The European Political Science Association, 2017 *doi:10.1017/psrm.2017.XX*

A Measurement Assessment Approach: Assessing The Varieties of Democracy Corruption Measures *

KELLY M. MCMANN, DANIEL PEMSTEIN, BRIGITTE SEIM, JAN TEORELL, AND STAFFAN I. LINDBERG

Social scientists face the challenge of assessing the quality of their measures, yet flexible and rigorous standards to do so remain elusive. This paper presents a three-component approach to measurement assessment, each component incorporating multiple tools: 1) assessing content validity by using face validity and factor analysis tools; 2) assessing the validity and reliability of the data generating process; and 3) assessing convergent validity with case studies and comparisons across coders and measures. We apply our approach to corruption measures from the Varieties of Democracy (V-Dem) Project, concluding the article by delineating the V-Dem Corruption Index's comparative strengths and limitations, including areas where its use may present different findings from extant corruption measures.

*Kelly M. McMann is Associate Professor of Political Science, Case Western Reserve University, Cleveland, Ohio 44106, USA (kmm39@case.edu). Daniel Pemstein is Assistant Professor of Political Science, North Dakota State University, Fargo, ND 58102, USA (daniel.pemstein@ndsu.edu). Brigitte Seim, **corresponding author**, is Assistant Professor of Public Policy, University of North Carolina, Chapel Hill, NC 27599, USA (bseim@ad.unc.edu). Jan Teorell is Professor of Political Science, Lund University, Lund, 221 00, Sweden (jan.teorell@svet.lu.se). Staffan Lindberg is Director of the V-Dem Institute, Political Science, University of Gothenburg, Gothenburg, 413 27, Sweden (staffan.i.lindberg@pol.gu.se). The authors are grateful to Gerardo Munck and other participants in the 2015 V-Dem Internal Research Conference for their comments and to Talib Jabbar and Andrew Slivka for their research assistance. This research project was supported by the Riksbankens Jubileumsfond, grant M13-0559:1 and the Knut & Alice Wallenberg Foundation (PI: Staffan I. Lindberg); the Swedish Research Council (PIs: Staffan I. Lindberg, Jan Teorell); National Science Foundation, grant SES-1423944 (PI: Daniel Pemstein); Wenner-Gren Foundation and the European University Institute's Fernand Braudel Senior Fellowship (PI: Jan Teorell). To view supplementary material for this article, please visit http://X

Most social scientists are concerned about the quality of their measures. Yet, as Herrera and Kapur (2007) wrote, “Inattentiveness to data quality is, unfortunately, business as usual in political science” (p. 366). Many scholars rely on pre-existing datasets in their research, and it is vitally important for us to understand how to responsibly use the measures provided by these datasets: how to diagnose measures’ strengths and limitations and implement strategies to mitigate concerns. A limiting factor is that there is no accepted, comprehensive approach in the field of political science for assessing a measure’s quality; the field also does not offer guidance about what to do with information gleaned from assessments.

The measurement assessment literature is extensive, but there are still several gaps. Many works examine only a single assessment tool, which leaves scholars to evaluate a given measure very narrowly or to puzzle over how to combine tools. Much of the literature implicitly or explicitly employs satisficing standards, asking: “What does a researcher have to do to show a measure is “valid *enough*” or “reliable *enough*”? Many existing approaches establish merely whether a particular measure is better (based on some criterion) than an extant alternative measure. Few researchers take what they learn in assessing the measure and actually incorporate their findings in analysis. Instead, the assessment serves only to put a rubber stamp on the measure.

To address these gaps in the measurement assessment literature, this paper proposes a set of complementary, flexible, practical, and methodologically rigorous tools for assessing the quality of a measure. Rather than recommending use of one tool over another, we advocate a comprehensive approach to assessment. Our approach is to assess content validity, the validity and reliability of the data generating process, and convergent validity, using a variety of tools. Particular innovations include a road map for evaluating the validity and reliability of the data generating process, a careful discussion of how one can use inter-coder agreement to assess both reliability and validity, and an expanded discussion of convergent validity assessment that advocates for the use of regression analysis and “blind” case studies to help explain where, and why, measures diverge.

We demonstrate use of our measurement assessment approach by evaluating a set of new corruption measures from the Varieties of Democracy (V-Dem) dataset. Corruption is a particularly difficult concept to measure, in part because of incentives for those engaged in it to hide the behavior and because of contextual variation in illicit practices. These challenges make assessment of corruption measures a particularly illuminating application of our guide.

