, reliable, confidential, and timely datasets?

Methods

Two large-scale efficacy studies inform the framework presented here. *Study A* was a three-year, nationwide study of revisions to a popular mathematics curriculum involving over 10,000

middle school students and 180 mathematics teachers. Students completed between five and eight end-of-unit assessments on paper, two attitude surveys, and two mathematics assessments either on computers or on paper. Teachers completed weekly logs and two teaching knowledge assessments, all electronically. School districts provided demographic data and state test scores for participating students. *Study B* is a two-year statewide study of the use of a computerized interactive learning platform in community college elementary algebra courses. The first year of the college-level study involved approximately 400 students and 89 instructors across the state; the second year of the study is underway. In this study, all data is collected electronically. Students complete two mathematics assessments, a background questionnaire, an academic motivation questionnaire, and an end-of-semester survey. Log data of student interactions in a web-based activity and testing system are also collected. Instructors complete assessments of knowledge of technology in teaching and content knowledge for teaching, background questionnaires, and weekly logs.

An initial data cleaning process was created for Study A based on prior experiences with small-scale research and evaluation projects. Across the three years of Study A, the data team documented their data related challenges and associated resolutions, revising the Study A data cleaning process as they went along. The revised process was also compared against rigorous research standards in the What Works Clearinghouse Procedures and Standards Handbook (2016), ethnical research guidelines around privacy and confidentiality (OHRP, 1993), general data modeling rules from the field of computer science, and data management practices used in educational survey research (e.g., Schleicher & Saito, 2005) and in statistics (e.g., de Jonge & van der Loo, 2013). The modified data cleaning process is being implemented in Study B.

Researcher's Role and Background

Describing one's background and one's role in research allows readers to understand the perspective researchers bring to their work (Creswell, 2012). Drawing on my undergraduate training in computer science, I applied my knowledge of data modeling and databases to the framework described in this paper. My research training in learning sciences provided me with the domain knowledge needed to understand the contexts of Study A and Study B, to understand how data could be organized for useful analysis, and to identify anomalies that might signal problems at other stages of the research cycle.

I was involved in multiple aspects of Study A including participant management, data collection, data cleaning, data analysis, instrument creation, classroom observations, meetings with the research team, and professional development workshops for participating teachers. This level of involvement gave me a chance to confront data issues directly at many points across the study. For example, answering participant phone calls about the study provided insights into the ways teachers implemented study data collection tasks, which sometimes conflicted with researcher expectations (e.g., administering student assessments across multiple days). Working with researchers to conduct item-response theory (IRT) analyses and teams to score constructed response items highlighted the importance of distinguishing the reasons for which data were sometimes missing. My involvement in Study B was constrained to data management, working with a team to clean data files, and interacting with the project staff to stay abreast of study developments. This narrower role allowed me to focus more on the refinement of the data cleaning process.

Results

Data Errors and Challenges

Despite standardized procedures for administering and gathering data, data collection in large scale educational studies often result in a host of data cleaning errors that are, to some extent, unavoidable. Errors can include duplicated records, illegal values, missing values, or misspellings (Rahm & Do, 2000). In the studies described here, errors in the gathered data created the need to make decisions about issues such as handling duplicate records, the validity of an assessment completed on an incorrect form, and how to link records in a hierarchical research design when participant identifiers changed. Challenges in study implementation hindered the timely collection and cleaning of study data. Common sources of error and challenges in data cleaning for both studies are described below.

Variations in assessment administration. Schools and colleges differed in their schedules and access to computers. For example, many middle school class periods lasted 45 minutes while other schools operated on block schedules where class periods lasted 90 minutes. Some teachers with shorter class periods would administer paper-based assessments across two days, having students complete selected response items on the first day and constructed response items on the second day. The educational institutions within which Study A and Study B were conducted varied widely in their computer availability. Teachers with computers inside their classrooms easily administered online assessments for the study. However, teachers who had to reserve a computer lab often lost time in transitioning to the lab room or they held multiple administration sessions due to an insufficient number of computers for their students. A handful of teachers had no computer access and administered study assessments on paper. Lastly, some teachers requested Spanish versions of study assessments, which were only available on paper.

