
# Strategies of Validation: Assessing the Varieties of Democracy Corruption Data

Kelly McMann, Daniel Pemstein,
Brigitte Seim, Jan Teorell,
and Staffan I. Lindberg

February 2016

Working Paper
SERIES 2016:23
THE VARIETIES OF DEMOCRACY INSTITUTE

UNIVERSITY OF GOTHENBURG
DEPT OF POLITICAL SCIENCE

**Varieties of Democracy (V-Dem)** is a new approach to the conceptualization and measurement of democracy. It is co-hosted by the University of Gothenburg and University of Notre Dame. With a V-Dem Institute at University of Gothenburg that comprises almost ten staff members, and a project team across the world with four Principal Investigators, fifteen Project Managers, 30+ Regional Managers, 170 Country Coordinators, Research Assistants, and 2,500 Country Experts, the V-Dem project is one of the largest-ever social science research-oriented data collection programs.

Please address comments and/or queries for information to:

V-Dem Institute

Department of Political Science

University of Gothenburg

Sprängkullsgatan 19, PO Box 711

SE 40530 Gothenburg

Sweden

E-mail: contact@v-dem.net

V-Dem Working Papers are available in electronic format at www.v-dem.net.

# Strategies of Validation:
## Assessing The Varieties of Democracy Corruption Data

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

[1]Political Science, Case Western Reserve University, Cleveland, Ohio 44106, USA
[2]Political Science, North Dakota State University, Fargo, ND 58102, USA
[3]Public Policy, University of North Carolina, Chapel Hill, NC, USA
[4]Political Science, Lund University, Lund, 221 00, Sweden
[5]V-Dem Institute, Political Science, University of Gothenburg, Gothenburg, 413 27, Sweden

**Abstract**

Social scientists face the challenge of determining whether their data are valid, yet they lack practical guidance about how to do so. Existing publications on data validation provide mostly abstract information for creating one's own dataset or establishing that an existing one is adequate. Further, they tend to pit validation techniques against each other, rather than explain how to combine multiple approaches. By contrast, this paper provides a practical guide to data validation in which tools are used in a complementary fashion to identify the strengths and weaknesses of a dataset and thus reveal how it can most effectively be used. We advocate for three approaches, each incorporating multiple tools: 1) assessing content validity through an examination of the resonance, domain, differentiation, fecundity, and consistency of the measure; 2) evaluating data generation validity through an investigation of dataset management structure, data sources, coding procedures, aggregation methods, and geographic and temporal coverage; and 3) assessing convergent validity using case studies and empirical comparisons among coders and among measures. We apply our method to corruption measures from a new dataset, Varieties of Democracy. We show that the data are generally valid and we emphasize that a particular strength of the dataset is its capacity for analysis across countries and over time. These corruption measures represent a significant contribution to the field because, although research questions have focused on geographic differences and temporal trends, other corruption datasets have not been designed for this type of analysis.

**Acknowledgments:** The authors are grateful to Nancy Bermeo, Ellen Lust, Gerardo Munck, Andreas Schedler and V-Dem colleagues for their comments on an earlier version of this paper and to Talib Jabbar and Andrew Slivka for their research assistance. This research project was supported by Riksbankens Jubileumsfond, grant M13-0559:1, PI: Staffan I. Lindberg, V-Dem Institute, University of Gothenburg, Sweden; by the Knut & Alice Wallenberg Foundation to Wallenberg Academy Fellow Staffan I. Lindberg, V-Dem Institute, University of Gothenburg, Sweden; by the Swedish Research Council, PI: Staffan I. Lindberg, V-Dem Institute, University of Gothenburg, Sweden & Jan Teorell, Department of Political Science, Lund University, Sweden; and by the National Science Foundation under Grant No. SES-1423944, PI: Daniel Pemstein. Jan Teorell also wishes to acknowledge support from the Wenner-Gren Foundation. We performed simulations and other computational tasks using resources provided by the Notre Dame Center for Research Computing (CRC) through the High Performance Computing section, and by the Swedish National Infrastructure for Computing (SNIC) at the National Supercomputer Centre in Sweden. We specifically acknowledge the assistance of In-Saeng Suh at CRC and Johan Raber at SNIC in facilitating our use of their respective systems.

# 1 Introduction

Most political science researchers explicitly or implicitly deal with data validity concerns in their work. Yet, there is still no cohesive answer to the question: "How do I know whether my data are valid?" While methods of data analysis have advanced substantially over the last decade, few scholars rigorously examine data validity and reliability prior to data analysis. As Herrera and Kapur [2007] wrote,"Inattentiveness to data quality is, unfortunately, business as usual in political science" (p. 366).

Past validation literature has made valuable contributions, but there is room for improvement. Validation advice typically focuses on satisficing standards: What does a researcher have to do to show a dataset is "valid enough" or "reliable enough"? Many of these approaches lack practical tools for evaluating the strengths and weaknesses of a dataset, instead relying on the ability to show that a particular dataset is simply better than the next best alternative. One reason for this inherent competition may be that "the truth" - the ultimate rubric against which to judge dataset validity - is unknowable, meaning validity assessments often rely more on a scholar's ability to argue than a dataset's alignment with reality. Moreover, some scholars seem to pit one validity approach against another rather than allowing approaches to serve as robustness checks for each other. Finally, few researchers take what they learn in assessing the dataset and use it when analyzing the dataset; the assessment serves to put a rubber stamp on the dataset, but rarely does a researcher draw on the strengths and weaknesses of a dataset in analysis. Yet, considering the majority of scholars use pre-existing datasets in their research, it is vitally important for us to understand how to responsibly use these datasets: how to diagnose dataset strengths and limitations, map their implications in empirical work, and implement mitigation strategies to address them.

To address these gaps in the validity literature, this paper proposes a set of complementary, flexible, practical, and methodologically rigorous tools for assessing dataset validity and reliability. Rather than recommending use of one tool over another, we advocate a comprehensive approach to assessment. We draw heavily on Schedler [2012], who stated that, "For measurement to be scientific, it must be grounded in shared concepts, shared realities, and shared rules of translation" (p. 21). The tools we propose assess the degree to which a dataset aligns with shared concepts (content validation), shared rules of translation (data generation validation), and shared realities (convergent validation). Collectively, the assessment tools proposed here are intended to help scholars more thoroughly and transparently evaluate both the process of generating data and the data themselves.

We use these tools to evaluate a new dataset on corruption, the Varieties of Democracy (V-Dem) corruption data.[1] The V-Dem corruption data cover all countries of the world, extend from 1900 to the

[1]The analysis completed in this paper used v4 of the V-Dem dataset.

present, and are based on expert surveys. Our purpose here is twofold: to illustrate how these tools might be used; and to assess the validity and reliability of the V-Dem data so that other scholars can use them more effectively. These purposes are well-suited to each other because corruption is particularly difficult to measure: the definition commonly used in the literature - the use of public office for private gain - is challenging to translate into measures (operationalize), and people's sensitivity about corruption hinders data collection.

Our tests for content, data generation, and convergent validity provide considerable evidence for the quality of the V-Dem corruption measures. We do find data from certain country-years, where the V-Dem coders disagree the most, less reliable and thus caution others to proceed carefully when using those data. Our analysis also reveals that female coders rate countries, on average, as more corrupt on a few of the V-Dem indicators. We encourage users of the dataset to consider two different implications of this finding. First, V-Dem data may reflect more corruption than other datasets, all else equal, especially for those country-years with more female coders. Second, to the extent that coder representativeness is inherently valuable, this is a strength of the V-Dem dataset. Finally, applying our data validity tools to V-Dem corruption measures emphasizes their validity for analyses across countries and over time. This is a significant strength considering that these analyses are common in corruption research, yet many of the existing datasets have not been designed to undertake this type of work.

In sum, the paper provides a practical guide to data validation, demonstrates the value of these tools as applied to the V-Dem corruption measures, and identifies weakness and strengths of the V-Dem data to facilitate their use by others. The next section examines the existing literature on data validation. We then acquaint the reader with our proposed approach and tools and the logic behind them. The section that follows introduces the V-Dem measures. We then apply our tools to those measures. The conclusion summarizes our data validation approach and findings about the V-Dem indicators.

## 2 Past Approaches to Validating Measures

How do I know whether my data are valid? In response to this question, many scholars have helpfully prescribed what valid data *should* look like. Yet, the standards in this literature are often highly abstract, which makes mapping them to real-world datasets challenging. Two of these works - Coppedge [2012] and Gerring [2001] - are designed to orient students of social science to the research process. Both of these books usefully examine criteria that one should employ when conceptualizing a theory and then translating that theory into measurable indicators. Similarly, Adcock [2001] walks through the steps to form a systemized concept and establish equivalence in measurement across domains of observation. The end of the article provides considerable detail, discussing content validation, convergent/discriminant

validation, and nomological/construct validation. All three works stop short, however, of providing a practical guide for data validation. Coppedge [2012] and Gerring [2001] provide example situations that are more simple and straightforward than what we typically encounter in our research. Adcock [2001] does not describe how to assess the types of validity he considers, focusing instead on isolated examples of what to emulate or avoid.

The literature has also tended to treat approaches to validation as competing camps, rather than using them as complements of one another. This is exemplified by Seawright and Collier [2014]. They usefully build on Collier's past work by posing four alternative approaches to validation: levels-of-measurement tradition, structural-equation modeling with latent variables, pragmatic, and case-based. Yet, as implied by the use of the word "rival" in the title of the article, this piece argues that these approaches all have strengths and weaknesses and that researchers should choose their approach carefully, rather than integrate them.

An additional oversight in the literature is that many scholars approach data validity challenges as if all researchers are building datasets from scratch. If they are examined at all, existing datasets are typically torn down without advice about how to proceed with conducting research. The implication of much of the literature is that a slightly less valid measure is as bad as a completely invalid measure: neither is worth using. Only the dataset that wins out as the most valid should remain in circulation: a "survival of the fittest" approach to dataset maintenance. There is very little attention to how to mitigate, or at least acknowledge, inherited problems such as poor mapping from a construct to a measure or obvious sources of measurement error. This is true of even one of the more comprehensive and nuanced analyses of data validation: Herrera and Kapur [2007] approach data collection "as an operation performed by *data actors* in a supply chain" (p. 366). They delineate these actors, their incentives, and their capabilities. They urge scholars to focus on data validity, coverage, and accuracy, and they offer several examples of datasets that have failed on these dimensions. But, they fall short of telling the reader how to assess existing datasets for their strengths and weaknesses and incorporate this information into their use (also see Mudde and Schedler [2010] and the mini-symposium in Political Research Quarterly).

While we acknowledge these weaknesses in the data validation literature, we also draw ideas from it in order to develop our own approach and tools below: concepts from Adcock, Gerring, and Herrera and Kapur, among others, play key roles. Likewise, we build on works that have examined the validity of a specific measure or set of measures. For example, our approach is informed by the large literature on the validity of democracy measures. Contributors to this literature have emphasized the importance of aligning measures with the higher-level conceptualization [Munck and Verkuilen, 2002, Bowman et al., 2005, Coppedge et al., 2011] and considering differences in coverage, the use of different sources, and the

use of different scales across measures [Bowman et al., 2005]. These works have also highlighted how critical transparency is. If a data source fails to transparently publicize its coding rules, aggregation choices, and measurement model details, assessing data validity is challenging [Coppedge et al., 2011, Munck and Verkuilen, 2002]. Works on democracy, as well as human rights, also provide useful insight for us into the validity of aggregation procedures, as these measures are typically indices constructed from lower-level indicators [Munck and Verkuilen, 2002, Bowman et al., 2005, Coppedge et al., 2011, Fariss, 2014, Pemstein et al., 2010]. To develop tools for convergent validity, we draw heavily on the work of Steenbergen and Marks [2007] and Martinez i Coma and van Ham [2015], who represent two additional validation literatures, those on party positions and election integrity, respectively. Finally, we use insights from the literature on the validity of corruption data, both because we apply our approach to corruption measures but also because the literature raises issues about data validation in general [Knack, 2007, Treisman, 2007, Knack, 2007, Hawken and Munck, 2009a,b, Galtung, 2006].

# 3 Data Validity Evaluation Approach and Tools

Responding to gaps in the existing literature, we provide a data validation strategy that offers practical, complementary tools useful to identifying strengths and weaknesses of datasets and employing that information when using those datasets. As a starting point, we define validity as the absence of systematic measurement error and reliability as the absence of unsystematic (or random) measurement error. A dataset that offers valid and reliable estimates of reality is preferable to a dataset that offers unreliable unbiased estimates or one that offers reliable biased estimates. And of course, the least usable dataset is one that unreliably offers a biased estimate of reality. Throughout this paper "validity" refers to this technical meaning provided here, as well as simply quality data that are free of both systematic and random measurement error (valid and reliable). Each context will make the meaning clear to the reader.

We recommend evaluating the level of error by considering a dataset's alignment with shared concepts, rules of translation, and realities, to use Schelder's terms [Schedler, 2012]. Past literature on data validity has focused particularly on comparing data to reality, in other words aiming to assess validity by evaluating alignment with shared realities. This is the approach employed in works such as Donchev and Ujhelyi [2014], Pemstein et al. [2010] and Fariss [2014]. The fundamental challenge in assessing validity in this manner is that we do not know the "true scores." If we had access to the "dataset of truth," we could evaluate existing datasets on their alignment with reality. In the absence of such a "dataset of truth," we are left comparing the datasets we have to case studies and existing datasetss.

In addition to assessing alignment with shared reality, we advocate assessing a dataset given its alignment with a shared concept and shared rules of translation. This latter piece has been largely

overlooked in the validation literature. We propose to evaluate whether the data generating process is unbiased and reliable. The assumption behind this approach is that an unbiased and reliable data generating process results in unbiased and reliable data. The strength of this approach is that it allows us to focus on something we *can* evaluate (i.e., the nature of a process) rather than something we *cannot* (i.e., a dataset's alignment with the truth). For example, though we cannot prove that a coder selected the "true" answer when coding Argentina's level of civil society freedoms in 1950, we can show that the process to recruit, engage, and synthesize data from that coder was unbiased and reliable.

Rather than assert that our approach to assessing validity should replace another, or that, in fact, any one approach is supreme, we put forth the idea that data validity should be assessed using all possible tools. These tools are complements, not substitutes.

Along with this, we encourage a departure from the "good enough" standards for validity. Validity is a continuous trait, not a binary one. On the spectrum of validity, there is no threshold a dataset must reach in order to be put into use. All datasets offer strengths and weaknesses, and it is important to understand and acknowledge them. A discipline that rejects datasets that cannot meet a standard of validity will be one that "teaches to the test," encouraging the proliferation of datasets that look like each other, that use previously vetted measurement approaches, or that align with the most well-known cases. We undertake this validity assessment to generate information, not rubber stamp approval.

