How to Improve Data Validation in Five Steps

Danilo Freire

MERCATUS WORKING PAPER

MERCATUS CENTER
George Mason University
3434 Washington Blvd., 4th Floor, Arlington, Virginia 22201
www.mercatus.org

Danilo Freire, "How to Improve Data Validation in Five Steps," Mercatus Working Paper, Mercatus Center at George Mason University, Arlington, VA, March 2021.

Abstract

Social scientists are awash with new data sources. Though data preprocessing and storage methods have developed considerably over the past few years, there is little agreement on what constitutes the best set of data validation practices in the social sciences. In this paper I provide five simple steps that can help students and practitioners improve their data validation processes. I discuss how to create testable validation functions, how to increase construct validity, and how to incorporate qualitative knowledge in statistical measurements. I present the concepts according to their level of abstraction, and I provide practical examples on how scholars can add my suggestions to their work.

JEL Codes: A20, C80, C88

Keywords: computational social science, computer programming, content validity, data validation, tidy data

Author Affiliation and Contact Information

Danilo Freire
Independent researcher
https://danilofreire.github.io
danilofreire@gmail.com

This paper can be accessed at https://www.mercatus.org/publications/regulation/how-improve-data-validation-five-steps.

How to Improve Data Validation in Five Steps

Danilo Freire

Introduction

Social scientists are awash with data. According to a recent estimate, humans produce about 2.5 billion gigabytes of information every day, and 90 percent of the global data currently in existence were created in only two years (IBM 2016). Governments also generate more data than ever before, with a growing number of agencies opening their archives and allowing users to access public records directly or via application programming interfaces, or APIs (e.g., Al-Ubaydli and McLaughlin 2017; McDonnell, Duarte, and Freire 2019).

In this context, researchers have developed a series of tools to obtain, clean, and store data files. R and Python, two open-source programming languages, have become the de facto standards for downloading and manipulating data (Magoulas and Swoyer 2020; Perkel 2018). Reproducible scripts are now a common feature in academic studies, so researchers can easily share and verify their analyses (Höffler 2017; Key 2016). Scholars can also store data and code in public repositories, such as GitHub or the Harvard Dataverse, which guarantee that academic materials will be preserved for future reference (King 2007).

Although there has been significant progress in data analytics, data validation techniques have received little attention in academia. Data validation is defined as "*an activity verifying whether or not a combination of values is a member of a set of acceptable combinations*" (ESSnet ValiDat Foundation 2018, 8; italics in original). One reason for this omission is that data quality procedures are not as standardized as other statistical methods, so users often need to create ad hoc semantic rules to compile new data (McMann et al. 2016). Moreover, scholars have to engage with several sources of information to establish conceptual validity and translate abstract concepts into plausible numeric values (Munck and Verkuilen 2002; Schedler 2012).

In this paper, I present five steps to help scholars improve their data validation process. My idea is to offer a short checklist that is useful to both data developers and data reviewers, so all parties involved in the validation procedures have a common understanding of what constitutes good practices in the field. My suggestions are based on the standards set by Eurostat, an agency that provides statistical information to the European Union, and on recommendations by Gerring (2001), McMann et al. (2016), and Schedler (2012). I discuss how to create testable validation functions, how to increase construct validity, and how to incorporate qualitative knowledge in statistical measurements. I present the concepts according to their level of abstraction, with the last three demanding more theorization than the first two, and I provide practical examples for how scholars could add my suggestions to their work.