Understandably, teachers were responding to the realities of their school environments. However, some administration choices led to students completing part of an assessment on the wrong form or completing the assessment more than once. Differences in administration also complicated data cleaning processes. An assessment completed on paper, where students could write what they wanted, required different cleaning checks than the same assessment completed online, where computer-based forms restricted the possible answers permitted. Also, it became more difficult to account for test completion, both at the student level and the class level, when some items were received on paper and others online.

Participant mobility and late joiners. Some participants joined the studies after baseline data were collected and others changed institutions during the studies. This mobility was explained by various factors. First, families moved and in doing so placed their children into new school zones. Student mobility is common at the K-12 level in the U.S., particularly amongst students from urban areas, lower income families, or migrant, military, or immigrant families (Welsh, 2016). Student mobility occurred in Study A, particularly in schools near the U.S. border with Mexico that served large numbers of migrant families. Second, many community college instructors work across multiple institutions. In California, the location of Study B, 36% of associate faculty (e.g., part-time or adjunct instructors) teach at more than one institution in order to make a livable wage (Smith, 2013). While instructor mobility was not a significant issue to data collection for Study B, one participant unexpectedly taught the study-target course at different colleges each semester of the study. Third, research teams in both studies were confronted with low enrollment numbers and participant dropouts. Recruitment and retention of study participants in large-scale educational studies is challenging because it requires a long-term commitment from teachers who already have busy schedules (Gallagher, Roschelle, & Feng,

2014). This issue required (a) additional rounds of recruitment to obtain participant counts that allowed for sufficiently powered statistical analyses and (b) an extension of administration windows to allow for greater data collection. Lastly, a teacher strike that occurred during Study A delayed data collection for one district containing several consented participants.

Participant mobility resulted in some student participants in Study A moving between treatment and control groups and completing pre- and post-intervention assessments under different experimental conditions. In other instances, records at a given level of the study design appeared to be missing a link to records at the other levels of the study design. For example, when a teacher in Study B changed institutions, she was assigned a new participant identifier. During the cleaning process, data from students in her first semester course appeared unconnected to any teacher. Issues resulting from participant mobility introduced the need to (a) create new versions of data files that included late joiners and (b) make decisions about how to resolve participants linked to multiple classes, teachers, or schools.

Multiple participant names. Some participants became associated with multiple names and e-mail addresses and some had names that changed. In Study A, we often saw the name a student wrote on assessments differed from the name on the teacher's roster, which differed from the name provided by districts when collecting demographic information. In Study B, students used both school-provided and personal email addresses when completing study tasks. At the teacher level, names occasionally changed when participants changed marital status during the study.

As a consequence, participants sometimes appeared to have missing records because their data could not be matched to the legal name or school email address provided to the research team. Connecting individuals to the correct names and e-mails, a process called identity resolution in the field of computing, was time consuming and usually required direct communication with participants. Identity resolution was further complicated by the fact that some study participants had the same or similar names, sometimes within the same classroom.

External Vendor Systems. Assessment vendors (e.g., a company that hosts a website through which participants complete a test) and school districts were *external vendors* who provided participant data and hosted instruments in both studies. Over 65 school districts across 22 states provided demographic information and state standardized test scores for middle school student participants in Study A. Assessment providers host copyrighted instruments on their own websites and had specific rules regarding how paper versions of their assessments could be administered. Study A and Study B both used assessments offered through the University of California at San Diego's Mathematics Diagnostic Testing Project (MDTP).

External vendors typically used varied and conflicting conventions for data values. For example, in Study A, the ethnicity definitions across elementary school districts were wide-ranging. The category of Black or African American had values such as: 1, 4, Black/African American, Bl, and African Am. We decided to use standard categories provided by the National Center for Education Statistics[1] and transformed data values into these conventions. Confidentiality was also an issue because external vendors needed to identify participants but could not be provided with the identifiers used by our research team. If external vendors saw our research identifiers, they could easily identify specific participants in our publicly released datasets. This necessitated an additional set of interim identifiers to allow our data cleaning team to map data received from external vendors with our own participant records.

[1] Common Education Data Standards (CEDS) also provides a data dictionary for information related to pre-school through post-secondary educational environments that can be used to establish conventions for an educational research study.

