

Cleaning and preparing data for analysis

Avery Hennigar, Rebecca Piatt, Sonia Alves, and Daniel Friend

Consistent data monitoring and regular data cleaning help evaluation teams quickly address errors and improve data quality for analysis and reporting (Osborne 2010). Without data monitoring and cleaning, staff and participants can unknowingly introduce errors that can lead to incorrect or biased conclusions. This brief describes how data errors can occur throughout a Healthy Marriage and Relationship Education (HMRE) evaluation and provides tips for how to avoid them.

Common sources of errors

Data errors can be introduced at various stages of an evaluation, including during data collection, entry and cleaning, and analysis. Regardless of the source, data errors are problematic if they are not identified and fixed. Such errors can introduce issues in the analysis and lead evaluators to incorrect conclusions.

- **Errors during data collection.** Data collection errors can occur when participants inadvertently select the wrong responses, misunderstand questions that may be confusingly worded, provide responses outside of the expected ranges, or enter values in the incorrect field. Logistical issues with survey administration also can cause errors, such as incorrectly programming a survey's skip logic. HMRE local evaluators and program staff may also introduce errors during data collection by administering surveys inconsistently.
- **Errors during data entry and cleaning.** Several mistakes can occur during data entry and data cleaning—including incorrectly transferring data from the survey, entering values in the wrong field, entering values incorrectly, or accidentally deleting or duplicating entries.
- **Errors during data analysis.** Evaluators can create data errors by incorrectly extracting data from a database or assigning the wrong values to a variable during data cleaning.

Box 6.1. Three examples of errors

1. Consider an HMRE program that uses an online survey for its evaluation. During survey development, the evaluator incorrectly programmed a skip pattern, which resulted in participants missing questions about their parenting skills. Because the data set includes incomplete data, the evaluator cannot fully answer the research questions at the end of the study.
2. An evaluator correctly entered the survey responses into the database for a question with a 5-point rating scale by coding responses 1 (strongly disagree) through 5 (strongly agree) according to the codebook's guidelines but mistakenly coded "I don't know" values as 6, instead of as a missing data value. This is particularly problematic if these data are later summed or averaged. The values will be inaccurate and not comparable to other published studies because they will include an additional value that potentially increases the sum or mean.
3. To create a summary score, some items might be coded as they are and others might be reverse-coded (with high values switched to low values and vice versa), so higher values always represent a favorable (or unfavorable) response. If an evaluator does not reverse-code items correctly during data cleaning, or the values are flipped when they should not be, the analysis will be incorrect.

The Administration for Children and Families (ACF) provides grants to fund healthy marriage and relationship education (HMRE) programs to strengthen and improve the quality of relationships. The programs offer a range of services from relationship education for high school students to marriage and relationship skills building for adult couples. Grant recipients may be funded to also conduct descriptive or impact evaluations of their funded programs. Independent local evaluators support grant recipients in conducting their local evaluations. This brief is part of a larger evaluation technical assistance (TA) toolkit developed by Mathematica to help HMRE local evaluators understand key program evaluation concepts, common evaluation challenges, and strategies to prevent or overcome challenges. The briefs are standalone documents that can be read in any order. The TA toolkit was developed with HMRE program staff, their local evaluators, and other partners in mind, but it is also relevant to other program areas and organizations.

Using strong data preparation and cleaning methods throughout the evaluation can reduce the risk of such errors. Preparing and checking data for analysis should occur on an ongoing basis—rather than after data collection is complete. HMRE local evaluators can use the following four tips to ensure that their data are accurate, free of errors, and ready for analysis.



Tip 1: Create a comprehensive plan for data collection and entry

An important first step when launching an evaluation is to create a comprehensive data collection plan. A thorough plan will help HMRE programs and evaluators achieve high-quality data collection that is consistent, complete, timely, and secure. The data collection plan should be a living document. Evaluators should update it and any associated materials regularly (HMRF Resource Site for 2020 Grantees 2022). The plan should include the following details (Box 6.2, Institute of Education Sciences 2021):

- Evaluation design and research questions
- Description of sample
- Data sources
- Data collection and analysis methods
- Person or persons responsible for data collection
- Security and storage method¹
- Evaluation timeline

Box 6.2. Data collection plan template

Evaluators can use the [Data Collection Plan Template](#)² on the Information, Family Outcomes, Reporting, and Management (nFORM) resources page to compile these details.

HMRE local evaluators should provide training to the program and evaluation staff who will be involved in data collection and data entry (Avellar et al. 2017). The training should provide guidance to staff on administering the surveys correctly and consistently across participants, which will reduce the risk of errors and bias. The training approach and mode should be based on the staff's evaluation experience—for example, less experienced staff may need more intensive in-person training, while more experienced staff may benefit from shorter virtual or prerecorded trainings to refresh their knowledge. Numerous nFORM resources, such as the user manual, training videos, and monthly office hours, are available to use when developing staff trainings (HMRF Resource Site for 2020 Grantees 2023).