Bearing all of this in mind, we propose three categories of tools to assess the validity of a dataset. They are outlined in Table 1 and described in detail in section five, where we also apply them to the V-Dem measures. Prior to those tasks, we provide a brief introduction to the V-Dem measures.

| CRITERION | CATEGORY | GUIDING QUESTIONS | TECHNIQUES |
|---|---|---|---|
| Alignment with Shared Concepts | 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 resonance, domain, differentiation, fecundity, consistency of the measure. |
| | | To what extent is the measure useful to research? | Evaluate the causal utility of the measure. |
| Alignment with Shared Rules of Translation | Data Generation Validity Assessment | Does the data generating process introduce any biases or analytical issues? | Evaluate dataset management structure, data sources, coding procedures, aggregation methods, and geographic and temporal coverage for each question. |
| | | How does it compare to the data generating process of alternative measures? | |
| Alignment with Shared Realities | Convergent Validity Assessment | Does the measure accurately capture actual cases? | Evaluate data against original or existing case studies. |
| | | Where multiple coders exist, to what extent do they converge? | Evaluate extent of disagreement among coders, whether disagreement varies systematically with level of difficulty, and extent to which coder characteristics predict their responses. |
| | | To what extent do the data produced by the measure correlate with data produced by other measures of the construct, and are areas of low correlation thoroughly understood? | Evaluate strength of correlations, any outliers, and the implications of these findings. |

Table 1: Tools to Assess Data Validity

# 4 V-Dem Corruption Measures

The V-Dem corruption measure is an index of six indicators of different forms of corruption based on original data from country experts.[2] Data have been collected for all countries of the world from 1900 to 2012 with updates to the current year in-progress. Multiple coders, at least three fifths of whom are native to the countries in question, were used to code each country-year observation, and the coder recruitment procedures and coding procedures were consistent over time and across countries. An item response theory (IRT) measurement model was used to aggregate the experts' responses.

The exact language of the six corruption indicators appears in Table 2. Additional details about the V-Dem data, including how the V-Dem Corruption Index was developed, are provided below in the evaluation of their validity.

[2]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.

| INDICATOR | QUESTION | CATEGORIES |
|---|---|---|
| **v2exbribe** Executive bribery and corrupt exchanges | How routinely do members of the executive (the head of state, the head of government, and cabinet ministers), or their agents, grant favors in exchange for bribes, kickbacks, or other material inducements? | 0: It is routine and expected.<br>1: It happens more often than not in dealings with the executive.<br>2: It happens but is unpredictable: those dealing with the executive find it hard to predict when an inducement will be necessary.<br>3: It happens occasionally but is not expected.<br>4: It never, or hardly ever, happens. |
| **v2exembez** Executive embezzlement and theft | How often do members of the executive (the head of state, the head of government, and cabinet ministers), or their agents, steal, embezzle, or misappropriate public funds or other state resources for personal or family use? | 0: Constantly. Act as though all public resources were their personal or family property.<br>1: Often. Responsible stewards of selected public resources but treat the rest like personal property.<br>2: About half the time. About as likely to be responsible stewards of selected public resources as they are to treat them like personal property.<br>3: Occasionally. Responsible stewards of most public resources but treat selected others like personal property.<br>4: Never, or hardly ever. Almost always responsible stewards of public resources and keep them separate from personal or family property. |
| **v2excrptps** Public sector corrupt exchanges | How routinely do public sector employees grant favors in exchange for bribes, kickbacks, or other material inducements? | 0: Constantly. Act as though all public resources were their personal or family property.<br>1: Often. Responsible stewards of selected public resources but treat the rest like personal property.<br>2: About half the time. About as likely to be responsible stewards of selected public resources as they are to treat them like personal property.<br>3: Occasionally. Responsible stewards of most public resources but treat selected others like personal property.<br>4: Never, or hardly ever. Almost always responsible stewards of public resources and keep them separate from personal or family property. |
| **v2exthftps** Public sector theft | How often do public sector employees steal, embezzle, or misappropriate public funds or other state resources for personal or family use? | 0: Constantly. Act as though all public resources were their personal/family property.<br>1: Often. Responsible stewards of selected public resources but treat the rest like personal property.<br>2: About half the time. About as likely to be responsible stewards of selected public resources as they are to treat them like personal property.<br>3: Occasionally. Responsible stewards of most public resources but treat selected others like personal property.<br>4: Never, or hardly ever. Almost always responsible stewards of public resources and keep them separate from personal or family property. |
| **v2lgcrrpt** Legislature corrupt activities | Do members of the legislature abuse their position for financial gain? | 0: Commonly. Most legislators probably engage in these activities.<br>1: Often. Many legislators probably engage in these activities.<br>2: Sometimes. Some legislators probably engage in these activities.<br>3: Very occasionally. May be a few but the vast majority do not.<br>4: Never, or hardly ever. |
| **v2jucorrdc** Judicial corruption decision | How often do individuals or businesses make undocumented extra payments or bribes in order to speed up or delay the process or to obtain a favorable judicial decision? | 0: Always.<br>1: Usually.<br>2: About half of the time.<br>3: Not usually.<br>4: Never. |

Table 2: V-Dem Corruption Indicators

# 5 Demonstration of the Tools and Validation of V-Dem Corruption Measures

To what extent are the V-Dem measures of corruption valid? More specifically, to what extent do they have sufficiently low levels of systematic measurement error and sufficiently low levels of unsystematic measurement error? To answer these questions we apply our data validation tools and in doing so provide additional guidance about how to use them.

## 5.1 Content Validity Assessment

As a first step in assessing the validity of a measure, we propose evaluating the degree to which a measure maps to a theoretical construct or its content validity. We break this task down into consideration of six properties: resonance, domain, differentiation, fecundity, consistency, and causal utility [Gerring, 2001]. Gerring proposes these as important attributes of concepts; we suggest that they can also be helpful in assessing measures. We find that the first five are useful to checking whether a measure captures the higher-level theoretical construct it is intended to capture while excluding irrelevant elements. The sixth property, causal utility, is a more instrumental version of this, ensuring that the measures captures the construct in a way that is actually helpful to conducting research. In this section we also demonstrate that the V-Dem corruption measures exhibit each of these characteristics and that because of this the V-Dem indicators offer some advantages over corruption measures from other datasets.[3]

For resonance, a measure should reflect how the underlying concept is used. The V-Dem corruption measures resonate with the accepted academic usage of the term corruption as the use of public office for private gain. Each of the six indicators refers to a public officeholder, specifically members of the executive, members of the legislature, public sector employees, and, indirectly, members of the judiciary. Each of these indicators also refers to private gain, either with explicit language, such as "material inducements," "personal or family use," "financial gain," or implicitly with terms such as "undocumented extra payments or bribes" (in the case of the judicial indicator). The public office and private gain are linked together with phrases that connote a trade off of public for private welfare, including "grant favors in exchange," "steal, embezzle, or misappropriate public funds," or "abuse their position." The judicial indicator again differs in this case by focusing on the citizen as actor, but nonetheless links public officials

[3]Information regarding other corruption datasets gathered from Afrobarometer, Rounds 2-6, 2002-2015, Arab Barometer, 2006-2007, 2010-2014, Asiabarometer, 2003-2006, Caucasus Barometer, Caucasus Research Resource Centers, 2008 and 2010-2013, Corporacion Latinobarmetro, Latinobarmetro, 2001-2011, 2013, and 2015, Former Yugoslavia Barometer, 2005, New Baltic Barometer, 2001, 2004, New Europe Barometer, 2001, 2004, 2005, New Russia Barometer, 2005-2012, European Bank for Reconstruction and Development, Business Environment and Enterprise Performance Survey, 1999-2014, Kaufmann et al., Transparency International, Corruption Perceptions Index, 2012-2014, Kaufmann and Stone, PRS Group, ICRG Methodology, Transparency International, Global Corruption Barometer, 2004-2013, United Nations Interregional Crime and Justice Research Institute, 1992, 1996, and 2000, World Values Survey Association, World Values Survey, 1995-1998, 2010-2014.

to the corrupt act by asking to what extent people use undocumented extra payments or bribes "to speed up or delay the process or to obtain a favorable judicial decision."[4]

The idea of domain takes resonance one step further to consider whether the measure captures the meaning for the relevant audiences. The domain of the V-Dem corruption measures is the meaning of corruption as discussed by scholars and policy makers. This domain does not include language referring to corruption as it is discussed in the lay community, which often applies a broader definition; we are not interested in capturing all instances where non-specialists, media, and pundits use the word corruption as a label for a behavior of which they do not approve. By including different types and forms of corruption, the V-Dem measures capture varied conceptualizations of corruption within the domain, thus strengthening the index's content validity [Trochim and Donnelly, 2001]. The V-Dem Corruption Index includes executive, legislative, judicial, and bureaucratic corruption, as well as grand and petty. It also covers different corruption forms, specifying bribes, undocumented extra payments, kickbacks, contracts for personal gain, future employment, theft, embezzlement and misappropriation, while also including the catch-all term "material inducements."

The comprehensive nature of the V-Dem measures is one of its strengths, relative to other datasets. By their own descriptions, many of the other corruption datasets gather information about "public sector" or bureaucratic corruption, excluding executive, legislative, and judicial corruption. This includes Transparency International's Corruption Perception Index (CPI), the World Bank's Business Environment and Enterprise Performance Survey (BEEPS), and nearly all the Barometers.[5] Ambiguously, Transparency International's Global Corruption Barometer (GCB) combines data on the public sector with private "big interests." By contrast, International Country Risk Guides Political Risk Services (ICRG) focuses on the "political system." The World Values Survey (WVS) offers a more expansive conceptualization, including petty and grand corruption and capture of government institutions by private interests. Problematically, some data used in studies as general measures of corruption actually capture a very narrow slice of the universe of all corruption forms falling under the definition, "the use of public office for private gain." For example, the International Crime Victims Survey asks only about exposure to bribery [Kennedy, 2014]. A narrow measure used as an indicator of overall level of corruption in a country will provide inaccurate results because different countries are marred by corruption in different forms or sectors [Knack, 2007, Gingerich, 2013]. The V-Dem Corruption Index minimizes this problem because it is considerably more comprehensive.

[4]Resonance is comparable to face validity, the idea that the measure seems like a good translation of the underlying idea [Trochim and Donnelly, 2001].

[5]The Afrobarometer is the exception. Depending on the year, it examines corruption among government officials generally or among particular groups of officials and civil servants. The other Barometers are the Arab Barometer, Asiabarometer, Caucasus Barometer, Former Yugoslavia Barometer, Latinobarometer, New Baltic Barometer, New Europe (post-communist) Barometer, and the New Russia Barometer.

While capturing relevant meanings, a valid measure must also exclude irrelevant ones. In other words, it must exhibit differentiation. The V-Dem corruption indicators clearly differentiate corruption from other similar behaviors. By specifying government officeholders, we do not include use of nongovernmental positions for private gain, such as the university admissions officer who takes a bribe in return for an admission acceptance. Likewise we exclude cases where the position might be public or private. For example, both independent and state media outlets might accept payments in return for favorable coverage. By specifying types of personal gain, we also exclude behaviors where there is no evidence of direct, immediate material gain. For example, buying votes and distributing government jobs can help an individual secure and remain in a government position but not necessarily enrich himself or herself. The detailed nature of the survey questions excludes other unethical behaviors, such as adultery, that do not involve the use of public office for private gain. This is also a relative strength of the V-Dem Corruption Index. Indicators from other datasets are used as measures of the use of public office for private gain, but they include superfluous information. For example, the World Governance Indicators' Control of Corruption (WGI) mixes electoral corruption, which does not necessarily involve private gain, along with public sector corruption.

By being both comprehensive and distinguishing, a measure demonstrates fecundity, meaning the characteristic of "reducing the infinite complexity of reality into parsimonious concepts that capture something important – something 'real' – about that reality" [Gerring, 2001]. To achieve this also requires coherence. The V-Dem Corruption index does well in this regard. The V-Dem Corruption Index is a factor index of the six indicators of corruption, which after simple imputation for missing values provides data for 173 country units for more than 95 years, on average. The six indicators that compose our index are linked to each other through the common ideas of public office, private gain, and the use of the former for the latter. The coherence of the index is evident from a Bayesian factor analysis of the six corruption indicators at the level of country-year. The results appear in Table 3.

| INDICATOR | LOADINGS ($\Lambda$) | UNIQUENESS ($\Psi$) |
|---|---|---|
| Executive bribery (v2exbribe) | .923 | .148 |
| Executive embezzlement (v2exembez) | .935 | .127 |
| Public sector bribery (v2excrptps) | .933 | .129 |
| Public sector embezzlement (v2exthftps) | .934 | .128 |
| Legislative bribery/theft (v2lgcrrpt) | .789 | .378 |
| Judicial bribery (v2jucorrdc) | .832 | .308 |

*Note:* Entries are factor loadings and uniqueness from a normal theory Bayesian factor analysis model, run through the MCMCfactanal() command in the MCMC package for R [Martin et al., 2011]. $n$=12,128 country years.

Table 3: Measuring Corruption with V-Dem Data (BFA Estimates)

These results provide some evidence that the V-Dem indicators map to the same underlying construct. All six indicators strongly load on a single dimension, although the fit for both legislative and judicial corruption is somewhat weaker. This could, however, simply be an artifact of over-representation this set of indicators gives to executive corruption. While a promising avenue for future research would be to discern the extent to which there are meaningful differences in types of corruption in particular countries or sets of countries, for present purposes these results lend support to the notion of corruption as a largely unidimensional phenomenon at the country level.

A measure should also exhibit consistency. To do so, it must capture the same meaning in multiple contexts. The V-Dem corruption indicators score well in terms of consistency; they are applicable to different empirical contexts, both across place and time. The indicators include numerous sufficient attributes depicting corruption, rather than necessary and sufficient attributes, so that they apply to more places and eras [Gerring, 2001]. For example, bribes might be a common form of corruption in one location whereas future employment opportunities are a common form elsewhere. Including numerous possible forms of corruption makes our indicators more broadly applicable and therefore less restrictive in their application across contexts.[6]

Finally, the V-Dem corruption indicators and resulting index offer casual utility. The index as a whole is useful in testing causal claims in which corruption is the primary cause or effect. Individual indicators can be used for more specific theories. By offering measures at different levels of aggregation and both specific and general conceptualizations, the V-Dem corruption data are particularly useful for corruption researchers.

This section shows that the V-Dem corruption data are valid from a theoretical standpoint. The

[6] We tested this logic in the pilot phase of V-Dem data collection, where country experts from two dissimilar countries in each region of the world answered our questions and provided feedback. The pilot served as a face validity test of our corruption construct as well [Trochim and Donnelly, 2001]. We also confirmed the consistency of our indicators through our own research experiences in Africa, Europe, the former Soviet Union, and North America, investigating both contemporary and historical periods.

indicators collectively and individually align with definitions of corruption in the literature, offering various measures of how public office can be used to achieve private gain. In other words, this section demonstrates that the V-Dem corruption indicators align with shared concepts, to use Schedler's term. [Schedler, 2012].

## 5.2 Data Generation Validity Assessment

As a next step in data validation, we propose assessing the degree to which the data generating process aligns with shared rules of translation [Schedler, 2012]. We interpret the "rules of translation" broadly to mean any component of the data generating process. We use two criteria to evaluate this process at each step: whether it is unbiased and whether it is reliable. In particular, we recommend assessing dataset management structure, data sources, coding procedures, aggregation methods, and geographic and temporal coverage. As we walk through these potential sources of bias and unreliability, we evaluate the V-Dem measures and highlight their strengths and weaknesses relative to other corruption datasests.

### 5.2.1 Dataset Management Structure

An often overlooked issue regarding data sources and the potential bias they introduce is the leadership and funding source for the dataset. With regards to corruption datasets, Hawken and Munck [2009b] find significant differences across data sources based on who is doing the evaluating. They compare different corruption datasets to mass surveys about corruption, which they consider the "gold standard", an assumption we challenge below. They find commercial datasets are the least correlated with mass surveys, followed by NGO-managed datasets, followed by surveys of businesses. Expert surveys run by multilateral development banks are those most highly correlated with the experience-based measures of corruption found in mass surveys. While this approach might not be the best nor apply to other topics, the point to be made is that the management of a dataset can affect the data produced.

