CHAPTER 2

Validity and Self-Reported Data

In this chapter we provide a discussion of the concept of validity in the social sciences. We first highlight the history of validity and how it has been conceptualized and measured over time. Next, we discuss a type of social science data that is often overlooked in the validity measurement and assessment literature – data that are based on self-reporting. Despite the widespread use of self-reported data in various social science disciplines such as economics, political science, and sociology, there are still few reported attempts to check data accuracy. As examples, we overview self-reported data in four areas, namely US prison population data, COVID-19 case data, toxic releases, and fish landings. We then discuss the need for a tool and established workflow for assessing the accuracy and validity of quantitative self-reported data in the social sciences. We suggest that applying Benford's law to these types of social science data can provide a measure of validity for data that would otherwise not be assessed for accuracy; then we briefly introduce the concept of "Benford validity." We conclude the chapter with a short review of existing studies that have applied Benford's law to social science data in some manner and a brief anticipatory view of Chapter 3. This is all part of showing, in the remainder of the book, how we intend to improve and systematize the use of Benford's law in the social sciences to assist in examining the validity of data used in empirical research.

Validity in Social Science

Validity of the concepts and data employed in a study is a necessary condition for research to successfully accomplish what it sets out to do. Most discussions of validity in the social sciences begin with the suggestion that concepts or data have validity if they measure what the researcher is suggesting they measure. While this is a broad and perhaps dated definition of validity (see Bandalos, 2018), we believe that it has utility in the

https://doi.org/10.1017/9781009127950.003 Published online by Cambridge University Press

case of self-reported data. Even if the analysis and assessment of measurement and data validity have become quite technical, ultimately a study has validity if the measures and data employed in it measure what the researcher indicates that they measure.

Early work on validity often focused on defining the different types of validity that existed and on how to assess whether the measurement of the concept and that of the resulting data had the same type of validity. Indeed, this fascination with uncovering and labeling the different types of validity was widespread. In an article that is over 20 years old, political scientists Adcock and Collier (2001) found 37 different adjectives attached to the word "validity," in a search on the conceptualization and measurement literature. We suspect that a similar search done today would add even more "types" of validity to the canon. Despite the many different types of validity, they are most frequently reduced to the following five types: content or face validity, predictive or criterion validity, concurrent or convergent validity, divergent or discriminant validity, and construct validity. We will now provide a brief review of these types of validity (for extended discussions of these types of validity and related areas, e.g. Adcock & Collier, 2001; Bandalos, 2018; Carmines & Zeller, 1979; Maxim, 1999).

Content or face validity usually refers to an assessment, made by researchers and experts, that a measurement construct and the resultant data successfully measure what they intended to measure. There is no actual "test" of this form of validity; rather it is a qualitative assessment that is made by researchers familiar with the literature in the substantive area being investigated. Face validity is focused on how well the construct and the measure used in the research actually represent the concept intended to be studied. In general, face validity should always be established. For instance, a researcher studying the predictors of quantitative literacy among social science students would likely measure the relevant knowledge using a set of questions that test a student's ability to correctly carry out calculations as well as communicate the results of calculations to readers. If, for instance, the researcher used instead scores from a spelling test as a measure of quantitative literacy, that measure would likely be judged as lacking face validity. That is, while there may be underlying reasons why the two measures could be correlated, it is obvious that spelling itself does not adequately measure the ability to perform calculations or interpret statistics correctly.

A measure is said to have predictive or criterion validity when it sufficiently predicts an important criterion. For example, universities often

use the SAT (Scholastic Assessment Test) in the United States or the GCSE (General Certificate of Secondary Education) in the United Kingdom as a way to predict a student's performance. If those tests have criterion validity, the scores on them should be substantively correlated with the students' grade point average during their university studies.