Tip 2: Develop procedures for regularly monitoring, cleaning, and checking data for errors



Data cleaning is the process of detecting (for example, identifying strange or unexpected patterns), diagnosing (for example, identifying true extreme values vs. true normal values), and editing (for example, correcting, deleting, or leaving data unchanged) faulty data. The process of data cleaning can help create a final data set that is as accurate, complete, and ready for analysis as possible (Institute of Education Sciences 2021; Van den Broeck et al. 2005).

Box 6.3. Remember to preserve original data files

Maintain an original copy of the data. Never alter, trim, or recode data in the original file. Before data cleaning begins, create a copy of the data set and then a clean data file from the copy.

Evaluators often create a codebook to outline how variables will be cleaned (but not replaced) (Box 6.3). A codebook is a useful document for establishing clear coding procedures for survey instruments (Institute of Education Sciences n.d.). The nFORM resources include a [data dictionary](#),³ which can be a useful starting place for creating a codebook. HMRE evaluators may consider incorporating additional

¹ It is important to include proper security protections in an evaluation's application to an institutional review board and to follow them throughout the study. For example, evaluators should create guidelines for storing hard-copy files of survey instruments in locked filing cabinet drawers and for using password protection for spreadsheets that include participant data.

² HMRF Resource Site for 2020 Grantees, nFORM 2.0 Team. "Data Collection Plan Template." Administration for Children and Families, 2022. <https://www.hmrfgrantresources.info/resource/data-collection-plan-template>.

³ HMRF Resource Site for 2020 Grantees. "Data Dictionary." Administration for Children and Families, 2022. <https://www.hmrfgrantresources.info/resource/data-dictionary>.

variables from their local evaluation into a copy of the nFORM data dictionary to ensure consistency in participant IDs and demographic information. Relevant staff should be trained on how to navigate the codebook so that they feel comfortable using it.

HMRE local evaluators can use a variety of data cleaning techniques to screen, diagnose, and edit identified data errors. In addition to checking for errors, data cleaning can include standardizing variable types, such as changing all dates to follow the same format, and running a spell-check to correct misspellings that could cause coding or frequency count errors. Cleaning data regularly can identify patterns in the data errors so that they can be fixed early in the data collection process. In addition, developing a clean, organized data file can make the data archiving process more efficient. Many data archiving repositories require files to be accessible and interpretable, so data cleaning is an important step in preparing files for submissions.

Table 6.1 provides examples of techniques for checking errors during the data cleaning process and how to remedy the errors during or after data cleaning.

Table 6.1. How to catch and fix common data errors

Technique	Description	Example error	Example remedy
<i>Descriptive analysis</i>	Evaluators can calculate descriptive statistics—such as mean, range, or the distribution of variables—to make sure they seem reasonable and theoretically possible.	An evaluator of an HMRE program that serves youth in high school checks the age variable on the local evaluation survey and notices that there are several maximum values of 34 years old. However, the program only serves youth up to age 24, so these values are implausible.	The evaluator decides to restrict the range in the age variable on the survey to a maximum value of 24 years old.
<i>Double entry</i>	When conducting data entry, evaluators can ask two or more staff to enter the same survey, then check for discrepancies. Double entering each survey is the gold standard. However, if resources are limited, consider selecting a small subset of surveys to double-enter.	After comparing double-entered surveys, the lead evaluator of an HMRE program that serves Hispanic couples notices that one of the data entry staff members entered the ethnicity variable for all couples as a value of 1 in the data file, while another staff member entered the ethnicity variable for the same couples as 2 in the data file.	The lead evaluator checks the codebook for how data should be entered and verifies that the ethnicity variable for Hispanic couples should be coded as 1. The lead evaluator instructs the staff member who entered erroneous data to correct past data entered and use 1 moving forward.
<i>Logic checks</i>	Evaluators can review whether answers to related questions make logical sense. With online surveys, logical skip patterns or checks can be built into the survey to expedite data cleaning.	When reviewing survey data, an evaluator staff member notices that participants who have not had past experiences with intimate partner violence (IPV) are being asked to complete the section on IPV.	The staff member notifies their supervisor to correct the survey logic.

Technique	Description	Example error	Example remedy
<i>Data visualizations</i>	Evaluators can use graphs, such as bar charts or scatter plots, to identify improbable values or outliers.	After conducting a cohort of workshops, an evaluator creates a scatter plot of the relationship between workshop attendance and demographic characteristics, a key research question in the program's evaluation. The evaluator notices that attendance recorded in nFORM is low among mothers, even though written attendance records show that most mothers participated in make-up sessions and completed 90 percent of the workshop content.	The evaluation team convenes a meeting with HMRE program staff to compare nFORM and written attendance records and make sure that nFORM data entry is timely and accurate for this and future cohorts.

Source: Adapted from Institute of Education Sciences (2021).

Tip 3: Interpret unexpected findings alongside the program team

Collecting feedback from people with a variety of experiences - including program staff, evaluators and participants - can promote stronger research and program improvement.