Data Cleaning Process

My experiences with Study A highlighted the need to attend to data cleaning at all phases of the research cycle. I developed a list of tasks to accomplish during data planning, data collection, and data cleaning to minimize the issues likely to occur in educational data sets (see Table 1). My goal was not to eliminate issues from occurring, but rather, to create a reproducible process to improve the identification and handling of data errors in the cleaning process. Below I describe these tasks and their rationales.

Table 1

Data Cleaning Tasks

Task
Data Planning Phase
Create visually distinct instrument forms
Clarify data requirements and timeline with external vendors
Set administration windows and a last enrollment date
Determine decision rules for handling duplicate records
Develop codebooks for each instrument to describe variables and their possible values
Data Collection Phase
De-identify study data as early as possible in the data collection process
Collect details on how assessments were administered and any anomalies that occurred
Create three sets of identifiers for participants: one used by data collectors, one used by researchers, and one used by external vendors
Make sure identifiers do not depend on malleable participant characteristics
Verify administration dates on completed study instruments are valid
Data Cleaning Phase
Use tools that allow you to log and retrace your data cleaning steps
Establish a review process so data cleaning work can be checked by another person
Make a copy of your raw data file and only work with the copy
Check data files for missing and extra data columns
Apply codes to indicate types of missing data (e.g., not completed, not administered, optional)
Transform categorical values into pre-determined standard values
Check identifier columns for duplicate values
Flag records with errors
Indicate administration format in final data sets (e.g., completed on paper or online)

Data Planning Phase. In preparing to launch a study, research teams can facilitate future data cleaning by carefully planning the design of instrument forms, data collection timelines, decision rules for rejecting data, and instrument codebooks. Although these tasks are presented sequentially, in practice they can occur in any order and even concurrently. First, instrument forms should be visually distinct to help participants, administrators, and data collectors notice when an instrument was completed on the wrong form. This can be accomplished, for example, by using different colors, specifying instrument names and dates, and customizing templates to display the number of items and answer choices corresponding to the instrument form (see Figure 1). Second, research teams need to establish timelines for data collection and participant

enrollment. This will require working with external vendors to understand the time and information they need to set up instruments or gather data. During this phase, research teams should familiarize themselves with the school calendars of study participants. For example, it would be extremely difficult for a middle school teacher to administer a study assessment during the week of state testing. Lastly, research teams need to consider the structure of their data and rules for rejecting data when issues occur. This can be accomplished by creating codebooks for each study instrument. Decisions for rejecting data may need to be made on a case-by-case base, but during the data planning phase research teams can consider the following question:

- When is it too late to accept data?
- How much of an instrument needs to be completed to be included in the dataset?
- How do we handle duplicate responses?

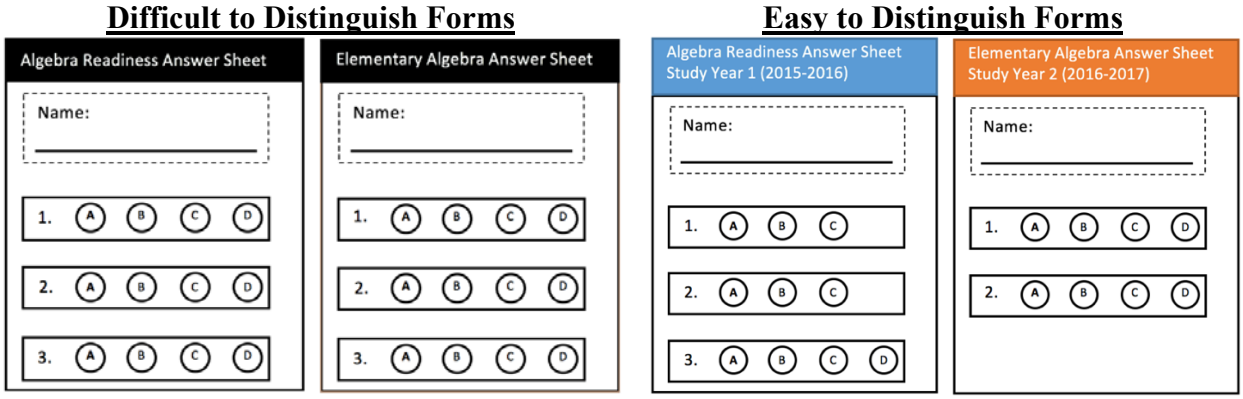

Figure 1. Sample Instrument Forms.