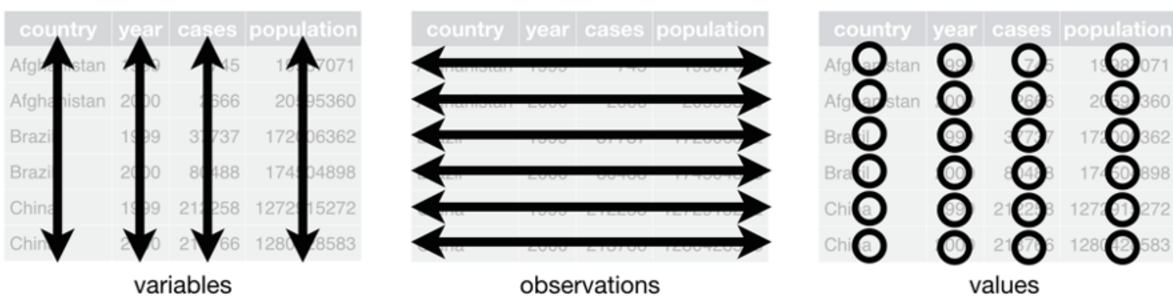
Five Steps toward Better Data Validation Processes

Step 1: Technical Consistency

Ensuring technical consistency is perhaps the easiest task in the data validation process, yet it is often overlooked even by experienced scholars. Technical consistency means that the data should be machine readable and as intuitive as possible to humans. More specifically, scholars have to ensure that their data do not produce parsing errors, that values corresponding to variables and observations are clearly identified, and that other researchers can analyze the data as soon as they receive them.

The computer science literature has some important suggestions in this regard. First, data should be "tidy"—that is, in a format where each column represents one variable, each row represents one observation, and each observational unit forms a table (Wickham 2014, 4). Although the definition seems intuitive, scholars sometimes break these rules when building a new dataset. It is not uncommon to see multiple variables stored as a single column (e.g., race and gender together) or one variable divided into many columns (e.g., one column per level of income). Fortunately, most datasets can be tidied up with simple operations, such as pivoting tables, splitting or combining columns, or filtering values (see figure 1). Please refer to Wickham and Grolemund (2016) for more information on how to clean messy data.

Figure 1. A Tidy Dataset



Source: Wickham and Grolemund (2016).

It is important to add an identifying variable (primary key) that is unique across all records (van der Loo and de Jonge 2019, 1). Doing so is particularly relevant when scholars need to merge different datasets, because the primary keys have to be the same across all tables. Also, variables should have appropriate cell formats, or the machine may not be able to parse the values correctly. For instance, users should store strings as character vectors, numbers as numeric values (floating points or integers), binary indicators as Boolean vectors (TRUE or FALSE), and categorical variables as factors (ESSnet ValiDat Foundation 2018, 10). Whenever possible, one should save tabular data in pure-text format, as any software can read that format.

Comma-separated values files (.csv) have wide support across many operating systems and can be easily opened in Excel, R, and Python with no errors.

Variables containing textual data or times and dates need special attention as they are prone to encoding issues. There are dozens of character encoding standards, but a recent survey shows that Unicode Transformation Format, 8-bit (UTF-8), is used on 96 percent of all internet websites (W3 Techs 2020). In this sense, scholars should use UTF-8 to store text variables because it currently is the world's most popular text encoding. UTF-8 is able to convert characters from many alphabets, and its recent version also stores emojis. Regarding times and dates, scholars are advised to record time in the ISO 8601 standard. The ISO standard uses the Gregorian calendar and a 24-hour timescale (van der Loo and de Jonge 2018, 52). Dates are stored in the YYYY-MM-DD format, in which YYYY represents the year, MM the month, and DD the day. Time of day should be expressed as hh:mm:ss, for hour, minute, and second, respectively.

Although not strictly required, it is recommended that researchers use version control to keep track of changes and ensure the technical consistency of their data files. Git is the most popular version control system. Though it was first designed to manage computer code, Git allows users to record and recall any particular version of a given document, and it works well with pure text files like .csv. This increases transparency in academic research, as others can trace all steps of the data management process and serve as a reliable backup in the case of data loss (Ram 2013). Moreover, contributors can modify the files and merge them to the main Git repository, which facilitates collaboration and review at every stage of the research project. Since users can quickly revert to previous versions of a file, the process is risk free. For more information about Git, please visit https://git-scm.com.