Concurrent or convergent validity implies that a measure is correlated with other measures that the literature suggests it should be correlated with, while, conversely, divergent or discriminant validity is present when the measure does not correlate with measures it should not be correlated with. For instance, if a researcher is interested in understanding levels of environmental concern among UK citizens and therefore develops a set of survey questions that can be used to create a scale of concern, that researcher would be wise to employ concurrent validity to help validate their scale. In particular, the researcher may compare their scale to the Revised New Ecological Paradigm scale (known as the NEP-R) created by Dunlap et al. (2000). For concurrent validity to exist, the researcher's scale would need to be sufficiently correlated with the Dunlap scale.

Construct validity, a concept first introduced by Cronbach and Meehl (1955) as sort of a last resort when the previous forms of validity could not be adequately assessed, has been expanded by others (see Loevinger, 1957; Messick, 1989 and, for a detailed discussion, Bandalos, 2018) and made to subsume alternative forms of validity previously discussed. In some ways, what researchers refer to when they speak of "construct validity" is similar to the original conceptualization of validity: does the value arrived at adequately and accurately measure what the researcher suggests that it measures? One way to think about construct validity is to assess your concepts through theoretical statements that can be tested. For instance, suppose you are interested in understanding the causes and consequences of children's food security. In order to achieve content validity, you may adopt a measure of food insecurity that focuses on the various aspects that make up food security. You may decide that access to and availability and use of food are, all, important components of food security. You may also decide that access to food depends on different variables that can be directly observed and measured, including the financial resources of the households in which the children reside.

Thus, one of your hypotheses about food security may be that food-secure children are more likely to live in households with relatively high levels of financial resources than children who are not food-secure. It would contradict your theoretical hypothesis about food security if other variables that measure components of access to food, for example parental

behavior (you could ask parents, perhaps, if they ever skip meals so that children in the household have enough to eat), were unrelated to the variables that measure financial resources. If the data you observed on financial resources were unrelated to the fact that parents skip meals so that children can have something to eat, you might then question whether your measure of food security lacks construct validity or whether your theory (and perhaps your definition of food security) needs to be modified based on the context in which your measure is employed – or both.

Contemporary conceptualizations of validity in social sciences tend to focus less on the type of validity that a measure or a dataset has or does not have and more on the overall validity of tests, latent variables, and the like. Composite measures of concepts need to be assessed for validity. Researchers resort to tests and latent variables to accurately measure multifaceted concepts such as depression in psychology, alienation in sociology, or government type in political science – to take a few examples. Single questions on a survey or psychological assessment are often inadequate to capture the full breadth of many important social science concepts. Therefore a vast psychometric literature has developed on the validity assessment of composite social science indicators (e.g., Chan & Idris, 2017). A common approach is to use confirmatory factor analysis (CFA) to assess how well numerous indicators that are based on the theoretical understanding of a concept combine into a latent variable that has construct validity – in other words, accurately measures the entire breadth of the construct (Goodwin, 1999; Kline, 2014). This is an important literature, as increasingly many social science measures and data are constructed from multiple indicators.

The literature on the validity of tests and latent variables constructed on data frequently focuses on how well these tests measure or represent what they are meant to measure when administered properly and no errors are made. Yet there are often reasons to question the validity of responses to the questions that make up the tests and, more frequently, the validity of what we will refer to as "self-reported data." As a reminder, we define self-reported data as data that are not based directly on the social scientist's observations of characteristics or behaviors. Instead, self-reported data rely on an individual's descriptions of the sorts of things that social scientists are interested in studying: characteristics, beliefs, attitudes, and behaviors. Individuals may self-report their own characteristics, beliefs, attitudes, and behaviors; or they may report such things on behalf of groups or organizations, in an official or unofficial capacity.

Many types of secondary data used in social science research can also be classified as self-reported data. Current psychometric approaches to validity assessment that are focused on the validity of the measurement of test scores and latent variables are not equally applicable to quantitative self-reported data. In the case of self-reported data, the accuracy of the data themselves is the primary concern. Rather than worrying about capturing the entire breadth of a complicated concept, researchers who plan on analyzing self-reported data are often concerned with what we will refer to as "misreporting."

We use this umbrella term to refer to any case of inaccurate self-reported data. Misreporting, then, can result from either intentional inaccurate reporting, as in cases of fraud, or unintentional inaccurate reporting, which may be due to data entry error or poor data collection instruments and facilities, which hinder accuracy. We now turn to a more detailed discussion of self-reported data in the social sciences.

