Apples to Oranges: Using Surveys to Measure Concepts Across Cases *(Proposed Title)*

Kyle L. Marquardt*, Daniel Pemstein†, and Brigitte Seim‡

1 Overview

Surveys are a cornerstone tool for the collection of data across cases in the social sciences. Data scientists regularly rely on respondent impressions of political institutions, economic trends, and social institutions to inform academic research; and surveys of citizens, consumers, or service users are a popular tool outside of academia as well. Notably, online ratings systems of consumer goods, services, and media—all essentially surveys—are now crucial to the modern economy. In survey research, questions of comparability arise when respondents understand questions, or response options, differently from one another, either because they operate in different contexts, or because of idiosyncratic individual differences. While not ubiquitous in survey research, comparability issues are extremely common. Yet we often ignore key questions of comparability when using surveys to produce indicators, though it is impossible to produce high-quality data without ensuring and verifying comparability across respondents, cases, and over time. Scholars of comparative politics often warn that context matters, and that survey questions mean different things to different people, in different places, at different times. However, researchers routinely treat survey responses as exchangeable, and ignore this fundamental threat to the veracity of the conclusions that they draw from survey data. Crucially, we use surveys to provide consistent measures across cases and respondents, but then simply assume—because respondents answer the same questions—that responses are comparable across contexts, even though we know that this assumption is often wrong! Lack of attention to this key issue potentially undermines many impressive findings in the academic literature, reduces the value of product and service ratings in online markets, and threatens the quality of our most valuable data sources. Yet standard training largely overlooks strategies for identifying and mitigating threats to comparability across surveys and neither survey producers, nor their users, commonly employ the battery of comparability-enhancing techniques available to them.

In this book, we will discuss how to achieve comparability in survey data across cases, across respondents, and over time, explicating the importance of these issues and delineating solutions. The book will answer questions that almost *all* survey enterprises confront: What kind of data are different kinds of respondents best-suited to produce? How does context affect the comparability of survey responses? Do respondents differ in their validity and reliability, and what are the determinants of this variation? How do

*National Research University Higher School of Economics
†North Dakota State University
‡University of North Carolina at Chapel Hill

respondent selection and motivation affect comparability, validity, and reliability? Under what circumstances are respondents most likely to disagree about latent concepts or how to interpret questions? When respondents disagree, what methods can resolve disagreement and align respondent survey data to maximize comparability across cases? What measurement models are best suited address problems of cross-respondent comparability, and how can these models be adjusted to fit different assumptions about the nature of the data and our knowledge of the respondents that provided it? Both producers and consumers of surveys grapple with these questions. Survey producers must attempt to answer them when designing their surveys, and survey consumers must be able to understand the answers provided by the producers so they can evaluate datasets and use them appropriately. The aim of the book is to arm both producers and consumers with the information and mitigation strategies necessary to overcome these problems and provide comparable survey data across cases, respondents, contexts, and time.

2 Niche of the Book

There are many impressive texts addressing how to design surveys—from question design to sampling approaches. However, these works typically pay scant attention to ensuring comparability across cases and coders, and over time, especially in post-collection analysis. For example, Groves, Fowler Jr, Couper, Lepkowski, Singer & Tourangeau (2011) offer one of the canonical works on survey methodology, initially published in 2004 and updated in 2011. Their book devotes one chapter (out of 12) to "postcollection processing" and four pages (out of 400) to "sampling variance estimation" (the closest they come to discussing cross-respondent or cross-case comparability adjustments). Similarly, Moser & Kalton (2017) devote one chapter (out of 18) to "processing of the data" and another to "analysis, interpretation and presentation." Roy, Acharya & Roy (2016) provide a more recent book devoted to the design of surveys, but post-hoc comparability methods are not discussed. Zeller & Carmines (1980) provide one of the most oft-cited books on measurement, but adjustments for comparability are also not discussed, though the importance of comparability is discussed in abstract. The more recent book by Kellstedt & Whitten (2013) includes a relevant chapter, aptly titled "Getting to Know your Data: Evaluating Measurement and Variations," but this chapter briefly discusses many topics and is 37 pages long, precluding thorough discussion and comprehensive presentation of solutions for achieving comparability. Books focusing less on survey methods and more on post-data-collection analysis are no more comprehensive in addressing how to achieve comparability with survey data, with even the most cited books for time series analysis (e.g., Gelman & Hill (2006), Box-Steffensmeier, Freeman, Hitt & Pevehouse (2014), and McCleary, McDowell & Bartos (2017)) failing to discuss these topics. The book with the most attention to comparability over time, across respondents, and across cases may be Atkeson & Atkeson (forthcoming). This forthcoming Oxford University Press book contains chapters entitled "Expert Surveys as a Measurement Tool" (by Cherie Maestas), "Cross-National Surveys and the Comparative Study of Electoral Systems" (by Jeffrey A. Karp and Jack Vowles), "Aggregating Survey Data to Estimate Subnational Public Opinion" (by Paul Brace), and "Public Opinion at the State and Local Level" (by Chris Warshaw). However, as this book is not yet in print, it is difficult to assess the degree of overlap between these chapters and the book content proposed here. The mere presence of this book on the market, however, indicates the issue of ensuring comparability is a

growing area of attention among producers and consumers of survey data.