Step 2: Logical Consistency

The second step involves the elaboration of logical validation rules to assess the consistency of recorded values. In contrast with technical integrity, logical consistency requires a priori knowledge from a particular scientific field, so these validation methods are domain specific by design. However, these specific rules can be evaluated using a general framework based on Boolean statements. In other words, scholars may use validation functions that produce a set of TRUE or FALSE responses to check whether the variables are in line with their theoretical expectations (van der Loo and de Jonge 2019, 3).

For instance, if a researcher measures the average age and GDP per capita of a population, the data should have no negative values or zeros. An *if* statement can verify whether the observations conform to that rule:

IF $age \leq 0$ OR $gdp\_per\_capita \leq 0$, $==$ FALSE

ELSE $==$ TRUE

Scholars can use similar validation functions to check the logical validity of any variable for which prior information is available. Users can also combine conditional statements and check the quality of related variables with a single function. As an example, if a dataset contains information about the location of the subjects, such as city and street name, one can assume that the postal code is the same if the two values are identical (van der Loo and de Jonge 2019, 12). Translating the rules into conditional expressions, the code would be as follows:

IF $city_{[i]} == city_{[j]}$ AND $street_{[i]} == street_{[j]}$

THEN $postal_{[i]} \equiv postal_{[j]}$

There are cases for which no exact information on the true values exist, but social scientists can still verify the consistency of their data using conditional rules. Instead of focusing on a particular Boolean outcome, researchers should estimate upper and lower bounds for the variable they want to check. One way to do so is via Monte Carlo simulations. The ESSnet ValiDat Foundation (2018, 72) proposes an easy yet effective method to estimate logical consistency bounds for a given variable. First, create a set of values, S, that seems plausible according to the related literature. Second, add disturbances to the dataset S by simulating cases with measurement error, missing values, mistakes in the data entry process, or other statistical issues that are common to that type of data. This yields a variable, S'. Then, apply the consistency test to S'. In the final step, repeat this process several times and change the number of wrong observations to create a distribution of rule statistics from the simulated S'_n variables. This method produces the lower and upper bounds required for the data, assuming that S correctly approximates the true values of the chosen variable.

Although in the section I discuss how rules can be applied to new datasets, they may also provide interesting insights when used to evaluate existing data. One possible avenue for research is to compare whether certain conditional statements produce different outcomes in datasets that measure the same phenomena, such as the level of democracy or changes in political regimes. Observations that do not conform to the rules should be contrasted and explained, and if that is the case, imputed to credible values. Indeed, de Waal (2017) suggests that imputation methods that both preserve statistical properties of the variables and conform to rule restrictions are the best way to fill missing data, even if they have been difficult to estimate so far.

Testing data with conditional expressions provides another benefit for researchers. Because logical rules can be defined in advance, they may be included in a preregistration plan. Registering the study design before data collection or analysis reduces the chances of "data dredging," where researchers release only the analyses that support their hypotheses (Klein et al. 2018). Preregistration increases the credibility of the findings and may even be submitted directly for publication in the form of registered reports (Chambers 2013). Journals can evaluate the research design before scholars know the results, and the manuscript's acceptance is independent of the final data. As of late 2020, 275 journals use registered reports as a regular or one-off submission option (Center for Open Science 2020).

Step 3: Content Validity

Content validity refers to whether the variables correspond to the theoretical concepts they intend to measure. This is a difficult task because there are no hard-and-fast rules on how to map concepts into values. Here I follow Gerring (1999, 2001, 2011) and suggest that researchers should check if their variables meet six criteria: resonance, domain, differentiation, fecundity, consistency, and causal utility. I explain each of them in further detail below.

Resonance means that the variable name brings to mind the core idea of a concept. A good name is one that includes a simple word that is used in common language and quickly conveys the point the authors are trying to make. Terms that resonate are akin to mnemonic devices, something that helps readers to remember what the variable means long after they see it (Gerring 1999, 370). Terms such as "social capital" or "civic culture" might be imprecise, but they do invoke an intuitive understanding of the concepts they refer to (Bjørnskov and Sønderskov 2013).