Our assessment reveals both strengths and limitations of the V-Dem corruption measures. Specifically, assessing content validity shows that the V-Dem measures do not capture the “revolving door” phenomenon of corruption well, where public sector positions are used to secure private sector jobs and vice versa. Our assessments of the data generating process reveal that V-Dem coders disagree significantly on particular subsets of country-year observations, indicating potentially lower levels of validity and reliability on some data points. The analysis also shows that female coders rate countries, on average, as more corrupt on several of the V-Dem corruption measures, a possible source of deviations in the data. However, a primary strength of the V-Dem corruption measures is that, unlike other corruption measures, they are particularly well-suited for analysis across countries and over time. With the information presented here, scholars and practitioners can more effectively use the measures.

In sum, the paper delineates a practical measurement assessment approach and demonstrates its utility. Secondarily, it identifies strengths and limitations of the V-Dem corruption measures to facilitate their use by others. We proceed by showing how our approach both differs from and builds on the existing measurement assessment literature, describing our proposed approach, introducing the V-Dem corruption measures, and then applying our approach to those measures.

A Departure from and a Refinement of Previous Work

Some of the most valuable work on measurement assessment, such as Adcock and Collier (2001), provides advice, primarily or exclusively, about *developing* valid measures. Yet, with a proliferation of cross-national datasets and global indices, political and other social scientists are in dire need of advice on how to assess *existing* measures. Our guide focuses on that task.

The literature also generally overlooks practical, step-by-step guidance. Some of the most enlightening works, such as Seawright and Collier (2014), make us attentive to measurement assessment debates, inform us of different tools, and illustrate them. However, they are not assessment road maps, but rather an abstract presentation of assessment concepts. We extend this body of work. We provide a catalog of tools, apply them in a detailed manner, and demonstrate how this holistic approach reveals substantive insights important to conducting research with existing measures.

Further, offering a comprehensive approach is a helpful complement to publications that examine only a single tool (see, for example, Thomas (2010), who focuses entirely on assessing content validity). Our approach underscores the value of combining different measurement assessment tools, including harnessing the advantages of both qualitative and quantitative tools. We do not claim that our catalog of tools is exhaustive, but rather that it can serve as a relatively flexible foundation for assessing the quality of a measure.

Further, many prior works on measurement assessment also provide narrower guidance because they focus exclusively on validity, generally defined as the alignment between a measure and the underlying concept. For example, some of the most oft-cited articles on measurement in political science do not even mention reliability (Adcock and Collier 2001; Collier, LaPorte, and Seawright 2012; Seawright and Collier 2014). Similarly, in Chapter 7 of his canonical book, "Social Science Methodology," John Gerring acknowledges that validity and reliability are the "two overall goals" in "pursuing the task of measurement," but subsequently only discusses how to assess reliability for half of a page, concluding that inter-coder reliability tests should be performed whenever multiple coders are used (Gerring 2012, p. 158-159). The approach we lay out in this article illustrates the benefits of jointly assessing validity and reliability.

Finally, even the most insightful works on measurement do not take the critical post-assessment step of discussing how the assessment's findings can be incorporated into analysis. Existing measures are typically torn down without advice about how to use imperfect measures; much of the literature implies that a less-than-perfect measure is not worth using (Mudde and Schedler 2010). There is very little attention to how to mitigate, or at least acknowledge, inherited problems. This is true of even one of the more comprehensive and nuanced analyses of validation; Herrera and Kapur (2007) approach data collection "as an operation performed by data actors in a supply chain," delineating these actors, their incentives, and their capabilities (p. 366). They urge scholars to focus on validity, coverage, and accuracy, offering several examples of measures that have failed on these dimensions. They stop short, however, of explaining how to use information about measures' advantages and disadvantages in research. We do so.

We build on this prior research to develop our approach to measurement assessment. Our work is informed by the large literature on the quality of democracy measures, particularly its emphasis on aligning measures with higher-level conceptualization; considering differences in coverage, sources, and scales across measures; and transparency in coding and aggregation procedures (Munck and Verkuilen 2002; Bowman, Lehoucq, and Mahoney 2005; Coppedge et al. 2011; Fariss 2014; Pemstein, Meserve, and Melton 2010). To develop tools for assessing both the data generating process and convergent validity, we draw heavily on the work of Steenbergen and Marks (2007) and Martinez i Coma and Ham (2015), who represent literature on party positions and election integrity, respectively. Finally, we extensively borrow insights from the literature on corruption measurement, both because we apply our approach to corruption measures and because the literature raises general issues about measurement assessment (Knack 2007; Treisman 2007; Knack 2007; Hawken and Gerardo L Munck 2009; Hawken and Gerardo L. Munck 2009; Galtung 2006).

A Practical Guide to Measurement Assessment

We advocate for assessing the quality of a measure in three ways: content validation; data generating process assessment; and convergent validation. Collectively, these considerations illuminate the degree to which the measure is valid and reliable.