The Varieties of Democracy Project (V–Dem), has spent the last ten years refining methods to achieve cross-case, cross-context, cross-respondent, and over time comparability. This project executes a large-scale panel survey engaging expert respondents in over 180 countries, coding over 300 political phenomena for over 200 years, producing more than 10 million data points. The scale of the enterprise has compelled the methodology team within V–Dem to reflect extensively on how to ensure cross-case, cross-context, cross-respondent and over time comparability, and cross-respondent validity and reliability. As a result of this undertaking, they have developed solutions to key methodological obstacles. This book introduces the suite of methods that V-Dem has developed to address fundamental problems in achieving cross-context, cross-respondent, and over time comparability in surveys to both consumers and producers of such data.

While this book was inspired by work at V–Dem and thus provides substantial guidance for the creation and consumption of expert surveys, it has wide applicability beyond expert surveys. Specifically, the insights of this book will apply to all surveys where different respondents reveal their attitudes about the traits of different cases[1] and where the traits being measured are latent.[2] Even outside of academia, the insights in this book could also be applied to consumer "surveys," such as online platforms Yelp and Amazon Reviews, or to any other corporate survey enterprises. All surveys that rely on different respondents situated within different contexts, and/or points in time, and evaluating traits of different cases—be it citizens of different countries, individuals in different areas of one country being surveyed over time, or individuals continuously submitting reviews on an online platform—should engage the issues we discuss here.

The book will therefore use V–Dem as a case study to convey the importance of considering these issues, to demonstrate how to apply and understand novel tools, and to impart recommended solutions. However, in the course of doing so, the book will also engage the broader data science community by explaining how to apply V–Dem's lessons to a variety of survey and consumer rating applications.

3 Target Audiences

This book targets a broad prospective audience spanning three primary groups: scholars, policymakers, and private sector analysts.

Within the scholarly community, there are three groups that we anticipate to be particularly interested in the points we raise in the book. Scholars who are using existing survey data from projects such as V-Dem and ANES (the "consumers" of survey data) are likely to use this book as a reference to understand the strengths and weaknesses of survey data and the methods used to produce the data. In addition, scholars who publish in various methodology literatures—Bayesian IRT models, survey methods, data validation experiments—are likely to cite this book and engage with its discussions. Beyond serving as a useful self-training text and desk reference, this book would fit into a variety of graduate classes in the social sciences, ranging from comparative politics or sociology,

[1] A "case" could be, e.g., a geographic unit (e.g., country), individual (e.g., politician), organization (e.g., restaurant) or consumer item (e.g., book)

[2] In this book, we will consider the set of "latent traits" that are difficult to validly observe and difficult to reliably measure. Our discussion will include traits that are challenging to capture via survey questions, and traits that could be perceived differently by different respondents, or even those that could be perceived differently by the same respondent over time.

to broad courses in research design, and more specialized methods courses on applied Bayesian statistics, survey design and analysis, and latent variable modeling. This book also serves academics involved in other survey projects (other "producers" of survey data), who grapple with the issues we discuss in their own projects. Such projects—for example, Bright Line Watch and the Electoral Integrity Project—are proliferating. People involved in these projects will be interested in the solutions V–Dem has developed, and are likely eager to build on the methods we present.

A second audience are the practitioners in international organizations, development agencies, and non-governmental organizations who increasingly use survey data to guide their work and important decisions. Policy practitioners, in particular, have posed many of the questions we answer in this book, as they attempt to assess the usability and flexibility of survey data in policy-centric applications. This audience, crucially, includes the wide array of practitioners—such as program officers at USAID—who now routinely use V-Dem data to make programming and funding decisions. These groups also increasingly conduct in-house data collection projects that would benefit from an understanding of the tools that this book presents. Indeed, the authors often consult with policymakers on how to effectively collect program-relevant data across disparate cases.

Finally, data scientists in the private sector routinely work with survey and rating data, both as consumers and producers, that suffers from the same issues of comparability that confront academics. The tools that we present in this book are relevant to both traditional consumer surveys and the online ratings and recommendation systems that are now endemic to modern commerce. Beyond rating products and aggregating consumer preferences, multi-national firms also rely on cross-national evaluations of latent traits like corruption and political stability to assess investment risk. Indeed the authors have experience consulting with firms about how to effectively collect data for risk assessment. Given the broad applicability of these methods in the private sector, this book could become part of the standard training and reference collection for private sector data scientists and serve as a textbook in the many new and emerging university programs offering degrees to aspiring data scientists.