No unreliability or bias stemming from the V-Dem organizational structure is evident. V-Dem is an academic venture, led by four professors as PIs and 12 scholars from leading universities in different countries, assisted by 37 (mostly) scholars from all parts of the world as regional managers, and the V-Dem Institute at University of Gothenburg, Sweden, as the organizational and management headquarters. Funding comes from research foundations and donor countries, mostly in Northern Europe. This seems to be an unbiased and reliable organizational structure capable of generating unbiased and reliable data.

### 5.2.2 Data Sources

A key question to consider when evaluating potential bias and unreliability due to data sources is whether the data are original or aggregated from different sources. Datasets that aggregate information from different sources multiply biases and measurement errors by including those from each source in their own index [Treisman, 2007, Herrera and Kapur, 2007, Hawken and Munck, 2009b]. This "polls of polls" approach also precludes analysis of the correlation across these datasets; the alignment between them is mechanical, not a sign of independent but aligning data [Hawken and Munck, 2009b]. V-Dem avoids this problem because it produces original corruption data. This is a strength, relative to many corruption datasets. Three other datasets aggregate information from different sources. WGI uses household and firm surveys and information from businesses, non-governmental organizations, and public-sector organizations. CPI relies on multiple governance and business climate ratings and surveys. ICRG's documentation indicates only that final ratings are determined by its staff, but those writing about ICRG have noted that it relies on surveys and a "network of correspondents with country-specific expertise" [Knack, 2007, p. 258].[7] The nine barometers and three of the other datasets – WVS, GCB, and BEEPS – collect their own data through surveys.

### 5.2.3 Coding Procedures

When data are generated by coders, it is important to consider 1) the qualifications and potential biases of the coders themselves, 2) the transparency and thoroughness of the coding guidelines, and 3) the procedures for combining coder ratings into a single indicator or index [Treisman, 2007, Martinez i Coma and van Ham, 2015]. We consider each of these below.

Regarding the coders themselves, several scholars have argued that expert-coded data on corruption are inferior to citizen-coded or "experience" data [Treisman, 2007, Hawken and Munck, 2009a,b, Donchev and Ujhelyi, 2014]. Rather than privilege one type of coder over another, we recommend considering what type of coder is a good match for generating the data of interest and what techniques can reduce bias and increase reliability. For example, with respect to corruption data, citizen coders offer certain disadvantages. Citizen perceptions of corruption are fundamentally a censored indicator of the latent variable we actually care about: the level of corruption in a society. Citizens interact with only certain kinds of officials and observe certain kinds of corruption. Not only are corruption data obtained from citizens noisy, they are systematically biased: countries with established institutions, stable incentive structures, and experienced public officials will likely have corruption that is more removed from citizens lives, which means that they will be unable to report observing it on surveys, which in turn means

[7] For more information about ICRG, see Knack [2007] and Razafindrakoto and Roubard [2010].

that cross-national measures of corruption based on citizen reports will over-estimate corruption in consolidating democracies and under-estimate it in stable democracies. 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.

To what extent do V-Dem coding procedures produce valid corruption data? V-Dem relies on expert perceptions of corruption. The stringent selection criteria for experts included in V-Dem could offset some of the inherent biases in other datasets based on corruption perceptions. Our experts have been recruited based on their academic or other credentials as field experts in the area for which they code, on their seriousness of purpose and impartiality [Coppedge et al., 2014]. Impartiality is not a criterion to take for granted in political science research. Unsurprisingly, Martinez i Coma and van Ham [2015] noted that variance in estimates of election integrity (Perceptions of Electoral Integrity dataset) was significantly higher when one of the coders was a candidate in the election. Understanding who the coders are and where they may provide biased data is critically important in evaluating data validity.

For V-Dem, at least five experts code each question-year observation for a total of more than 2000 experts assisting us in gathering the data. As a rule, at least three fifths of the experts coding a particular country either are nationals of or reside in the country in question. We thus tap into a local source of expertise and knowledge on corruption, avoiding the problem of far-removed experts and also the problem of citizens within limited experience and information. In order to aggregate up from coders to the level of country-years, we apply an IRT measurement model that models variation in coder reliability and allows for the possibility that raters apply ordinal scales differently from one another [Pemstein et al., 2016]. In other words, coders may have varying error rates and be more or less strict than one another when making ordinal rating decisions. The model also uses bridge raters—who rate multiple countries for many years—and lateral coders—who, in addition to providing a time-series for one country, provide single-year ratings for a number of other countries—to calibrate estimates across countries [Pemstein et al., 2014]. Of course, such statistical techniques do not guarantee that our corruption data are unbiased. Therefore, below, we perform convergent/divergent validity checks and examine how individual-level information on experts and structural country characteristics explain variation in corruptions ratings, both within V-Dem, and across multiple datasets.

#### 5.2.4 Aggregation Model

Many datasets offer low-level indicators that they combine into higher-level indices. To assess the validity of the resulting data, it is important to consider a) the choice of indicators to aggregate and b) the aggregation rule.

With respect to selecting indicators to aggregate into a corruption index, V-Dem was guided by the

conceptualization of corruption, as described in the content validity section above. This is a relative strength of V-Dem. Both the WGI and CPI choose indicators that reduce missingness [Hawken and Munck, 2009b]. V-Dem does not have such a constraint, as the level of missingness does not vary greatly from one indicator to another.

V-Dem aggregates corruption indicators using a two-stage approach. First, as we briefly describe above, we use IRT methods to aggregate individual codes into low-level indicators. At the second stage we use Bayesian factor analysis to aggregate individual measures into a higher-level indices, using the method of composition [Tanner, 1993] to propagate estimation uncertainty in the first stage into the resulting indices. In particular, we construct the executive corruption index (**v2x_execorr**) by fitting a factor analysis model to the indicators for executive bribery (**v2exbribe**) and executive embezzlement (**v2exembez**). The model estimates the posterior distribution of the latent factor score for each observation (country-year); we use these posterior distributions to produce index point estimates (posterior averages) and estimates of uncertainty (standard deviations and highest posterior density regions). We build the public corruption index (**v2x_pubcorr**) similarly, and base the index on the estimated latent factor scores from a model incorporating low-level indicators of public sector bribery (**v2excrptps**) and embezzlement (**v2exthftps**). Finally, to construct the overarching corruption index (**v2x_corr**), we average (a) the executive corruption index (**v2x_execorr**), (b) the public sector corruption index (**v2x_pubcorr**), (c) the indicator for legislative corruption (**v2lgcrrpt**), and (d) the indicator for judicial corruption (**v2jucorrdc**). In other words, we weight each of these four spheres of government equally in the resulting index.

### 5.2.5 Coverage Across Countries and Time

It is important to consider potential biases introduced by limited geographic or temporal coverage of a dataset [Treisman, 2007]. Particularly with sensitive topics, such as corruption, missing data are likely not missing at random. Therefore, it is valuable to consider any selection bias introduced in the process of deciding which cases to include or exclude. When these concerns are allayed, we can be more confident in conducting analyses across countries and over time. To further assess the validity of cross-country analysis with a particular dataset, it is important to assess the controls put in place to anchor country ratings to the same scale, according to Treisman [2007].

By these criteria, the V-Dem corruption measures are highly valid. V-Dem includes every single country in the dataset, avoiding the bias in datasets of only on a subset of countries - those easiest to code or those for which coders are readily available. An additional way that V-Dem ensures reliable and unbiased data is by using the same coder recruitment procedures and data coding methods across countries and over time. The validity of using V-Dem measures for analysis over time is further en-

hanced by relying on the same set of coders throughout the coded history of each country. To facilitate cross-national research, V-Dem is in the process of implementing anchoring vignettes across all coders. Anchoring vignettes serve to reduce cross-coder differences by assessing each coder's rating scale and then adjusting the ratings of each coder to be consistent.

The validity of V-Dem corruption measures for analysis across countries and over time is one of the dataset's key strengths. By asking the same questions of each coder for each country for each year, V-Dem allows over-time and cross-country comparisons of corruption levels in the world going all the way back to 1900. This is an important contribution to the corruption literature in and of itself, since existing measures of corruption are not designed for panel analysis and yet existing measures are often used this way.

Most other techniques for generating corruption data tend to be on a smaller scale, their coverage typically ranging from a single community over a year to a small number of countries for a few years at most. These sources include 1) measures based on observing corruption (e.g., McMillan and Zoido [2004], Barron and Olken [2007], or Sequeira and Djankov [2010]); 2) measures based on estimating corruption by subtracting budget allocations from actual receipts (e.g., Reinikka and Svensson [2003], Golden and Picci [2005],Olken [2005], Niehaus and Sukhtankar [2013]); 3) measures based on audits of government accounts (e.g., Ferraz and Finan [2008]); 4) measures based on criminal records (e.g., Teorell and Rothstein [forthcoming], Peters and Welch [1980], Chang et al. [2010], Goel and Nelson [1998]; and 5) measures based on media reports (e.g., Cox and Kousser [1981], Nyblade and Reed [2008]). With such limited data, global analysis over time is not possible.

With those datasets that provide (near) global time series data, the validity of the measures for such analysis is questionable. Measures of corruption are typically taken at the country level, where comparisons across countries often come at the expense of comparisons over time [Christiane et al., 2006, Galtung, 2006, Knack, 2007]. For example, WGI is calculated such that the global average is the same every year, meaning that 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 the WGI's index over time is the product of changes in the sources and coding rules used rather than actual changes in corruption levels. Treisman [2007] notes that the aggregation strategies and sources have changed over time for CPI as well. Finally, the WGI forces a consistent global average over time, preventing by construction an understanding of trends over time.

To further illustrate the V-Dem Corruption Index's relative strength, we consider what we can learn about trends in corruption levels – something other datasets are not designed to do. In Figure 1, we depict the global trends building on the V-Dem Corruption Index, also including a version where we have

imputed missing data on the regression line (mostly needed for all country years with no legislature). The trend depicted in the figure could be somewhat surprising: according to the V-Dem data, corruption levels have been on the rise globally since at least the 1960s, with a peak just around the time when corruption started to be on the global agenda for reform. The world thus looks much more corrupt today than it did 100 or 60 years ago. Since 2000, on the other hand, the level of corruption in the world has been in slight decline.

We think there are several reasons to believe a global surge in corruption over the latter half of the 19th century makes intuitive sense. For one thing, the world economy is much more monetized than it was half a century ago. More bank notes exchange hands. That in itself should lead one to expect higher levels of corruption. Relatedly, the collapse of the Soviet economies in the early 1990s, as well as a global rise of libertarian values, has led to a flurry in privatization reforms, also known to increase levels of corruption. Finally, the number of hybrid regimes has been on the rise [Teorell and Hadenius, 2007], and we know from previous studies that corruption levels tend to be highest in countries at the crossroads between authoritarianism and democracy [Montinola and Jackman, 2002, Sung, 2004, Bäck and Hadenius, 2008, Rock, 2007, Treisman, 2007, Charron and Lapuente, 2010].

There is however one alternative interpretation of the trend that needs to be taken seriously: reporting bias. Maybe the perceived increase in corruption is a reflection of the fact that the media reported more about corruption by the mid-1990s than they did previously. That could imply that the increase is more a reflection of hearsay than actually increased levels of corruption. More generally, concerns over media bias are at the forefront of the V-Dem data validation agenda.

In the case of the V-Dem corruption data, two pieces of evidence speak against this interpretation, or at least makes it unlikely that is the whole story behind the trends. The first is the downward trend we observe from around the year of 2000. Since there is no reason to expect the media to report less on corruption over the last decade (if anything more), this speaks directly against the notion that the increase in the decades preceding this decline should merely be a reflection of reporting bias. Second, in Figure 2 we present the aggregate trend in corruption levels for a sub-sample of all country years where there, according to the Whitten-Woodring and Van Belle [2014] measure, was no media freedom. If reporting bias was driving the upward trend, we should expect a flat line (or perhaps even decline) in countries where reporting was severely restricted. This is however not the case. Although a more corrupt sub-sample overall, and only covering the post-WWII period, the upward trend in corruption across the second half of the 20th century is clearly present also under these conditions. The trajectories only differ after 2000; specifically the decline is not present in countries void of a free press. That should come as no surprise however.

Evaluating the dataset management structure, data sources, coding procedures, aggregation methods,

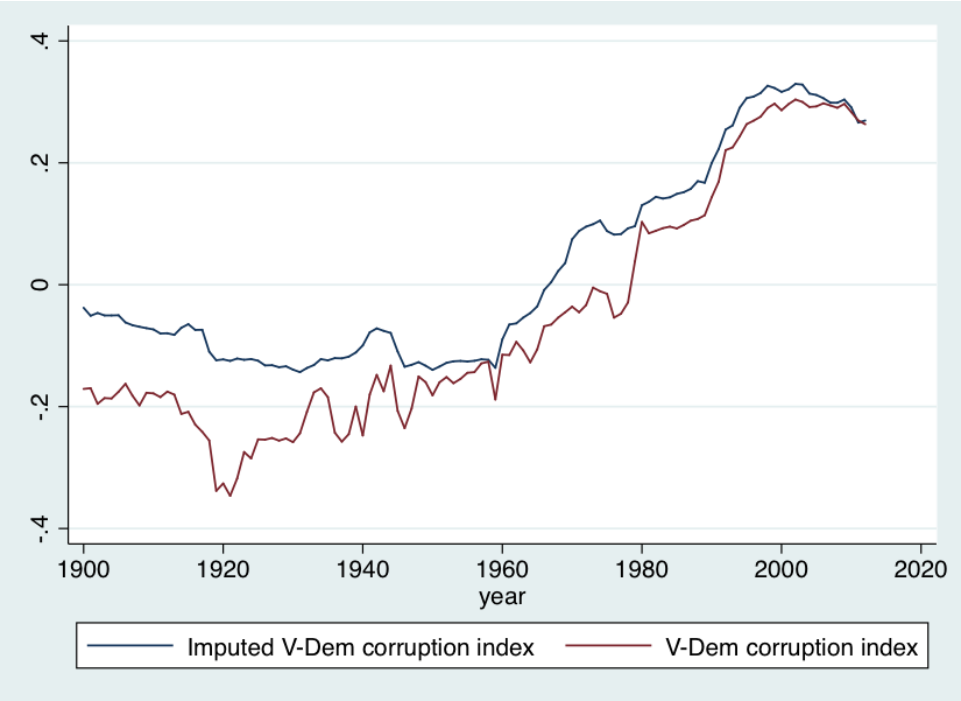


Figure 1: Global Levels of Corruption, 1900-2012

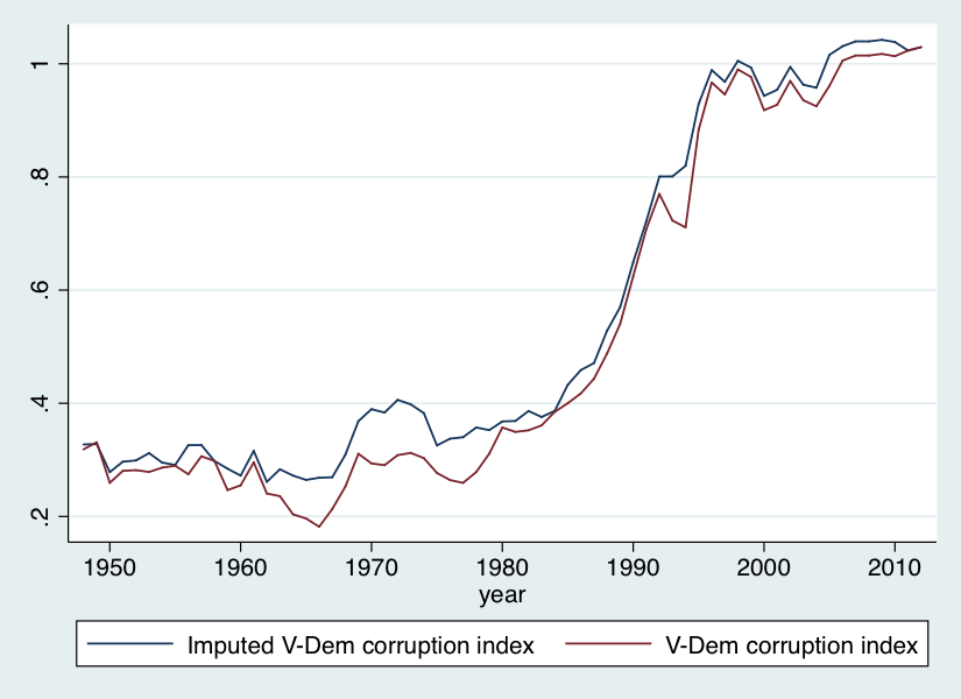


Figure 2: Levels of Corruption in Countries with no Free Media, 1948-2012

and geographic and temporal coverage of the V-Dem corruption measures provides additional evidence that they are valid. In particular, it highlights the V-Dem Corruption Index's value as measure for comparison across countries and over time. The validity of data generation process is typically not considered by scholars, but it does illuminate the quality of indicators.