Validity can be thought of as the absence of systematic measurement error. Reliability can be thought of as precision, or the absence of unsystematic (or random) measurement error. Reliability should not be overlooked when assessing the quality of a measure; while precision is not useful on its own, neither is a well-conceptualized-but-imprecise-measure. A measure that offers a valid and reliable estimate of reality is preferable to a measure that offers unreliable, unbiased estimates or one that offers a reliable, biased estimate. And of course, the least usable measure is one that unreliably offers a biased estimate of reality.

First, it is helpful to examine the extent to which the measure captures the higher level theoretical concept. This can be done through a content validity assessment, where measures are mapped to abstract theoretical concepts and evaluated with face validity checks and factor analysis. In addition, we suggest assessing the content validity of the measure relative to other available measures.

Second, it is important to examine the data generating process for evidence of bias, unreliability, and aggregation inconsistency. An unbiased and reliable data generating process results in unbiased and reliable measures. The appeal of including this component in a measurement assessment approach is that it compels a focus on something that can be evaluated (i.e., the nature of a process) rather than something that cannot (i.e., a measure's alignment with the truth). For example, though we cannot prove that a 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. In evaluating the data generating process, we recommend scrutinizing the data management structure, data sources, coding procedures, aggregation procedures, and geographic and temporal coverage. Where multiple coders are used, we encourage examining the level of convergence across coders to evaluate the reliability of the data generating process and to expose potential determinants of systematic bias. In particular, considering the individual coder traits that predict disagreement, rather than simply the aggregate level of convergence in codes, allows researchers to identify threats to validity that are a function of the composition of their coder pools. As in the first component of our measurement assessment approach, the measure can be evaluated against objective standards, as well as assessing its strengths and limitations relative to other measures.

Third, we advocate for evaluating the quality of the measure in terms of whether it matches existing knowledge – an expanded convergent validity assessment. We use two tools in our convergent validity assessment: comparing the measure to existing comparable measures; and comparing the measure to actual cases. With regard to the former, it is important to acknowledge that the quality of other measures might not be certain. So, the task at hand is to evaluate the strength of correlations and any outliers in order to more completely understand the advantages and disadvantages of the measure of interest. A useful tool is to analyze the predictors of differences across measures, rather than only the aggregate correlation level. Qualitatively, original or existing case studies can be used for comparisons, to assess whether the measure "converges" with case history. However, in completing this case study analysis, we encourage the researcher to recode the cases independently, prior to examining alignment across the cases and the measure.

The three components, guiding questions and specific tools are outlined in Table 1. We now illustrate their utility and elaborate on them by applying the approach to assess the quality of the V-Dem corruption measures.

TABLE 1 *Measurement Assessment Approach*

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

Corruption Measures

First, we introduce the V-Dem corruption measures to be assessed. V-Dem provides six measures of corruption based on survey responses from country experts: two measures each for the executive and public sector – one for bribery and other corrupt exchanges, and another for theft and embezzlement – and a single measure each for legislative and judicial corruption. The online appendix includes the exact language from the survey instrument. The V-Dem Corruption Index then aggregates these low-level measures to produce an overall measure of corruption.[1] The V-Dem dataset covers all countries of the world, except micro-states, from 1900 to 2012.

As our approach to measurement assessment involves both stand-alone and comparative evaluations, Table 2 introduces the other corruption measures that we consider in our analysis.

[1] 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.

TABLE 2 *Alternative Corruption Measures*

Measure Name	WGI Control of Corruption (WGI)	TI Corruption Perceptions Index (CPI)	International Country Risk Guide	World Business Environment Survey	Global Corruption Barometer	Barometers	World Values Survey
Parent Organization	World Bank	Transparency International	Political Risk Services	World Bank	Transparency International	*N/A*	*N/A*
Data Sources	Surveys of households and firms, data from NGOs, public data	Other governance and business climate ratings and surveys	ICRG correspondents and staff	Survey of firms	Survey of households	Survey of households	Survey of households
Years of Data	1996-present	1995-present, but not comparable over time pre-2012	1984-present	1999-2000	2003-2007, 2009, 2010-2011, 2013	varies by region	1995-1998, 2010-present
Corrupt Actors	Government officials, elites, private interests	Public sector	“Political system”	Bureaucracy	Public sector and private “big interests”	Public sector; included public offices vary by year and region	Public sector, elections

Applying the Measurement Assessment Approach

We now apply our approach for assessing the validity and reliability of measures to the V-Dem corruption measures.