4 Outline

The book's proposed content will serve the needs of these different audiences. We begin with an introduction that bounds the set of survey applications considered in the book, defines the terms we will use, presents an accessible summary of the comparability challenges that confront producers and consumers of these sorts of surveys, and previews current approaches for addressing these challenges. Chapter 2 asks how one chooses or recruits respondents to survey, what self-reported data to collect to best model cross-respondent variability, and how one can best motivate respondents to produce high quality data. Chapter 3 examines the fundamental question of how to deal with the fact that respondents disagree. This chapter develops the core components of a survey measurement modeling approach, and elucidates a rigorous, and general, framework for quantifying, and communicating, uncertainty. Chapters 4 and 5 argue that one of the greatest strengths of cross-contextual surveys—that they leverage diverse respondents to provide case and domain knowledge—is also one of their biggest weaknesses. These chapters explore how we can ensure that disparate respondents provide us with information that is comparable across cases, and introduce data collection designs, and statistical

modeling techniques, to achieve this goal. Chapter 6 introduces advanced statistical techniques for survey projects that, like V–Dem, try to measure traits across space and over time. This chapter should prove especially useful to designers of cross-national surveys in the social sciences, and to data scientists working on related problems, such as evaluating product appeal across markets or doing cross-national risk analysis. Chapter 7 asks three vital questions related to the presentation of survey results. First, how do we simultaneously maximize data quality by applying cutting-edge statistical tools to survey data,while producing information that policymakers, journalists, web-commerce consumers, and even schoolchildren can understand? Second, how can we leverage statistical models and data visualization techniques to meet the needs of disparate audiences? Third, how do we effectively communicate measurement uncertainty and comparability adjustments to consumers with little training in probability and statistics? Throughout this discussion, the chapter incorporates a frank portrayal of the strengths and weaknesses of the approaches for achieving comparability we developed within V–Dem and present in the book.

5 Authorship

This book will be co-authored by three members of the V–Dem methods team: Kyle Marquardt (National Research University Higher School of Economics), Daniel Pemstein (North Dakota State University); and Brigitte Seim (University of North Carolina, Chapel Hill). Marquardt joined V–Dem in 2015, first as a Research Fellow and now as a Project Manager for Measurement Methods. Pemstein, who joined V–Dem in 2013, is also a V–Dem Project Manager for Measurement Methods, and serves on the V–Dem Steering Committee. Seim was initially a Research Fellow at the V–Dem Institute in 2014, but in 2015 became the Project Manager for Experiments when she joined the faculty at UNC, Chapel Hill.

6 Table of Contents and Chapter Abstracts

We present the abstracts for seven proposed chapters in the following sections. The chapters presented here constitute six substantive chapters, plus an introduction. There will also be a 5,000 word conclusion summarizing the key points of the book, which is not outlined here. Some of these chapters could possibly be combined, and we have also discussed two potential additional chapters: one attempting to apply some of the techniques described below to other survey datasets (e.g., the Chapel Hill Expert Survey, the Afrobarometer); and one in which we discuss more advanced methods for estimating coder reliability. While we have not fleshed out these additional chapters, if there is interest, we would be happy to do so.

6.1 Surveys: What are they good for?

In the introduction, we describe the scope of surveys for which variation in cases to be rated, context or timing of the survey, or respondents to the survey may undermine response comparability. We provide a set of exemplars—e.g., V–Dem's survey of democratic institutions, the American National Election Studies survey among U.S. voters after each

election, and Yelp restaurant reviews[3]—to illustrate the range of applications for the methods that we present. In so doing, we both focus on situations where comparability is an obvious concern (V-Dem and Yelp) and illustrate that even well-worn methods for analyzing traditional public opinion surveys (ANES) can be improved by directly engaging questions of comparability in measurement. Next we introduce a typology of possible survey respondents—experts, coders, crowd workers, citizen respondents—and consider the survey topics, questions, and cases that make one type of respondent more appropriate than another. We conclude by previewing the challenges with surveys the book addresses and the solutions we recommend, using our three core exemplar applications to briefly illustrate each issue.

5,000 words

6.2 Choosing Respondents

A fundamental task in any survey project is determining the theoretical population of respondents and the questions this population is capable of answering. This task is complicated because respondents possess knowledge of the issue areas affecting their lives, or in which they have received formal training, but their level of knowledge is both relative and dependent on the cases and concepts the enterprise investigates. A citizen may be aware of bribery patterns in his or her community, but be ill-equipped to evaluate corruption in the national legislature, assess bribery patterns in a neighboring district, or to provide a response that one can compare directly to citizens in other contexts. A historian who studies the development of civil society in Azerbaijan is an expert on both civil society and Azerbaijan, but may not be an expert on either of these topics in other contexts: she may know little about civil society in Botswana, or dynamics of the resource curse in Azerbaijan. This issue extends across survey domains. For example, Yelp restaurant reviews rely on self-selected lay raters, but might also plausibly incorporate information from professional restaurant critics, chefs, and other experts. The dining habits of lay raters themselves may also lead them to have substantial "expertise" in certain types of food, which could substantially affect their cross-domain judgments and the comparability of their responses with those of other raters. Survey enterprises must devote substantial attention to appropriately matching respondents and their expertise to concepts and cases, and weigh the advantages and disadvantages of recruiting different types of respondents. Similarly, they must be aware of the trade-offs involved in asking respondents to provide information beyond the areas in which they possess knowledge. This is a general problem in survey research, but is of especial importance to concerns about comparability. Some samples of respondents will plausibly answer questions in comparable ways, while others will not. Issues of comparability raised by one group of participants may be assuaged, or at least mitigated, by changing the sampling frame. Understanding these trade-offs is crucial to identifying sampling strategies that will reduce the need to improve comparability with post-hoc adjustments. It also provides the grounding necessary to effectively deploy the tools for improving comparability that we describe in the remainder of the book.