Data Collection Phase. Once a study has started, research teams can implement processes that make it easier to track data and to share real-time information with external vendors. First, de-identify study data as early as possible. The rationale behind this task is that people closer to participants will find it easier to identify who completed an instrument. When using paper instruments, we send teachers a list of barcode stickers that they distribute to specific students who then place the barcodes on their own assessments. These barcodes contain the student's study identifier and the instrument name which can be quickly entered into a tracking system using a barcode scanner. Second, research teams should collect information on instrument administration and any anomalies that occurred during administration. This can take the form of a feedback sheet administrators complete and send back with study data. When using online systems, this might require communicating with external vendors about any issues that appeared during administration (e.g., servers going down). Third, create unique identifiers for each study participant. Identifiers should not depend on characteristics that might change during a study. For example, a student identifier should not include the identifier of the student's teacher, because the student may change teachers or schools during the study. Also, if working with external vendors or external analysts, then an additional set of identifiers should be created to share with these audiences. Figure 2 provides an example of a data file shared with external vendors and analysts. Lastly, data collectors should also check administration dates to make sure they are within acceptable ranges. This can help identify anomalous data (e.g., students entering birthdates).

Original Data File

Internal_ID	Vendor_ID	Analyst_ID	Item_1	Item_2	Item_3
STU023	VSTU010	ASTU771	A	A	B
STU055	VSTU222	ASTU823	A	B	B
STU023	VSTU987	ASTU802	C	E	B

External Vendor Data File

Vendor_ID	Item_1	Item_2	Item_3
VSTU010	A	A	B
VSTU222	A	B	B
VSTU987	C	E	B

Analyst Data File

Analyst_ID	Item_1	Item_2	Item_3
ASTU771	A	A	B
ASTU823	A	B	B
ASTU802	C	E	B

Figure 2. Sample Data Files.

Data Cleaning Phase. After data has been collected, it should undergo a series of validation checks to identify and repair anomalies. Prior to beginning this process, data cleaners should identify a tool that will log steps used to transform data files into their final format. This will make it easier to reproduce files and to retrace steps in case errors are identified later. Our teams have worked with R (https://www.r-project.org/) and OpenRefine (http://openrefine.org). Second, data cleaners should establish a review process so that their work can be verified by another person. When working with many data files, it is easy to make simple mistakes (e.g., assigning a string value like 'treatment' the wrong numeric value). We implement this review process in two ways. First, we have each data cleaning script reviewed by someone who did not author the script. Second, we ask our participant managers and researchers to compare our file counts against their records to identify missing or extra participant records. Next, data cleaners should create copies of their raw data files and only work with the copies. This allows you to return to the original version if needed. To distinguish these files, we add the suffix *_raw* to our original files. Once data cleaners have setup these processes, they can begin working on the data cleaning checks listed in Table 1.

Figures 3 to 6 provide an example of data cleaning checks applied to a student questionnaire. First, the data file is compared against a codebook to identify any missing or extra data columns (Figure 3). Extra columns arise frequently in data files provided by external vendors or captured through online survey tools. Columns can go missing due to software failures or because they were simply overlooked. Second, missing data values are replaced with codes indicating the reason for their absence (Figure 4). I distinguish three types of missing data: data I expected to receive that was not provided, data I did not expect to receive, and data that were optional. Next, categorical values from an open-ended response item are transformed into a limited number of standard values (Figure 5). Demographic data often require mapping to standard values. Where possible, I use values common in educational work (e.g., NCES conventions) to support external researchers in comparing their own data against the data files I produce. Lastly, the unique identifier column is reviewed for duplicate values (Figure 6).

Check for missing and extra columns

By comparing the data file against the codebook, we see *Math Club* is missing and *School ID* was added. We need to work with data collectors to retrieve the missing Math Club information. The School ID column can be deleted.