Content Validity Assessment

As a first component in our measurement assessment approach, we propose evaluating the degree to which an instrument generates a measure that maps to a theoretical construct – in other words, to determine the extent to which the measure captures all relevant meanings while excluding ones irrelevant to the "systematized" concept (Adcock and Collier 2001). We propose making this determination by using the tools of face validity and factor analysis. Face validity is a judgment call, made by either measure designers or users, that there is a correspondence between the measure and the systematized concept. Bayesian factor analysis is a statistical tool examining how closely different measures relate to the same underlying concept. A measure's quality can also be assessed comparatively, so we examine content validity in relation to other measures.

Applying this tool to the V-Dem measures, we find that the systematized concept for the V-Dem Corruption Index is the "use of public office for private gain," the common academic definition of corruption. This high-level measure of corruption captures a wide variety of participants in corruption and a large number of illicit practices, including both top officials and public sector employees to capture both grand and petty corruption. Each of the six V-Dem corruption measures refers to a particular public officeholder. And, they use specific language to indicate numerous, particular corrupt practices, such as "bribes" and "steal, embezzle, or misappropriate public funds or other state resources for personal or family use," as well as more general language to capture other forms of illicit behavior. This language enables the survey questions to generate measures that cover a wide range of meanings of the use of public office for private gain. However, the V-Dem measures do *not* capture "revolving door" corruption, where public sector positions are used to secure private sector jobs and vice versa, only for legislators, not other government officials.

The V-Dem measures exclude meanings of corruption that are irrelevant to the systematized concept. By specifying government officeholders, the instruments do not include use of nongovernmental positions, such as university admissions posts, for private gain. Likewise, they exclude cases where the position might be public or private, as is the case with media outlets. By specifying types of personal gain, the instruments also exclude behaviors where there is no evidence of direct, immediate material gain; this includes vote-buying that might not necessarily enrich oneself. 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.

Bayesian factor analysis also provides evidence that the six V-Dem corruption measures represent meanings relevant, and not irrelevant, to the systematized concept. The results

TABLE 3 *Measuring Corruption with V-Dem Data (BFA Estimates)*

Measure	Loadings (Λ)	Uniqueness (Ψ)
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, Quinn, and Park 2011). *n*=12,128 country years.

appear in Table 3. All six V-Dem corruption measures 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 the over-representation this set of measures gives to executive corruption.

Assessing content validity relative to other measures, we find that the V-Dem corruption measures are relatively comprehensive. By their own descriptions, many of the other corruption measures include information about "public sector" or bureaucratic corruption, excluding executive, legislative, and judicial corruption. This includes Transparency International's Corruption Perceptions Index (CPI), the World Bank's Business Environment and Enterprise Performance Survey (BEEPS), and nearly all the Barometers.[2] It is more difficult to determine the corruption construct the other measures capture because of the ambiguous language in associated documentation: Transparency International's Global Corruption Barometer (GCB) combines data on the public sector with private "big interests," and International Country Risk Guides' Political Risk Services (ICRG) focuses on the "political system." The World Values Survey (WVS) offers a more transparent and expansive conceptualization, including petty and grand corruption and capture of government institutions by private interests. Problematically, some measures used in studies as general measures of corruption actually capture a very narrow slice of "the use of public office for private gain." For example, the International Crime Victims Survey asks only about exposure to bribery (Kennedy, 2014). Overly narrow measures 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.

The exclusion of irrelevant meanings is also a strength of V-Dem corruption measures. Measures from other sources often include superfluous information. For example, the

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

Worldwide Governance Indicators' Control of Corruption (WGI) mixes electoral corruption, which does not necessarily involve private gain, along with public sector corruption.

In sum, this section demonstrated the utility of face validity tests, (Bayesian) factor analysis, and comparison with other measures to assess content validity. V-Dem corruption measures, in particular, have strong content validity in that they are comprehensive yet specific and align with the predominant definition of corruption, although they do not provide a measure of revolving-door corruption.

Data Generating Process Assessment

The second component in our measurement assessment approach is to evaluate whether each step of the data generating process is both unbiased and reliable. These steps include the dataset management structure, data sources, coding procedures, aggregation procedures, and geographic and temporal coverage. When measures draw upon the contributions of multiple coders, we also recommend leveraging information about coder disagreement to assess both the validity and reliability of the data generating process. We illustrate this by evaluating the V-Dem data generating process and highlighting its strengths and limitations relative to other corruption data sources.

Dataset Management Structure. Often overlooked sources of bias are the leadership and funding for a dataset. Hawken and Gerardo L. Munck (2009) find significant differences across corruption datasets, based on who is doing the evaluating. V-Dem itself is an academic venture, led by four professors as principal investigators and 12 scholars from 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, North America, and South America, but also from global organizations with members from all regions of the world. Leadership that is academic, rather than political or for-profit, and funding that is from diverse regions of the world, rather than from a single region or country, help to ensure that the organizational structure generates unbiased and reliable measures.