Validity of Self-Reported Data

It is no easy task to summarize the vast literature on self-reported data. This literature spans numerous disciplines and various topics, and the answer to the question whether self-reported data are valid can depend on the subject matter that is examined in a study. There are numerous kinds of studies that employ self-reported data – our short list contains more than 50 easily identifiable categories of research in the academic literature that rely on this kind of material: absenteeism from work or school; age; alcohol use, types and amounts; attitudes toward individuals from different racial, ethnic, religious, or gender groups; automobile accident record; automobile repairs; birth control use, generally and by specific populations; chronic illnesses; cell phone usage; chronic pain frequency; church membership and attendance; crime and delinquency, types, amount, and prior record; depression, anxiety, or other psychological conditions; drinking while pregnant; driving while impaired; drug use, types and amount; exercise types and frequency; eating out or eating fast food behaviors; gun ownership; grade point average; hearing aid use; height and weight; injuries and illnesses; internet use and access; IQ and other standardized test scores; learning disabilities; mental health status; physical disabilities; preventative medical or dental care routines; seatbelt use; sexual activities, types and amounts; sexually risky behaviors and diseases; smoking behavior; speeding and other traffic violations; suicide attempts; sunscreen use; use of prescribed medications; use of 'swear

words' or obscene language; visits to dental or doctors' offices for annual checkups; vitamin use; work habits and work hours.

Generally, when we use the term "self-reported data," most people – even most researchers – think of information reported by individuals on a survey. We will refer to that as "traditional self-reported data." But not all self-reported data come in that traditional form; there are other forms in addition to it. For example, many kinds of environmental performance data are self-reported. Consider the information on toxic releases that is contained in the US Environmental Protection Agency (US EPA) database and derives from the Toxics Release Inventory (TRI). The TRI shows the number of pounds of the different chemicals that individual facilities emit into the environment and are required to report to US EPA. Those reports also show how the emission occurred (e.g., through release into the air, the water, or a land site, or through placement in a storage). Those data are measured and self-reported by company employees who are charged with auditing and reporting these chemical releases at each individual facility.

Self-reported data are widely employed in many social science disciplines, and there is a related literature that examines their validity in disparate subject areas that use different methodologies. Thus studies on self-reported validity in different subjects include antibiotic use (Zanichelli et al., 2019), chemotherapy side effects (Pearce et al., 2017), grade point average (Kuncel, Credé, & Thomas, 2005), height and weight (Wen & Kowaleski-Jones, 2012), illegal drug use (Garg et al., 2016), illegal income (Nguygen & Loughran, 2017), periodontal disease (Blicher, Joshipura, & Eke, 2005), smoking behavior (Gorber et al., 2009), suicide attempts (Kokkevi, Arapaki, & Richardson, 2012), weight (Sherry, Jefferds, & Grummer-Strawn, 2007). While these studies attempt to assess self-reported data validity in some manner, there is no general approach, methodology, and workflow for these types of analyses.

As researchers, we must take care to ensure that we use valid data, and use them in appropriate ways. Part of doing so is asking how valid the data are. Thus, whether the self-reported data are traditional (i.e., generated by survey research) or the kind generated by reporting practices that may be required of an organization, it is still necessary for researchers to ask questions about, and to investigate, the validity of those data. This step should be taken before performing any statistical analysis, attempting to make any knowledge claims, or recommending policy changes generated from those data. We now give a few examples of the different types of self-reported data that social scientists regularly analyze and highlight how they can be affected by misreporting.