School ID	Study ID	Grade	Grade 9 GPA	Grade 10 GPA	Elective
111-11-1111	STU01	9	2.0		Engineer
222-22-2222	STU01	9	3.2		Choir
333-33-3333	STU02		3.5	3.0	French

Codebook

Study ID	Unique ID
Grade	9-12
Grade 9 GPA	0.0 – 4.0
Grade 10 GPA	0.0 – 4.0
Elective	STEM Non-STEM
Math Club	Yes, No

Figure 3. Checking questionnaire data for missing and extra columns.

Apply codes to indicate types of missing data

The students in the first two rows are missing a Grade 10 GPA. Since they are in 9th grade, we expect their Grade 10 GPA values to be missing. We indicate missing values that are expected with 888888. The student in the last row is missing a value in the grade column, but we expect all students to have a grade level. We indicate missing values that are unexpected with 999999. Missing codes should stand out from other values in the same column.

Study ID	Grade	Grade 9 GPA	Grade 10 GPA	Elective	Math Club
STU01	9	2.0	888888	Engineer	Yes
STU01	9	3.2	888888	Choir	Yes
STU02	999999	3.5	3.0	French	No

Figure 4. Applying missing codes to questionnaire data.

Transform categorical values into standard values

According to the codebook, the *Elective* column should only contain the values of STEM or Non-STEM. We map fields such as engineering to STEM and other fields to Non-STEM.

Study ID	Grade	Grade 9 GPA	Grade 10 GPA	Elective	Math Club
STU01	9	2.0	888888	STEM	Yes
STU01	9	3.2	888888	Non-STEM	Yes
STU02	999999	3.5	3.0	Non-STEM	No

Figure 5. Transforming categorical values in questionnaire data.

Check identifier columns for duplicate values

The first two rows contain the same value in the *Study ID* column, but our codebook indicates this column should be unique. Looking back to our original data file, we see these students had different values in *School ID*. We would need to work with data collectors to identify if these records represent two different students. In the meantime, we add a column to flag that there is an error with these records.

Study ID	Grade	Grade 9 GPA	Grade 10 GPA	Elective	Math Club	Has Error
STU01	9	2.0	888888	STEM	Yes	1
STU01	9	3.2	888888	Non-STEM	Yes	1
STU02	999999	3.5	3.0	Non-STEM	No	0

Figure 6. Checking questionnaire data for duplicate values.

Communication Processes

Given the large scopes of Study A and Study B, data cleaning was completed by several individuals and required working with staff involved in other aspects of the projects. In Study B, for example, staff were divided into (a) a participant team responsible for recruitment and participant management, (b) a data management team responsible for data collection and cleaning, (c) a research team responsible for study design and analysis, and (d) a management team responsible for project planning and coordination. Working within and across teams to accomplish data cleaning work was not straightforward nor free from error.

As an example, several participating teachers were erroneously excluded from Study A data files because of differences between the participant team and the research team's understanding of *participant*. For the participant team, a 'participant' was someone who started study tasks and had communicated with the project staff at some point. However, for the research team, a 'participant' was any person who enrolled in the study and was randomized into a study condition, regardless of the number of study tasks completed. The error was uncovered when a research team member compared a data file record count against the randomization record count. As a bridge between participant teams and research teams, our data teams needed to navigate across different group norms and discourses to accomplish our work. Next, I briefly summarize the communication and collaboration processes I now use with other project staff to facilitate data cleaning tasks.

Prior to the data cleaning phase, I meet with both the research team and the participant team to discuss their study plans. For the research team, I review their list of data sources and prepare a low-tech sketch of each data file that includes a file name, data columns, and sample values (see Figure 7). Reviewing this sketch with the research team provides confirmation on the data files to be produced, helps to identify if additional files are needed (e.g., a master data file combining information from multiple files), and establishes a shared terminology. With the participant team, I gather information on recruitment deadlines, administration windows they have shared with participants, and school calendars, which inform data validity checks and the data cleaning timeline. During these conversations, I also attend to the ways in which the research team and the participant team discuss their work, being vigilant for possibly confounding terminology. After I meet with both teams, I produce an accounting spreadsheet listing each data file to be created and an estimated due date (see Figure 8).