Data Sources. A key question to consider when evaluating potential bias and unreliability due to data sources is whether the measures 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 composite measure (Treisman 2007; Herrera and Kapur 2007; Hawken and Gerardo L. Munck 2009). V-Dem avoids this problem because it produces original corruption measures. This is a strength, relative to many existing corruption measures. Three of the most commonly used corruption measures – WGI, CPI, and ICRG – aggregate information from different sources.

Coding Procedures. When coders generate data, 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 measure (Treisman 2007; Martinez i Coma and Ham 2015). We consider each of these below.

Several scholars have argued that expert-coded measures of corruption are inferior to citizen-coded or "experience" measures (Treisman 2007; Hawken and Gerardo L Munck 2009; Hawken and Gerardo L. Munck 2009; Donchev and Ujhelyi 2014). Rather than privileging one type of coder over another, we recommend considering which type of coder is a good match for generating the measure of interest. For example, with respect to corruption measure, citizen coders offer certain disadvantages. Citizen perceptions of corruption are fundamentally limited because they interact with only certain kinds of officials and observe certain kinds of corruption. Moreover, corruption measures obtained from citizens are systematically biased because corruption is likely more removed from citizens' lives in countries with established institutions, stable incentive structures, and experienced public officials. As a result, cross-national measures of corruption based on citizen reports will over-estimate corruption in consolidating democracies and under-estimate it in stable democracies. Alternately, the potential disadvantage of far-removed experts coding conditions in a country can be addressed by relying on experts who are residents or nationals of the countries – effectively serving as both expert coders and citizen respondents.

Given the use of expert coders, then, to what extent do V-Dem coding procedures produce valid corruption measures? V-Dem relies on expert evaluation of corruption. The stringent selection criteria for experts could offset some of the biases common to other expert-coded measures. The experts have been recruited based on their academic or other credentials as field experts in the area for which they code and on their seriousness of purpose and impartiality (Coppedge et al. 2015). Impartiality is not a criterion to take for granted in political science research. Martinez i Coma and Ham (2015) noted that variance in estimates of election integrity in the Perceptions of Electoral Integrity dataset was significantly higher when one of the coders was a candidate in the election. Understanding the background, incentives, and biases of the V-Dem coders is critically important in evaluating the validity and reliability of V-Dem measures.

A key feature of the V-Dem enterprise is that no one coder's background or biases will drive the estimates for a given country. At least five V-Dem experts code each question-country-year observation for a total of more than 2000 experts involved to produce the dataset. As a rule, at least three-fifths of the experts coding a particular country either are nationals of or reside in the country. V-Dem thus taps into a local source of expertise and knowledge on corruption, avoiding the problems of far-removed experts and of citizen coders. Further, using multiple coders facilitates inter-coder reliability tests.

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

To aggregate up from coders to the level of country-years, V-Dem use statistical item response theory (IRT) techniques to model variation in coder reliability while allowing for the possibility that raters apply ordinal scales differently (Pemstein et al. 2015). The model uses bridge raters, who rate multiple countries for many years, to calibrate estimates across countries. It also uses lateral coders. These, in addition to providing a time-series for one country, provide single-year ratings for a number of other countries (Pemstein, Tzelgov, and Wang 2014). Thus the model explicitly attempts to adjust for DIF across raters. Furthermore, these procedures allow V-Dem to produce estimates of uncertainty around measures that are available to users and that can assist researchers in weighing the relative quality of measures across cases.

Aggregation Model. Many extant datasets, including V-Dem, offer low-level measures that they combine into higher-level measures. To assess the validity and reliability of the resulting measures, it is important to consider a) the choice of measures to aggregate and b) the aggregation rules.

V-Dem chose measures for its Corruption Index based on the conceptualization of corruption, as described in the content validity section. V-Dem aggregates corruption measures using a two-stage approach. First, V-Dem uses IRT methods to aggregate individual codes into low-level measures. Second, V-Dem uses Bayesian factor analysis to aggregate individual measures into a higher-level measures, using the method of composition (Tanner 1993) to propagate estimation uncertainty in the first stage into the resulting high-level measures, providing users with estimates of measure reliability. Although results in Table 3 would support a simple additive measure or measure based on factor analysis, to weight the measures, V-Dem employed a theory-based strategy. In particular, the executive corruption measure (**v2x_execorr**) was constructed by fitting a factor analysis model to the measures for executive bribery (**v2exbribe**) and executive embezzlement (**v2exembez**). The model estimates the posterior distribution of the latent factor score for each observation (country-year); one can use these to produce point estimates (posterior averages) and estimates of uncertainty (standard deviations and highest posterior density regions). V-Dem builds the public corruption measure (**v2x_pubcorr**) similarly, basing the measure on the estimated latent factor scores from a model incorporating low-level measures of public sector bribery (**v2excrptps**) and embezzlement (**v2exthftps**). Finally, to construct the overarching Corruption Index (**v2x_corr**), V-Dem averages (a) the executive corruption measure (**v2x_execorr**), (b) the public sector corruption measure (**v2x_pubcorr**), (c) the measure for legislative corruption (**v2lgcrrpt**), and (d) the measure for judicial corruption (**v2jucorrdc**). In other words, V-Dem weighs each of these four spheres of government equally in the resulting V-Dem Corruption Index. Further, both WGI and CPI choose measures that reduce missingness (Hawken and Gerardo L. Munck 2009). V-Dem does not have such a constraint, as the level of missingness does not vary greatly from one measure to another.