## 5.3 Convergent Validity Assessment

Our final data validation approach draws from Schedler's idea of shared realities [Schedler, 2012]. To what extent do the data to validate correspond to existing knowledge? To answer this question, we recommend three techniques. First, compare the data to original or existing case studies. Second, with data aggregated from multiple experts, conduct statistical analyses to compare coders trying to measure the same thing. Third, conduct statistical analyses to compare the measure being assessed with other measures. The following section demonstrates the first technique while evaluating the validity of the V-Dem data. The section after that does the same for the second and third techniques.

### 5.3.1 Convergent Validity Testing with Case Studies

We applied the first technique to V-Dem data for four countries. Case studies of contemporary Georgia and Zambia show that the V-Dem data, relative to other corruption datasets, better mirror detailed descriptions from published accounts of corruption in the countries. Two historical case studies, of Spain and the U.S., help validate our historical data and thus increase our confidence in our data collection methods, especially V-Dem efforts to recruit experts with historical knowledge. Finally, a comparison of individual V-Dem corruption indicators and the U.S. case show the value of our approach to disaggregate the concept of corruption, rather than providing a single measure as other datasets do.

To develop the case studies, a research assistant used scholarly articles and 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 its datasets that have corruption measures, WGI and BEEPS. Importantly, the research assistant did not view the quantitative data from V-Dem or these other datasets prior to, or while writing, the case studies.

In presenting V-Dem data for the four countries throughout this section, each of the country graphs includes only the portion of the scale where a country's corruption scores fall. So that the reader does not exaggerate changes in corruption in each country—for example, that a decrease indicates an end to corruption—Figure 3 illuminates the absolute values of corruption in the countries. Specifically, it includes all four countries with their quite different levels of corruption for the time periods examined.

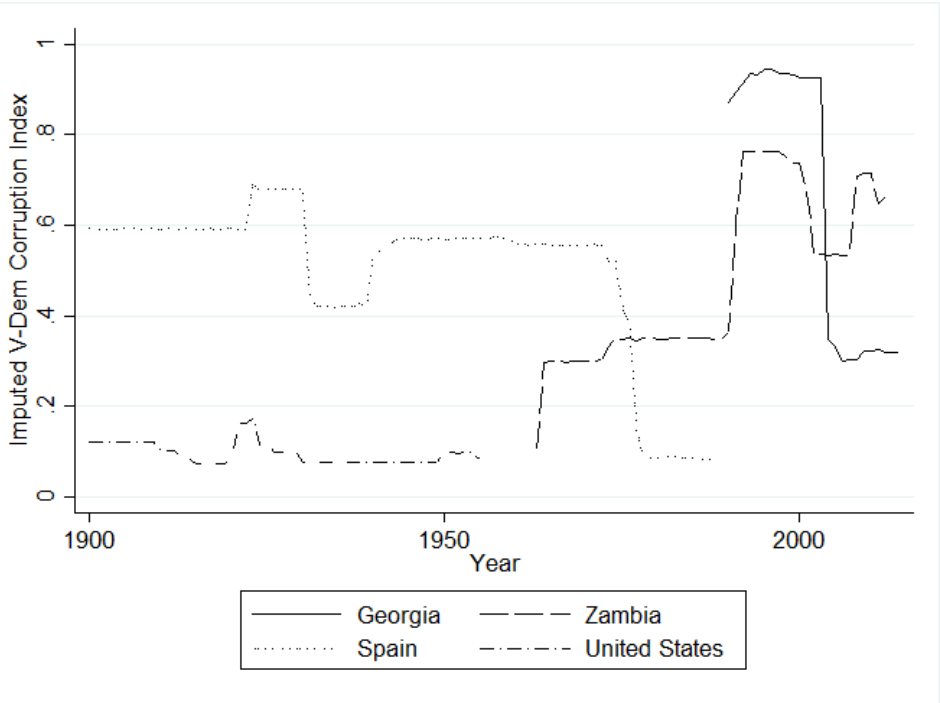


Figure 3: V-Dem's Corruption Measure for Gambia, Spain, the United States, and Zambia

#### 5.3.1.1 Contemporary Cases: Georgia and Zambia

We focused on Georgia and Zambia from their points of independence, 1991 and 1964 respectively, to the present. We selected these countries because V-Dem data for these countries differ significantly from those produced by other corruption measures, specifically the WGI and CPI data. For Georgia (Figure 4), V-Dem shows a steep, isolated drop and then a leveling off in corruption, whereas WGI and CPI mostly portray a gradual, less significant decline with increases during some periods.[8] The contrast among the indices is even greater for Zambia, as Figure 5 demonstrates. For a period, V-Dem and CPI move in opposite directions with V-Dem also 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 decreases and an increase in corruption. (Figure 5, for Zambia, also uses a normalized version of each index.)

For Georgia we find that V-Dem data mirror the thick description from the published accounts, suggesting that they offer a more accurate depiction of corruption in the country [Chene, 2011, Engvall, 2012, Huber, 2004, Kukhianidze, 2009, Mitchell, 2009, Shahnazarian, 2012, Alam and Southworth, 2012]. In the early independence period, corruption was rampant as officials engaged in schemes, often in collaboration with organized crime, to enrich themselves and their clients during the socialism-to-market transformation, according to published accounts. Public sector corruption flourished as economic up-

[8] Figure 4 uses a normalized version of each index so that they are comparable to each other.

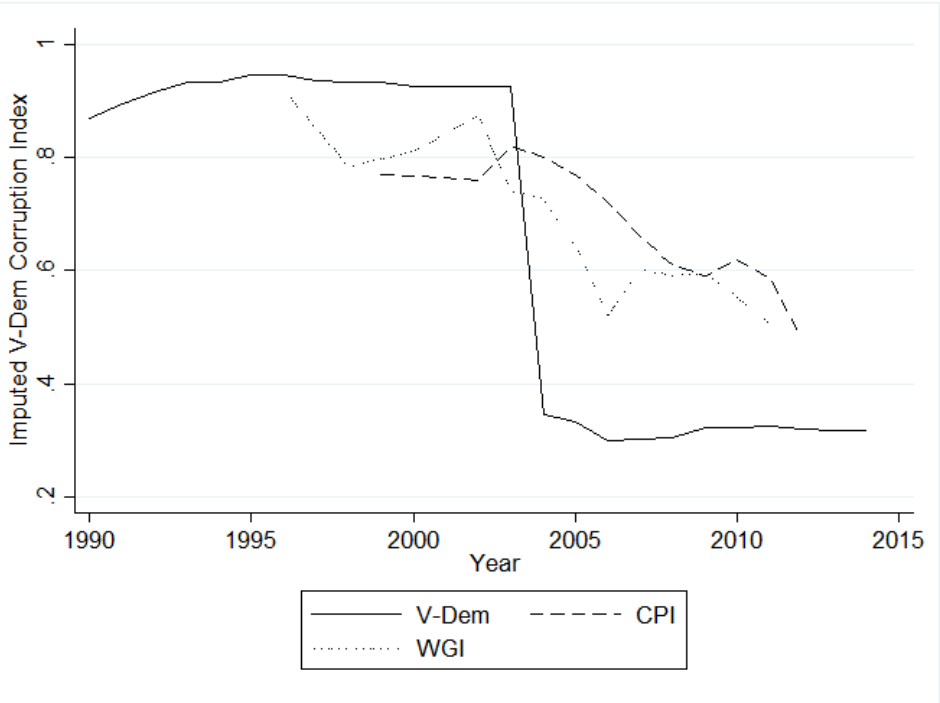


Figure 4: Corruption in Georgia

heaval encouraged civil servants to take bribes to supplement their meager salaries. The V-Dem data show high levels of corruption during this period too, as Figure 4 illustrates. In part, the public's frustration with this corruption sparked the Rose Revolution of 2003, which resulted in the ouster of the president and a dramatic drop in corruption beginning that year, the publications recount. During his first year the new president, Mikhail Saakashvili, implemented extraordinary anti-corruption measures, which included firing all traffic police and large numbers of other civil servants and which accounts indicate were successful. The V-Dem data show a corresponding significant decline in corruption from 2003 to 2004. Sources describe how prosecutions against former government officials beginning in 2004 and 2005 organized crime legislation reduced corruption. The V-Dem trend line drops again from 2004 and 2006. No significant anti-corruption efforts have been undertaken since 2005 and high-level corruption remains a problem, according to the publications. The V-Dem data reflect this with a leveling off of the line at a relatively high value in 2006.

The V-Dem data for Zambia also match published accounts of corruption in that country [Chikulo, 2000, Van Donge, 2009, Mbao, 2011, Szeftel, 2000], although not as well as in the case of Georgia. During Zambia's First and Second Republic, from independence until 1990, corruption was pervasive in the country, according to published accounts. Both low-ranking and high-ranking officials engaged in bribery or theft of public resources. In particular, quasi-governmental enterprises offered employees many opportunities to steal resources. The relatively high score on the V-Dem scale reflects this, as seen

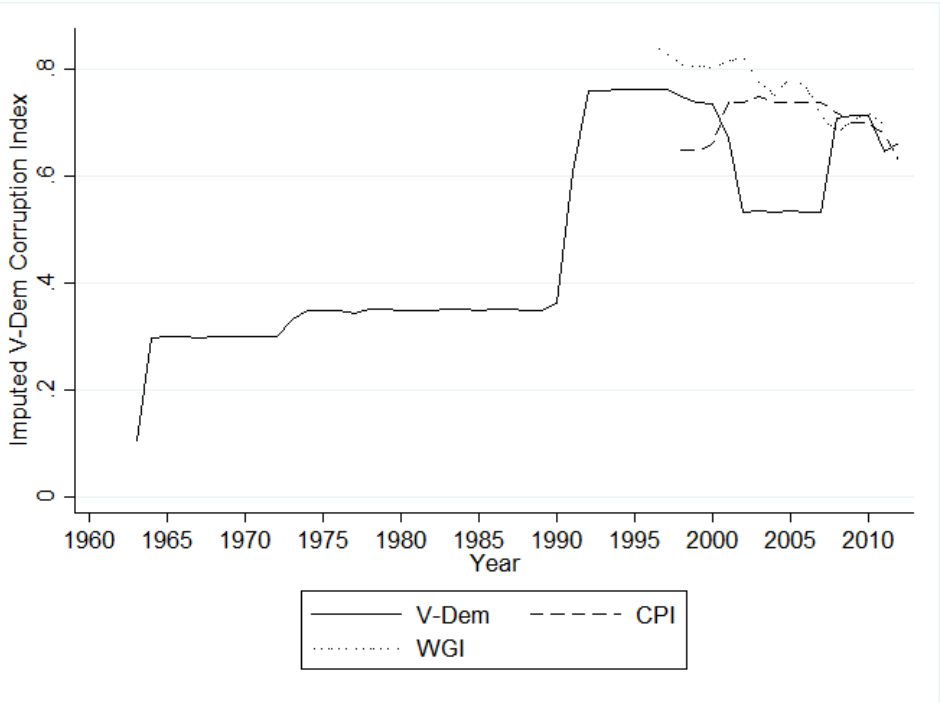


Figure 5: Corruption in Zambia

in Figure 5. As the economy worsened in the early 1970s, civil servants increasingly turned to theft of state resources to augment their salaries; the V-Dem data capture this increase. Since then growth in corruption can mainly be attributed to the informal practices of government elites. Corruption increased dramatically in the first years of the Third Republic, according to published reports. Government officials used the privatization campaign in this era as an opportunity to further enrich themselves. Thick descriptions do not mention the decrease in the late 1990s that the V-Dem data depict (as does WGI with a smaller magnitude, but not CPI). Otherwise, the publications and V-Dem data 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 top 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 then and also in 2012. The new presidents abandoned genuine anti-corruption efforts enabling practices like public theft to again increase. Essentially, both these later administrations strengthened anticorruption rules while selectively enforcing those rules, primarily against political opponents. The V-Dem data mirror this pattern except for showing a small drop in 2011, which the publications do not mention (but the other indices depict).

Although only two cases, the analysis of Georgia and Zambia boosts our confidence in the V-Dem data. For these countries, V-Dem data outperform the other indices, which do not capture the trends revealed in the thick descriptions. The comparison between the V-Dem data and the thick descriptions

also demonstrates V-Dem's precision: like thick descriptions, the V-Dem data are able to capture an increase or decrease within a short time period, such as a year or two.

#### 5.3.1.2 Historical Cases: Spain and the U.S.

We also selected Spain and the U.S. to check the validity of our historical data. Examining historical periods for Spain and the U.S. allows us to check our data collection methods, which aim to draw on not only contemporary, but also historical, knowledge of experts. These countries make for challenging validity tests because the V-Dem data show that corruption levels changed numerous times in the past. We examine both countries from 1900 and stop with 1988 for Spain and 1955 for the U.S. in order to capture periods of dramatic change while also keeping the amount of in-depth research manageable. We do not compare these V-Dem data with other corruption indicators because the others do not provide historical coverage.

For Spain, we find that the V-Dem data match the detailed accounts scholars have provided (e.g., Ben-Ami [1983], Cabrera and del Rey Reguillo [2007], Carr [1980], Heywood [1996], Jiménez [1998], Moreno-Luzón [2012], Preston [1994], Pujas and Rhodes [2002], Towson [2012]). In the beginning of the 20th century during Primo de Rivera's rule, corruption increased because his economic plan, involving the development of industrial monopolies and significant state economic intervention, introduced many opportunities for illicit personal gain, according to published accounts. Monopoly concessions were granted and public investments were made to financially benefit the government decision-makers or their relatives. Government officials also received bribes in exchange for granting regulatory exceptions or deciding for a particular party in an economic dispute. With the end to this dictatorship, corruption levels fell to include only infrequent scams involving smaller numbers of government officials, published reports indicate. The V-Dem data correspond with this account, showing that the corruption level jumped between 1922 and 1923, when Primo de Rivera took power, and fell from its high in 1930, when he left office. With the start of the Franco regime in 1939, corruption began to rise again, according to scholarly sources. The regime's economic policies facilitated bribe-taking and theft of government resources by officials. The Franco government's autarchic postwar reconstruction plan required officials to ration resources to industries. This resulted in government officials enriching themselves in a black market for industrial supplies. To staff reconstruction, the government rapidly hired thousands of people, and this is thought to have also increased corruption among officials. Corruption continued through the end of Franco's rule in 1975. The corruption resulting from the reconstruction plan subsided slightly beginning in the 1960s, but government officials continued to find new ways to enrich themselves. Franco's regime did almost nothing to combat corruption: in fact, "Franco seems to have regarded corruption as a necessary lubrication for the system that had the advantage of compromising many with the regime and


Figure 6: Corruption in Spain

binding them to it" [Payne, 1987, p. 399]. The V-Dem data are consistent with this thick description; they depict a sudden rise in corruption in 1939 and then a steady amount of corruption through the Franco period with the slight decrease from 1960. Franco's death and the subsequent transition to democracy in Spain changed the nature of political corruption in the country. Corruption shifted from government officials enriching themselves to political parties engaging in illegal schemes to secure funds for campaigning. The V-Dem Corruption Index does not include measures of campaign fraud so it correctly shows a sudden decline through the period of Franco's death and the democratic transition, reflecting the reduction in government officials use of public office for private gain. In sum, V-Dem data match detailed descriptions of historical corruption in Spain well.