Domain considers whether relevant parties agree with the concept being measured. The idea of domain is similar to that of resonance, but it describes how particular audiences, mainly area specialists, interpret the concept one intends to describe (McMann et al. 2016, 9). For example, the domain of "democracy," as measured by political indices, may differ substantially from what the lay community generally understands as the “government of the people” (Munck and Verkuilen 2002). In that respect, the concept should embrace as many domains as possible, although it should strive first for internal validity and consistency (Tortola 2017, 241). Thus, it is essential that researchers have a firm idea of what their target public expects from a concept.

Differentiation means that the variable should include the unique aspects of a given concept. In other words, it entails that one should find what makes a concept distinct from related terms; the sharper those boundaries are, the stronger the validity of the concept. As Gerring (1999, 375) notes, a useful definition of “state” has to single out what characteristics are particular to states and do not appear in other forms of social organization, such as tribes or empires. Many concepts in the social sciences cause confusion precisely because they include traits that are not exclusive to these categories. For instance, if one defines “armed conflict” as “the use of force by the state or civilians against other groups,” it is not possible to differentiate such cases from episodes of lynching or genocide.

Similarly, *fecundity* means that the variable is parsimonious and excludes all information that is not related to the concept. It refines the differentiation attribute and highlights that it is required not only for scholars to affirm what the concept is, but also to show what *it is not* (Gerring 2001, 92). This exercise involves counterfactual thinking, and it is not always clear which unrealized outcome scholars should focus on. One suggestion may be to start with competing definitions and remove characteristics that conflate the original concept with similar behaviors. For instance, a corruption variable should exclude private benefits that do not originate from someone’s governmental position (McMann et al. 2016, 10).

Consistency is an attribute that signals whether a concept retains its validity across different settings (McMann et al. 2016, 11). An indicator of liberal democracy should be able not only to explain Western regimes, but also to identify liberal traits in countries that do not share a similar background and to retain its consistency over time. In this sense, the variable should preferably measure sufficient attributes of the concept, which can be easily identified in other cases.

Lastly, *causal utility* means that researchers are able to test hypotheses in which the concept described is either the main cause or the expected effect (Gerring 1999, 367). In most cases, concepts designed to be employed as independent variables require fewer attributes to be causally useful than those created to be dependent variables (Gerring 2011, 130). Although it is hard to ensure that a concept is completely exogenous from other theoretical constructs, authors should avoid a definition where known confounders connect the concept to background factors (Gerring 2011, 130).

Step 4: Data Generation Validity

This step corresponds to the relationship between the concepts researchers want to translate into numeric values and the data generating process. In particular, scholars need to be aware that their data generating process may be biased or unreliable (McMann et al. 2016, 12). Problems may arise either when gathering new data or when using secondary sources. Here I focus on two common threats to data generation validity: low intercoder reliability and data aggregation challenges.

With regard to novel datasets, scholars should address whether coders have inadvertently introduced their own views during the data collection stage. Although intercoder reliability does not guarantee that the data are correct, disagreements between coders raise a red flag about the validity of the recorded values (Kolbe and Burnett 1991, 248). A first step is to calculate intercoder agreement using two popular statistics, Cohen’s Kappa and Krippendorff’s Alpha (Lombard, Snyder-Duch, and Bracken 2002). Although these tests have their limitations, they are easy to estimate and are available in many statistical languages. R users have the irr package (Gamer et al. 2019), which provides functions to estimate those statistics for any dataset. Scholars are also expected to report in their materials the results of these tests as well as a justification for the minimum acceptable level of intercoder reliability that they have adopted (Lombard, Snyder-Duch, and Bracken 2002, 600).

There are a few recommendations on how to proceed when intercoder agreement is low. First, one should provide training and clear guidelines to the raters. Allowing them to discuss and explain to each other where they disagree can also bring substantial increases to intercoder reliability (O'Connor and Joffe 2020). Second, if time allows, one should implement multiple coding rounds and modify the coding frame accordingly until intercoder agreement reaches a desired level (MacPhail et al. 2016). A final suggestion would be to apply item-response theory models to convert ordinal data to latent variables, as the latter allow for intercoder variation in skill and in perceived scale differences (Marquardt and Pemstein 2018).