Coverage Across Countries and Time. It is important to consider potential biases introduced by limited geographic or temporal coverage of a measure. Particularly with sensitive topics, such as corruption, choosing cases can introduce selection bias. It is also important to assess how a measure anchors cases to a consistent scale (Treisman 2007). Thus, maximizing case coverage also improves measurement validity. The V-Dem corruption measures perform well on the question of coverage. V-Dem covers 173 countries across the globe, avoiding the bias implicit in measures that cover only a subset of countries (those easiest to code or those for which coders are readily available).[3] V-Dem also helps ensure reliable and unbiased measures by using the same coder recruitment procedures and coding rules across countries and time. By asking the same questions of each coder for each country-year, V-Dem allows over-time and cross-country comparisons of corruption levels in the world back to 1900. V-Dem uses an IRT-based measurement modeling strategy to anchor ratings to a consistent scale and, to further facilitate cross-national research, V-Dem is in the process of implementing anchoring vignettes across all coders.

The quality of V-Dem corruption measures for analysis across space and time is one of their key strengths. 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. Measures of corruption are typically taken at the country level, where comparisons across countries often come at the expense of comparisons over time (Christiane and Oman 2006; Galtung 2006; Knack 2007). For example, WGI is calculated such that the global average is the same every year; changes in

[3] The countries omitted currently from V-Dem are micro-states. Among the countries covered by V-Dem, there is only one case of missing data in the V-Dem Corruption Index: East Timor prior to independence.

the level of corruption within a country are not revealed unless the change is so great as to move it up or down in the comparative rankings (Lambsdorff 2007). Kaufmann and Kraay (2002) estimate that half the variance in WGI over time is the product of changes in the sources and coding rules used, rather than actual changes in corruption levels. Treisman (2007) notes that CPI's aggregation procedures and data sources have changed over time. Finally, WGI forces a consistent global average over time, preventing by construction an understanding of trends.

We consider what we can learn about trends in corruption levels, given this comparative strength of V-Dem to facilitate over-time analysis. According to the V-Dem Corruption Index, corruption levels have risen globally since at least the 1960s, with a peak just around the time when corruption appeared on the global reform agenda (Figure 1). We discuss this trend and what can be learned from it more extensively in the online appendix. Here, we simply note that this kind of analysis is possible given standards within V-Dem for coders, coding procedures, and aggregation procedures.

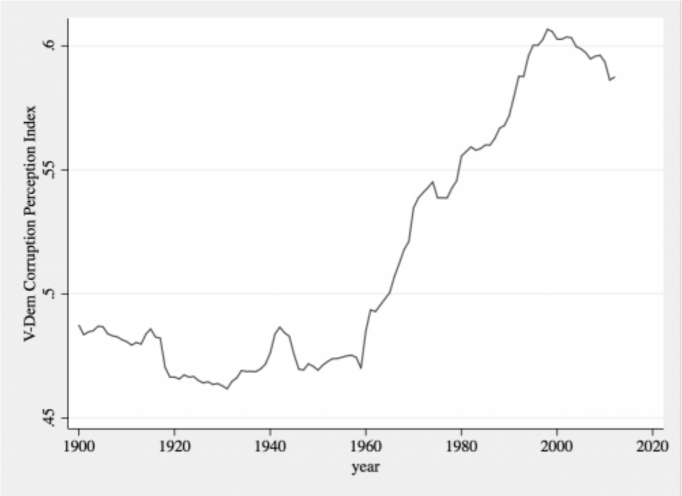

Figure 1. Global Levels of Corruption, 1900-2012

Examining Coder Disagreement. Conducting an analysis of coder (dis)agreement allows for one to empirically examine the data generating process. Unlike Steenbergen and Marks (2007) and Martinez i Coma and Ham (2015), who primarily compare ratings across coders as a test of validity, we argue that inter-coder agreement provides insight into both validity and reliability, another advantage of multi-coder measures. Clearly, a measure is more reliable when inter-coder disagreement is low. Inter-coder agreement can also be