We begin this chapter with a typology of actors—experts, trained coders, crowd workers, citizen respondents—who provide survey data. Though the chapter will more com-

[3]Yelp provides an open dataset for use in academic work that includes over 8 million reviews of more than 200 thousand businesses, provided by almost two million users.

prehensively delineate the definitions and characteristics of these categories, general definitions are as follows: Experts are individuals who who have devoted a significant portion of their lives to developing specialized knowledge on a topic. Trained coders often do not have an advanced degree in the survey topic, but through the act of completing the task, these respondents may develop a level of knowledge equivalent to expertise. Individuals who constitute crowd workers are not a random selection of individuals across the world. Instead, they are individuals associated with online enterprises such as Amazon Mechanical Turk (MTurk) or Crowdflower. They are recruited to participate in surveys because of their ability to acquire knowledge or technical expertise, but instead because they are relatively cheap and numerous. Finally, citizen respondents are recruited to participate in a survey because the topics addressed in the survey pertain to their experiences living their day-to-day lives; they are experts in what it is like to be individuals with their traits, and they are asked to draw on this expertise in responding to the survey.

After developing this typology, we then theorize about the types of questions and tasks that are best answered or completed by each category of respondents, and the characteristics of the questions/tasks, respondents, and the incentives they face that may condition data quality and cross-case and respondent comparability in responses. Drawing on a series of experiments conducted among Amazon Turk and Crowdflower workers, as well as V–Dem experts, we test our expectations regarding the determinants of data quality, both within and across the respondent groups. We also compare one common survey question type —answering Likert-scale questions—to another type of question that has been employed in other surveys: paired comparisons of country-year cases. We assess whether asking respondents to compare two cases rather than answer a Likert-scale question about a case in isolation could be used to improve survey data validity and reliability, and especially with respect to comparable cross-case concept measurement.

Next, we provide a thorough discussion of how the V–Dem project both defines an expert, and the steps the project takes to ensure that it recruits respondents who fit that definition. In the process, we compare and contrast the V–Dem project's definition and recruitment strategies with that of other prominent survey enterprises, discussing how both the concepts and constraints in these projects affects these strategies, and derive general lessons about respondent identification and selection for survey design. We also provide concrete examples of the implications of these strategies using V–Dem data. First, we analyze variation in expert confidence, the level of certainty that experts self-report regarding their codings at the observation level. This analysis provides evidence regarding the degree to which experts are confident in their expertise; further regression analyses provide insight into what makes some experts more confident than others. Second, we analyze variation how error-prone experts are, and discuss how to identify "bad" respondents and specific coding instances.

Together, the analyses presented in this chapter provide empirical insight into how respondent selection and recruitment affects downstream data quality and measure comparability. In doing so, it will be of use for both projects involved in identifying respondents and users of existing datasets who are attempting to determine how respondent group characteristics and survey question structure may have affected the resulting data.

15,000 words

6.3 Making Respondents Comparable

Survey respondents differ from one another, and these differences can cause them to disagree about how to respond to particular questions. In some contexts, such as a public opinion survey, this disagreement may be exactly what researchers designed the survey to measure, although explicitly modeling this in terms of question comparability can pay dividends even in this situation. In other contexts, such as an expert survey of political parties' ideological positions, or a website that aggregates consumer good or service ratings, analysts must adjudicate across respondent disagreement to produce valid measurements of concepts. In general, cross-respondent disagreement on surveys stems from numerous sources, and whether or not these sources of disagreement affect measurement quality depends on the goals of the survey in question. First, respondents may disagree about the true value of a difficult-to-observe and difficult-to-measure (or "latent") concept, or experience the world differently from one another. Sometimes this disagreement is the quantity that the analyst wants to measure, but often it introduces systematic error. Second, they may perceive question scales differently, and thus provide incomparable responses, regardless of their perception of the thing in question. Third, there is almost certainly variation in knowledge across respondents, especially as the number of recruited respondents increases. These problems are clearly relevant to surveys, such as V-Dem, that leverage respondent expertise to measure hard-to-observe concepts. But these forms of disagreement can undermine the quality of data gathered on any survey. For example, surveys of public attitudes towards public figures generally assume that respondents both conceptualize "feeling thermometers" similarly, and have similar levels of relevant knowledge. When these assumptions are wrong, consumers of attitudinal surveys will draw biased conclusions, even though one form of incomparability—variance in attitudes—is the quantity of interest, and not something the analyst wants to "correct."

Standard approaches to aggregating and summarizing respondent data do not adequately address any of these sources of disagreement. In this chapter, we describe how V–Dem and other survey enterprises have improved on standard approaches by incorporating these sources of disagreement into their measurement strategy. To do so, we discuss the implications of these different types of disagreement in detail, providing illustrative examples of both concerns and solutions. We conclude with a description of the V–Dem measurement model, explain how it adjusts for all three sources of disagreement, and discuss how the tools we developed for V-Dem's expert survey can be applied in other contexts ranging from the measurement of collective identity to restaurant quality.

