PA

# Assessing Data Quality: An Approach and An Application

**Kelly McMann[1], Daniel Pemstein[2], Brigitte Seim[3], Jan Teorell[4] and Staffan Lindberg[5]**

[1] *Political Science, Case Western Reserve University, Cleveland, OH 44106, USA. Email: kelly.mcmann@case.edu*
[2] *Political Science and Public Policy & Challey Institute, North Dakota State University, Fargo, ND 58102, USA. Email: daniel.pemstein@ndsu.edu*
[3] *Public Policy, University of North Carolina, Chapel Hill, NC27599, USA. Email: seimbri@gmail.com*
[4] *Political Science, Lund University, Lund, 221 00, Sweden. Email: jan.teorell@svet.lu.se*
[5] *V-Dem Institute & Political Science, University of Gothenburg, Gothenburg, 413 27, Sweden. Email: staffan.i.lindberg@pol.gu.se*

## Abstract

Political scientists routinely face the challenge of assessing the quality (validity and reliability) of measures in order to use them in substantive research. While stand-alone assessment tools exist, researchers rarely combine them comprehensively. Further, while a large literature informs data *producers*, data *consumers* lack guidance on how to assess existing measures for use in substantive research. We delineate a three-component practical approach to data quality assessment that integrates complementary multimethod tools to assess: (1) content validity; (2) the validity and reliability of the data generation process; and (3) convergent validity. We apply our quality assessment approach to the corruption measures from the Varieties of Democracy (V-Dem) project, both illustrating our rubric and unearthing several quality advantages and disadvantages of the V-Dem measures, compared to other existing measures of corruption.

*Keywords:* validity, reliability, Bayesian IRT, corruption

Most political scientists are concerned about the quality of the measures they use, yet few scholars rigorously evaluate a chosen measure's quality prior to incorporating it in substantive research (Herrera and Kapur 2007). This tendency stems not from a lack of tools: the social sciences are replete with methods for evaluating measure validity and reliability (see, e.g., Shadish, Cook, and Campbell, 2002 or Seawright and Collier, 2014). The challenge is twofold. First, there is no accepted, comprehensive, practical approach for assessing a measure's quality. Second, there is no broadly adopted strategy for scholars to incorporate information from such assessments in substantive research or to amend their research designs accordingly.

We argue that it is critical to assess the quality of measures in advance of using them, both to diagnose a chosen measure's strengths and limitations and to implement strategies to mitigate quality concerns. Forty years ago, Zeller and Carmines (1980) elucidated a strategy for measurement in the social sciences that prioritized evaluating measure validity and reliability in terms of theoretical construct validity and patterns of internal and external association between indicators. They argued that "...the *auxiliary theory* specifying the relationship between concepts and indicators is equally important to social research as the substantive theory linking concepts to one another (Carmines and Zeller 1979, p. 11)." This article builds on this tradition, but emphasizes the need for these two types of theories—measurement and substantive models—to speak to one another, and for analysts to rigorously apply both sorts of models in tandem to inform the substantive conclusions that they draw from research using measures of social concepts.

***Political Analysis (2021)***

**DOI:** 10.1017/pan.2021.27

**Corresponding author**
Brigitte Seim

**Edited by**
Jeff Gill

The data quality assessment[1] literature is extensive,[2] but it lacks practical applicability in several key ways. First, as we elaborate in more detail below, many works provide guidance for *developing* a measure, rather than guidance for assessing and effectively *using* a measure someone else developed. In other words, the literature is primarily geared toward data *producers*, even though many researchers are primarily data *consumers*.[3] Second, much of the literature implicitly or explicitly employs satisfying standards, asking: What does a researcher have to do to show a measure is "valid enough" or "reliable enough?" This all-or-nothing approach is more useful to producers than to consumers, who often must rely on existing imperfect measures to answer their substantive questions. Third, works typically focus on one assessment tool (or a select few), rather than synthesizing the tools to provide a comprehensive and easily implemented assessment process. Fourth, many prior works on data quality assessment also have a narrow focus on validity, overlooking or giving little attention to reliability. Finally, few researchers take what they learn in assessing a measure and incorporate it when conducting analysis associated with substantive research. Instead, the assessment serves only to "rubber stamp" the measure.

To address these five gaps in existing literature, this paper synthesizes a set of complementary, flexible, practical, and methodologically diverse techniques for assessing data quality.[4] Rather than recommending the use of one technique over another, we advocate a comprehensive approach to assessment. In our proposed approach, we first assess content validity, which we define as the alignment between the higher-level theoretical concept under study, the measure being assessed that is designed to capture that concept, and any alternative available measures. Second, we assess the validity and reliability of the data generation process, including evaluating the bias and error introduced in the dataset management structure, data sources, respondent coding procedures, aggregation models, and case coverage, as well as analyzing the predictors of inter-respondent disagreement and intra-respondent biases.[5] Third, we assess convergent validity, or the alignment between the measure being considered, alternative available measures, and qualitative case studies. Throughout the assessment, we employ a variety of qualitative and quantitative tools.

While much of what we recommend synthesizes existing tools, we also innovate in developing three new assessment strategies. First, we delineate a road map for evaluating the validity and reliability of the *data generation process* as a signal of resulting data quality. Second, we analyze how respondent and case characteristics affect inter-respondent[6] disagreement and intra-respondent biases. Third, we build on existing convergent validity assessments to combine regression analysis and qualitative "blind" case studies to identify the determinants of measure divergence and the set

1 Throughout this paper, we use the word "assessment" to include what is often referred to as a "validation" exercise (Adcock and Collier 2001; Seawright and Collier 2014), as well as a broader evaluation of both validity and reliability.

2 To name a few key works: Seawright and Collier (2014); Collier, LaPorte, and Seawright (2012); Adcock and Collier (2001); Zeller and Carmines (1980); Sartori (1970); Campbell and Fiske (1959). There is also a strand of research that examines the validity of particular measures. See, for example, Donnelly and Pop-Eleches (2018); Jeong (2018); Marcus, Neuman, and MacKuen (2017); Morin-Chassé *et al.* (2017), and Schnakenberg and Fariss (2014).

3 In this article, we focus on providing practical guidance for those who *use* pre-existing measures in substantive research (data "consumers"). Providing particular guidance for those who *produce* measures (data "producers") is outside the scope of this article, though much of our advice for consumers would be relevant for producers as well.

4 We use the word "quality" to encompass both a measure's reliability and its validity, an inclusive term that is particularly useful when we discuss evaluating the data generation process.

5 By "intra-respondent biases" we mean biases specific to particular (subsets of) respondents. While we use the term respondent, these tools are generally applicable whenever measures draw on multiple sources—coders or otherwise.

6 Throughout the article, we use the terms "respondent" and "coder" interchangeably to refer to a survey respondent who provides responses to questions about a particular case. The assessment tools in this article regarding respondent recruitment, coding procedures, and inter-respondent reliability pertain to all types of survey respondents. For the data source we examine in this article—V-Dem—"expert coders" are the respondents, and they answer questions on a survey about a particular country. These answers are then aggregated to provide measures at the level of the country-year. As an example of another type of respondent, in a household survey such as the Afrobarometer, household members are the respondents, and they answer questions on a survey about a particular household. These answers are then (often) aggregated to provide data across geographic units (villages, districts, countries, etc.).

of affected cases. Throughout the article, we highlight how researchers can use these assessment tools to evaluate how appropriate a measure is for a particular research question or a particular set of cases, to identify potential limitations in substantive conclusion, and to adapt substantive conclusions to measurement concerns.

We demonstrate our proposed quality assessment approach by applying it to the Varieties of Democracy Project (V-Dem) corruption measures. Corruption is a particularly difficult concept to measure, so assessing the quality of corruption measures is critical. In line with one of the gaps in existing literature highlighted above, we focus our assessment on points useful for *consumers* of the V-Dem corruption measures, not for the V-Dem *producers*: we assume attributes of the V-Dem enterprise such as managerial structure, respondent recruitment, and aggregation procedures are fixed, and we simply assess these attributes to understand their implications for those using the V-Dem corruption measures in substantive research. To address another of the gaps highlighted above, our assessment of the V-Dem corruption measures is comparative. As corruption is challenging to measure, we assume no corruption measure is of perfect quality. We aim to unearth areas of *relative* strength and weakness to, in turn, determine the set of research questions, cases, and areas of corruption that best align with the measures' strengths.

Our assessment reveals both strengths and limitations of the V-Dem corruption measures. The V-Dem corruption measures are relatively high-quality in capturing exchange-based material corruption among government officials, corruption in non-Western countries, corruption in contexts with relatively high levels of corruption and more expansive freedom of expression, and historical corruption (i.e., in years prior to the starting year for other corruption measures). In cases where V-Dem employed a greater proportion of women respondents, or respondents with doctoral degrees, the V-Dem corruption measures diverge from other corruption measures. We encourage considering these areas of relative strength (and corresponding areas of weakness) when evaluating whether the V-Dem corruption measures can be employed to answer a particular research question for a given set of cases in substantive research. Further, we propose the data quality assessment approach we present can be applied to evaluate the quality of existing measures more generally.

## 1 A Departure from and a Refinement of Previous Work

Some of the most valuable work on data quality assessment, including Adcock and Collier (2001); Seawright and Collier (2014); Hayes and Krippendorff (2007), and Zeller and Carmines (1980), provides advice, primarily or exclusively, for data *producers*—those who develop datasets. They refer to their readers as individuals producing data (e.g., "help scholars develop measures"; Adcock and Collier 2001, p. 534). Yet, with a proliferation of publicly available cross-national datasets and global indices, social scientists are in dire need of advice for data *consumers*—those who use these publicly available datasets produced by others in substantive research. Consumers must conduct active quality assessments of the measures that they use precisely because no producer can adequately anticipate a wide array of consumer needs. Our guide focuses on that task.