United States Prison Data

It may be strange to think of prison population data as self-reported. Normally, when we talk about self-reported prison data, we have in mind survey data gathered from inmates, say, through the Survey of Prison Inmates (SPI), which is handed to a sample of state and federal prison inmates and conducted periodically (it has been conducted in 1974, 1979, 1986, 1991, 1997, 2004, and 2016). As described by the Bureau of Justice Statistics of 2016, the SPI's

> primary objective is to produce national estimates for the state and sentenced federal prison populations across a variety of domains, including but not limited to demographic characteristics, current offense and sentence, incident characteristics, firearm possession and sources, criminal history, socioeconomic characteristics, family background, drug and alcohol use and treatment, mental and physical health and treatment, and facility programs and rule violations. From January through October 2016, data were collected through face-to-face interviews with prisoners using computer-assisted personal interviewing (CAPI).[1]

Many kinds of prison data can be considered self-reported. While the procedure may not be the same in every US state, every day the states produce inmate counts designed to track the prison inmates and to ensure that they are in the proper locations. For security reasons, each state counts its inmates multiple times during one day and reports those counts to a centralized state agency. That agency then checks the counts against the number of inmates assigned to each institution, including those who were released, received, and transferred during the day. These are self-reported data about the prison population insofar as they are not generated by a third party but are counted by representatives who are assigned these duties in each state.

Virtually nothing has been written about the validity of prisoner population counts. Indeed, if one searches the phrases "validity prison population data" and "reliability prison population data" in Google Scholar, one will find that there are zero valid returns. There is a wide range of studies on the validity of self-reported studies and much information that can be gathered from inmates (e.g., inmate social identity), or even about inmates

[1] Visit Survey of Prison Inmates at https://bjs.ojp.gov/data-collection/survey-prison-inmates-spi#:~:text=Its%20primary%20objective%20is%20to,and%20sources%2C%20criminal%20history%2C%20socioeconomic.

https://doi.org/10.1017/9781009127950.003 Published online by Cambridge University Press

(e.g., reliability of ADHS diagnosis among prison inmates), but none about the counts of inmate populations.

Counting COVID-19 Cases

Chapter 6 examines COVID-19 data in detail. Here, we discuss a few general validity issues that might be of concern when employing COVID-19 data. Since the beginning of the global COVID-19 pandemic in March of 2020, health providers and government officials moved swiftly to comprehend the nature of the health threat posed by the virus, the general and specific kinds of policies and procedures that would need to be implemented to contain its spread, and especially the volume of deaths caused by the disease. Little was known about COVID-19 specifically before the pandemic, although information was available about the virus class to which COVID-19 is related – SARS or severe acute respiratory syndrome. Understanding the spread of COVID-19 within and across countries, and also the severity of outcomes associated with contracting it, would require access to valid counts of the occurrence and distribution of COVID-19 cases.

There are a number of areas of COVID-19-related research in which validity is a concern. For example, if we wanted to study the trend in, or the distribution of, COVID-19-related deaths, we would require an accurate count of those kinds of deaths. Obtaining an accurate count of COVID-19 deaths could be linked to numerous factors; here our intention is not to review all those factors, but rather to give some relevant examples. We begin with the issue of identifying whether the cause of a person's death was COVID-19. To do so, both the doctors who treat the patients and those who perform autopsies must know what tests are appropriate to determine that a person was infected with COVID-19. Those doctors must also make some assumptions to certify that a particular individual died from COVID-19 and not from some other, underlying cause. Among countries, or even within the same country, accurate counts of COVID-19 cases and deaths may be adversely affected in places where there is less medical expertise or fewer resources devoted to the detection of this virus. The counting process could also be affected by other factors – such as the pressure to count – or not to count – certain deaths as the result of COVID-19. That pressure might have stemmed from government officials, hospitals, or even the insurance industry, which is reported to have battled with businesses and healthcare providers over the designation of cases as COVID-19-related, and even over whether pandemic diseases

would be covered by business interruption insurance policies or not (French, 2020).