The first source of disagreement is that surveys within social and data science often focus on concepts that are difficult or impossible to directly observe. As a result, there is generally no unambiguously correct answer to the questions that these sorts of surveys ask. Data on corruption provide an illustrative example. Some aspects of corruption are relatively unambiguous and easy to observe (e.g., the exposure to bribery among citizens in the respondent's network), ensuring accurate and replicable coding. In contrast, other aspects of corruption require more detailed understanding and information to complete the task (e.g., the prevalence of nepotism among local government officials), creating the potential for inconsistent answers across respondents. For that reason, many scholars who work on the measurement of latent concepts argue that it is best to conceptualize them as a distribution of possible values, as opposed to a "true value" that is measured with error. This epistemological belief lends itself to a Bayesian approach to measurement, which we describe in detail in the chapter. Using the V–Dem measurement model to

explicate, we demonstrate best practices for using surveys to measure latent variables, emphasizing the importance of effectively modeling uncertainty around point estimates.

The second source of disagreement involves scale perception, and is referred to as "differential item functioning" (DIF) (Aldrich & McKelvey 1977, King, Murray, Salomon & Tandon 2004, LeBreton & Senter 2007, Hare, Armstrong, Bakker, Carroll & Poole 2015, Lindstaedt, Proksch & Slapin 2016). Surveys often rely on Likert scales for coding: they present respondents with a question, and then ask them to assign a score that corresponds to different ordinal scale levels. Ideally, scales would be unambiguous, and respondents would immediately know the correct response category for each case. However, ambiguity is difficult to avoid when mapping a continuous latent concept to an ordinal scale. Respondents translate perceptions into ordinal answers based on their interpretation of the answer categories and their personally held thresholds about what justifies moving from one category to another, and these interpretations and thresholds often differ across respondents. DIF has great implications for measurement. DIF confounds point estimates and survey consumers who ignore DIF often work with invalid measures. DIF also masquerades as disagreement between respondents in a way that can cause investigators to overestimate measurement uncertainty. Yet survey users routinely ignore DIF. We describe modeling techniques that account for this source of variation by assuming that respondents have different thresholds for their scores and illustrate the use of these models. We demonstrate the application of these techniques with the three primary examples that we use to motivate the book in the introduction: V-Dem's expert surveys of political institutions, ANES questions designed to measure collective identity, and Yelp restaurant reviews.

The final source of disagreement is the least-studied and most difficult to fix. The assumption that underlies survey enterprises is that all individuals they recruit are qualified to respond to the domain of the survey. However, as the scale of the survey enterprise increases—in number of respondents, spread of cases, or breadth of questions—it becomes more likely that this assumption is false. Returning to the corruption example, perhaps most citizens can complete a survey about their experiences with bribery and perceptions of corruption within their local community, but as they are asked to comment on more types of corruption or other levels of government, they begin to vary significantly in their ability to do so. As a result, modeling all respondents as being equally reliable is problematic, in that it undoubtedly increases estimate uncertainty and possible introduces estimate bias. V–Dem addresses this issue by leveraging the scores of other respondents to determine individual respondents' level of random error—their "reliability." In essence, this means that respondents who provide information dissimilar from other respondents associated with their sector of the population contribute less to the estimation procedure. Ideologically, this procedure is in line with the survey ethos: it assumes that the majority of respondents are providing "good" responses, while allowing for the possibility that some respondents answer less well. In the chapter, we will provide a detailed overview of the method V–Dem uses to estimate respondent reliability, as well as a description of applications in which this method could go awry. We also demonstrate the application of this approach to modeling reliability to surveys of lay respondents (ANES) and consumer ratings (Yelp).

15,000 words

6.4 Bridging Cases

As previewed in Chapter 3, cross-coder comparability is a critical issue for survey projects to address. This chapter highlights one key approach for improving cross-coder comparability within V-Dem. We focus on bridging in three domains. First, we highlight the importance of bridging for expert surveys, using examples from V-Dem, the Electoral Integrity Project, and the Chapel Hill Survey to motivate our discussion. Second, we show how the techniques that we develop in the expert survey domain generalize to a key measurement problem in political science, the construction of ideological "common spaces," which use a combination of survey and voting data to compare ideology across citizens, politicians, and jurists. Third, we use the Yelp open data to demonstrate how fundamental bridging issues are to comparability in consumer ratings data.

Chapter 3 highlights two key stumbling blocks to measurement through survey. First, respondents disagree with one another because each respondent perceives the world differently, exhibiting systematic biases both in the information to which she has access, and how she evaluates such information. Second, respondents also disagree because they make random errors, and error rates vary across respondents. Thus, two respondents often provide inconsistent information about the same cases, and some respondents are more biased, and less reliable, than others. However, both of these concerns are exacerbated when different respondents evaluate different cases.