The literature also generally overlooks practical, step-by-step guidance. Some of the most informative works, such as Seawright and Collier (2014), make us attentive to data quality assessment debates, inform us of different tools, and illustrate them. However, they are not practical data quality assessment road maps, but rather a theoretical presentation of assessment concepts. The classic work of Carmines and Zeller (1979) and Zeller and Carmines (1980) serves as a model for our approach here. We update and extend this body of work to synthesize tools—several provided by other scholars as well as several we develop—into a practical assessment process, apply the process to a particular case, and demonstrate how this holistic approach reveals insights useful in conducting substantive research with a chosen measure.

Further, offering a comprehensive approach is a helpful complement to publications that examine only a single tool (see, e.g., Bollen 1980; Campbell and Fiske 1959; Thomas 2010, and Sartori (1970)). Our approach underscores the value of combining different data quality assessment tools, including harnessing the advantages of both qualitative and quantitative approaches. We do not claim that our catalog of tools is exhaustive, but rather that it can serve as a relatively flexible foundation for assessing the quality of a measure.

In addition, many prior works on data quality assessment provide narrower guidance because they focus exclusively on validity, generally defined as the alignment between a measure and the underlying concept. For example, some of the most oft-cited articles on measurement in political science do not even mention reliability (Adcock and Collier 2001; Collier, LaPorte, and Seawright 2012; Seawright and Collier 2014). Similarly, in Chapter 7 of his canonical book, *Social Science Methodology*, John Gerring acknowledges that validity and reliability are the "two overall goals" in "pursuing the task of measurement," but subsequently only discusses how to assess reliability for half of a page, concluding that inter-respondent reliability tests should be performed whenever multiple respondents are used to generate each data point (Gerring 2012, p. 158–159). Reliability is, of course, central to data quality assessment, as a large literature emphasizes (Reise, Widaman, and Pugh 1993; Hayes and Krippendorff 2007). The approach we lay out in this article illustrates how jointly assessing validity and reliability (referred to collectively as "quality") leads to concrete findings that can be incorporated in substantive research. Here, we push readers to return to the balanced assessment of validity and reliability emphasized by Zeller and Carmines (1980).

Finally, even the most insightful works on measurement do not take the critical post-assessment step of discussing how the assessment's findings can be incorporated into substantive research. Measures are typically torn down without advice about how to use imperfect measures: much of the literature implies that a less-than-perfect measure is not worth using (Mudde and Schedler 2010). There is very little attention to mitigating, or at least acknowledging, limitations.[7] In many ways our work is in the spirit of, and an update to, the 40-year-old book by Zeller and Carmines (1980), yet even that classic does not illustrate how the assessment's findings can affect substantive research. Likewise, Herrera and Kapur (2007) approach data collection "as an operation performed by data actors in a supply chain," delineating these actors, their incentives, and their capabilities (p. 366). They urge scholars to focus on validity, coverage, and accuracy, offering several examples of measures that have failed on these dimensions. They stop short, however, of explaining the implications of measures' strengths and limitations for substantive research. We provide suggestions on how to do so.

In addition to the data quality assessment literature, our work is also informed by the large literature on the quality of democracy measures, particularly its emphases on aligning measures with higher-level conceptualization; considering differences in coverage, sources, and scales across measures; and transparency in coding and aggregation procedures (Bollen 1980; Bollen 1990; Munck and Verkuilen 2002; Bowman, Lehoucq, and Mahoney 2005; Pemstein, Meserve, and Melton 2010; Coppedge *et al.* 2011; Fariss 2014). To develop tools for assessing both the data generation process and convergent validity, we draw heavily on the work of Steenbergen and Marks (2007), Dahlström, Lapuente, and Teorell (2012), and Martinez i Coma and van (2015), who represent literature on party positions, public administration, and election integrity, respectively. Finally, we extensively borrow insights from the literature on corruption measurement, both because we apply our approach to the V-Dem corruption measures and because the literature on measuring corruption raises general issues about data quality assessment more generally (Knack 2007; Treisman 2007; Hawken and Munck 2009b,Hawken and Munck 2009a; Galtung 2006).

7 Some works that examine specific topics, such as robust dynamic models for modelling latent traits that change quickly (Mislevy 1991; Bolck, Croon, and Jagenaars 2004; Reuning, Kenwick, and Fariss 2019), do incorporate assessment findings, but these works do not provide general guidance.

**Table 1.** Data quality assessment approach.

| Category | Guiding questions | Tool |
|---|---|---|
| Content validity assessment | To what extent does the measure capture the higher-level theoretical construct it is intended to capture and exclude irrelevant elements? | Evaluate the inclusion of relevant meanings and exclusion of irrelevant meanings using qualitative assessment and quantitative factor analysis. |
| | How does it compare in content to alternative measures? | |
| Data generation process assessment | Does the data generation process introduce any biases, reliability problems, or analytic issues? | Evaluate dataset management structure, data sources, coding procedures, aggregation procedures, and geographic and temporal coverage of the measure. |
| | How does it compare to the data generation process of alternative measures? | |
| | Where multiple respondents exist, to what extent do they generate consistent and converging information? | Evaluate extent of disagreement among respondents, and whether respondent and case characteristics predict disagreement and patterns of responses. |
| Convergent validity assessment | Does the measure accurately capture actual cases? | Evaluate measure against original or pre-existing case studies. |
| | To what extent does the measure correlate with alternative measures of the construct, and are areas of low correlation thoroughly understood? | Evaluate predictors of difference, any outliers, and the implications of differences across measures. |

## 2 A Practical Approach to Assessing Data Quality

We propose a practical approach for assessing a measure's quality that involves three components: a content validity assessment; a data generation process assessment; and a convergent validity assessment (see Table 1). Collectively, these considerations illuminate the degree to which the measure is valid and reliable.[8] Validity is the absence of systematic measurement error. Reliability is the absence of unsystematic (or random) measurement error.[9] Reliability should not be overlooked when assessing the quality of a measure; while it is not useful on its own, neither is a well-conceptualized but unreliable measure.

First, one should examine the extent to which the measure captures the higher level theoretical concept. Content validity assessment, where the analyst maps indicators to theoretical concepts (Carmines and Zeller 1979; Zeller and Carmines 1980; Bollen 1989; Adcock and Collier 2001; Seawright and Collier 2014) is the primary way to approach this question.[10] In addition to assessing the measure against a theoretical construct, we suggest assessing the content validity of the measure relative to other available measures of that construct. It is important to note that a measure's content validity is specific to a particular theoretical concept, so research projects

8 As noted in Collier and Levitsky (1997) and Adcock and Collier (2001), there has been a proliferation of terms to describe different types of validity. Our goal is not to evaluate types of validity or validation tools, but rather to *synthesize* from the literature a set of practically useful tools that facilitate a comprehensive assessment of a chosen measure. We aim to reference relevant works throughout and delineate how the types of validity we discuss and the validation tools we propose map to other works in the literature, though we hasten to point out that a full reconciliation of this literature is an article in and of itself.

9 Seawright and Collier (2014) consider validity and reliability to be distinct properties from measurement error, but we view these three properties as related.

10 Schedler (2012) refers to this as assessing alignment with shared concepts.

that use different theoretical concepts will likely find different measures are strongest in terms of content validity.

Second, it is important to assess the validity and reliability of the data generation process. An unbiased and reliable data generation process results in unbiased and reliable measures. The appeal of including this component in a data quality assessment approach is that it compels a focus on something that can be evaluated (i.e., the nature of a process) rather than something that cannot (i.e., a measure's alignment with the truth). For example, though we cannot prove that a respondent selected the "true" answer when answering a question about Argentina's level of civil society freedoms in 1950, we can show that the process to recruit, engage, and synthesize data from that respondent was unbiased and reliable. In evaluating the data generation process, we recommend scrutinizing the dataset management structure, data sources, respondent coding procedures, aggregation models, and case coverage. Where multiple respondents are used, we encourage analyzing the predictors of inter-respondent disagreement and intra-respondent biases to evaluate the reliability of the data generation process and to expose potential determinants of systematic bias. In particular, considering the individual respondent and case characteristics that predict disagreement or response patterns allows researchers to identify threats to validity and reliability that are driven by the composition of respondent pools. As in the first component of our data quality assessment approach, researchers should consider their particular theory as they evaluate the data generation process of a measure, and should assess a measure's strengths and limitations relative to other measures.

The third component in our approach is to assess convergent validity, or the alignment between the measure being considered, alternative available measures, and qualitative case studies.[11] We use two tools in our convergent validity assessment: comparing the measure to alternative comparable measures; and comparing the measure to actual cases. With regard to the former, it is important to acknowledge that the quality of other measures might not be certain. So, the task at hand is to evaluate the strength of correlations and any outliers in order to more completely understand the advantages and disadvantages of the measure of interest. A useful approach is to analyze the *predictors* of differences across measures, rather than only the aggregate correlation level. One can use original or pre-existing case studies for qualitative comparisons, to assess whether the measure "converges" with case history. However, it is critical that the researcher considering the case material is "blind" to the measures; she must effectively recode the cases independently, using only the case material. She can examine the alignment between the coded cases and the measure after completing this blind recoding.

## 3 V-Dem Corruption Measures

The V-Dem dataset[12] covers nearly all countries of the world from 1900 to 2012.[13] V-Dem provides six indicators of corruption based on six survey questions: two each for the executive and public sector on: (a) bribery and other corrupt exchanges and (b) theft and embezzlement. Then, there is a single indicator for corruption in the legislature and another for corruption in the judiciary. The

11 This type of validation is often referred to as either convergent validity (Bollen 1989) or convergent/discriminant validity (Campbell and Fiske 1959; Adcock and Collier 2001; Seawright and Collier 2014), though Schedler (2012) deviates from the norm by evaluating a measure's "alignment with shared realities."

12 Replication code for this article has been published in Code Ocean and can be viewed interactively at https://doi.org/10.24433/CO.0269024.v1 (McMann *et al.* 2021a). A preservation copy of the same code and data can also be accessed via Dataverse at https://doi.org/10.7910/DVN/BXV4AT (McMann *et al.*, 2021b).

13 The analysis in this paper is based on v4 of the V-Dem dataset (Coppedge *et al.* 2015a,b; Pemstein *et al.* 2015). The primary update to each version is to extend the time period covered by the dataset, and the use of different versions is highly unlikely to affect the results presented here.