The U.S case serves two purposes: to illustrate the quality of the V-Dem historical data and to demonstrate in detail the value of V-Dem's individual indicators of corruption. Both the V-Dem Corruption Index and its individual indicators match the details provided by published accounts [Benson et al., 1978, Grossman, 2003, Menes, 2003, Reeves, 2000, Woodiwiss, 1988].

At the turn of the century, U.S. government bureaucrats were engaged in the theft of state resources and the exchange of state goods and service for personal material gain. Pocketing money for the awarding of contracts or lax police enforcement were two such schemes. Beginning in the early 1900s, however, the country experienced a decline in corruption due to successful reform efforts of the Progressive Movement. During the first years of the new century journalists reported on municipal corruption and many cities

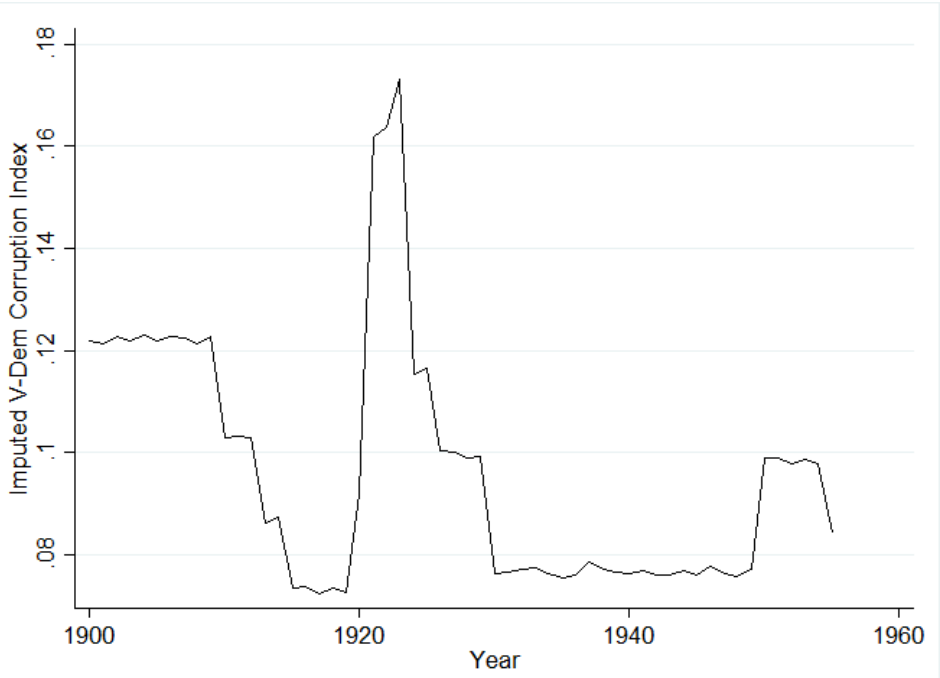


Figure 7: Corruption in the USA

responded by restructuring local government, although some of the largest cities continued to be ruled by political machines until the 1950s. At the state level Progressive-era reformers pushed through changes, such as civil service rules, that helped reduced bribery and theft. The V-Dem Corruption Index depicts this decrease in corruption, as Figure 7 shows.

After this period of positive change, corruption increased significantly in 1921 with the administration of Warren Harding. Considered one of the most corrupt U.S. administrations in the 20th century, Harding's government, particularly cabinet members and bureaucrats, engaged in bribery and embezzlement. Most infamous was Secretary of Interior Albert B. Fall's role in the Teapot Dome Scandal, where bribes were paid for secret leases on government oil fields. In the public sector, bribery and theft were common among personnel in the Veterans Bureau and the Department of Justice. The U.S. ban on alcoholic beverages during Prohibition helped fuel the corruption during Harding's administration. Law enforcement personnel, from staff in the Federal Prohibition Unit to Attorney General Harry Daugherty himself, took money from liquor smugglers in exchange for lax enforcement and pardons, respectively. Following Harding's death in 1923, efforts, such as prosecutions, reduced corruption. The V-Dem Corruption Index approximates this account well, as the previous graph indicates. The data show a small increase in 1920 but then, like the thick description, a significant increase in 1921 and a dramatic decrease in 1924. For this period the individual V-Dem indicators diverge, reflecting the published accounts. It is evident in Figure 8 that most of the increase is attributable to executive and public sector bribery and

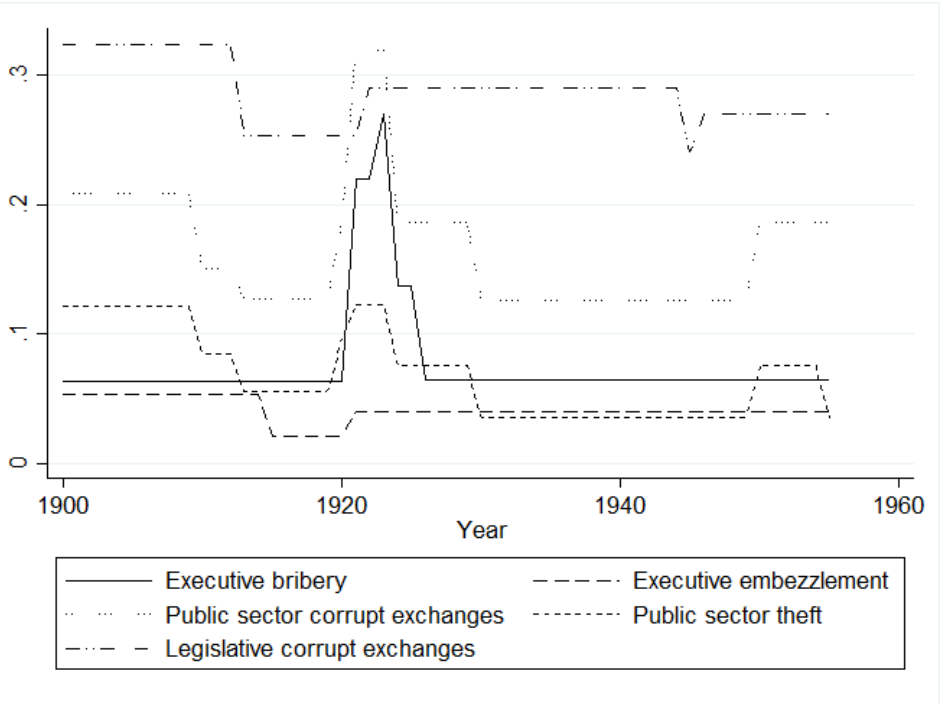


Figure 8: Disaggregated Corruption in the USA

then embezzlement. This period is not characterized by a dramatic increase in legislative corruption, as is clear from the published reports.[9]

Illicit practices by U.S. legislators were central to corruption during World War II. New opportunities for corruption during the war sustained corruption levels. With the end of the war and prosecutions for the schemes these opportunities subsided. U.S. legislators were the government officials typically profiting from these illegal practices. For example, Representative Andrew Jackson May accepted bribes in exchange for the favorable awarding of military contracts. The V-Dem legislative corruption indicators captures the dip in corruption that coincided with the end of the war in 1945, evident in Figure 8.

Corruption substantially increased with the Truman administration, which is considered the other most corrupt government of the 20th century, besides Harding's. Bureaucrats in numerous agencies used their positions for personal gain. Tax collectors with the Internal Revenue Service (IRS) took bribes and reduced or ignored tax liabilities in return. High-level bureaucrats, including the IRS commissioner and general counsel, were involved in this corruption as well as embezzlement schemes. The corruption extended beyond the IRS to other agencies, including the Federal Housing Agency, whose employees received monetary inducements to approve illegal activity, and the Defense Department and Maritime Commission, whose staff benefited from kickbacks on contracts and other agreements. The V-Dem

[9]The judicial corruption indicator is not included in this analysis of the U.S. because it does not vary during this period, although it does in later eras.

data show that corruption increased during the Truman administration, which lasted from 1945 to 1953. Corruption levels jump in 1950 during his leadership and drop in 1955 after he left office. An examination of individual V-Dem indicators support the scholars' accounts, showing that public sector bribery and theft, rather than executive or legislative corruption, were the problem. This is evident from Figure 8. Overall, the V-Dem data present a picture similar to thick descriptions of corruption in the U.S. historically.

The cases of Spain and the U.S. increase our confidence in our approach to collecting historical data and the validity of that data. Moreover, the U.S case demonstrates the value of not only the V-Dem Corruption Index but individual V-Dem corruption indicators. Overall, the validity assessment with case studies suggests that V-Dem data correspond to existing knowledge well.

### 5.3.2 Convergent Validity Testing with Statistical Analysis

With expert coded data with multiple experts, there are two fundamentally different approaches for assessing convergent validity statistically: (i) the first is to compare different coders trying to measure the same thing (i.e., their responses to the same question); (ii) the other is to compare the measure being assessed with other measures (typically but not necessarily from other datasets). In both cases, the analysis could consider: (a) the disagreement among coders/measures; or (b) the average differences between two coders/measures.

#### 5.3.2.1 Comparing Coders

Comparing coders is the approach taken in Steenbergen and Marks [2007], focusing on expert survey data on party positions for multiple parties in multiple countries, and Martinez i Coma and van Ham [2015], focusing on expert survey data on election integrity across multiple indicators in multiple countries (both at a single time point). Both seem to take as their point of departure that by comparing expert survey responses we can capture validity, the general idea being that the less inter-expert disagreement, or the less systematic sources of variation in inter-expert judgments, the more valid the data.

Following their lead, we start in Table 4 with a simple variance decomposition of the raw ratings provided by the V-Dem coders, relying on fixed country- and time-effects (since we are not sampling either countries or time periods), but with random effects at the coder level. Overall, given that each of the corruption indicators is on a 0-4 scale, these estimates do not suggest egregious levels of coder disagreement (recall that they are variances, so the standard deviations would be slightly higher). They do suggest slightly more coder disagreement than Steenbergen and Marks [2007] (Table 2) estimate for party positions in Western Europe (0.917 on a 7-point scale equals 0.131 on average), and even more so compared to the Martinez i Coma and van Ham [2015] (Table 1) estimate for election integrity in a

| | Exec. Bribery | Exec. Theft | Pub. Bribery | Pub. Theft | Legisl. Corr. | Jud. Bribery | Pooled |
|---|---|---|---|---|---|---|---|
| Variance components | | | | | | | |
| Expert | 0.814*** | 0.909*** | 0.779*** | 0.851*** | 0.756*** | 0.626*** | 0.495*** |
| | (0.038) | (0.044) | (0.037) | (0.041) | (0.035) | (0.030) | (0.023) |
| Indicator | | | | | | | 0.358*** |
| | | | | | | | (0.008) |
| No of experts | 1086 | 1091 | 1091 | 1089 | 1105 | 1061 | 1585 |
| No of observations | 79098 | 79522 | 79763 | 79891 | 60601 | 76051 | 454926 |

Variance decomposition with country- and year-fixed effects, coder- and indicator-random effects.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 4: Variance Decomposition of Raw Expert Ratings

| | Exec. Bribery | Exec. Theft | Pub. Bribery | Pub. Theft | Legisl. Corr. | Jud. Bribery | Pooled |
|---|---|---|---|---|---|---|---|
| Variance components | | | | | | | |
| Expert | 0.031*** | 0.044*** | 0.041*** | 0.053*** | 0.051*** | 0.029*** | 0.027*** |
| | (0.002) | (0.003) | (0.002) | (0.003) | (0.003) | (0.002) | (0.002) |
| Indicator | | | | | | | 0.030*** |
| | | | | | | | (0.001) |
| No of experts | 924 | 600 | 903 | 847 | 877 | 872 | 1346 |
| No of observations | 57290 | 28843 | 52976 | 44614 | 56989 | 32000 | 272711 |

Variance decomposition with country- and year-fixed effects, coder- and indicator-random effects.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 5: Variance Decomposition of Expert Perception Estimates

global sample (0.217 on average across 49 5-point scales equals 0.043 on average). When pooling across the six indicators, adding indicator-random effects to the model, the estimate is 0.495 across the experts, or 0.099 if standardizing by the number of points on the scale. Given that corruption is in principle a non-observable and also hard to collect information on when going back in time as far as 1900 and in low-information environments such as authoritarian regimes, we find this level of coder disagreement low. In the case of Steenbergen and Marks [2007], the coders only look at highly industrialized countries where there is vast information available about the parties to be rated. In Martinez i Coma and van Ham [2015], the coding task is also arguably easier given that the surveys were sent out relatively close to the elections being assessed.

With the IRT measurement model used by V-Dem to aggregate ratings across coders, described above, we can make a more direct comparison of the extent to which our coders agree or disagree, by examining their underlying perceived level of corruption (the perceived latent trait) rather than their actual (raw) ratings. In other words, from the measurement model we can back out the coder-level estimates once variability in coder thresholds and reliability (error variance) has been controlled for. In Table 5, we perform the exact same variance decomposition on these adjusted corruption perceptions, scaled to vary from 0 to 1.[10] Again, we do not find the level of coder disagreement to be particularly high.

Next, again following Steenbergen and Marks [2007] (Table 3), and also Martinez i Coma and van Ham [2015] (Tables 2 and 4), we suggest testing the assumption that inter-expert disagreement varies systematically with the level of difficulty of the coder task. Thus, apart from concluding that overall

[10] In order to discount extreme outliers resulting from measurement model uncertainty, all estimates have been weighted by the inverse of the standard error of the perception estimate.

expert disagreement is not substantial, Steenbergen and Marks [2007] (p. 354) find that experts agree more when parties take more differentiated stands, when the issue position is more salient, and the party is relatively united on the issue – in other words, under conditions that make the judgment task of experts less difficult. Similarly, Martinez i Coma and van Ham [2015] find that experts agree more on factual than on evaluative questions, and when asked to assess more observable election day fraud as compared to pre-election integrity.

In the case of corruption validation, there are first and foremost two potential sources of "level of difficulty" we could expect to affect the degree of coder disagreement. The first relates to the amount of information on corruption available (e.g., Bollen [1986], Bollen [1993] or Bollen and Paxton [2000], although on the measurement of human rights abuses and democracy). With V-Dem data, there are (at least) two ways to proxy for the availability of information on corruption. The first proxy is time: we would, *ceteris paribus*, expect the experts to have more information about present-day than historical corruption, both due to their own lived experience and through the availability of other studies of corruption, which has arguably sky-rocketed in the last two decades. The second proxy is media freedom or, more generally, freedom of expression. In closed authoritarian systems, all else being equal, there should be less information available about corruption, since one of the goals of censorship or government violations of the freedom of expression is to conceal the extent and nature of corruption in the country. A second potentially systematic source of variation in coder-level disagreement on corruption is the level or frequency of corruption in the country in question. Arguably, non-corrupt and outrageously corrupt settings should elicit less disagreement among coders. The most difficult countries to assess, by contrast, should be countries that have intermediate levels of corruption.