We illustrate in Chapter 3 that latent variable modeling techniques allow researchers to adjust for these key reasons for disagreement, to estimate respondent bias and reliability, and generally outperform traditional procedures for adjudicating respondent disagreement and quantifying confidence in latent trait estimates. However, these tools rely on overlap in the cases that respondents rate to estimate bias and reliability; when respondents rate disparate cases, such tools will perform poorly. Intuitively, if one set of respondents only rates case A, while another rates only case B, then, while we might produce decent estimates of relative bias and reliability within each set of respondents, we have no information about relative bias and reliability across sets. For example, say one set of hikers rates trail difficulty in the Smoky Mountains while another scores difficulty in Yellowstone, but neither set of hikers provides ratings for both parks. We might be able to use this information to learn a lot about how hikers in each park differ in how they rate trail difficulty, but we will know nothing about how the general perceptions of hikers vary across parks. In turn, we will have no way to align our estimates of difficulty in one park to the other, although we would be able to rank trails within parks. Indeed, while it may be likely that a "difficult" trail in the Smoky Mountains corresponds to a "moderate" trail in Yellowstone, our lack of *bridging* across parks makes it impossible to know for sure. This simple example illustrates a problem that is endemic in surveys, but which receives amazingly little attention from most survey dataset producers, potentially undermining data quality across the social and behavioral sciences. A similar problem afflicts web-commerce rating engines, which rarely, if ever, attempt to standardize ratings to reflect differences in how sets of raters—say diners in New York and Fargo—evaluate products.

This chapter focuses on strategies to overcome such bridging problems, and emphasizes the need to design surveys, from the ground up, to explicitly bridge cases. We use simulations, data from the Varieties of Democracy project, and examples from multiple other surveys, to demonstrate how badly things break when researchers fail to bridge effectively. We then show how sparse bridging networks can provide a practical, and

cost-effective, solution to this fundamental issue. While it is typically impossible for respondents to rate a wide variety of cases, it may often be practical for them to score a handful. For example, a resident of Gatlinburg, TN might find time to hike a few trails in Yellowstone. We show how to efficiently select bridges, and how to evaluate if one's network of bridges is sufficiently dense to produce robust estimates of latent traits. Finally, we discuss how these techniques apply to other survey topics, such as common space ideal point estimation across voters or consumer ratings on e-commerce platforms such as Yelp.

15,000 words

6.5 Anchoring Respondents

The chapter begins with a literature review of anchoring vignettes and their use in surveys. Anchoring vignettes are a widely accepted approach for addressing cross-respondent differences (King et al. 2004, King & Wand 2007, Bakker, Jolly, Polk & Poole 2014). Anchoring vignettes are descriptions of specific, but fictional—or at least unnamed—cases that provide the information required to answer a certain question. For example, in a cross-national expert survey of political institutions, like V–Dem, they are descriptions of imaginary country-years that focus on the quality of a given democratic institution, such as the degree of vote-buying. In a feeling thermometer towards a public policy proposal, they might be concrete descriptions of positive or negative feelings. Respondents rate these vignettes; patterns of variation in how raters evaluate these fictional cases provide information about differences in how they translate concrete aspects of cases into ordinal ratings.

There are several reasons vignettes are a powerful and efficient tool for addressing DIF. First, respondents have all the information about the case in question at their fingertips. Coding vignettes, therefore, requires substantially less effort than evaluating actual cases. Respondents can thus provide many vignette responses in a given set of time. Second, vignettes require no case knowledge, so even respondents who are not qualified to evaluate certain actual cases can code vignettes. Third, vignettes can also provide perfect overlap, because every respondent answers the same questions. Fourth, vignettes have the potential for high threshold variability, because their implementers control their content and strive to maximize that variability. Further, because every respondent considers the same information when they rate a vignette, it is potentially safe to assume low random error in the rating process and treat all cross-respondent variation as evidence of threshold differences. In contrast, in other bridging methods there remains variation in respondent reliability because of idiosyncratic variation in the information available to each respondent, reducing what we can learn about thresholds from each provided rating. Finally, in asking all respondents to code vignettes, we address the potential selection bias by having only those who opt to respond to certain questions—i.e., those who are either most knowledgeable about those questions or those who just think they are—provide data to adjust for cross-case comparability.

Engaging a growing literature on when vignettes prove effective, the remainder of the chapter uses V-Dem's vignettes data to examine three related questions about anchoring vignettes. First, we evaluate vignette data: whether respondents are able to distinguish vignettes and order them (via their coding) as we expected. We model vignette coding disagreement with question, answer, and vignette characteristics. We also explore the

somewhat puzzling instances in the V–Dem vignettes data where we see inconsistent coding by one respondent of the same vignette over time, a fascinating instance of within-respondent DIF based on the timing of the survey. Second, we consider how to incorporate the vignettes data in the sorts of measurement models that we develop throughout the book. Finally, we evaluate the quality and performance of V–Dem's anchoring vignettes empirically, drawing conclusions about how to write effective anchoring vignettes and the scope conditions on the use of vignettes in survey projects.

15,000 words

6.6 Modeling Longitudinal Responses

Throughout this book, we show how to bring a standard toolkit of statistical latent variable models to bear on the problem of producing valid and reliable data from surveys. These tools have found broad application across the social sciences; one main contribution of the book is to demonstrate their value for analyzing survey data, and to show how to best design survey instruments, improve comparability, and incorporate uncertainty and information into the measurement model.