**Table 2.** Conceptual alignment across V-Dem corruption indicators (BFA Estimates).

| Measure | Loadings ($\Lambda$) | Uniqueness ($\Psi$) |
|---|---|---|
| Executive bribery (**v2exbribe**) | 0.829 | 0.313 |
| Executive embezzlement (**v2exembez**) | 0.827 | 0.316 |
| Public sector bribery (**v2excrptps**) | 0.846 | 0.285 |
| Public sector embezzlement (**v2exthftps**) | 0.848 | 0.281 |
| Legislative bribery/theft (**v2lgcrrpt**) | 0.693 | 0.496 |
| Judicial bribery (**v2jucorrdc**) | 0.753 | 0.434 |

exact language of each question appears in Table S1 of the Supplementary Appendix. The V-Dem Corruption Index aggregates these six indicators to produce an overall measure of corruption.[14]

## 4 Applying the Data Quality Assessment Approach

To demonstrate our proposed assessment approach, we apply it to the V-Dem corruption measures. Our approach to data quality assessment involves both stand-alone and comparative evaluations. Therefore, throughout the paper, we provide head-to-head tests with alternative corruption measures that make the required information available (e.g., details about the data generation process). The two alternative corruption measures we consider most often are the Worldwide Governance Indicators' Control of Corruption (WGI) and Transparency International's Corruption Perceptions Index (CPI). For readers who are not familiar with alternative corruption measures, we provide Table 2 in the Appendix as a reference.

### 4.1 Content Validity Assessment

As a first component in our data quality assessment approach, we propose evaluating the alignment between the higher-level theoretical concept under study and the measure being assessed that is designed to capture that concept—in other words, to determine the extent to which the measure captures all relevant meanings while excluding ones irrelevant to the "systematized" concept (Adcock and Collier 2001). We propose making this determination by using qualitative evaluations and quantitative factor analysis. In line with Seawright and Collier (2014), the qualitative evaluations involve assessing the level of correspondence between the data that the measure will generate and the systematized concept.[15] Factor analysis is a statistical tool that examines how closely different indicators relate to the same underlying concept.[16] It is often used as a test of convergent validity; we propose that it also helps illuminate whether a set of measures forming an index represent relevant meanings and exclude irrelevant meanings.

Our theoretical construct of interest in assessing the V-Dem Corruption Index is the "use of public office for private gain," a widely accepted academic definition of corruption (Rose-Ackerman 1999; Treisman 2000). We find that the V-Dem Corruption Index includes most, but not all, relevant meanings, and excludes irrelevant meanings. The V-Dem Corruption Index captures a wide variety of participants in corruption, including both top officials and public sector employees. It also captures a large number of corrupt practices, including both grand and petty corruption. Each of the six V-Dem corruption measures refers to a particular public officeholder. And, they

14 The V-Dem Corruption Index uses all the corruption variables available from V-Dem except for one, which pertains to corruption in the media rather than corruption in government.

15 This is sometimes called a "face validity" assessment (Seawright and Collier 2014), though there is ambiguity over the definition of such an assessment (Adcock and Collier 2001), so we avoid using the term here.

16 Basic factor analysis assumes that indicators are "reflective" of—or "caused" by–an underlying construct. If the indicators are instead "formative," there is no expectation that they should be inter-correlated. More complex structural equation models (e.g., MIMIC) may provide useful content validity assessment tools when the analyst believes that some indicators are reflective of, while others cause, the construct in question, but such models are beyond the scope of this paper.

use specific language to indicate numerous, particular corrupt practices, such as "bribes" and "steal, embezzle, or misappropriate public funds or other state resources for personal or family use," as well as more general language to include other forms of corrupt behavior. This language enables the survey questions to generate measures that cover a wide range of behaviors that fall under "the use of public office for private gain." However, the V-Dem corruption indicators are weaker in capturing "revolving door" corruption, where public sector positions are used to secure private sector jobs and vice versa, as this form of corruption is included only in the question about legislator corruption, not in the questions about other government officials (i.e., the executive, bureaucracy, or judiciary).

The V-Dem measures also exclude meanings of corruption that are irrelevant to the systematized concept. By specifying government officeholders in the questions, the measures exclude cases where the corruption involves only those outside government. By specifying types of personal gain in the questions, the measures also exclude behaviors where there is no evidence of direct, immediate material gain. For example, vote buying is not captured in any of the V-Dem corruption measures.

Next, we use Bayesian factor analysis[17] to assess content validity—specifically whether the six V-Dem corruption measures reflect one underlying systematized concept (i.e., corruption). As shown in Table 2, all six indicators strongly load on a single dimension, although the fit for both legislative and judicial corruption is somewhat weaker. Despite the strong factor loadings produced by this exploratory analysis, it is not theoretically clear that we should assume that these indicators reflect only a single underlying factor. Out of the six V-Dem corruption indicators, the four executive and public sector indicators have separate questions on bribery and embezzlement, the judicial corruption indicator focuses on bribery, and the legislative corruption indicator asks about both bribery and embezzlement in one question.[18] That said, an unconstrained two-factor model explains less than 2% more variance in the manifest variables than the unidimensional model that we report here, providing observational justification for the one-factor assumption.[19] In sum, the factor analysis provides empirical support for the broad conclusions of the qualitative construct validity assessment: the indicators largely reflect a single underlying systematized concept.

This application of factor analysis illustrates that its value for evaluating data quality rests fundamentally on combining it with theory-driven content validity assessment, which generates expectations about factor structure that one can explore empirically. Here, we used exploratory factor analysis tools to examine a simple question: Do the indicators plausibly reflect a single coherent latent construct? The model allows us both to evaluate our expectations and to explore inconsistencies. Indeed, the weaker loadings for legislative and judicial corruption potentially inform subsequent analysis. One can use confirmatory factor analysis to examine the same question, as we demonstrate in the Appendix, and this approach may be especially useful when one has strong, or particularly nuanced, *a priori* assumptions about factor structure (see Harrington,

17 Bayesian factor analysis assumes the same likelihood function as the traditional frequentist approach. It uses prior information to overcome the rotational invariance problem, typically by assuming, a priori, that at least one loading is strictly positive (or negative). Factor loadings generated by Bayesian factor analysis can be interpreted in the same way that one interprets loadings estimated through frequentist methods. We use the Bayesian approach here because it allows us to incorporate measurement error in the manifest variables, which themselves were estimated using Bayesian methods, into the model. Specifically, we start with hundreds of simulated draws from the posterior distributions of the manifest variables. Using vague priors, we fit the Bayesian factor model to each draw (i.e., one draw from the posterior of each manifest variable, covering every country-year), using Markov chain Monte Carlo (MCMC) methods to simulate a few hundred draws from the posterior of the factor model for that set of manifest variable draws. We then combine the posterior draws from every run of the factor model, and compute parameter estimates from this full set of draws, essentially averaging across uncertainty in the manifest variables.

18 This inconsistency reflects the fact that these six indicators are included on various V-Dem surveys, each designed by different scholars. They reflect a weakness in the V-Dem data generating process.

19 We further examine questions of model fit in the Supplementary Appendix, using frequentist confirmatory methods. Our one-factor model appears to fit the data well, although there is some disagreement across model fit statistics. There is little evidence that the indicators reflect two distinct underlying factors.

2008 for an accessible introduction to this massive literature). In general, factor analysis enables the researcher to explore the contours of the underlying latent structure of the data, which is a complement to—not substitute for—a theory-driven content validity assessment.

Finally, a measure's content validity can also be assessed comparatively. In our application, assuming that the corruption researcher seeks a measure that captures "the use of public office for private gain," we evaluate the content validity of the V-Dem corruption measures compared to alternative corruption measures. By their own descriptions, many of the alternative corruption measures include information about "public sector" or bureaucratic corruption, excluding executive, legislative, and judicial corruption. This includes CPI, the World Bank's Business Environment and Enterprise Performance Survey (BEEPS), and nearly all the Barometers.[20] In contrast, some alternatives are not clear about which public offices are included in their measures. For example, Transparency International's Global Corruption Barometer (GCB) combines data on the public sector with private "big interests," and International Country Risk Guides' Political Risk Services (ICRG) focuses on the "political system." The World Values Survey (WVS) offers a more transparent and expansive conceptualization, including petty and grand corruption as well as the perversion of government institutions by private interests. In contrast, some alternative corruption measures capture a very narrow slice of "the use of public office for private gain." For example, the International Crime Victims Survey asks only about exposure to bribery (Kennedy 2014). Of course, if a narrower (or broader) conceptualization of corruption is held by the researcher, one of these alternative corruption measures may be appealing. However, given a corruption definition of "the use of public office for private gain," the multi-form, multi-sectoral nature of the V-Dem corruption measures is useful because different countries are marred by corruption in different forms or sectors (Knack 2007; Gingerich 2013). Again, for substantive researchers, there is no objective "best" measure when it comes to content validity. The foremost consideration is the researcher's theory and theoretical constructs, and each available measure will likely offer relative strengths and limitations.

## 4.2 Data Generation Process Assessment

The second component in our data quality assessment approach is an assessment of the validity and reliability of the data generation process. While these attributes of the data generation process have been discussed in other works, we synthesize these discussions and illustrate their implications for a quality assessment. The steps in this portion of the assessment include assessing the dataset management structure, data sources, coding procedures, aggregation procedures, and geographic and temporal coverage. Lack of transparency about the data generation process will make it difficult to assess these aspects for some measures, in which case the researcher may have to skip these steps of the assessment.[21] Data consumers should demand this information from producers when it exists and privacy concerns do not preclude its publication.

When measures draw upon the contributions of multiple respondents, we also recommend analyzing inter-respondent disagreement and intra-respondent biases to assess both the validity and reliability of the data generation process.[22] We illustrate this component of the assessment by evaluating the V-Dem data generation process and highlighting its strengths and limitations relative to alternative corruption data sources.