TABLE 4 *Variance Decomposition of Adjusted Expert Ratings*

	Exec. Bribery	Exec. Theft	Public Bribery	Public Theft	Leg. 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)
Measure							0.030*
							(0.001)
No. Experts	924	600	903	847	877	872	1346
No. Observations	57290	28843	52976	44614	56989	32000	272711

Note: Variance decomposition with country- and year-fixed effects, coder- and measure-random effects. * $p < 0.05$.

seen as a measure of validity if one is willing to assume that multiple coders are unlikely to exhibit identical biases.[4] When coder traits predict disagreement systematically, this provides insight into potential sources of bias. We demonstrate each of these tools below.

The V-Dem measures rely on a measurement model that corrects for systematic threshold bias, or the tendency of coders' to be more or less strict in their application of ordinal scales. Using the model, we can estimate coders' "perceptions" of corruption, after controlling for fixed threshold bias (Johnson and Albert 1999). Table 4 displays a variance decomposition of these adjusted ratings, scaled to vary from 0 to 1.[5] After adjusting for differences in how coders use ordinal scales, we find little coder disagreement overall.

Following Steenbergen and Marks (2007) and Martinez i Coma and Ham (2015), we suggest testing if inter-expert disagreement varies systematically with the level of difficulty of the coder task. In the case of corruption validation, two potential sources of difficulty stand out. The first is the availability of information (e.g., K. A. Bollen (1986), K. Bollen (1993) or Bollen and Paxton (2000)). There are (at least) two ways to proxy for this. The first proxy is time; we would, *ceteris paribus*, expect the experts to have more information about present-day than historical corruption. The second proxy is media freedom, or freedom of expression.

A second potentially systematic source of variation in coder-level disagreement about corruption is the level of corruption itself. Non-corrupt and outrageously corrupt settings will elicit less disagreement than those with intermediate levels of corruption. This point

[4] The plausibility of this assumptions will vary across applications and requires careful consideration.

[5] Discounting extreme outliers resulting from measurement model uncertainty, we weighted estimates by the inverse of the standard error of adjusted estimates. A decomposition of raw scores yields similar results.

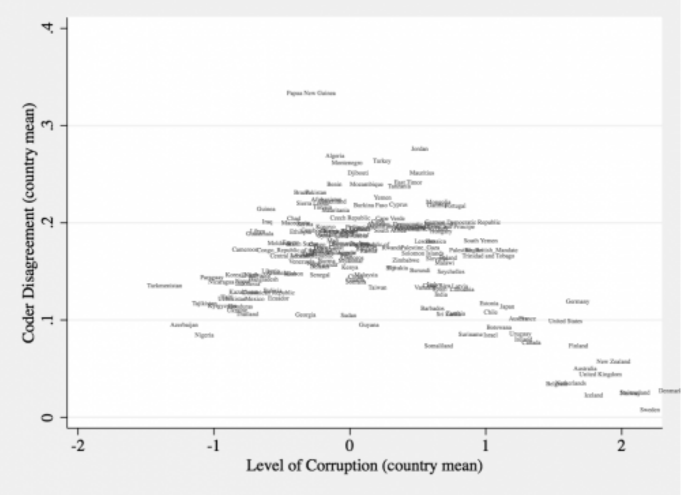

Figure 2. Coder Disagreement by Corruption Level

is illustrated in Figure 2, which shows that mean coder disagreement by country is highest at middle levels of corruption.

We test these assertions in Table 5,[6] controlling for the number of coders. Excepting time, our expectations are mostly borne out by the findings, and our findings are consistent for the raw and measurement model-adjusted ratings.[7] Coder disagreement is statistically significantly lower in countries with widespread freedom of expression for three of six corruption measures, and in the pooled model, indicating that for some measures, limited access to information influences coders' evaluations. The quadratic term for the level of corruption is negative and statistically significant, indicating that the most disagreement occurs in countries with intermediate levels of corruption.

The time variable produces a more mixed pattern across individual low-level corruption measures (online appendix), and the coefficient for the pooled model is statistically insignificant. This result qualifies the notion that the distant past is harder to code than the present.

Overall, we conclude that coder disagreement is not critically high and that it does not vary with difficulty in a meaningful way. These findings help to establish the reliability of

[6] Table 5 displays results from a pooled model including all V-Dem corruption measures. The online appendix provides results for each individual measure.

[7] Table 5 presents the results for raw scores.