We test these assertions in Table 6, also controlling for the number of coders, which in itself could be expected to drive up disagreement (in particular when taking lateral coding into consideration). Except for time, our expectations are mostly borne out by the data, and our findings are consistent for the raw ratings and perceptions data (hence, for presentational purposes we omit the results for raw scores). Coder disagreement is significantly lower in countries with widespread freedom of expression for three out of six corruption indicators, and also in the pooled analysis. Almost perfectly consistent across all indicators, the quadratic term for the level of corruption is negative and statistically significant, signifying that the largest amount of disagreement occurs in countries with intermediate levels of corruption.

The time-variable produces a more mixed pattern. For both raw scores and perceptions, it is negative and significant for judicial corruption, in line with the expectation that disagreement goes up when coding the distant past. In the perceptions measure, however, the coefficient is positive and significant for executive embezzlement, meaning that disagreement actually increases with time. And for the other four indicators, as well as for the pooled analysis, there is simply no significant effect of time. This result

| | (1) Exec. Bribery | (2) Exec. Theft | (3) Pub. Bribery | (4) Pub. Theft | (5) Legisl. Corr. | (6) Jud. Bribery | (7) Pooled |
|---|---|---|---|---|---|---|---|
| Year | 0.000 | 0.000** | -0.000 | 0.000 | 0.000 | -0.000** | -0.000 |
| | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) |
| Freedom of expression index | -0.052*** | -0.042 | -0.054*** | -0.055** | -0.012 | -0.017 | -0.039*** |
| | (0.015) | (0.030) | (0.018) | (0.023) | (0.015) | (0.022) | (0.009) |
| Level | 0.010** | -0.001 | -0.001 | -0.013 | 0.000 | -0.015* | -0.003 |
| | (0.004) | (0.010) | (0.006) | (0.008) | (0.005) | (0.008) | (0.003) |
| Level$^2$ | -0.051*** | -0.013 | -0.053*** | -0.031*** | -0.033*** | -0.048*** | -0.042*** |
| | (0.005) | (0.008) | (0.004) | (0.005) | (0.005) | (0.007) | (0.003) |
| No of coders | 0.006* | -0.000 | -0.002 | -0.000 | 0.002 | 0.003 | 0.001 |
| | (0.003) | (0.007) | (0.004) | (0.004) | (0.003) | (0.003) | (0.002) |
| Adjusted R-squared | 0.321 | 0.050 | 0.247 | 0.184 | 0.129 | 0.344 | 0.234 |
| No of countries | 173.000 | 118.000 | 168.000 | 163.000 | 161.000 | 169.000 | 173.000 |
| No of observations | 15653 | 6099 | 13901 | 10303 | 8748 | 15235 | 69939 |

Entries are regression coefficients, with standard errors, clustered on countries, in parentheses.
Indicator-fixed effects included in the pooled model but omitted from the table.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 6: Predicting Coder Disagreement

should caution the otherwise intuitive notion that the distant past is necessarily harder to code than the present. On the one hand, one could expect that with very little information available about an historical period, each expert's rating contains more guesswork – that is, a larger random component. In that sense, less information should lead to more disagreement. On the other hand, more information about the contemporary era also entails more potential for conflicting information, which could lead experts to disagree more. With respect to time at least, these two tendencies could cancel each other out, leading to no obvious patterns with respect to how much our coders disagree.

Overall, we conclude that coder disagreement is not critically high and that the disagreement does vary by level of difficulty in a meaningful way. This lends additional support to the validity of our data.

As a final examination of coders, we can also model the extent to which coder characteristics bias the coders away from the "true score." By including country- and year-fixed effects, so we fix the comparisons to the same countries and time periods, we can model the coder point estimates directly as a function of coder characteristics. This is the exact approach taken in Dahlström et al. [2012], who model expert survey responses to questions on bureaucratic recruitment patterns as a function of gender, age, education, state employment and whether the expert was born or resides in the country coded.[11]

Table 7 focus on the exact same coder characteristics as Dahlström et al. [2012], plus three attitudinal measures that might tap into ideological biases when coding a country's level of corruption: support for a free market; support for the principle of electoral democracy; and support for the principle of liberal democracy.

These results are overall good news. With few exceptions, coder characteristics do not predict our coders' actual ratings or perceived level of executive bribery, holding the frame of comparison (country and year) constant. In the raw score (results not shown), there is a tendency that female coders rate countries systematically lower than their male correspondents (meaning more corrupt), but when variable

[11] It is also very similar in spirit to Martinez i Coma and van Ham [2015] (Table 3), although they look at deviations from the coder mean with country random effects.

| | (1) Exec. Bribery | (2) Exec. Theft | (3) Pub. Bribery | (4) Pub. Theft | (5) Legisl. Corr. | (6) Jud. Bribery | (7) Pooled |
|---|---|---|---|---|---|---|---|
| Gender | -0.012 | -0.068** | -0.040** | -0.017 | -0.011 | -0.028* | -0.029*** |
| | (0.016) | (0.032) | (0.016) | (0.017) | (0.014) | (0.017) | (0.011) |
| Age | -0.000 | -0.004 | -0.004 | -0.006 | -0.005 | -0.005 | -0.004 |
| | (0.005) | (0.006) | (0.005) | (0.006) | (0.004) | (0.006) | (0.003) |
| $Age^2$ | -0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) |
| PhD education | 0.005 | 0.017 | 0.016 | -0.009 | -0.015 | 0.016 | -0.004 |
| | (0.021) | (0.024) | (0.017) | (0.017) | (0.017) | (0.018) | (0.012) |
| Government employee | -0.064** | 0.066 | 0.021 | 0.030 | -0.022 | -0.002 | 0.001 |
| | (0.028) | (0.043) | (0.029) | (0.033) | (0.031) | (0.029) | (0.021) |
| Born in country | 0.009 | 0.031 | 0.051*** | 0.044** | 0.024* | 0.001 | 0.024* |
| | (0.018) | (0.027) | (0.019) | (0.020) | (0.013) | (0.024) | (0.012) |
| Resides in country | 0.019 | 0.026 | -0.002 | -0.004 | -0.012 | 0.038* | 0.018 |
| | (0.018) | (0.034) | (0.019) | (0.019) | (0.014) | (0.021) | (0.013) |
| Supports free market | -0.002 | 0.014** | 0.013** | 0.007 | 0.013** | 0.005 | 0.005 |
| | (0.006) | (0.006) | (0.006) | (0.006) | (0.006) | (0.007) | (0.004) |
| Supports electoral democracy | 0.003 | 0.010 | 0.005 | 0.003 | -0.007 | -0.005 | -0.003 |
| | (0.006) | (0.008) | (0.007) | (0.008) | (0.007) | (0.010) | (0.005) |
| Supports liberal democracy | -0.011 | -0.004 | -0.003 | -0.013* | -0.013* | 0.001 | -0.007 |
| | (0.007) | (0.011) | (0.008) | (0.007) | (0.008) | (0.009) | (0.005) |
| Mean coder discrimination (beta) | 0.012 | -0.005 | -0.011 | -0.028** | -0.022* | 0.004 | -0.010 |
| | (0.012) | (0.013) | (0.008) | (0.011) | (0.012) | (0.013) | (0.007) |
| R-squared | 0.627 | 0.587 | 0.644 | 0.625 | 0.617 | 0.570 | 0.529 |
| No of countries | 173.000 | 119.000 | 168.000 | 163.000 | 161.000 | 169.000 | 173.000 |
| No observations | 70471 | 28258 | 64223 | 47982 | 41395 | 66937 | 319266 |

Entries are regression coefficients, with standard errors, clustered on countries, in parentheses. Year- and country-fixed effects included, indicator-fixed effects in the pooled model, but omitted from the table.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 7: Predicting Coder Ratings with Coder Traits

thresholds and coder reliability has been taken into consideration, the gender differences (presented) are only significant for three out of six indicators (although also in the pooled model). There is also a slight tendency among coders born in the country to rate that country as less corrupt than coders born in other countries, but this tendency only reaches conventional levels of statistical significance for two out of six indicators (and not in the pooled model).

There is also a tendency among free market believers to both rate and perceive countries as less corrupt, but not in the pooled model. And, interestingly, there is no inherent "democratic" bias in our coders' perceptions of corruption (neither of the electoral or liberal type). As a final check on ideological bias, we also test 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 could test whether ideological bias is geared towards certain types of countries. One could for example imagine that a strong believer in free markets would have no general tendency to rate countries as more or less corrupt, but a more specific tendency to rate countries with free markets as less corrupt. In other words, we might want to assess ideological bias conditional on country characteristics of the country being rated.

With the three measures of ideological bias available, there are three such interaction effects we might assess: one between support for free markets and a proxy for openness to trade (data from the Correlates Of War project); the other two between support for electoral/liberal democracy and the level of electoral/liberal democracy (data from V-Dem).[12]

[12]However, this test comes at a price: we cannot control for country-fixed effects. This means we will either have to revert to random effects, or simply assume that the country characteristics control for the relevant differences between countries in terms of average ideological bias of the coders. The latter approach is taken here. (Results for raw scores,

| | (1) Exec. Bribery | (2) Exec. Theft | (3) Pub. Bribery | (4) Pub. Theft | (5) Legisl. Corr. | (6) Jud. Bribery | (7) Pooled |
|---|---|---|---|---|---|---|---|
| Supports free market | 0.011 | 0.035** | 0.022** | 0.024* | 0.008 | 0.018* | 0.019** |
| | (0.014) | (0.014) | (0.010) | (0.012) | (0.014) | (0.011) | (0.009) |
| Openness to trade | 0.000** | 0.000 | 0.000*** | 0.000** | 0.000 | 0.000*** | 0.000*** |
| | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) |
| Supports free market×Openness to trade | -0.000 | -0.000 | -0.000 | -0.000 | 0.000 | -0.000 | -0.000 |
| | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) |
| Supports electoral democracy | -0.042 | -0.027 | -0.019 | -0.026 | -0.078** | -0.027 | -0.032** |
| | (0.029) | (0.022) | (0.017) | (0.021) | (0.032) | (0.023) | (0.014) |
| Electoral democracy index | -0.083 | 0.165 | 0.234 | -0.064 | -0.412* | -0.246 | -0.038 |
| | (0.220) | (0.231) | (0.179) | (0.247) | (0.249) | (0.232) | (0.155) |
| Supports electoral democracy×Electoral democracy index | 0.053 | -0.016 | 0.015 | 0.035 | 0.119** | 0.100** | 0.041 |
| | (0.048) | (0.037) | (0.036) | (0.056) | (0.055) | (0.048) | (0.028) |
| Supports liberal democracy | -0.008 | 0.048 | 0.023 | 0.022 | 0.000 | 0.016 | 0.015 |
| | (0.031) | (0.036) | (0.021) | (0.024) | (0.032) | (0.024) | (0.018) |
| Liberal component index | 0.582*** | 0.521* | 0.552*** | 0.748*** | 0.620*** | 0.463** | 0.605*** |
| | (0.198) | (0.298) | (0.147) | (0.162) | (0.191) | (0.193) | (0.144) |
| Supports liberal democracy×Liberal component index | 0.011 | -0.055 | -0.019 | -0.031 | -0.032 | -0.006 | -0.023 |
| | (0.043) | (0.063) | (0.032) | (0.031) | (0.045) | (0.034) | (0.025) |
| R-squared | 0.466 | 0.274 | 0.527 | 0.444 | 0.313 | 0.401 | 0.408 |
| No of countries | 148.000 | 108.000 | 144.000 | 143.000 | 140.000 | 149.000 | 149.000 |
| No of observations | 40548 | 22057 | 38177 | 34840 | 30567 | 38495 | 204684 |

Entries are regression coefficients, with standard errors, clustered on countries, in parentheses.
Year-fixed effects included, indicator-fixed effects in the pooled model, but omitted from the table.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 8: Predicting Coder Ratings with Coder and Country Traits

Results omitting all other coder characteristics are presented in Table 8. Again, these results are quite reassuring, and differ very little between raw scores and coder perceptions. That more "liberal" countries, which among other things mean that they abide by the rule of law, are considered as less corrupt on average should hardly come as a surprise. More importantly, there is no tendency among coders who strongly support this "liberal" principle to code or perceive more liberal countries differently than coders who do not exhibit such support. Similarly, more open economies are considered less corrupt, but this has no effect on how free market ideological bias affects coder perceptions. The one coder-country interaction we find that is statistically significant for both raw scores and perceptions is for electoral democracy and legislative corruption, where there is a tendency that strong believers in the principle of electoral democracy rate countries with high levels of electoral democracy as having less corruption in the legislature. There is also a similar tendency in the perceptions for judicial corruption. With these exceptions noted, there seems to be no overall ideological bias induced by the context of the country being coded.

#### 5.3.2.2 Comparing Measures

We now turn to a systematic comparison of how countries are rated according to the V-Dem Corruption Index on the one hand, and the WGI and CPI on the other. Assessing this second type of convergent validity is fundamentally about assessing comparative advantage: When embarking on using a dataset or measure for the first time, what are its strengths and weaknesses compared to existing datasets? What is gained by using this dataset instead of others?

where they can be tested, are very similar in spirit when run with country-random effects).

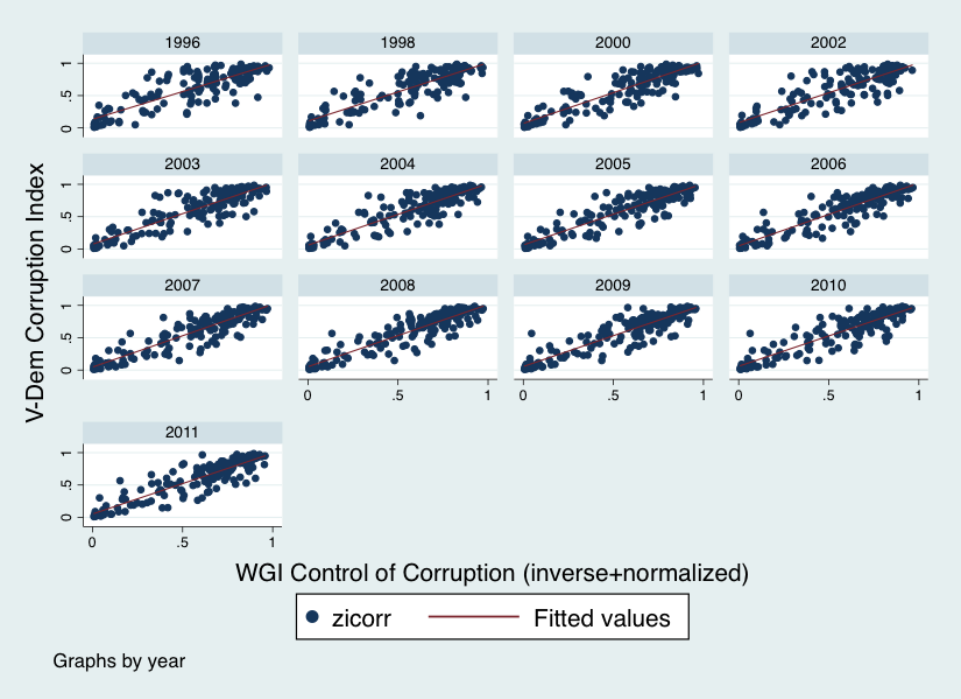


Figure 9: Comparing the V-Dem and WGI Measures of Corruption

Since the non-V-Dem data sources explicitly discourage comparisons over time, we perform these comparisons on a year-by-year basis, starting in the first year of measurement for the corresponding measure (1995 and 1996). As the simple bivariate scatterplots of Figure 9 and 10 make clear, the V-Dem measure overall aligns well with the other two. The pooled correlation coefficients are around .90 in both cases. In other words, countries rated as more corrupt by extant measures of corruption also to a large extent tend to be rated as more corrupt by the V-Dem index. This is thus clear evidence of convergent validity. Despite differences in exact question wording and methodology, the V-Dem measure of corruption paints an overall picture very similar to the ones from WGI and CPI.