*Dataset Management Structure.* Often overlooked sources of bias are the leadership and funding for a dataset. This is documented by Hawken and Munck (2009a), who find significant differences

20 The Afrobarometer is the exception, examining corruption among government officials generally or among particular groups of officials, depending on the year.

21 Lack of transparency surrounding the data generation process may also be a proxy for data quality, but we leave evaluating this untested assertion for future research.

22 We acknowledge that this technique applies to only those measures that do, in fact, rely on multiple respondents to generate a value for a given geographic unit at a point in time.

across corruption datasets, based on who is generating the data. In terms of data quality, leadership that is academic, rather than political or for-profit, and funding that is from diverse regions of the world, rather than from a single region or country, help to ensure that the organizational structure generates unbiased and reliable measures. In the case of V-Dem, it is an academic venture, led by scholars from universities in different countries with the V-Dem Institute at the University of Gothenburg, Sweden, as the organizational headquarters. Funding comes from research foundations and donor countries, mostly in Northern Europe, North America, and South America.

*Data Sources.* A key question to consider when evaluating potential bias and unreliability due to data sources is the number of data sources involved in generating the indicators and indices. As others have pointed out, datasets that aggregate information from different sources multiply biases and measurement errors by including those from each source in their composite measure, particularly if measurement errors across data sources are correlated (Herrera and Kapur 2007; Treisman 2007; Hawken and Munck 2009a). V-Dem avoids this problem because it uses one data collection process to generate all corruption indicators and indices rather than synthesizing multiple data sources. In contrast, three of the most commonly used corruption measures—WGI, CPI, and ICRG—aggregate information from different sources.

V-Dem's data generation process has the potential to generate a correlated errors problem of its own. In particular, because many of the same experts respond to the various corruption questions across the V-Dem surveys, there is the potential for rater error to correlate across indicators. Such correlated errors could undermine other aspects of our quality assessment, such as the factor analysis in our content validity analysis. This potential issue also implies that researchers should generally avoid predicting one V-Dem corruption indicator with another in applied work (Coppedge *et al.* 2020).[23]

*Respondent Coding Procedures.* When respondents generate data, it is important to examine: (1) the qualifications and potential biases of the respondents themselves and (2) the procedures for combining respondent answers into a single measure (Treisman 2007; Martinez i Coma and van 2015). We consider both of these below for the case of the V-Dem corruption measures.

Before evaluating the first area of coding procedures regarding respondent qualifications and biases, we first consider the appropriateness of the respondent pool. Several scholars have argued that expert-coded measures of corruption are inferior to citizen-coded or "experience" measures (Treisman 2007; Hawken and Munck 2009a,b; Donchev and Ujhelyi 2014). Rather than privileging one type of respondent over another, we recommend considering which type of respondent is best from a content validity perspective. For example, if a researcher is defining corruption as "the use of public office for private gain," citizen respondents present certain disadvantages. Citizen perceptions of corruption are fundamentally limited because they interact with only certain kinds of officials and observe certain kinds of corruption. Alternatively, the potential disadvantage of far-removed experts coding conditions in a country can be addressed by relying on experts who are residents or nationals of the countries—effectively serving as both expert coders and citizen respondents. If, instead, a researcher defines corruption narrowly to mean bribery, these disadvantages of citizen-coded measures of corruption transform into advantages. Once again, the choice of most appropriate respondent (and therefore most appropriate measure) should be based on the theory underpinning the research.

Given the definition of corruption employed elsewhere in this article—"the use of public office for private gain"—we assume an expert perspective is useful, and we move on to considering

23 In the Supplementary Appendix, we show that while raw residuals correlate highly across expert ratings of multiple corruption indicators, this appears to stem largely from differential item functioning (DIF), and we find little evidence of such cross-indicator correlations in rater errors after correcting responses for DIF (see the next section).

whether the particular experts within V-Dem are unbiased. The stringent selection criteria for experts within V-Dem could offset possible sources of bias. V-Dem experts have been recruited based on their academic or other credentials as field experts in the area for which they code and on their seriousness of purpose and impartiality (Coppedge *et al.* 2017). Impartiality is not a criterion to take for granted in political science research. Martinez i Coma and van (2015) noted that variance in estimates of election integrity in the Perceptions of Electoral Integrity dataset was significantly higher when one of the respondents was a candidate in the election. Further, no one respondent's background or biases can drive the estimates for a given country in the V-Dem dataset. At least five V-Dem experts code each question-country-year observation for a total of more than 3,000 experts involved to produce the dataset.[24]

We now turn our attention to the second area of coding procedures regarding combining respondent ratings into a single measure. When measures are based on ratings from multiple respondents, we can evaluate the process for combining information across respondents and use this information to provide estimates of the reliability of the measure. Researchers can adjust their inferences accordingly for measurement error. In assessing this, we ask if the process accounts for both systematic biases in how respondents answer questions and non-systematic variation in respondent reliability. For example, if respondents provide ordinal ratings and they vary in how they map those ratings onto real cases—perhaps, for example, one respondent has a lower tolerance for corruption than another—then a process that models and adjusts for this issue will outperform a more naive process. This is known as a differential item functioning (DIF) and affects most survey-based data collection processes. Similarly, it might be justifiable to weight more highly the contributions of more reliable respondents. Most multirespondent measures are generated by taking the average of the responses and, if reliability estimates are provided, they are in the form of standard deviations. These simple estimation procedures implicitly assume that there are no systematic differences in the way respondents produce ratings, treating respondents as equally reliable. When these assumptions are wrong, such procedures will generate flawed point estimates and measures of reliability (Pemstein, Meserve, and Melton 2010; Lindstaedt, Proksch, and Slapin 2016).

To combine respondent answers to generate country-year observations, V-Dem use statistical item response theory (IRT) techniques to model variation in respondent reliability while allowing for the possibility that respondents apply ordinal scales differently (Pemstein *et al.* 2020). The model uses bridge respondents, who rate multiple countries for many years, to calibrate estimates across countries. The model also uses lateral coding, which involves coding many countries for only 1 year, a technique which facilitates calibration across respondents. Finally, the model employs anchoring vignettes to further improve the estimates of respondent-level parameters and thus the concepts being measured. Anchoring vignettes are descriptions of hypothetical cases that provide all the necessary information to answer a given question. Since there is no contextual information in the vignettes and all respondents evaluate the same set of vignettes, they provide information about how individual respondents understand the scale and how they systematically diverge from each other in their coding.

In general, this discussion provides some reassurance that the V-Dem respondents are relatively unbiased and there is a comprehensive approach to mitigate DIF across respondents. The more general insight here, however, is that no respondent is free of bias and no respondent pool is free of DIF. High-quality measures come from producers who attempt to minimize biases, including DIF, and provide transparent information about how they do so.

24 We conducted a pilot study in 2010–2011 in part to estimate the minimum *N*-experts needed. In practice, *N* is often higher than five for a given country-year. Using the pattern of expert coverage as a starting point, Marquardt and Pemstein (2018) and Marquardt (2019) use simulation techniques to examine the robustness of V-Dem's expert aggregation methods and find that, except when experts are extremely biased or unreliable, these methods do a good job of recovering true values.

*Aggregation Model.* Many datasets, including V-Dem, offer low-level measures (indicators) that they combine into higher-level measures (indices). To assess the validity and reliability of the resulting higher-level measures, it is important to consider: (a) the choice of measures to aggregate and (b) the aggregation rules. There are no objective standards for selecting low-level measures for an index or developing aggregation rules. When a researcher evaluates these decisions as part of a measure quality assessment, the most important consideration is the researcher's theory. Similar to the evaluation of content validity discussed above, a relatively high-quality index will be one with constituent indicators that capture all the dimensions of the theoretical construct, that is formed using aggregation rules that align with the researcher's theory regarding how these indicators interact and weight relative to each other.

In the case of the V-Dem corruption measures, the V-Dem dataset includes six corruption indicators. The first four capture bribery in the executive (**v2exbribe**), in the legislature (**v2lgcrrpt**), in the judiciary (**v2jucorrdc**), and in the public sector (**v2excrptps**). The last two capture embezzlement in the executive (**v2exembez**) and in the public sector (**v2exthftps**). V-Dem aggregates these indicators into the V-Dem Corruption Index using a two-stage approach. In the first stage of aggregation, V-Dem fits a Bayesian[25] factor analysis model to the two indicators capturing executive branch corruption (**v2exbribe** and **v2exembez**) and, separately, to the two indicators capturing public sector corruption (**v2excrptps** and **v2exthftps**). In the second stage, to construct the high-level V-Dem Corruption Index (**v2x_corr**), V-Dem averages the executive corruption index (**v2x_execorr**), the public sector corruption index (**v2x_pubcorr**), the indicator for legislative corruption (**v2lgcrrpt**), and the indicator for judicial corruption (**v2jucorrdc**). In other words, V-Dem weighs each of these four spheres of government equally in the resulting index.[26]

From a comparative standpoint, both the WGI and CPI choose indicators for aggregation to minimize missingness (Hawken and Munck 2009a). V-Dem does not have such a constraint, as the level of missingness does not vary greatly from one indicator to another.

*Coverage Across Countries and Time.* It is important to consider potential biases introduced by limited geographic or temporal coverage of a measure. Particularly with sensitive topics, such as corruption, choosing cases can introduce selection bias. Thus, maximizing case coverage also improves measurement validity.

The V-Dem corruption measures perform well on the question of coverage. V-Dem covers nearly all countries, avoiding the bias in datasets of only a subset of countries (those easiest to code or those for which respondents are readily available).[27] By asking the same questions of each respondent for each country-year, V-Dem allows over-time and cross-country comparisons of corruption levels in the world back to 1900.

The quality of V-Dem corruption measures for analysis across space and time is one of their key strengths. Alternative measures of corruption are typically taken at the country level, where comparisons across countries often come at the expense of comparisons over time (Arndt and Oman 2006; Galtung 2006; Knack 2007). For example, WGI is calculated such that the global average is the same every year; changes in the level of corruption within a country are not revealed unless the change is so great as to move it up or down in the comparative rankings (Lambsdorff 2007). Kaufmann and Kraay (2002) estimate that half the variance in WGI over time is the product

25 The Bayesian approach allows for the incorporation of estimation uncertainty into the resulting indices, providing users with estimates of index reliability. Specifically, V-Dem uses the method of composition (Tanner 1993).