HMRE local evaluators can share preliminary analyses through reports, presentations, or data visualizations on a regular cadence, such as monthly or quarterly, or during technical assistance (TA) meetings with family assistance program specialists (FPSs) and evaluation technical assistance partners (ETAPs). During these calls, evaluators can share preliminary findings and ask program staff to provide additional context for them.

For example, consider an evaluation that finds that more youth are reporting relationship violence at program exit compared to baseline—even after gaining skills to recognize unhealthy relationship patterns or abusive partnerships through the program. Discussing such findings with program staff may suggest ways to change the data collection procedures or instruments to improve data quality (Box 6.4). If evaluators make tweaks to existing processes or procedures, they should clearly document their decisions and report them in the study's final report.

Box 6.4. Reflect on the data collected with program and evaluation staff

Program staff may offer important insights into program operations that could explain unexpected findings and inform tweaks to data collection. Suppose the evaluation team notices findings trending in the opposite direction than expected, such as decreases in relationship satisfaction after program completion. Evaluators can engage staff in discussions about why this is happening. These conversations might reveal that variables are not being reverse-coded correctly or that the program served a greater number of couples in unstable relationships—who break up after receiving more information about healthy relationships. The evaluator can use this context to (1) correct the way data are coded; or (2) describe the study's unexpected findings in their final report.



Tip 4: Cautiously correct data errors

If evaluators find data errors, they can address them by removing duplicates, correcting entry errors, and removing outliers when appropriate. More advanced techniques include imputing data (that is, replacing missing values or incomplete data with estimated values) or transforming variables (that is, changing the scale, format, or distribution of data to make it more consistent or suitable for a particular analysis). Determining how to correct data errors is a nuanced process. Evaluators should carefully select an approach, as it can influence the analyses and types of conclusions that can be drawn (Box 6.5).

It is important to document how each data error is handled so that it can be explained in final reports and other publications. This will increase transparency and allow other researchers to accurately reproduce results. Decisions on how to handle data errors should not be made in isolation, but rather discussed as a program and evaluator team with input from TA providers, FPSs, and other staff or experts who are familiar with the program and data.

Box 6.5. Not all outliers should be removed

Finding an outlier in the data does not necessarily mean it is incorrect and should be removed. Outlier values should only be removed if they are determined to be an error after careful scrutiny.

References

- Avellar, Sarah, Kelley Borradaile, and Debra Strong. "Evaluation Technical Assistance Brief: Tips for Enrolling and Retaining Evaluation Participants." Washington, DC: Mathematica Policy Research, November 2017.
- Healthy Marriage and Responsible Fatherhood. "nFORM Performance Measures and Data Collection Logistics Manual." March 2023. https://hmrfggrantresources.info/sites/default/files/2023-03/PMDC%20Logistics_Mar2023.pdf. Accessed July 7, 2023.
- Healthy Marriage & Responsible Fatherhood (HMRF) Resource Site for 2020 Grantees. "Propelling Forward to Grant Year 3." September 2022. <https://www.hmrfggrantresources.info/resource/propelling-forward-grant-year-3>. Accessed October 27, 2023.
- Institute of Education Sciences. "Guidelines for a Codebook." n.d. <https://ies.ed.gov/ncee/edlabs/regions/central/resources/pemtoolkit/pdf/module-7/CE5.3.2-Guidelines-for-a-Codebook.pdf>. Accessed October 30, 2023.
- Institute of Education Sciences. "Program Evaluation Toolkit." October 2021. <https://ies.ed.gov/ncee/rel/Products/Region/central/Resource/100644>. Accessed July 7, 2023.
- Osborne, Jason W. "Data Cleaning Basics: Best Practices in Dealing with Extreme Scores." *Newborn and Infant Nursing Reviews*, vol. 10, no. 1, March 2010, pp. 37–43. <https://www.sciencedirect.com/science/article/pii/S1527336909001779?via%3Dihub>.
- Van den Broeck, J., S. Argeșeanu Cunningham, R. Eeckels, and K. Herbst. "Data Cleaning: Detecting, Diagnosing, and Editing Data Abnormalities." *PLoS Medicine*, vol. 2, no. 10, 2005, pp. 966–970. <https://doi.org/10.1371/journal.pmed.0020267>.

Suggested citation: Hennigar, Avery, Rebecca Piatt, Sonia Alvez, and Daniel Friend (2024). "Cleaning and Preparing Data for Analysis." OPRE Report #2024-144. Washington, DC: Office of Planning, Research, and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.

Acknowledgements: Many people contributed to this toolkit. First, we acknowledge staff at the OPRE in the Administration for Children and Families at the U.S. Department of Health and Human Services. We are particularly grateful for the direction and feedback from Samantha Illangasekare, Rebecca Hjelm, and Kathleen McCoy. We want to extend our gratitude to Sarah Avellar and Angela D'Angelo for reviewing drafts of the briefs. We also extend our appreciation to Effie Metropoulos and Bridget Gutierrez for editing and Yvonne Marki-Korosec and Gwyneth Olson for designing the graphics in this toolkit.