When data come from secondary sources and scholars want to combine them into an index, it is good practice to describe how aggregation rules may change the results. Loss of information is inevitable when aggregating attributes into an index, so scholars need to first clarify which of the attributes will be combined and for what reason. This is where theory comes in, as aggregation rules should always be theoretically motivated. For instance, it is unclear whether additivity, the default method for merging low-level attributes, is the best aggregation rule for most indices and why scholars do not assign unequal weights to their attributes more often (Munck and Verkuilen 2002, 24). As long as the steps are theoretically consistent, researchers can use very different methods for index construction.

There are cases, nevertheless, where one has no a priori knowledge about what features to include in an aggregate measure. This can be either because the current literature offers little guidance on the topic or because features are multicollinear. One suggestion is to use Bayesian factor analysis (BFA) as a technique to model latent traits (Conti et al. 2014). Though researchers have long used principal component analysis (PCA) to reduce the dimensionality of a dataset, BFA has several advantages over PCA. BFA propagates uncertainty in the estimates, allows for correlated factors and control variables, and can either decide automatically or let users include as many factors as they see fit into the index. The R package BayesFM (Piatek 2020) performs the analysis presented here.

Step 5: Convergent Validity

The last step concerns how well the variables compare with well-documented cases. McMann et al. (2016) suggest that researchers should evaluate their dataset against other sources that cover the same topics, including qualitative studies. One can borrow the idea of validation functions discussed earlier and apply the same logic here. But as I note earlier, as the statements come from domain knowledge, authors and reviewers should be familiar with the specialized literature in order to check the consistency of the convergence rules.

Consider a dataset with two variables, gender and salary. Suppose that a previous case study in the same location states with high confidence that 20 percent of males have earnings lower than \$2,000 per month. One can formulate a validation rule set as follows:

IF $gender == \text{“}male\text{”}$

THEN $\frac{count(salary > 2{,}000)}{count(salary)} = 0.2$

The example shows how researchers can integrate previous information to their estimates. Authors can select the case studies they want to analyze using different criteria. Seawright and Gerring (2008) offer an interesting comparison between seven case selection methods and the corresponding large-N statistical reasoning behind the choices. The first method is the selection of typical cases, ones that best represent the intended relationship. This is the equivalent of analyzing low-residual observations, those that are very close to the fitted statistical curve. Second, authors can choose diverse cases if they are interested in obtaining a range for their variables. Doing so is similar to checking the spread of a statistical distribution. Third, extreme cases describe observations that lie at the tails of a distribution. Deviant cases are akin to outliers in statistical modeling. Influential cases are those that are often not representative but have a particular effect caused by independent variables. The last two kinds of cases are most similar cases, which parallel matching techniques in large-N studies, and most different cases, their opposite. Please refer to Seawright and Gerring (2008) for more information on case selection strategies.

Finally, authors should explain the reasons behind eventual divergences between their measurements and current theoretical expectations. That said, it is recommended that authors investigate why outliers occur and test whether coder characteristics or measurement construction explain these differences. McMann et al. (2016, 27–36) provide an empirical case study on how to assess convergent validity.

Conclusion

Data validation is a crucial yet undertheorized topic in the social sciences. Although estimation methods have made significant progress over the past few decades, data validation procedures remain largely absent from university courses and academic textbooks. This absence is at odds with the famous saying that "80% of data analysis is spent on the process of cleaning and preparing the data" (Wickham 2014, 1) and with the growing number of datasets social scientists have amassed recently. My goal in this short paper has been not to provide an abstract view of the data validation process, but to offer a few pieces of practical advice that social scientists may find useful when creating a new dataset or assessing the properties of existing ones. I divide the data validation process in five steps, from technical consistency to convergent validity, and suggest additional reading for authors who are interested in the topics discussed here.