26 In our substantive work with the V-Dem Corruption Index (McMann *et al.* 2020), we explore different aggregation approaches: specifically, combining all indicators in one stage instead of two; and using principal components analysis instead of Bayesian factor analysis. The results are invariably robust to any aggregation approach, which gives us reason to believe the aggregation decisions discussed in this section are more theoretically meaningful than substantively important.

27 The countries that this version of V-Dem omits are microstates. Among the countries covered by V-Dem, there is only one case of missing data on the V-Dem Corruption Index: East Timor prior to independence.

of changes in the sources and coding rules used, rather than actual changes in corruption levels. Treisman (2007) notes that CPI's aggregation procedures and data sources have changed over time, which may mean substantive results using CPI data are release-dependent.

*Analyzing Respondent Disagreement and Biases.* Conducting an analysis of respondent disagreement is another tool to assess the validity and reliability of the data generation process.[28] Unlike Steenbergen and Marks (2007) and Martinez i Coma and van (2015), who primarily compare ratings across respondents as a test of validity, we argue that inter-respondent disagreement provides insight into both validity *and* reliability. Clearly, a measure is more reliable when inter-respondent disagreement is low. Inter-respondent disagreement can also be seen as a measure of validity if one is willing to assume that multiple respondents are unlikely to exhibit identical biases.[29] When respondent or country characteristics predict disagreement systematically, this suggests potential sources of bias in the data.

For the V-Dem corruption measures, We assess systematic determinants of respondent disagreement in a regression framework in Table 3, where the dependent variable is the standard deviation of measurement model-adjusted ratings among respondents for each country and year.[30] To ease interpretation of the results in Table 3, note that the mean of the standard deviation of model-adjusted ratings across respondents is 0.158 (the median is 0.148) and the standard deviation is 0.112. In interpreting the coefficient of –0.039 for freedom of expression, for example, it implies that a unit shift in freedom of expression - tantamount to a comparison between North Korea and Switzerland (the index being scaled to vary from 0 to 1)—on average implies a 0.039 decrease in respondent disagreement, which amounts to around a third of the variation.

Controlling for the number of respondents, we find that respondent disagreement varies predictably.[31] For three of the six V-Dem corruption measures (Supplementary Appendix Table S6) and in the pooled model (Table 3), respondent disagreement is statistically significantly lower in countries with widespread freedom of expression, indicating that limited access to information influences respondents' evaluations. The quadratic term for the level of corruption is negative and statistically significant, indicating that the greatest disagreement occurs in countries with the lowest levels of corruption. The time variable (Century) produces a more mixed pattern across the disaggregated corruption measures (Supplementary Appendix Table S6), and the coefficient for Century in the pooled model is statistically insignificant. This result qualifies the notion that the distant past is harder to code than the present. Overall, we conclude that respondent disagreement is not critically high and that it varies with the level of information and the level of corruption in a way that might be expected.

We next evaluate the quality of the V-Dem corruption measures by testing for a potential form of bias that Bollen and Paxton (2000) call "situational closeness," or the idea that "judges will be influenced by how situationally and personally similar a country is to them" (p. 72). In other words, we test whether there is ideological bias among respondents geared toward certain types of countries. The V-Dem postsurvey questionnaire allows us to evaluate three such respondent-country characteristic interactions: whether respondents who support free markets provide different corruption ratings for free trade economies (using a measure for trade openness from the Correlates of War project); whether those who support the principles of electoral democracy tend to provide different corruption ratings for electoral democracies; and whether those who

28 This analysis relies on the use of multiple respondents and the availability of observable respondent-level covariates.
29 The plausibility of this assumption will vary across applications and requires careful consideration.
30 Measurement-model adjusted ratings are model-generated estimates of rater "perceptions" of a latent scores after adjusting for DIF. Our regression models therefore examine expert disagreement, net of DIF. We follow Johnson and Albert (1999) and Pemstein, Meserve, and Melton (2010), who explain how to estimate these scores in detail.
31 Table 3 displays results from a pooled model including all V-Dem corruption measures. Supplementary Appendix Table S6 provides the results when modeling disagreement separately for the six corruption measures.

**Table 3.** Predicting respondent disagreement.

| | DV: Respondent disagreement |
|---|---|
| Century | −0.001 |
| | (0.007) |
| Freedom of expression | −0.039 |
| | (0.009) |
| Level of corruption | −0.003 |
| | (0.003) |
| Level of corruption $^2$ | −0.042 |
| | (0.003) |
| Number of respondents | 0.001 |
| | (0.002) |
| Adjusted R-squared | 0.234 |
| No. Countries | 173 |
| No. Observations | 69939 |

Entries are regression coefficients, with standard errors, clustered on countries, in parentheses. Measure-fixed effects are included in the model but omitted from the table.

support the principles of liberal democracy tend to provide different corruption ratings for liberal democracies.

The results of the analysis considering how respondent and country characteristics might interact in Table 4 are again quite reassuring.[32] Unsurprisingly, respondents consider more "liberal" countries less corrupt. More importantly, respondents who strongly support this "liberal" principle do not code or perceive more liberal countries differently than respondents who do not exhibit such support. Respondents also consider more open economies less corrupt, but this has no effect on how free market ideological bias affects ratings. With no interactions being statistically significant, there seems to be no overall ideological bias or "situational closeness" introduced by the context of the country being coded.

Beyond the assessment of the V-Dem corruption measures, this section illustrates how researchers can use information provided about the data generation process and data generators (e.g., respondents) to analyze disagreement between data generators and determinants of generator biases. Often, as here, data generators will be respondents, but analysts can apply these techniques whenever data is generated by multiple sources. This analysis, in turn, suggests the types of cases, points in time, and research questions where a measure may provide higher or lower quality estimates, information that can shape substantive conclusions from data analysis.

## 4.3 Convergent Validity Assessment

Our final data quality assessment component asks: To what extent do the measures correspond to alternative data sources? First, we suggest conducting a traditional convergent validity analysis, visually and statistically comparing the correlation between the chosen measure and possible alternatives. Second, we recommend statistically examining the extent to which

32 The results presented here are for a pooled model of measurement model-adjusted ratings. Models for each individual measure, appear in Table A7 in the Supplementary Appendix.

**Table 4.** Predicting respondent ratings with respondent and country characteristics.

| | DV: Respondent ratings |
|---|---|
| Respondent supports free market | 0.019 |
| | (0.009) |
| Country openness to trade (rescaled) | 0.015 |
| | (0.005) |
| Respondent supports free market × | −0.001 |
| country openness to trade (rescaled) | (0.001) |
| Respondent supports electoral democracy | −0.032 |
| | (0.014) |
| Country electoral democracy level | −0.038 |
| | (0.155) |
| Respondent supports electoral democracy × | 0.041 |
| country electoral democracy level | (0.028) |
| Respondent supports liberal democracy | 0.015 |
| | (0.018) |
| Country liberal democracy level | 0.605 |
| | (0.144) |
| Respondent supports liberal democracy × | −0.023 |
| Country liberal democracy level | (0.025) |
| R-squared | 0.408 |
| No. Countries | 149 |
| No. Observations | 204684 |

Entries are regression coefficients, with standard errors, clustered on countries, in parentheses. Year-fixed effects, respondent characteristics, and measure-fixed effects are included in the model but omitted from the table.

observable aspects of the data generation process predict systematic divergence between the chosen measure and the alternatives. Finally, we recommend examining the convergence between the measure and original or pre-existing qualitative cases studies.

*Basic Quantitative Convergent Validity.* A typical convergent validity test aims to use statistical tools (e.g., correlation coefficients) to assess whether various measures appear, on aggregate, to tap into the same concept. However, a broader convergent validity assessment can identify a measure's relative strengths and limitations compared to other measures the researcher might choose. In terms of substantive research, the goal of this exercise is to answer the question: When might the findings of research be sensitive to using this measure instead of others?

Before embarking on this more detailed analysis of convergent validity, however, considering aggregate correlation coefficients is a useful first step. Since the measures most comparable to the V-Dem Corruption Index—WGI and CPI—explicitly discourage comparisons over time, we assess aggregate convergent validity on a year-by-year basis. In Figure 1, we present the association between the V-Dem Corruption Index and WGI and CPI for 1 year (2010), but the patterns remain similar across all years. Both pooled correlation coefficients are around 0.90: clear evidence of convergent validity. Divergence between V-Dem and WGI or CPI is particularly limited when considering the most corrupt countries. However, there are differences in how V-Dem compares

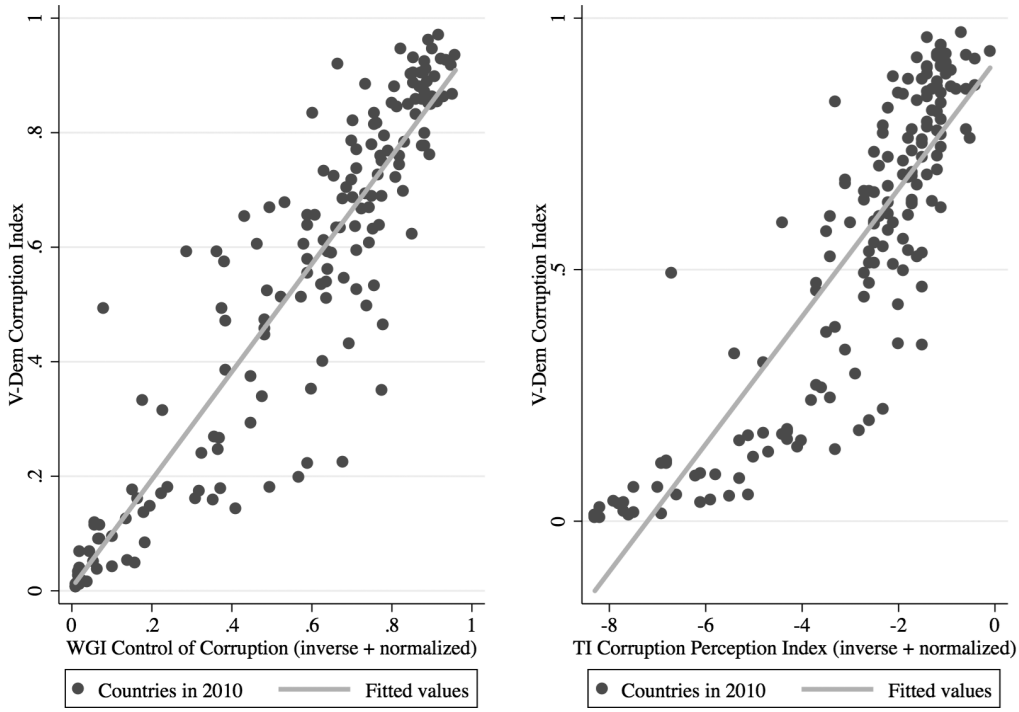



**Figure 1.** Comparing the V-Dem Corruption Index with the WGI and CPI Corruption Indices.

to WGI versus CPI. The deviations from WGI are more uniformly distributed over the range of the V-Dem Corruption Index, whereas the V-Dem Corruption Index is systematically lower than CPI for countries with a moderate level of corruption, and systematically higher for countries with extreme levels of corruption.