However, the figures also make it clear that there are exceptions to the rule. Some countries, such as Malaysia and Qatar, are systematically rated as more corrupt by the V-Dem index than by other extant measures. Others, such as Latvia and Lithuania, are consistently considered less corrupt. This opens the question of what can explain the deviations. Explaining areas that lack convergence is as, or more, important as demonstrating strong correlations [Adcock, 2001, Bowman et al., 2005]. Put in technical terms, what can explain the residuals – the vertical distance to the regression line – in the year-by-year comparisons in Figure 9 and Figure 10? As Hawken and Munck [2009b] notes, "Consensus is not necessarily indicative of accuracy and the high correlations by themselves do not establish validity."

In Table 9 and Table 10, we extend the analysis of the effect of coder-level determinants to explaining deviations from the alternative corruption measure with the broadest coverage: WGI. We thus still

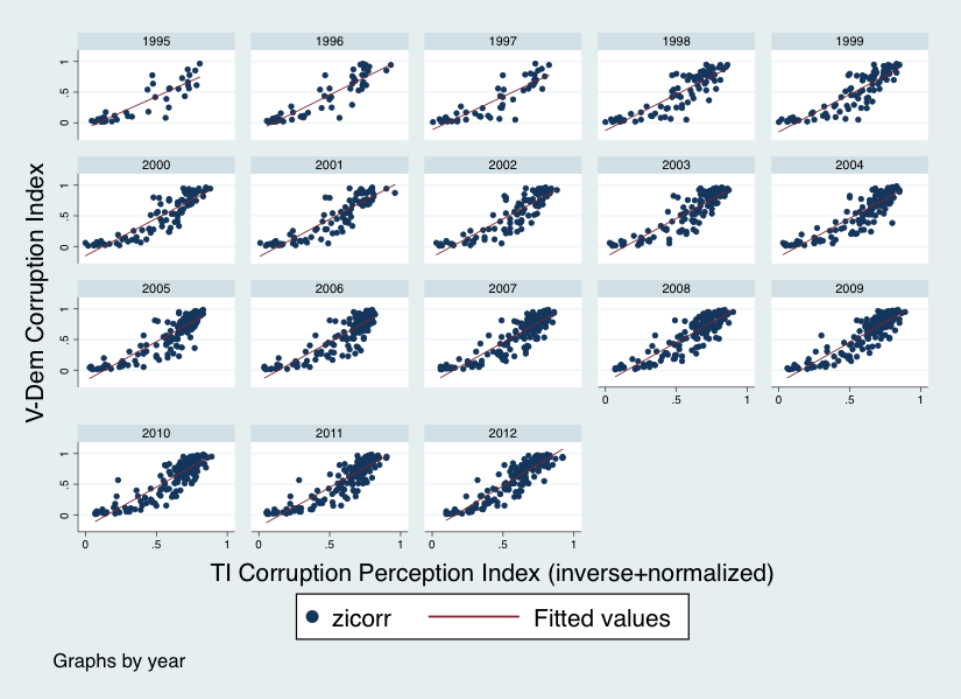


Figure 10: Comparing the V-Dem and CPI Measures of Corruption

operate at the individual coder-level of analysis, in order to avoid the ecological fallacy that would have resulted from running regressions on average coder characteristics across countries and years. We look at individual-level associations between coder characteristics and the tendency to code each V-Dem corruption indicator differently than the overall WGI measure, while controlling for average coder characteristics (or, as it were, the "composition" of V-Dem coders for each country). We also control for the aggregate country and year-level of disagreement among coders, the number of coders, their individual reliability (as measured by beta from the measurement model), and year-fixed effects.

As should be expected, there is a close correspondence between the WGI measure and our V-Dem corruption indicators. A unit increase in the WGI measure, roughly corresponding to a standard deviation, maps to an average increase in our corruption estimates by 0.215, roughly half a standard deviation. But beyond this, there is also a statistically significant tendency among V-Dem female coders to perceive the countries they are coding as scoring lower (meaning more corrupt) than the WGI measure on three of our six indicators, including the pooled one. It should be noted that this systematic pattern is not necessarily a sign of bias of the V-Dem measure. To the best of our knowledge, corruption perception measures have not been gendered before, but to the extent that incorporating more female views of corruption brings a less distorted picture of the true nature of corruption in the world, the female "bias" in the V-Dem data might be seen as a virtue. This is a topic worthy of further study.[13]

[13]Interestingly, the gender composition is signed in the opposite direction, meaning that the larger the share of female

| | (1) Exec. Bribery | (2) Exec. Theft | (3) Pub. Bribery | (4) Pub. Theft | (5) Legisl. Corr. | (6) Jud. Bribery | (7) Pooled |
|---|---|---|---|---|---|---|---|
| WGI control of corruption | 0.245*** | 0.144*** | 0.258*** | 0.238*** | 0.189*** | 0.240*** | 0.215*** |
| | (0.015) | (0.022) | (0.012) | (0.013) | (0.015) | (0.015) | (0.009) |
| Gender | -0.043** | -0.119*** | -0.040** | -0.004 | -0.025 | -0.016 | -0.041*** |
| | (0.019) | (0.035) | (0.019) | (0.021) | (0.022) | (0.021) | (0.013) |
| Share female coders | 0.168** | 0.114 | 0.100 | 0.168** | -0.021 | 0.111 | 0.129** |
| | (0.069) | (0.118) | (0.077) | (0.070) | (0.079) | (0.073) | (0.056) |
| Age | 0.007 | -0.005 | 0.001 | 0.007 | 0.010 | -0.004 | 0.002 |
| | (0.007) | (0.009) | (0.006) | (0.006) | (0.007) | (0.006) | (0.003) |
| $Age^2$ | -0.000 | 0.000 | -0.000 | -0.000 | -0.000 | 0.000 | -0.000 |
| | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) |
| Average age of coders | -0.004 | 0.013 | 0.033* | 0.004 | 0.041 | 0.045 | 0.031 |
| | (0.019) | (0.041) | (0.020) | (0.022) | (0.031) | (0.027) | (0.021) |
| Average age of coders×Average age of coders | 0.000 | -0.000 | -0.000 | 0.000 | -0.000 | -0.000 | -0.000 |
| | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) |
| PhD education | -0.018 | 0.021 | -0.007 | -0.055** | -0.007 | 0.039 | -0.007 |
| | (0.021) | (0.031) | (0.024) | (0.024) | (0.026) | (0.025) | (0.013) |
| Share of PhD coders | 0.062 | -0.173* | -0.053 | 0.036 | -0.036 | -0.101 | -0.023 |
| | (0.064) | (0.097) | (0.069) | (0.061) | (0.076) | (0.064) | (0.045) |
| Government employee | -0.020 | -0.021 | 0.055 | -0.004 | -0.011 | 0.039 | 0.009 |
| | (0.031) | (0.050) | (0.039) | (0.037) | (0.038) | (0.027) | (0.017) |
| Share of coders employed by government | -0.062 | -0.139 | -0.091 | -0.122 | 0.167 | -0.114 | -0.042 |
| | (0.111) | (0.214) | (0.116) | (0.116) | (0.135) | (0.103) | (0.078) |
| Born in country | -0.036 | 0.005 | 0.034* | 0.029 | 0.017 | -0.014 | -0.001 |
| | (0.027) | (0.036) | (0.021) | (0.024) | (0.025) | (0.027) | (0.014) |
| Share of coders born in country | 0.124* | 0.059 | 0.065 | 0.092 | -0.195** | -0.108 | -0.002 |
| | (0.066) | (0.121) | (0.065) | (0.072) | (0.090) | (0.091) | (0.060) |
| Resides in country | 0.013 | 0.008 | -0.007 | 0.016 | -0.022 | 0.049* | 0.014 |
| | (0.029) | (0.039) | (0.025) | (0.023) | (0.024) | (0.027) | (0.016) |
| Share of coders residing in country | -0.054 | 0.129 | -0.122* | -0.082 | 0.202*** | 0.054 | 0.036 |
| | (0.073) | (0.126) | (0.069) | (0.073) | (0.075) | (0.086) | (0.060) |
| Supports free market | 0.010 | 0.014 | 0.018** | 0.018* | 0.009 | 0.009 | 0.012* |
| | (0.012) | (0.013) | (0.009) | (0.009) | (0.011) | (0.010) | (0.007) |
| Average free market support | -0.043 | -0.047 | -0.045 | -0.021 | -0.071** | -0.001 | -0.034 |
| | (0.029) | (0.035) | (0.031) | (0.026) | (0.032) | (0.027) | (0.023) |
| Supports electoral democracy | 0.005 | -0.031** | 0.004 | -0.008 | -0.008 | 0.008 | -0.006 |
| | (0.009) | (0.013) | (0.010) | (0.009) | (0.013) | (0.012) | (0.006) |
| Average electoral democracy support | 0.007 | -0.012 | -0.004 | -0.011 | -0.022 | -0.029 | -0.016 |
| | (0.032) | (0.057) | (0.038) | (0.040) | (0.053) | (0.042) | (0.033) |
| Supports liberal democracy | -0.012 | 0.006 | 0.000 | -0.002 | -0.012 | 0.001 | -0.006 |
| | (0.012) | (0.015) | (0.010) | (0.011) | (0.012) | (0.013) | (0.006) |
| Average liberal democracy support | 0.002 | 0.033 | -0.000 | 0.015 | 0.034 | 0.037 | 0.018 |
| | (0.033) | (0.043) | (0.034) | (0.032) | (0.045) | (0.035) | (0.026) |
| Mean coder discrimination (beta) | -0.043*** | -0.042** | -0.031*** | -0.061*** | -0.045*** | 0.012 | -0.033*** |
| | (0.013) | (0.019) | (0.011) | (0.012) | (0.015) | (0.020) | (0.008) |
| Coder disagreement | 0.263* | 0.406** | 0.419*** | 0.251** | 0.538*** | 0.292* | 0.238** |
| | (0.142) | (0.198) | (0.155) | (0.113) | (0.206) | (0.162) | (0.100) |
| No of coders | -0.000 | -0.017 | 0.009 | 0.014 | -0.005 | -0.025** | -0.002 |
| | (0.008) | (0.011) | (0.008) | (0.009) | (0.011) | (0.013) | (0.006) |
| r2_a | 0.607 | 0.343 | 0.656 | 0.590 | 0.482 | 0.548 | 0.518 |
| N_clust | 163.000 | 114.000 | 159.000 | 154.000 | 151.000 | 160.000 | 164.000 |
| Observations | 10032 | 6619 | 9684 | 9333 | 8998 | 9569 | 54235 |

Entries are regression coefficients, with standard errors, clustered on countries, in parentheses.
Year-fixed effects included, indicator-fixed effects in the pooled model, but omitted from the table.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 9: Explaining Deviations from the WGI Corruption Measure with Coder Traits

| | (1) Exec. Bribery | (2) Exec. Theft | (3) Pub. Bribery | (4) Pub. Theft | (5) Legisl. Corr. | (6) Jud. Bribery | (7) Pooled |
|---|---|---|---|---|---|---|---|
| WGI control of corruption | 0.216*** | 0.094*** | 0.196*** | 0.182*** | 0.156*** | 0.204*** | 0.174*** |
| | (0.020) | (0.034) | (0.019) | (0.021) | (0.020) | (0.022) | (0.015) |
| Supports free market | 0.009 | 0.021 | 0.016 | 0.027** | -0.001 | 0.008 | 0.012 |
| | (0.017) | (0.018) | (0.013) | (0.012) | (0.014) | (0.013) | (0.009) |
| Openness to trade | -0.000 | -0.000 | -0.000 | 0.000 | -0.000 | -0.000 | -0.000 |
| | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) |
| Supports free market ×Openness to trade | 0.000 | -0.000 | 0.000 | -0.000* | 0.000 | 0.000* | 0.000 |
| | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) |
| Supports electoral democracy | -0.037 | -0.027 | -0.015 | 0.004 | -0.057* | -0.014 | -0.036** |
| | (0.031) | (0.038) | (0.024) | (0.035) | (0.034) | (0.026) | (0.017) |
| Electoral democracy index | -0.104 | -0.960 | -0.668 | -0.946 | -1.569** | -1.558** | -0.753 |
| | (0.605) | (1.001) | (0.732) | (0.773) | (0.614) | (0.657) | (0.573) |
| Supports electoral democracy×Electoral democracy index | 0.071 | -0.019 | 0.015 | -0.032 | 0.093* | 0.048 | 0.047* |
| | (0.046) | (0.064) | (0.039) | (0.057) | (0.056) | (0.047) | (0.028) |
| Supports liberal democracy | -0.037 | 0.029 | 0.002 | -0.049* | -0.027 | 0.029 | -0.016 |
| | (0.034) | (0.047) | (0.023) | (0.027) | (0.033) | (0.024) | (0.020) |
| Liberal component index | 0.079 | -0.466 | 0.549* | 0.493 | 0.404 | -0.201 | 0.438 |
| | (0.353) | (0.662) | (0.324) | (0.354) | (0.417) | (0.394) | (0.316) |
| Supports liberal democracy×Liberal component index | 0.043 | -0.028 | 0.007 | 0.080** | 0.032 | -0.055 | 0.019 |
| | (0.044) | (0.070) | (0.034) | (0.039) | (0.048) | (0.040) | (0.028) |
| Adjusted R-squared | 0.628 | 0.381 | 0.682 | 0.608 | 0.506 | 0.598 | 0.540 |
| No of countries | 144.000 | 106.000 | 140.000 | 139.000 | 135.000 | 145.000 | 145.000 |
| No of observations | 7500 | 5072 | 7257 | 7056 | 6686 | 7207 | 40778 |

Entries are regression coefficients, with standard errors, clustered on countries, in parentheses.
Year-fixed effects included, indicator-fixed effects in the pooled model, but omitted from the table.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 10: Explaining Deviations from the WGI Corruption Measure with Coder and Country Traits

Apart from this, there are very few systematic patterns pertaining to coder-level characteristics (when running so many statistical tests, we should of course expect some coefficients to be significant by chance). There is however a tendency across almost all indicators that country years where the V-Dem coders disagree end up being coded as less corrupt than the WGI measure. This could be something to keep in mind for future users of our data. Yet overall, the pattern is clear: there are few indications of systematic bias in our deviations from the WGI measure of corruption.

This also goes, finally, for the country-coder interactions tested in Table 10, that seek to determine whether deviations between V-Dem and WGI measures of corruption accrue from "situational closeness" as per above. Focusing on only the interaction terms, there is just one instance in which we find an effect that reaches conventional levels of statistical significance. This is for public sector theft or embezzlement (v2exthftps), where the V-Dem coders have a tendency to rate more liberal countries as less corrupt than WGI the stronger they themselves believe in liberal democracy. This is however a rare exception to a very strong rule: that there are no evident signs of context- or "similarity"-induced ideological biases in the V-Dem perceptions of corruption.

In sum, the results from these statistical analyses speak in favor of the validity of the V-Dem indicators. First, the degree of coder disagreement is comparatively small; and, when disagreement does exist, it varies meaningfully by context. Second, with the infrequent exception of gender, coders are not systematically affected by background or ideological factors. Finally, our V-Dem indicators correlate strongly with other measures of corruption, and the deviations from those measures cannot, by and large,

coders, the less corrupt the country (as compared to WGI). While highlighting the importance of controlling for ecological correlations, we interpret this as reflecting the fact that we have more female coders in developed, and hence less corrupt, countries.

be explained by the background traits or composition of V-Dem coders.