I encourage authors to pay special attention to issues of logical consistency and content validity, which are particularly difficult parts of the data validation process. Translating concepts into numeric values is more art than science, especially in areas where many foundational ideas remain contested. Thus, careful theoretical considerations are key to better measurement. Another important part of the data validation process is reproducibility. All steps of the data collection process should be documented and shared along with the final results. Reproducible research leads to timely feedback, quality reviews, and stronger academic collaborations, so scholars have an incentive to adopt reproducible methods in their work. Computer science and its many successful open-source projects provide good evidence in favor of greater research transparency. As stated by Eric Raymond (2001, 30), a software developer, "given enough eyeballs, all bugs are shallow." Maybe the same is true in social sciences as well.

References

Al-Ubaydli, Omar, and Patrick A. McLaughlin. 2017. "RegData: A Numerical Database on Industry-Specific Regulations for All United States Industries and Federal Regulations, 1997–2012." *Regulation & Governance* 11, no. 1 (March): 109–23.

Bjørnskov, Christian, and Kim Mannemar Sønderskov. 2013. "Is Social Capital a Good Concept?" *Social Indicators Research* 114 (3): 1225–42.

Center for Open Science. 2020. “Registered Reports: Peer Review before Results Are Known to Align Scientific Values and Practices.” Accessed January 27, 2021. https://www.cos.io/initiatives/registered-reports.

Chambers, Christopher D. 2013. “Registered Reports: A New Publishing Initiative at Cortex.” *Cortex* 49, no. 3 (March): 609–10.

Conti, Gabriella, Sylvia Frühwirth-Schnatter, James J. Heckman, and Rémi Piatek. 2014. “Bayesian Exploratory Factor Analysis.” *Journal of Econometrics* 183, no. 1 (November): 31–57.

de Waal, Ton. 2017. “Imputation Methods Satisfying Constraints.” Paper presented at the Work Session on Statistical Data Editing, The Hague, Netherlands, April.

ESSnet ValiDat Foundation. 2018. *Methodology for Data Validation 2.0*, revised. ed..

Gamer, Matthias, Jim Lemon, Ian Fellows, and Puspendra Singh. 2019. “irr: Various Coefficients of Interrater Reliability and Agreement.” January 26. https://CRAN.R-project.org/package=irr.

Gerring, John. 1999. “What Makes a Concept Good? A Criterial Framework for Understanding Concept Formation in the Social Sciences.” *Polity* 31, no. 3 (spring): 357–93.

———. 2001. *Social Science Methodology: A Criterial Framework*. Cambridge, UK: Cambridge University Press.

———. 2011. *Social Science Methodology: A Unified Framework*. Cambridge, UK: Cambridge University Press.

Höffler, Jan H. 2017. “Replication and Economics Journal Policies.” *American Economic Review* 107, no. 5 (May): 52–55.

IBM. 2016. *10 Key Marketing Trends for 2017*. ftp://ftp.www.ibm.com/software/in/pdf/10_Key_Marketing_Trends_for_2017.pdf.

Key, Ellen M. 2016. “How Are We Doing? Data Access and Replication in Political Science.” *PS: Political Science & Politics* 49, no. 2 (April): 268–72.

King, Gary. 2007. “An Introduction to the Dataverse Network as an Infrastructure for Data Sharing.” *Sociological Methods & Research* 36, no. 2 (November): 173–99.

Klein, Olivier, Tom E. Hardwicke, Frederik Aust, Johannes Breuer, Henrik Danielsson, Alicia Hofelich Mohr, Hans IJzerman, Gustav Nilsonne, Wolf Vanpaemel, and Michael C. Frank.

2018. “A Practical Guide for Transparency in Psychological Science.” *Collabra: Psychology* 4 (1): 1–15.

Kolbe, Richard H., and Melissa S. Burnett. 1991. “Content-Analysis Research: An Examination of Applications with Directives for Improving Research Reliability and Objectivity.” *Journal of Consumer Research* 18, no. 2 (September): 243–50.