*Statistical Analysis of Measure Convergence.* Explaining patterns of convergence and divergence is as, or more, important as demonstrating strong correlations (Adcock and Collier 2001; Bowman, Lehoucq, and Mahoney 2005). As Hawken and Munck (2009a) note, "Consensus is not necessarily indicative of accuracy and the high correlations . . . by themselves do not establish validity" (p. 4). While one rarely has access to a "gold standard" against which to assess convergence, researchers can model systematic determinants of divergence. Therefore, the next step in our proposed convergent validity assessment is to identify the correlates of divergence and attempt to diagnose the cases where the use of one measure over another could be consequential.

In applying this analysis to the case of the V-Dem corruption indicators, we ask whether the composition of V-Dem respondents per country and year, measured with average respondent characteristics, affects the tendency for V-Dem to deviate from WGI.[33] In other words, what are the respondent composition predictors of the absolute residuals in Figure 1 (pooled across all years)?

We present the results of this analysis in Table 5.[34] The gender composition coefficient is positive and statistically significant; the larger the share of female V-Dem respondents, the larger the absolute difference between V-Dem and WGI. Moreover, WGI and V-Dem diverge less when V-Dem relies more heavily on PhD-holding respondents. Overall, however, the pattern is clear: there are few systematic predictors of the deviations between WGI and V-Dem Corruption Index.

33 As argued by Huckfeldt and Sprague (1993), the only way of avoiding both the ecological fallacy of making individual-level inferences from aggregated measures, and the "individual-level fallacy" of making aggregate-level inferences from individual-level measures, is to incorporate both individual- and aggregate (average) characteristics on the right-hand side of the equation.

34 These are the results for the V-Dem Corruption Index. We provide results for the individual corruption measures in Table S8 in the Supplementary Appendix.

**Table 5.** Explaining deviations from WGI control of corruption index with aggregate respondent characteristics

| | DV: Absolute residual WGI vs. V-Dem |
|---|---|
| Share female respondents | 0.052 |
| | (0.025) |
| Average age of respondents (in decades) | −0.017 |
| | (0.085) |
| Average age of respondents (in decades) $^2$ | 0.002 |
| | (0.009) |
| Share respondents with PhD | −0.084 |
| | (0.023) |
| Share respondents employed by government | −0.068 |
| | (0.042) |
| Share respondents born in country | −0.009 |
| | (0.028) |
| Share respondents residing in country | 0.010 |
| | (0.027) |
| Average support for free market among respondents | 0.006 |
| | (0.010) |
| Average support for electoral democracy among respondents | 0.001 |
| | (0.015) |
| Average support for liberal democracy among respondents | −0.005 |
| | (0.013) |
| Mean respondent discrimination ( $\beta$) | 0.004 |
| | (0.004) |
| Respondent disagreement | 0.345 |
| | (0.043) |
| Number of respondents | −0.008 |
| | (0.002) |
| R-squared | 0.099 |
| No. Countries | 164 |
| No. Observations | 54,235 |

Entries are regression coefficients, with standard errors, clustered on countries, in parentheses. The dependent variable is the absolute residuals from regressing each V-Dem measure on WGI. Year-fixed effects, respondent characteristics, and measure-fixed effects are included in the model but omitted from the table.

Predicting convergence between a measure and its alternative(s) as we have modeled it here relies on the availability of data about respondent traits. This information is not always available. However, there may be other available information about cases, respondents, or data sources to facilitate unpacking patterns in convergence and divergence. Our aim is not to prescribe the set of predictors, but rather to demonstrate the kind of insight that can be obtained by a detailed convergent validity assessment.

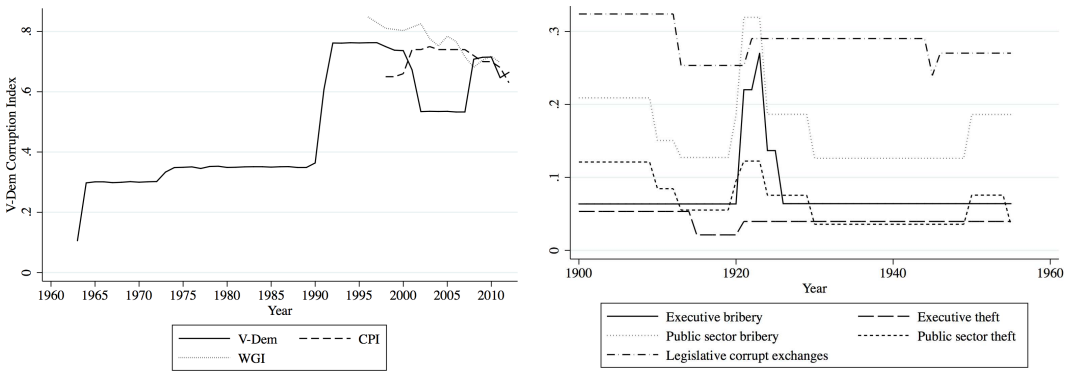
*Convergent Validity Assessment with Case Studies.* Even when a chosen measure generally converges with alternative measures, it is useful to examine convergent validity in the context of specific examples. Researchers can use case studies to scrutinize particularly salient examples of divergence and examine how the information presented by quantitative measures corresponds to actual cases. The case studies are labor-intensive, so it is important to select cases purposefully to assess the measures in question. It is also preferable to perform the exercise "blind," meaning that one reads the available case material about corruption in a country and writes her case study before seeing the measures (which possibly necessitates the involvement of a collaborator who has not yet examined the data).

In applying this component of the approach to assess the V-Dem corruption measures, an experienced research assistant completed the blind case studies prior to ever reviewing the V-Dem corruption data (or any of the alternative corruption measures) for the cases.[35] We selected four countries for case studies to evaluate V-Dem. We chose Georgia and Zambia, from their points of independence to the present, because V-Dem corruption measures for these countries differ significantly from those produced by alternative corruption measures, specifically WGI and CPI. We also selected historical Spain and the United States to check the quality of the V-Dem Corruption Index going back in time. We examine both countries from 1900 and stop with 1988 for Spain and 1955 for the United States to capture periods of dramatic change. In this case, we do not compare the V-Dem measures of corruption with alternative corruption measures because there are no alternative corruption measures with this level of historical coverage. Due to space constraints, we present Zambia and the United States here and Georgia and Spain in the Supplementary Appendix.

For Zambia, the contrast among the measures is substantial, as Panel (a) of Figure 2 demonstrates. For a period, V-Dem and CPI move in opposite directions, with V-Dem showing a greater magnitude of change. V-Dem also differs from WGI, which depicts a relatively steady decline in corruption, whereas V-Dem shows more sudden shifts in corruption. Yet, the V-Dem measure matches published accounts of corruption in Zambia more closely than alternative corruption measures (Chikulo 2000; Szeftel 2000; Van Donge 2009; Mbao 2011). During Zambia's First and Second Republics, from independence in 1964 until 1990, corruption was pervasive in the country, according to published accounts. The relatively high score on the V-Dem scale reflects this. As the economy worsened in the early 1970s, civil servants increasingly turned to theft of state resources to augment their salaries: the V-Dem measure captures this increase. Since then, the increase in corruption can mainly be attributed to the informal practices of government elites. In the first years of the Third Republic, government officials used the privatization campaign to enrich themselves, according to published reports. Thick descriptions do not mention the small dip in the late 1990s that the V-Dem measure depicts (as does WGI, but not CPI). Otherwise, the case material and V-Dem measure move in lockstep for this era. The published accounts allude to a decline in corruption with the 2001 exit of President Frederick Chiluba and other officials who were implicated in theft of state resources. Corruption in the country then began to increase in 2008 with the election of new presidents in 2008 and 2012, according to those accounts. The V-Dem measure mirrors this pattern, except for showing a small drop in 2011, which the case material do not mention (but the other measures depict).

For the United States, both the V-Dem Corruption Index and its constituent measures align with the details provided in the case material, increasing our confidence in the V-Dem measures going back in time and demonstrating the utility of providing disaggregated measures of corruption in addition to a high-level measure (Benson, Maaranen, and Heslop 1978; Woodiwiss 1988; Reeves

35 To develop the case studies, a research assistant used scholarly articles, books, and intergovernmental and nongovernmental reports to describe the extent and nature of corruption generally and, where possible, in each branch of government and the public sector. The reports he used included thick descriptions from the World Bank but not their data sources that include quantitative corruption measures—WGI and BEEPS.