Once data cleaning begins, I meet regularly with members of the research and participant teams to stay abreast of study progress that might impact data cleaning. For example, the participant team may decide to extend the administration window for a background questionnaire because a new class of students joined the study later than expected. Or, the research team may decide to include additional items on an attitude survey before the second administration of that instrument. These frequent meetings help to identify study plan changes that other project staff may not realize impact data cleaning procedures. When such changes occur, I record them in an appropriate documentation source such as the data accounting spreadsheet.

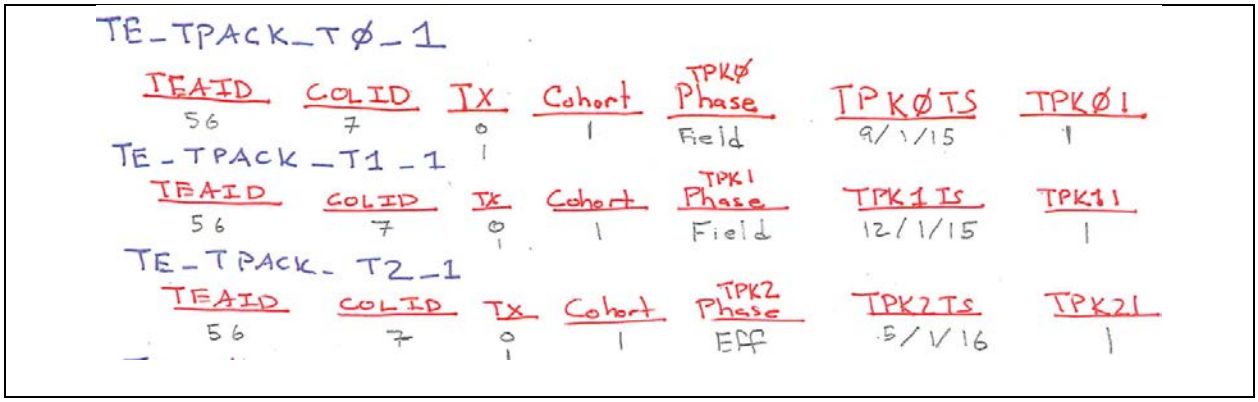

Figure 7. Low-tech sketch of a teacher (TE) assessment of teaching knowledge (TPACK). Teachers completed the assessment at the beginning (T0), middle (T1), and end (T2) of the school year.

Instrument	Pre/Post	Assigned to	Base File Name	Date Due	Status
TE_TPACK	Pre	Tammi	F.1.Tea.Tpk.0.V0	11/25/15	Finished
TE_TPACK	Mid	Tammi	F.1.Tea.Tpk.1.V0	01/15/16	Started
TE_TPACK	Post	Jean	F.1.Tea.Tpk.2.V0	06/30/16	Not started

Figure 8. Data accounting spreadsheet.

Discussion

While the topic of data cleaning may seem tangential to research in undergraduate mathematics education, I hope I have highlighted the critical role data cleaning plays in our research practices. The messiness of environments within which educational research studies are conducted necessitate attention to how we collect and prepare study data. The work presented here demonstrates that processes can be put in place to facilitate the efficient production of quality data sets. The data cleaning process, list of common data error sources, and communication processes offered here provide a framework for other researchers to evaluate their current data management strategies and to provide more comprehensive methods training for researchers.

For readers implementing the framework or embarking on their own data cleaning projects, I offer two recommendations. First, it is important to acknowledge that a comprehensive data cleaning process cannot be created a priori. Studies evolve, participants change, and unexpected events occur. It is impossible to predict all of these variations before a study begins. Implementing an initial data cleaning process that is flexible can help you plan for common errors while giving you the space to adopt procedures as needed in the future. Second, documentation of decisions, processes, and data files is essential. Recording such information helps with accountability and communicating across teams during the data cleaning phase. Documentation also helps researchers recall their procedures and reproduce their work months (even years) after studies have finished and data cleaning decisions are distant memories.