Lombard, Matthew, Jennifer Snyder-Duch, and Cheryl Campanella Bracken. 2002. “Content Analysis in Mass Communication: Assessment and Reporting of Intercoder Reliability.” *Human Communication Research* 28, no. 4 (October): 587–604.

MacPhail, Catherine, Nomhle Khoza, Laurie Abler, and Meghna Ranganathan. 2016. “Process Guidelines for Establishing Intercoder Reliability in Qualitative Studies.” *Qualitative Research* 16 (2): 198–212.

Magoulas, Roger, and Steve Swoyer. 2020. “5 Key Areas for Tech Leaders to Watch in 2020.” *Radar* (blog), O’Reilly. February 8. https://www.oreilly.com/radar/oreilly-2020-platform-analysis/.

Marquardt, Kyle L., and Daniel Pemstein. 2018. “IRT Models for Expert-Coded Panel Data.” *Political Analysis* 26 (4): 431–56.

McDonnell, Robert Myles, Guilherme Jardim Duarte, and Danilo Freire. 2019. “Congressbr: An R Package for Analyzing Data from Brazil’s Chamber of Deputies and Federal Senate.” *Latin American Research Review* 54 (4): 958–69.

McMann, Kelly M., Daniel Pemstein, Brigitte Seim, Jan Teorell, and Staffan I. Lindberg. 2016. “Strategies of Validation: Assessing the Varieties of Democracy Corruption Data.” Working Paper Series 2016:23, Varieties of Democracy, Gothenburg, Sweden, February.

Munck, Gerardo L., and Jay Verkuilen. 2002. “Conceptualizing and Measuring Democracy: Evaluating Alternative Indices.” *Comparative Political Studies* 35, no. 1 (February): 5–34.

O’Connor, Cliodhna, and Helene Joffe. 2020. “Intercoder Reliability in Qualitative Research: Debates and Practical Guidelines.” *International Journal of Qualitative Methods* 19: 1–13.

Perkel, Jeffrey M. 2018. “Why Jupyter is Data Scientists’ Computational Notebook of Choice.” *Nature* 563, no. 7732 (November): 145–47.

Piatek, Rémi. 2020. “BayesFM: Bayesian Inference for Factor Modeling.” October 19. https://CRAN.R-project.org/package=BayesFM.

Ram, Karthik. 2013. “Git Can Facilitate Greater Reproducibility and Increased Transparency in Science.” *Source Code for Biology and Medicine* 8 (1): 1–8.

Raymond, Eric S. 2001. *The Cathedral & the Bazaar: Musings on Linux and Open Source by an Accidental Revolutionary*, revised ed. Sebastopol, CA: O’Reilly.

Schedler, Andreas. 2012. “Judgment and Measurement in Political Science.” *Perspectives on Politics* 10, no. 1 (March): 21–36.

Seawright, Jason, and John Gerring. 2008. “Case Selection Techniques in Case Study Research: A Menu of Qualitative and Quantitative Options.” *Political Research Quarterly* 61, no. 2 (June): 294–308.

Tortola, Pier Domenico. 2017. “Clarifying Multilevel Governance.” *European Journal of Political Research* 56 (2): 234–50.

van der Loo, Mark, and Edwin de Jonge. 2018. *Statistical Data Cleaning with Applications in R*. Hoboken, NJ: Wiley.

van der Loo, Mark P. J., and Edwin de Jonge. 2019. “Data Validation Infrastructure for R.” Preprint, submitted December 20. https://arxiv.org/pdf/1912.09759.pdf.

W3 Techs. 2020. “Usage of Character Encodings Broken Down by Ranking.” Accessed January 27, 2021. https://w3techs.com/technologies/cross/character_encoding/ranking.

Wickham, Hadley. 2014. “Tidy Data.” *Journal of Statistical Software* 59, no. 10 (August): 1–23.

Wickham, Hadley, and Garrett Grolemund. 2016. R for Data Science (website). Accessed January 27, 2021. https://r4ds.had.co.nz.