# 6 Conclusion

Greater attentiveness to data quality can improve political science research. This practical guide to data validation is a step in that direction. Rather than abstract advice or suggestions relevant only to creating a dataset, this guide describes and demonstrates an approach and tools to assess the strengths and weakness of existing datasets so that researchers can judge which dataset to use and how. Specifically, our method helps reveal systematic and random measurement error, in order to judge the validity and reliability of measures, respectively. To do so, we advocate for three approaches, each incorporating multiple tools: 1) assessing content validity through an examination of the resonance, domain, differentiation, fecundity, and consistency of the measure, 2) evaluating data generation validity through an investigation of dataset management structure, data sources, coding procedures, aggregation methods, and geographic and temporal coverage, and 3) assessing convergent validity using case studies and comparisons among coders and among measures.

Our application of this method to the V-Dem corruption measures demonstrates that they have high validity. It is nonetheless good to be aware of the set of country-year observations where our data are less reliable, as evidenced by less coder convergence. Users should consider the implications of our finding regarding female coders for a few of the indicators: a finding that could suggest more accurate data relative to those data generation processes that are less representative. The results have shown that researchers can be confident in using the V-Dem corruption measures for analyses across countries and over time. This advantage is significant considering that research questions have required such analyses and yet other datasets have not been designed for this purpose. Thus, the V-Dem corruption data have the potential to advance our understanding of corruption trends and causes considerably.

## References


R. Adcock. Measurement validity: A shared standard for qualitative and quantitative research. In *American Political Science Association*, volume 95, pages 529–546. Cambridge Univ Press, 2001.

Afrobarometer, Rounds 2-6, 2002-2015. `http://www.afrobarometer.org`.

A. Alam and V. Southworth. Fighting corruption in public services: Chronicling georgias reforms. *World Bank*, 2012.

Arab Barometer, 2006-2007, 2010-2014. `http://www.arabbarometer.org`.

Asiabarometer, 2003-2006. `http://www.asiabarometer.org`.

H. Bäck and A. Hadenius. Democracy and state capacity: Exploring a j-shaped relationship. *Governance*, 21(1):1–24, 2008.

P. Barron and B. Olken. The Simple Economics of Extortion: Evidence from Trucking in Aceh. CEPR Discussion Papers, 2007.

S. Ben-Ami. *Fascism from Above: The Dictatorship of Primo de Rivera in Spain, 1923–1930*. Clarenden Press, Oxford, 1983.

G. C. Benson, S. A. Maaranen, and A. Heslop. 1978.

K. Bollen. Liberal democracy: Validity and method factors in cross-national measures. *American Journal of Political Science*, pages 1207–1230, 1993.

K. A. Bollen. Political rights and political liberties in nations: An evaluation of human rights measures, 1950 to 1984. *Human Rights Quarterly*, 8(4):567–591, 1986.

K. A. Bollen and P. Paxton. Subjective measures of liberal democracy. *Comparative Political Studies*, 33(1):58–86, 2000.

K. Bowman, F. Lehoucq, and J. Mahoney. Measuring political democracy case expertise, data adequacy, and central america. *Comparative Political Studies*, 38(8):939–970, 2005.

M. Cabrera and F. del Rey Reguillo. *The Power of Entrepreneurs: Politics and Economy in Contemporary Spain*. Berghahn Books, New York, 2007.

R. Carr. *Modern Spain 1875-1980*. Oxford University Press, Oxford, 1980.

Caucasus Barometer, Caucasus Research Resource Centers, 2008 and 2010-2013. `http://www.crcenters.org/caucasusbarometer`.

E. C. Chang, M. A. Golden, and S. J. Hill. Legislative malfeasance and political accountability. *World Politics*, 62(02):177–220, 2010.

N. Charron and V. Lapuente. Does democracy produce quality of government? *European Journal of Political Research*, 49(4):443–470, 2010.

M. Chene. Anti-Corruption Progress in Georgia, Liberia, Rwanda. *U4*, 2011.

B. C. Chikulo. Corruption and accumulation in zambia. *Corruption and Development in Africa. Lessons from Country*, 2000.

A. Christiane et al. *Development Centre Studies Uses and Abuses of Governance Indicators*. OECD Publishing, 2006.

M. Coppedge. *Democratization and research methods*. Cambridge University Press, 2012.

M. Coppedge, J. Gerring, D. Altman, M. Bernhard, S. Fish, A. Hicken, M. Kroenig, S. I. Lindberg, K. McMann, P. Paxton, et al. Conceptualizing and measuring democracy: A new approach. *Perspectives on Politics*, 9(02):247–267, 2011.

M. Coppedge, J. Gerring, S. I. Lindberg, J. Teorell, D. Altman, M. Bernhard, M. S. Fish, A. Glynn, A. Hicken, C. H. Knutsen, K. McMann, D. Pemstein, M. Reif, S.-E. Skaaning, J. Staton, E. Tzelgov, and Y.-t. Wang. Varieties of Democracy: Codebook v3. Technical report, Varieties of Democracy (V-Dem) Project, 2014.

Corporacion Latinobarmetro, Latinobarmetro, 2001-2011, 2013, and 2015. `http://www.latinobarometro.org/lat.jsp`.

G. W. Cox and J. M. Kousser. Turnout and rural corruption: New york as a test case. *American Journal of Political Science*, pages 646–663, 1981.

C. Dahlström, V. Lapuente, and J. Teorell. Public administration around the world. *Good Government. The Relevance of Political Science. Cheltenham: Edward Elgar*, pages 40–67, 2012.

D. Donchev and G. Ujhelyi. What do corruption indices measure? *Economics & Politics*, 26(2):309–331, 2014.

J. Engvall. Corruption and what it means. *Silk Road Studies*, 2012.

European Bank for Reconstruction and Development, Business Environment and Enterprise Performance Survey, 1999-2014. `http://ebrd-beeps.com`.

C. J. Fariss. Respect for human rights has improved over time: Modeling the changing standard of accountability. *American Political Science Review*, 108(02):297–318, 2014.

C. Ferraz and F. Finan. Exposing Corrupt Politicians: The Effects of Brazil's Publicly Released Audits on Electoral Outcomes. *The Quarterly Journal of Economics*, 123(2):703–745, 2008.

Former Yugoslavia Barometer, 2005. `http://www.cspp.strath.ac.uk/quest-index-seb.html`.

F. Galtung. Measuring the immeasurable: boundaries and functions of (macro) corruption indices. *Measuring corruption*, pages 101–130, 2006.

J. Gerring. *Social science methodology: A criterial framework*. Cambridge University Press, 2001.

D. W. Gingerich. Governance indicators and the level of analysis problem: empirical findings from south america. *British Journal of Political Science*, 43(3):505–540, 2013.

R. K. Goel and M. A. Nelson. Corruption and government size: a disaggregated analysis. *Public Choice*, 97(1-2):107–120, 1998.

M. A. Golden and L. Picci. Proposal for a new measure of corruption, illustrated with italian data. *Economics & Politics*, 17(1):37–75, 2005.

M. Grossman. *Political Corruption in America: An Encyclopedia of Scandals, Power, and Greed*. ABC-CLIO, Santa Barbara, 2003.

A. Hawken and G. L. Munck. Measuring corruption: A critical assessment and a proposal. *Perspectives on Corruption and Human Development, Macmillan*, 2009a.

A. Hawken and G. L. Munck. Do you know your data? measurement validity in corruption research. *Unpublished typescript, Pepperdine University and University of Southern California, Malibu, CA, and Los Angeles*, 2009b.

Y. M. Herrera and D. Kapur. Improving data quality: actors, incentives, and capabilities. *Political Analysis*, 15(4):365–386, 2007.

P. M. Heywood. Continuity and change: Analysing political corruption in modern spain. In W. Little and E. Posada-Carbó, editors, *Political Corruption in Europe and Latin America*. St. Martin's Press, New York, 1996.

M. Huber. State-building in georgia: Unfinished and at risk. *Clingendeal Institute*, 2004.

F. Jiménez. Political scandals and political responsibility in democratic spain. *West European Politics*, 21(4):80–99, 1998.

D. Kaufmann and A. Kraay. Growth Without Governance. *World Bank Policy Research Working Paper*, (No. 2928), 2002.

D. Kaufmann and A. H. W. Stone.

D. Kaufmann, A. Kraay, and M. Mastruzzi. Worldwide Governance Indicators (WGI) Project, 1996-2013. `http://info.worldbank.org/governance/wgi/index.aspx#home`.

J. Kennedy. International Crime Victims Survey. *The Encyclopedia of Criminology and Criminal Justice*, 2014.

S. Knack. Measuring corruption: A critique of indicators in eastern europe and central asia. *Journal of Public Policy*, 27(03):255–291, 2007.

A. Kukhianidze. Corruption and organized crime in georgia before and after the rose revolution. *Central Asian Survey*, 28(2):215–234, 2009.

J. G. Lambsdorff. The Methodology of the Corruption Perceptions Index 2007. Technical report, Transparency International and the University of Passau, 2007.

A. D. Martin, K. M. Quinn, and J. H. Park. Mcmcpack: Markov chain monte carlo in r. 2011.

F. Martinez i Coma and C. van Ham. Can experts judge elections? Testing the validity of expert judgments for measuring election integrity. *European Journal of Political Research*, 54(2):305–325, 2015.

M. Mbao. Prevention and combating of corruption in zambia. *Comparative and International Law Journal of Southern Africa*, 44(2):255–274, 2011.

J. McMillan and P. Zoido. How to Subvert Democracy: Montesinos in Peru. *Journal of Economic Perspectives*, 18(4):69–92, 2004.

R. Menes. Corruption in cities: Graft and politics in american cities at the turn of the twentieth century. *NBER Working Paper*, 2003.

L. A. Mitchell. Compromising democracy: state building in saakashvili's georgia. *Central Asian Survey*, 28(2):171–183, 2009.

G. R. Montinola and R. W. Jackman. Sources of corruption: a cross-country study. *British Journal of Political Science*, 32(01):147–170, 2002.

J. Moreno-Luzón. *Modernizing the Nation: Spain During the Reign of Alfonso XIII, 1902–1931*. Sussex Academic Press, Eastbourne, UK, 2012.

C. Mudde and A. Schedler. Introduction: rational data choice. *Political Research Quarterly*, pages 410–416, 2010.

G. L. Munck and J. Verkuilen. Conceptualizing and measuring democracy evaluating alternative indices. *Comparative political studies*, 35(1):5–34, 2002.

New Baltic Barometer, 2001, 2004. `http://www.balticvoices.org/nbb/surveys.php`.

New Europe Barometer, 2001, 2004, 2005. `http://www.cspp.strath.ac.uk/nebo.html`.

New Russia Barometer, 2005-2012. `http://www.cspp.strath.ac.uk/catlog1_0.html`.

P. Niehaus and S. Sukhtankar. The Marginal Rate of Corruption in Public Programs: Evidence from India. *Journal of Public Economics*, 104(C), 2013.

B. Nyblade and S. R. Reed. Who cheats? who loots? political competition and corruption in japan, 1947–1993. *American Journal of Political Science*, 52(4):926–941, 2008.

B. A. Olken. Monitoring Corruption: Evidence from a Field Experiment in Indonesia. National Bureau of Economic Research, 2005.

S. G. Payne. *The Franco Regime, 1936–1975*. University of Wisconsin Press, Madison, 1987.

D. Pemstein, S. A. Meserve, and J. Melton. Democratic compromise: A latent variable analysis of ten measures of regime type. *Political Analysis*, 18(4):426–449, 2010.

D. Pemstein, E. Tzelgov, and Y.-T. Wang. Evaluating and improving item response theory models for cross-national expert surveys. 2014.

D. Pemstein, K. Marquardt, E. Tzelgov, Y. Wang, and F. Miri. Latent variable models for the varieties of democracy project. *Varieties of Democracy Institute Working Paper Series*, Forthcoming, 2016.

J. G. Peters and S. Welch. The effects of charges of corruption on voting behavior in congressional elections. *American Political Science Review*, 74(03):697–708, 1980.

P. Preston. *Franco: A Biography*. Basic Books, New York, 1994.

PRS Group, ICRG Methodology. `http://www.prsgroup.com/wp-content/uploads/2014/08/icrgmethodology.pdf`.

V. Pujas and M. Rhodes. Party finance and political scandal: Comparing italy, spain, and france. In A. J. Heidenheimer and M. Johnston, editors, *Political Corruption: Concepts and Contexts*, pages 739–760. Transaction Publishers, New Brunswick, 2002.

M. Razafindrakoto and F. Roubard. Are international databases on corruption reliable? a comparison of expert opinion surveys and household surveys in sub-saharan africa. *World Development*, 38(08): 1057–1069, 2010.

T. C. Reeves. *Twentieth-Century America: A Brief History*. Oxford University Press, New York, 2000.

R. Reinikka and J. Svensson. *Survey Techniques to Measure and Explain Corruption*, volume 3071. World Bank Publications, 2003.

M. T. Rock. Corruption and democracy social affairs. 2007.

A. Schedler. Judgment and Measurement in Political Science. *Perspectives on Politics*, 10(1):22–36, 2012.

J. Seawright and D. Collier. Rival strategies of validation: Tools for evaluating measures of democracy. *Comparative Political Studies*, 47(1):111–138, 2014.

S. Sequeira and S. Djankov. An Empirical Study of Corruption in Ports. *Working paper*, 2010.

N. Shahnazarian. Police reform and corruption in georgia, armenia, and nagorno-karabakh. *Policy Memo*, 232, 2012.

M. R. Steenbergen and G. Marks. Evaluating expert judgments. *European Journal of Political Research*, 46(3):347–366, 2007.

H.-E. Sung. Democracy and political corruption: A cross-national comparison. *Crime, Law and Social Change*, 41(2):179–193, 2004.

M. Szeftel. eat with us: Managing corruption and patronage under zambia's three republics, 1964-99. *Journal of Contemporary African Studies*, 18(2):207–224, 2000.

M. A. Tanner. *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions*. 1993.

J. Teorell and A. Hadenius. Pathways from authoritarianism. *Journal of democracy*, 18(1):143–157, 2007.

J. Teorell and B. Rothstein. Getting to sweden, part i: War and malfeasance, 1720-1850. *Scandinavian Political Studies*, forthcoming.

N. Towson. A third way? centrist politics under the republic. In *The Spanish Second Republic Revisited: From Democratic Hopes to the Civil War (1931–1936)*. ussex Academic Press, 2012.

Transparency International, Corruption Perceptions Index, 2012-2014. `http://www.transparency.org/research/cpi/overview`.

Transparency International, Global Corruption Barometer, 2004-2013. `http://www.transparency.org/research/gcb/overview`.

D. Treisman. 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, 2007.

W. M. Trochim and J. P. Donnelly. Research methods knowledge base. 2001.

United Nations Interregional Crime and Justice Research Institute, 1992, 1996, and 2000. `http://www.unicri.it/services/library_documentation/publications/icvs/data/`.

J. K. Van Donge. 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, 2009.

J. Whitten-Woodring and D. A. Van Belle. *Historical Guide to World Media Freedom: A Country-by-country Analysis*. CQ Press, 2014.

M. Woodiwiss. *Crimes, Crusades, and Corruption: Prohibitions in the United States, 1900–1987*. Barnes and Noble, New Jersey, 1988.

World Values Survey Association, World Values Survey, 1995-1998, 2010-2014. `http://www.worldvaluessurvey.org/wvs.jsp`.