Lastly, some readers may wonder if the procedures presented here apply beyond large-scale, quantitative studies. I argue that all research data needs to undergo some level of cleaning before analysis. While the specific checks of the framework may not apply to all studies (e.g., duplicate records may not be an issue in a case study with one participant), its underlying ideas are relevant to other types of research. In the qualitative research I conduct, my data undergo condensation or "the process of selecting, focusing, simplifying, abstracting, and/or transforming the data that appear in the full corpus (body) of written-up field notes, interview transcripts, documents, and other empirical materials" (Miles & Huberman, 2013, p. 12). I still review my transformed qualitative data to ensure all records are accounted for and no erroneous values exist (e.g., a participant being mistakenly labeled as a teacher instead of a student). At the heart of data cleaning is the acknowledgement that data errors can occur in our studies and that rigorous research practices involve correcting them before analysis.

Acknowledgements

I would like to thank Bryan Matlen and Shandy Hauk for encouraging the writing of this paper and their reviews of my earlier drafts. I would also like to thank Yvonne Kao, Katie D'Silva, Kimkinyona Cully, Eunice Chow, Danielle Oberbeck, and Mykael Thompson for their collaborations on the Study A and Study B data teams.

This project is supported by grants from the U.S. Department of Education, Institute of Education Sciences (IESR305A140340; IESR305C100024). Any opinions, findings, conclusions or recommendations are those of the authors and do not necessarily reflect the views of the Federal Government.

References

Creswell, J. W. (2012). *Educational Research: Planning, Conducting, and Evaluating Quantitative and Qualitative Research*. Addison Wesley.

de Jonge, E., & van der Loo, M. (2013). *An introduction to data cleaning with R.* Statistics Netherlands, The Hague.

Gallagher, H. A. (2014). Recruiting Participants for Randomized Controlled Trials. *Society for Research on Educational Effectiveness.*

Leahey, E., Entwisle, B., & Einaudi, P. (2003). Diversity in Everyday Research Practice: The Case of Data Editing. *Sociological Methods & Research*, *32*(1), 64–89.

Miles, M. B., Huberman, A. M., & Saldaña, J. (2013). *Qualitative Data Analysis: A Methods Sourcebook* (3 edition). Thousand Oaks, California: SAGE Publications, Inc.

Munson, M. A. (2012). A Study on the Importance of and Time Spent on Different Modeling Steps. *SIGKDD Explor. Newsl.*, *13*(2), 65–71. https://doi.org/10.1145/2207243.2207253

Odom, L. R., & Henson, R. K. (2002). *Data Screening: Essential Techniques for Data Review and Preparation.* Paper presented at the Annual Meeting of the Southwest Educational Research Association, Austin, TX.

Osborne, J. W. (2013). *Best practices in data cleaning: A complete guide to everything you need to do before and after collecting your data.* SAGE Publications, Inc.

Rahm, E., & Do, H. H. (2000). Data Cleaning: Problems and Current Approaches. *Bulletin of the Technical Committee on Data Engineering*, *23*(4), 3–13.

Schleicher, A., & Saito, M. (2005). Data Preparation and Management. In K. N. Ross (Ed.) *Quantitative Research Methods in Educational Planning*. IIEP/UNESCO.

Smith, S. R. (2013). *Supporting California's Community College Teaching Faculty: Improving Working Conditions, Compensation and the Quality of Undergraduate Education*. Berkeley, CA: University Professional & Technical Employees Communications Workers of America. Retrieved from http://www.upte.org/cc/supportingfaculty.pdf

Welsh, R. O. (2016). School Hopscotch: A Comprehensive Review of K–12 Student Mobility in the United States. *Review of Educational Research*, 34654316672068. https://doi.org/10.3102/0034654316672068

Whalley, A. (2011). Education and Labor Market Risk: Understanding the Role of Data Cleaning. *Economics of Education Review*, *30*(3), 528–545.

Wilkinson, L. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*(8), 594–604. https://doi.org/10.1037/0003-066X.54.8.594

U.S. Department of Education, Institute of Education Sciences, What Works Clearinghouse. (2016). What Works Clearinghouse: Procedures and Standards Handbook (Version 3.0). Retrieved from http://whatworks.ed.gov

U.S. Department of Health and Human Services (HHS) Office for Human Research Protections (OHRP). (1993). *IRB Guidebook*. Retrieved from http://wayback.archive-it.org/org-745/20150930181805/http://www.hhs.gov/ohrp/archive/irb/irb_guidebook.htm