(a) Zambia: V-Dem Corruption Index, WGI, CPI (b) US: V-Dem Corruption Measures

**Figure 2.** Corruption over time: Zambia and the United States.

2000; Grossman 2003; Menes 2003). At the turn of the century, U.S. government bureaucrats stole state resources and exchanged state services for personal material gain. However, the Progressive Movement of the early 1900s discouraged and lessened this corruption. The V-Dem Corruption Index depicts this decrease in corruption. Corruption increased in 1921 with the administration of Warren Harding, fueled by Prohibition-era bribes from liquor smugglers, and declined upon his death in 1923. The V-Dem Corruption Index approximates this account well. The measure shows a small increase in 1920 but then, aligning with case material about the United States, a significant increase in 1921 followed by a dramatic decrease in 1924.

The value of the individual V-Dem measures becomes especially apparent with the Harding administration. The measures diverge, reflecting patterns described in the case material. As depicted in Panel (b) of Figure 2, there is an increase in executive and public sector bribery, and—to a lesser extent—embezzlement. However, this period is not characterized by a dramatic increase in legislative corruption, as is also discussed in the case material.[36] Legislative corruption, such as the awarding of military contracts in exchange for bribes, was central to corruption during World War II and sustained it during this period. With the end of the war and prosecutions for the schemes, these opportunities subsided. The V-Dem legislative corruption measures capture the dip in corruption at the end of the war in 1945. The individual V-Dem measures also match the published accounts of increased corruption by bureaucrats in numerous agencies during the Truman administration. The V-Dem measure shows that corruption increased during the Truman administration (1945–1953): corruption levels jump in 1950 and drop in 1955. Trends for individual V-Dem measures support the scholars' accounts, showing that public sector bribery and theft, rather than executive or legislative corruption, were driving this shift (panel (b), Figure 2). Overall, the V-Dem measures present a picture similar to qualitative case material regarding corruption in the United States historically.

In general, the analysis of Georgia, Zambia, Spain and the United States that we present here and in the Supplementary Appendix suggests that the V-Dem corruption measures generally converge with available case material. The six corruption measures capturing different forms of corruption in different sectors of government seem to converge and diverge in line with published reports on the cases. More generally, this application illustrates the value of using qualitative material to validate quantitative measures. In the case where the measure being assessed does not have a quantitative alternative, comparing it to case studies facilitates a form of convergent validity assessment that can still yield valuable information about the measures' areas of strength and limitation.

36 The judicial corruption measure is not included in this analysis of the United States because it does not vary during this period, although it does in later eras.

## 5 Discussion

As Herrera and Kapur (2007) wrote, "Inattentiveness to data quality is, unfortunately, business as usual in political science" (p. 366). To address this issue, we synthesize a set of complementary, flexible, practical, and methodologically diverse tools for assessing data quality into a comprehensive approach. This holistic approach updates early guidance that balanced careful attention to construct validity with the application of empirical tools for assessing both validity and reliability (Zeller and Carmines 1980).

Our proposed approach includes three components: a content validity assessment; a data generation process assessment; and a convergent validity assessment. Each component involves the use of qualitative and quantitative tools, developed by other scholars and synthesized by us. In addition, we innovate over existing validity assessment in three ways. First, our assessment includes a road map for evaluating the validity and reliability of the data generation process as a signal of resulting data quality. Second, it includes an analysis of the predictors of inter-respondent disagreement and intra-respondent biases to assess both reliability and validity. Third, we propose a qualitative case assessment using blind coding as one piece of a convergent validity assessment.

In a world of limited data, it is often tempting to conduct validation tests, mention they have been done in a footnote of a paper, and then say no more about it. The literature on validation has provided scant guidance about what to do with the findings of a validation exercise, nor how to use validation results to inform substantive research conclusions, beyond adopting or discarding a measure. Yet, validation exercises provide rich information about how strengths and limitations of a chosen measure might affect the findings of substantive research, or more specifically, the conditions under which substantive conclusions might be more or less robust. We therefore now provide five examples of how the findings of our data quality assessment approach applied to the V-Dem corruption measures might be incorporated by researchers conducting substantive analyses with these measures.

First, our content validity assessment reveals that V-Dem corruption measures are best suited to research on exchange-based, material corruption among public officials. The six low-level measures and the high-level corruption measure do not capture, or capture only minimally, other forms of corruption, including revolving door, vote-buying, and nepotism. Substantive research about these forms of corruption should not rely on the V-Dem corruption measures for data.

Second, our data generation process assessment underscored that V-Dem respondents and V-Dem management each represent diverse backgrounds. This finding suggests that the V-Dem corruption measures might be particularly useful when conducting substantive research in which the theory is most salient in non-Western societies or researchers expect heterogeneous effects across contexts.

Third, also from the data generation process assessment, we learned that V-Dem inter-respondent disagreement for a country-year observation is inversely related to the level of freedom of expression and the level of corruption. This in turn means there will be more uncertainty in V-Dem Corruption Index estimates for countries with low freedom of expression or low levels of corruption. This uncertainty has the potential to diminish the robustness of results when testing theories pertaining to less free societies or relatively low-corruption contexts.[37]

Fourth, the data generation process assessment highlighted the relative value of using V-Dem measures for time-series, cross-sectional research on corruption. The consistency of the V-Dem coding procedures and aggregation procedures across all years will enable researchers to use the V-Dem Corruption Index to examine corruption dynamics over time. Similarly, V-Dem's use

37 In addition to the point estimates for each country-year observation, the V-Dem dataset includes the confidence intervals surrounding the point estimates. These can be incorporated into robustness checks to ascertain how sensitive findings are to variations in estimates within the confidence intervals.

of a sophisticated measurement model, bridge respondents, lateral respondents, and anchoring vignettes facilitates cross-country comparison. The extensive temporal and geographic coverage of the measures also enables time-series, cross-sectional research. Researchers more focused on a particular time period or set of countries may not highly value these relative strengths of the V-Dem corruption measures.

Fifth, our convergent validity findings about respondent characteristics indicate it may be useful, when using the V-Dem corruption measures, to conduct additional measurement validation specific to one's research project. We found that as the percentage of female or non-PhD respondents increases, so does the difference between the V-Dem Corruption Index and WGI. Because recruiting either women or those with PhDs might be correlated with another characteristic of a country that is under study, researchers using V-Dem measures of corruption may be over- or under-inflating findings compared to using alternative corruption measures like the WGI Control of Corruption Index. For that reason, researchers would be wise to examine correlations between female and PhD respondents with their variables of interest to understand how use of the V-Dem corruption measures may affect their findings.

These five points highlight how researchers might begin to think about mitigating concerns and utilizing strengths in working with the V-Dem corruption measures. More generally, this discussion offers an example of how the findings of a data quality assessment could inform substantive research. The overarching point is that any given measure will be more or less appropriate depending on the theoretical concepts under study, the expected relationship, and the set of cases pertinent to the research question. There are no optimally valid and perfectly reliable measures, and data consumers would be wise to diagnose, acknowledge, and mitigate strengths and limitations regarding data quality proactively and transparently.

## Funding

This work was supported by the Riksbankens Jubileumsfond (grant number M13-0559:1); the Knut & Alice Wallenberg Foundation (to S.L.); the Swedish Research Council (to S.L. and J.T.); the National Science Foundation (grant number SES-1423944 to D.P.); and the Wenner–Gren Foundation and the European University Institute's Fernand Braudel Senior Fellowship (to J.T.).

## Acknowledgments

The authors are grateful to Gerardo Munck and other participants in the 2015 V-Dem Internal Research Conference for their comments and to Talib Jabbar and Andrew Slivka for their research assistance.

## Conflicts of Interest

The authors of this manuscript have no conflicts of interests to disclose.

## Data Availability Statement

The replication materials for this paper can be found at McMann *et al*. (2021a; 2021b).

## Supplementary Material

For supplementary material accompanying this paper, please visit https://doi.org/10.1017/pan.2021.27.

## References

Adcock, R., and D. Collier. 2001. "Measurement Validity: A Shared Standard for Qualitative and Quantitative Research." *American Political Science Review* 95(3):529–546.

Arndt, C., and C. Oman. 2006. *Uses and Abuses of Governance Indicators*. Paris: Development Centre Studies, OECD Publishing.