Regardless of the specific factors that might affect COVID-19 case counts, the question remains as to whether those counts could be considered valid and, in the case of data with validity concerns, whether the validity of COVID-19 case counts was a general problem or a specific problem. By "general" and "specific" we mean the following: as a "general" problem, we can think of the validity of COVID-19 case counts as being something that is prevalent everywhere across time and place. This means that, regardless of the specific time or place the data are from, there is a reason to be concerned with the validity of COVID-19 counts; such a reason could be for example the ability to identify COVID-19 (our research shows that this was not an actual or serious problem in COVID-19 reporting). By "specific" we mean that COVID-19 case counts may be generally valid, but for certain locations or for certain time periods this statement may not be true – that is, the data for those locations or periods may not be valid. For example, at the beginning of the pandemic, before the ability to identify COVID-19 became widely available, COVID-19 cases were likely to be undercounted. In some places COVID-19 may have been undercounted on purpose, for political reasons, or as a result of perceived political impacts (Adolph et al., 2021), or through failures of the administrative case-processing systems (Dubrow, 2021). Low case counts could also be unintentional, related to financial conditions in some nations, and due to the lack of COVID-19 testing resources. This means that, in some places or at certain points in time, COVID-19 case count data may not be valid.

Toxics Release Inventory

The Toxics Release Inventory (TRI) is a database maintained by the US Environmental Protection Agency (US EPA). It contains information about the types and amounts of pollutants emitted by a facility required to report that information to the US EPA. At the time of this writing there are 744 individual chemical pollutants listed under TRI reporting requirements. Not all facilities that produce those pollutants, however, must report to the US EPA. Only facilities that produce more than the specified threshold amounts for a given chemical or for the aggregation of their emissions and who employ ten or more people must report their emissions to the US EPA.

The concern with TRI emission data is that they are self-reported. Facilities can evade certain types of inspections, or consequences such as being found in violation of their US EPA emission permits, if they under-report their emissions. Hence there is some motivation for under-reporting emissions. If under-reporting was widespread, and if it was prevalent among large companies that produce a great deal of toxic waste, then the validity of TRI as an indicator of pollution would certainly be called into question.

Several studies have examined factors that affect the reporting of TRI emissions. Some studies have found, for example, that TRI self-reports have been affected by market conditions in general (Konar & Cohen, 1997) and by stock market prices in particular. This means that, as general market conditions for a firm or industry decline, or as the stock price of a company declines, so too does the volume of self-reported TRI emissions.

Another way to examine the validity of TRI emissions is to compare those emissions to pollution levels. De Marchi and Hamilton (2006) used this method by comparing TRI reported air emissions to air pollution monitor measurements taken by the US EPA. In their study, they noted that during the 1990s self-reported TRI emissions dropped by 48% (p. 58). This raised the question whether TRI emissions had actually decreased by that amount, or whether companies were under-reporting their emissions. As de Marchi and Hamilton note, the US EPA brings few cases for misreporting emissions under the TRI and, out of all such cases in the year they examined, fewer than 2% of the environmental compliance cases investigated by US EPA involved TRI self-reports. Often those cases are focused on facilities that fail to file any reports, and investigations do not address the accuracy of the reports filed by facilities.

To examine the validity of TRI facility-level self-reports, de Marchi and Hamilton collected air emissions-monitoring data from more than 5,000 US EPA air-monitoring stations. They then collected self-reported emissions data for 12 air pollutants emitted by facilities within a 50 km (31 mile) radius of the EPA monitors. Their analysis showed several differences between TRI self-reports and EPA air monitors for the 12 chemicals they examined. They concluded, however, that these differences were largest and most serious for two chemicals: lead and nitric acid, the two chemicals on their list considered most dangerous.

Fish Catch Reports

Fishing is an important global activity. According to the United Nations, more than 200 million people worldwide earn their livelihoods in the

marine fishing industry. Marine fisheries produce about 5% of global gross domestic product, which is approximately 3 trillion USD. In addition, more than 3 billion people depend on marine fisheries for some proportion of their dietary needs. Fishing industry and other fish catch data can be traced back for nearly 150 years and shows the importance and growth of fishing over time (Watson & Tidd, 2018).

Given the importance of fisheries to global well-being – both ecologically, in terms of ecosystem stability, and from a food subsistence perspective – the global fish take has been monitored since 1999 by the Sea Around Us research project. The data available from Sea Around Us are country- and year-specific. They measure fish takes and discards, are available for many countries, and go back as far as 1950. These data come from self-reported catch reports (see Chapter 5).