This chapter explores three complex, but important, extensions to the standard toolkit we propose. First, we show how to use Bayesian statistical techniques, particularly prior specifications, to leverage careful concept formation and theory when estimating latent traits from survey responses. Second, we show how to tailor the error specifications built into canonical models to better fit common concept measurement domains. Specifically, we show how modeling the temporally "sticky" errors common in panel surveys improves the accuracy and reliability of measures. Third, we demonstrate how to model a common comparability problem in social science: how to detect and adjust for changes in how respondents understand a given question over time.

The first two section in this chapter generalize the tools that we introduce earlier in the book, providing technical guidance for how to most effectively deploy the comparability-enhancing techniques described in the rest of the book to a common use-case. Numerous applications of latent variable models in the social sciences use Bayesian techniques and dynamic prior specifications to help overcome issues with bridging latent variable estimates across time periods (e.g., Martin & Quinn 2002, Schnakenberg & Fariss 2014). We emphasize that such prior specifications rely on theory to make restrictive assumptions about the dynamics of the latent trait in question. Further, it's critical to align prior assumptions with the data generating process, to avoid low accuracy and over-confidence in estimates. Using the V–Dem Project as an example, we argue that commonly used dynamic latent variable models provide a poor theoretical and empirical fit to the concepts that many panel surveys, especially in the social sciences, seek to measure. In particular, we argue that political institutions rarely follow smooth dynamic processes, but rather exhibit periods of stasis, or slow drift, followed by sudden change. But commercial data science applications can also benefit from careful theorizing about dynamics. For example, restaurant quality changes over time in a manner surprisingly consistent with institutions: quality tends to hold reasonably constant, or drift slowly, except when changes in management or other shocks cause large alterations to operations. We argue that change-point models (Chib 1998, Park 2011) provide a better theoretical match to institutional (and restaurant quality) dynamics and show that latent variable models built upon a theory of institutional change-points better fit such data than do standard

models of smooth dynamic evolution. We also demonstrate a series of robustness tests to check dynamic modeling assumptions, discuss the broad costs and benefits of relying on dynamic modeling assumptions when fitting latent variable models to panel data, and situate our contribution relative to recent work on robust dynamic latent variable models (Reuning, Kenwick & Fariss 2019).

Panel surveys violate one key assumption of standard latent variable models: that respondent evaluations of cases are independent across time. Standard models assume, for instance, that an expert's rating of judicial independence in Kenya in 2001 is independent of her rating of independence in 2000 and 2002, or that a lay respondent's evaluation of corruption incidence today is unrelated to what she reported last year. This assumption is clearly at odds with reality, and standard latent variable models therefore tend to overstate confidence in estimates when researchers apply them to panel surveys. Traditional dynamic latent variable models do not address this problem. We describe extensions to our core modeling framework that allow for the possibility that errors are "sticky," and use simulations and examples from the V–Dem project to show that modeling sticky errors appropriately affects measurement in substantively meaningful ways.

Finally, we ask how to best deal with situations in which one repeatedly deploys a question over time—either to the same panel of survey respondents or to a rotating cast of participants—even as the meaning of that question evolves. Using the measurement of respect for human rights (Fariss 2014) as a motivating example, we explain how to produce comparable measures of concepts, even as understandings of those concepts change.

In sum, this chapter addresses cutting edge questions about how to best measure dynamic latent traits from surveys, especially cross-national panel surveys such as V–Dem, but also unconventional "panels," such as continuously running e-commerce ratings systems. We will accompany this chapter with software and online tutorials to allow readers to fit the models described here to their own datasets.

15,000 words

6.7 Interpreting and Presenting Cross-Contextual Survey Data

Survey data are necessary to capture important hard-to-measure concepts across cases and over time. Output from models which aggregate cross-contextual survey data are of great interest to a wide range of audiences, from consumers, to policymakers, to quantitative social scientists. Such widespread appeal presents challenges for producers of survey datasets. The technical skills of audiences diverge widely, as do their desired applications. The goal of a survey project is to produce data that is suited to this diverse audience, maintaining methodological integrity while also ensuring that the data are readily interpretable. The V–Dem project has been an ideal laboratory for threading this needle, as its data have been used in projects that range from cutting-edge quantitative research, to policy-oriented country rankings on different metrics, to undergraduate classroom presentations. V-Dem data have even graced numerous glossy policy publications, like the World Bank's annual Development Report, and have been featured in visualizations produced by numerous news outlets. This chapter draws on this experience to help survey consumers and producers to leverage the sophisticated modeling framework that we develop in the book to produce high quality measures, while effectively communicating with diverse audiences.

Using V–Dem's experience as a motivating example, but also including examples based

on the Yelp dataset, this chapter demonstrates how to: 1) leverage the flexibility of the modeling framework to tailor outputs to different consumers, and 2) effectively communicate measurement uncertainty when presenting estimates and conducting downstream analyses.

The measurement modeling framework that we develop throughout the book has a key disadvantage relative to more traditional techniques: it is complicated. Most importantly, because the framework allows for DIF, it necessarily produces estimates that are on a different scale than the original questions. This scale is nonetheless amenable to some consumers, particularly academic researchers who are comfortable working in latent spaces, and for whom an interval-level scale is methodologically convenient. But other consumers, especially policymakers, are less comfortable working with this abstract output. We show how to work within the modeling framework to produce a variety of outputs—rankings, ordinal estimates on the original question scale, and probabilistic case comparisons—that meet the needs of a variety of audiences, and allow one to simultaneously produce quality estimates from surveys while communicating to audiences with varying needs, and different levels of statistical literacy. We show how to produce and graphically present these various outputs, describe how they meet different constituencies' needs, and introduce software that makes this process easy for readers.