Benson, G. C., S. A. Maaranen, and A. Heslop. 1978. *Political Corruption in America*. Lexington: D.C. Heath and Company.
Bolck, A., M. Croon, and J. Jagenaars. 2004. "Estimating Latent Structure Models with Categorical Variables: One-Step Versus Three-Step Estimators." *Political Analysis* 12(1):3–27.
Bollen, K. A. 1980. "Issues in the Comparative Measurement of Political Democracy." *American Sociological Review* 45:370–390.
Bollen, K. A. 1989. *Structural Equations with Latent Variables*. New York: Wiley.
Bollen, K. A. 1990. "Political Democracy: Conceptual and Measurement Traps." *Studies in Comparative International Development* 25:7–24.
Bollen, K. A., and P. Paxton. 2000. "Subjective Measures of Liberal Democracy." *Comparative Political Studies* 33(1):58–86.
Bowman, K., F. Lehoucq, and J. Mahoney. 2005. "Measuring Political Democracy Case Expertise, Data Adequacy, and Central America." *Comparative Political Studies* 38(8):939–970.
Campbell, D. T., and D. W. Fiske. 1959. "Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix." *Psychological Bulletin* 56(2):81.
Carmines, E. G., and R. A. Zeller. 1979. *Reliability and Validity Assessment*. Beverly Hills, CA: Sage.
Chikulo, B. C. 2000. "Corruption and Accumulation in Zambia." In *Corruption and Development in Africa: Lessons from Country Case Studies*, edited by K. R. Hope, Sr., and B. Chikulo, 161–182. London: Palgrave Macmillan.
Collier, D., J. LaPorte, and J. Seawright. 2012. "Putting Typologies to Work: Concept Formation, Measurement, and Analytic Rigor." *Political Research Quarterly* 65(1):217–232.
Collier, D., and S. Levitsky. 1997. "Democracy with Adjectives." *World Politics* 49(3):430–451.
Coppedge, M., et al. 2011. "Conceptualizing and Measuring Democracy: A New Approach." *Perspectives on Politics* 9(2):247–267.
Coppedge, M., et al. 2015a. V-Dem Country-Year Dataset v4. Varieties of Democracy (V-Dem) Project.
Coppedge, M., et al. 2020. *Varieties of Democracy: Measuring Two Centuries of Political Change*. New York: Cambridge University Press.
Coppedge, M., et al. 2017. *V-Dem Country-Year Dataset v7.1*. Varieties of Democracy (V-Dem) Project.
Coppedge, M., et al. 2015a. V-Dem Country-Year Dataset v4. Varieties of Democracy (V-Dem) Project.
Coppedge, M., et al. 2015b. Varieties of Democracy: Codebook v4. Varieties of Democracy (V-Dem) Project.
Dahlström, C., V. Lapuente, and J. Teorell. 2012. "Public Administration Around the World." In *Good Government. The Relevance of Political Science*, edited by S. Holmberg and B. Rothstein, 40–67. Cheltenham, UK: Edward Elgar.
Donchev, D., and G. Ujhelyi. 2014. "What Do Corruption Indices Measure?" *Economics & Politics* 26(2):309–331.
Donnelly, M. J., and G. Pop-Eleches. 2018. "Income Measures in Cross-National Surveys: Problems and Solutions." *Political Science Research and Methods* 6(2):355–363.
Fariss, C. J. 2014. "Respect for Human Rights has Improved Over Time: Modeling the Changing Standard of Accountability." *American Political Science Review* 108(2):297–318.
Galtung, F. 2006. "Measuring the Immeasurable: Boundaries and Functions of (Macro) Corruption Indices." In *Measuring Corruption*, edited by F. Galtung, and C. Sampford, 101–130. Aldershot, UK: Ashgate.
Gerring, J. 2012. *Social Science Methodology: A Unified Framework*, 2nd ed. Cambridge, UK: Cambridge University Press.
Gingerich, D. W. 2013. "Governance Indicators and the Level of Analysis Problem: Empirical Findings from South America." *British Journal of Political Science* 43(3):505–540.
Grossman, M. 2003. *Political Corruption in America: An Encyclopedia of Scandals, Power, and Greed*. Santa Barbara: ABC-CLIO.
Harrington, D. 2008. *Confirmatory Factor Analysis*. Oxford: Oxford University Press.
Hawken, A., and G. L. Munck. 2009a. "Do You Know Your Data? Measurement Validity in Corruption Research." *Working Paper, School of Public Policy, Pepperdine University*.
Hawken, A., and G. L. Munck. 2009b. "Measuring Corruption: A Critical Assessment and a Proposal." In *Perspectives on Corruption and Human Development*, vol. 1, edited by A. K. Rajivan and R. Gampat, 71–106. New Delhi, India: Macmillan India for UNDP.
Hayes, A. F., and K. Krippendorff. 2007. "Answering the call for a standard reliability measure for coding data." *Communication Methods and Measures* 1(1):77–89.
Herrera, Y. M., and D. Kapur. 2007. "Improving data quality: Actors, incentives, and capabilities." *Political Analysis* 15(4):365–386.
Huckfeldt, R., and J. Sprague. 1993. " Citizens, Contexts, and Politics." In *Political Science: The State of the Discipline II,* edited by A. W. Finifter, 281–303. Washington, DC: American Political Science Association.
Jeong, G.-H. 2018. "Measuring Foreign Policy Positions of Members of the US Congress." *Political Science Research and Methods* 6(1):181–196.
Johnson, V. E., and J. H. Albert. 1999. *Ordinal Data Modeling*. New York: Springer.
Kaufmann, D., and A. Kraay. 2002. "Growth Without Governance." *World Bank Policy Research Working Paper no. 2928*.

Kennedy, J. 2014. "International Crime Victims Survey." *The Encyclopedia of Criminology and Criminal Justice*.
Knack, S. 2007. "Measuring Corruption: A Critique of Indicators in Eastern Europe and Central Asia." *Journal of Public Policy* 27(3):255–291.
Lambsdorff, J. G. 2007. "The Methodology of the Corruption Perceptions Index 2007." *Transparency International (TI) and the University of Passau*.
Lindstaedt, R., S.-O. Proksch, and J. B. Slapin. 2016. "When Experts Disagree: Response Aggregation and Its Consequences in Expert Surveys." *Working paper*.
Marcus, G. E., W. R. Neuman, and M. B. MacKuen. 2017. "Measuring Emotional Response: Comparing Alternative Approaches to Measurement." *Political Science Research and Methods* 5(4):733–754.
Marquardt, K. L. 2019. "How and How Much Does Expert Error Matter? Implications for Quantitative Peace Research." *V-Dem Working Papers Series no. 84*.
Marquardt, K. L., and D. Pemstein. 2018. "IRT Models for Expert-Coded Panel Data." *Political Analysis* 26(4):431–456.
Martinez i Coma, F., and C. van Ham. 2015. "Can Experts Judge Elections? Testing the Validity of Expert Judgments for Measuring Election Integrity." *European Journal of Political Research* 54(2):305–325.
Mbao, M. 2011. "Prevention and Combating of Corruption in Zambia." *Comparative and International Law Journal of Southern Africa* 44(2):255–274.
McMann, K., D. Pemstein, B. Seim, J. Teorell, and S. Lindberg. 2021a. Replication Data for: Assessing Data Quality: An Approach and An Application. Version v1. https://doi.org/10.24433/CO.0269024.v1.
McMann, K., D. Pemstein, B. Seim, J. Teorell, and S. Lindberg. 2021b. Replication Data for: Assessing Data Quality: An Approach and An Application. https://doi.org/10.7910/DVN/BXV4AT, Harvard Dataverse, V1.
McMann, K., B. Seim, J. Teorell, and S. Lindberg. 2020. "Why Low Levels of Democracy Promote Corruption and High Levels Diminish It." *Political Research Quarterly* 73(4):893–907.
Menes, R. 2003. "Corruption in Cities: Graft and Politics in American Cities at the Turn of the Twentieth Century." *NBER Working Paper no. 9990*.
Mislevy, R. 1991. "Randomization-Based Inference about Latent Variables from Complex Samples." *Psychometrika* 56(2):177–196.
Morin-Chassé, A., D. Bol, L. B. Stephenson, and S. Labbé St-Vincent. 2017. "How to Survey About Electoral Turnout? The Efficacy of the Face-Saving Response Items in 19 Different Contexts." *Political Science Research and Methods* 5(3):575–584.
Mudde, C., and A. Schedler. 2010. "Introduction: Rational Data Choice." *Political Research Quarterly* 63(2):410–416.
Munck, G. L., and J. Verkuilen. 2002. "Conceptualizing and measuring democracy: Evaluating alternative indices." *Comparative Political Studies* 35(1):5–34.
Pemstein, D., et al. 2020. "The V-Dem Measurement Model: Latent Variable Analysis for Cross-National and Cross-Temporal Expert-Coded Data." *V-Dem Working Paper No. 21, 5th edition*.
Pemstein, D., K. Marquardt, E. Tzelgov, Y.-t. Wang, and F. Miri. 2015. "Latent Variable Models for the Varieties of Democracy Project." *Varieties of Democracy Institute Working Paper Series no. 21*.
Pemstein, D., S. A. Meserve, and J. Melton. 2010. "Democratic Compromise: A Latent Variable Analysis of Ten Measures of Regime Type." *Political Analysis* 18(4):426–449.
Reeves, T. C. 2000. *Twentieth-Century America: A Brief History*. New York: Oxford University Press.
Reise, S. P., K. F. Widaman, and R. H. Pugh. 1993. "Confirmatory Factor Analysis and Item Response Theory: Two Approaches for Exploring Measurement Invariance." *Psychological Bulletin* 114(3):552.
Reuning, K., M. R. Kenwick, and C. J. Fariss. 2019. "Exploring the Dynamics of Latent Variable Models." *Political Analysis* 27(4):503–517.
Rose-Ackerman, S. 1999. *Corruption and Government: Causes, Consequences, and Reform*. Cambridge: Cambridge University Press.
Sartori, G. 1970. "Concept Misformation in Comparative Politics." *American Political Science Review* 64(4):1033–1053.
Schedler, A. 2012. "Judgment and Measurement in Political Science." *Perspectives on Politics* 10(1):22–36.
Schnakenberg, K. E., and C. J. Fariss. 2014. "Dynamic Patterns of Human Rights Practices." *Political Science Research and Methods* 2(1):1–31.
Seawright, J., and D. Collier. 2014. "Rival Strategies of Validation: Tools for Evaluating Measures of Democracy." *Comparative Political Studies* 47(1):111–138.
Shadish, W., T. D. Cook, and D. T. Campbell. 2002. *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton Mifflin.
Steenbergen, M. R., and G. Marks. 2007. "Evaluating expert judgments." *European Journal of Political Research* 46(3):347–366.
Szeftel, M. 2000. "Eat With Us: Managing Corruption and Patronage Under Zambia's Three Republics, 1964–99." *Journal of Contemporary African Studies* 18(2):207–224.
Tanner, M. A. 1993. *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions*, 2nd Edn. New York, NY: Springer Verlag.

Thomas, M. A. 2010. "What Do the Worldwide Governance Indicators Measure?" *European Journal of Development Research* 22(1):31–54.
Treisman, D. 2000. "The Causes of Corruption: A Cross-National Study." *Journal of Public Economics* 76(3):399–457.
Treisman, D. 2007. "What Have We Learned About the Causes of Corruption from Ten Years of Cross-National Empirical Research?" *Annual Review of Political Science* 10:211–244.
Van Donge, J. K. 2009. "The Plundering of Zambian Resources by Frederick Chiluba and His Friends: A Case Study of the Interaction between National Politics and the International Drive Towards Good Governance." *African Affairs* 108(430):69–90.
Woodiwiss, M. 1988. *Crimes, Crusades, and Corruption: Prohibitions in the United States, 1900–1987*. Lanham, MD: Rowman & Littlefield Publishers.
Zeller, R. A., and E. G. Carmines. 1980. *Measurement in the Social Sciences: The Link between Theory and Data*. Cambridge: Cambridge University Press.