Sea Around Us argues that these data have important ecological policy implications, as well as relevance for managing fisheries and hunger worldwide. These data can show, for example, which fisheries are producing less and perhaps are being overfished or collapsing in response to other ecological conditions such as climate change. While these data have been widely employed in research, they have not been thoroughly examined for data validity concerns. To be sure, data issues that periodically threaten the validity of the Sea Around Us database have been noticed and addressed. For example, in the early 2000s, Sea Around Us became aware that the data coming from China were suspect. China, it seems, had been overreporting its fish catch data, making it appear as if the fisheries from which it was extracting resources were healthier than would be expected, especially if China was reporting declining fish catches from those fisheries (Watson, Pang, & Pauly, 2001). Leaving aside a few questions about specific instances in which fish catch reports from a country are suspect, the general validity of the Sea Around Us data has not been explored.

Benford's Law and the Validity of Self-Reported Data

Data on prison populations, on COVID-19 cases, on releases of toxic waste, and on fish catches are illustrative examples of the myriad types of self-reported data that social scientists employ in their research. Given the potential concerns with self-reported data validity, social scientists need a tool and a methodology for assessing the validity of self-reported data. Providing that tool and that methodology is the primary goal of this book. In the chapters that follow we overview Benford's law and its application to the validity assessment of self-reported data; we introduce a new measure

of conformity to the Benford distribution, namely one based on the statistical concept of agreement; and then we provide a workflow for analyzing the validity of self-reported data using the Benford distribution. We introduce the concept of "Benford validity" and argue that establishing it should be the first step in any empirical analysis of self-reported data.

Of course, we are not the first social scientists to use Benford's law in some way in assessing data accuracy and validity. For example, research on environmental outcomes (Beiglou et al., 2017; Brown, 2005; Cole, Maddison, & Zhang, 2020; Coracioni, 2020; de Marchi & Hamilton, 2006; de Vries & Murk, 2013), religion (Mir, 2012; Mir, 2014), crime (Badal-Valero, Alvarez-Jareño, & Pavía, 2018; Hickman & Rice, 2010), election and campaign financing fraud (Breunig & Goerres, 2011; Deckert, Myagkov, & Ordeshook, 2011; Tam Cho & Gaines, 2007), and survey research (Judge & Schechter, 2009; Kock & Okamura, 2020) has used Benford's law to evaluate the quality of the data in all these areas. While such studies employ Benford's law to assess data accuracy and validity, there is room for improvement. These studies use different measures of conformity to the Benford distribution and different methodological processes to ensure data accuracy. Throughout the remainder of this book, we attempt to improve the use of Benford's law in examining the validity of data (1) by introducing a new measure of conformity to the Benford distribution that is based on statistical agreement; and (2) by providing a workflow for Benford agreement analysis that ensures that agreement with the Benford distribution does not substantially vary between subgroups of the data created on the basis of variables that could potentially impact agreement.

Conclusion

In this chapter we provided a discussion of the concept of validity in the social sciences. We first highlighted the history of validity and how it has been conceptualized and measured over time. Next we discussed data that are based on self-reports. This is a type of social science data that is often overlooked in the validity measurement and assessment literature, despite its widespread use in various social science disciplines such as economics, political science, and sociology. As examples, we overviewed self-reported data in a few areas: US prisons, COVID-19 cases, toxic releases, and fish catches. We suggested that applying Benford's law to these types of social science data provides a measure of validity assessment for data whose accuracy would otherwise remain unassessed. We concluded the chapter

with a short review of existing studies that have employed Benford's law to social science data.

Chapter 3 describes and illustrates the Benford probability distribution. A brief summary of the origin and evolution of the Benford distribution is provided and the development and assessment of various measures of goodness of fit between an empirical distribution and the Benford distribution are described and illustrated. These measures include Pearson's chi-squared test, Wilks' likelihood ratio, Hardy and Ramanujan's partition theory, Fisher's exact test, Kuiper's V_n measure, Tam Cho and Gaines' d measure, Cohen's **w** measure, and Nigrini's MAD measure.

https://doi.org/10.1017/9781009127950.003 Published online by Cambridge University Press