Finally, this chapter provides an in-depth description of the importance of both presenting and using measurement uncertainty in a variety of applications. The tools that we develop throughout the book engage measurement uncertainty in a coherent way, and allow scholars and policy makers to be honest about what they learn from surveys. But most existing applications of surveys either provide only crude measures of uncertainty, or ignore it altogether, and both scholars and practitioners lack the tools and experience to effectively communicate uncertainty and incorporate it in analysis. We show the importance of taking measurement uncertainty seriously, using a series of examples to demonstrate the pitfalls inherent in ignoring it, or modeling it simplistically. We then show how to communicate uncertainty around a variety of model outputs, and to differing audiences. Finally, we show how to incorporate uncertainty into statistical analyses using survey measures. Again, we introduce software that readers can use to employ these methods in their own work.

In sum, this chapter shows how to leverage complex tools to produce broadly interpretable output and how to weigh scholarly considerations against public engagement in data production enterprises.

15,000 words

References

Aldrich, John H & Richard D McKelvey. 1977. "A Method of Scaling with Applications to the 1968 and 1972 Presidential Elections." *American Political Science Review* 71(1):111–130.

Atkeson, Lonna & Lonna Atkeson, eds. forthcoming. *The Oxford Handbook of Polling and Survey Methods*. Oxford University Press.

Bakker, Ryan, Seth Jolly, Jonathan Polk & Keith Poole. 2014. "The European Common Space: Extending the Use of Anchoring Vignettes." *The Journal of Politics* 76(4):1089–1101.

Box-Steffensmeier, Janet M., John R. Freeman, Matthew P. Hitt & Jon C. W. Pevehouse. 2014. *Time Series Analysis for the Social Sciences*. Cambridge University Press.

Chib, Siddharta. 1998. "Estimation and Comparison of Multiple Changepoint Models." *Journal of Econometrics* 86(2):221–241.

Fariss, Christopher J. 2014. "Respect for Human Rights has Improved Over Time: Modeling the Changing Standard of Accountability." *American Political Science Review* 108(2):297–318.

Gelman, Andrew & Jennifer Hill. 2006. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.

Groves, Robert M, Floyd J Fowler Jr, Mick P Couper, James M Lepkowski, Eleanor Singer & Roger Tourangeau. 2011. *Survey methodology*. Vol. 561 John Wiley & Sons.

Hare, Christopher, David A Armstrong, Ryan Bakker, Royce Carroll & Keith T Poole. 2015. "Using Bayesian Aldrich-McKelvey Scaling to Study Citizens' Ideological Preferences and Perceptions." *American Journal of Political Science* 59(3):759–774.

Kellstedt, Paul M. & Guy D. Whitten. 2013. *The fundamentals of political science research*. Cambridge University Press.

King, Gary, Christopher J. L. Murray, Joshua A. Salomon & Ajay Tandon. 2004. "Enhancing the Validity and Cross-Cultural Comparability of Measurement in Survey Research." *The American Political Science Review* 98(1):191–207.

King, Gary & Jonathan Wand. 2007. "Comparing Incomparable Survey Responses: Evaluating and Selecting Anchoring Vignettes." *Political Analysis* 15(1):46–66.

LeBreton, J. M. & J. L. Senter. 2007. "Answers to 20 questions about interrater reliability and interrater agreement." *Organizational Research Methods* 11(4):815–852.

Lindstaedt, Rene, Sven-Oliver Proksch & Jonathan B. Slapin. 2016. "When Experts Disagree: Response Aggregation and Its Consequences in Expert Surveys.".

Martin, Andrew D. & Kevin M. Quinn. 2002. "Dynamic Ideal Point Estimation via Markov Chain Monte Carlo for the U.S. Supreme Court, 1953–199." *Political Analysis* 10:134–153.

McCleary, Richard, David McDowell & Bradley Bartos. 2017. *Design and Analysis of Time Series Experiments*. Oxford University Press.

Moser, Claus Adolf & Graham Kalton. 2017. *Survey methods in social investigation*. Routledge.

Park, Jong Hee. 2011. *Modeling Preference Change via a Hidden Markov Item Response Theory Model*. Boca Raton: CRC Press.

Reuning, Kevin, Michael R. Kenwick & Christopher J. Fariss. 2019. "Exploring the Dynamics of Latent Variable Models." *Political Analysis* 27(4):503–517.

Roy, Tarun Kumar, Rajib Acharya & Arun Roy. 2016. *Statistical survey design and evaluating impact*. Cambridge University Press.

Schnakenberg, Keith & Christopher J. Fariss. 2014. "Dynamic Patterns of Human Rights Practices." *Political Science Research and Methods* 2(1):1–31.

Zeller, Richard A. & Edward G. Carmines. 1980. *Measurement in the social sciences*. Cambridge University Press.