TABLE 5 *Predicting Coder Disagreement*

	DV: Absolute Coder Disagreement
Year	-0.000
	(0.000)
Freedom of expression	-0.039*
	(0.009)
Level	-0.003
	(0.003)
Level^2	-0.042*
	(0.003)
No of coders	0.001
	(0.002)
Adjusted R-squared	0.234
No. Countries	173
No. Observations	69939

Note: Entries are regression coefficients, with standard errors, clustered on countries, in parentheses. Measure-fixed effects omitted from the table. $^* p < 0.05$.

the V-Dem data generating process, both across raters and cases. Given strong assumptions, they also help to establish validity.

We can extend this analysis by modeling the determinants of individual coder deviations from typical responses, illuminating sources of coder bias, and identifying potential threats to validity. Evidence of systematic disagreement will indicate bias. On the other hand, if patterns of disagreement are stochastic, we need not worry that certain types of coders are over or under-represented across cases.[8]

We model the extent to which coder characteristics bias the coders away from the "true," or typical, score. By including country- and year-fixed effects, we model the coder point estimates as a function of coder characteristics. Dahlström, Lapuente, and Teorell (2012) also employ this strategy[9]

Table 6 depicts a model that predicts adjusted coder ratings, pooled across measures,[10] and focuses on the same coder characteristics as Dahlström, Lapuente, and Teorell (2012), plus three attitudinal measures: support for a free market; support for the principle of

[8] Here we assume that the coder recruitment process is otherwise of high quality and draws from a cross-section of relevant types of coders, as described above.

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

[10] The online appendix provides models disaggregated by measure.

TABLE 6 *Predicting Coder Ratings with Coder Traits*

	DV: Coder Ratings
Gender	-0.029*
	(0.011)
Age	-0.004
	(0.003)
Age^2	0.000
	(0.000)
PhD education	-0.004
	(0.012)
Government employee	0.001
	(0.021)
Born in country	0.024
	(0.012)
Resides in country	0.018
	(0.013)
Supports free market	0.005
	(0.004)
Supports electoral democracy	-0.003
	(0.005)
Supports liberal democracy	-0.007
	(0.005)
Mean coder discrimination (beta)	-0.010
	(0.007)
R-squared	0.529
No. Countries	173
No. Observations	319266

Note: Entries are regression coefficients, with standard errors, clustered on countries, in parentheses. Year- and country-fixed effects, and measure-fixed effects are omitted from the table. * $p < 0.05$.

electoral democracy; and support for the principle of liberal democracy.

With a few exceptions, coder characteristics do not predict V-Dem coders' adjusted score for executive bribery, holding country and year constant. Female coders rate countries systematically lower (meaning more corrupt) than men, along with a slight tendency to rate one's own country as less corrupt. Interestingly, there is no "democratic" bias in V-Dem coders' adjusted ratings of corruption.

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.

TABLE 7 *Predicting Coder Ratings with Coder and Country Traits*

	DV: Coder Ratings
Supports free market	0.019*
	(0.009)
Openness to trade	0.000*
	(0.000)
Supports free market × Openness to trade	-0.000
	(0.000)
Supports electoral democracy	-0.032*
	(0.014)
Electoral democracy	-0.038
	(0.155)
Supports electoral democracy × Electoral democracy	0.041
	(0.028)
Supports liberal democracy	0.015
	(0.018)
Liberal democracy	0.605*
	(0.144)
Supports liberal democracy × Liberal democracy	-0.023
	(0.025)
R-squared	0.408
No. Countries	149
No. Observations	204684

Note: Entries are regression coefficients, with standard errors, clustered on countries, in parentheses. Year-fixed effects and the same coder traits as in Table 6 are included. Measure-fixed effects are omitted from the table. * $p < 0.05$.

The V-Dem post-survey questionnaire allows us to evaluate three such interactions: whether those who support free markets provide different corruption ratings for free trade economies (using a measure for trade openness from the Correlates of War project); whether those who support the principles of electoral democracy tend to provide different corruption ratings for electoral democracies; and whether those who support the principles of liberal democracy tend to provide different corruption ratings for liberal democracies.[11]

We present results of the analysis considering how coder and country traits might interact in Table 7.[12] Again, these results are quite reassuring. Unsurprisingly, coders

[11] However, these tests come at a price: we cannot control for country-fixed effects.

[12] The results presented here are for a model pooled across corruption measures. Models for each individual measure appear in the online appendix.

consider more “liberal” countries less corrupt. More importantly, coders who strongly support this “liberal” principle do not code or perceive more liberal countries differently than coders who do not exhibit such support. Coders consider more open economies less corrupt, but this has no effect on how free market ideological bias affects ratings. With these exceptions noted, there seems to be no overall ideological bias introduced by the context of the country being coded. This kind of analysis relies on the availability of observable coder-level covariates. It is useful for measurement assessment when datasets provide this coder-level information.

Convergent Validity Assessment

Our final measurement assessment component asks: To what extent do the measures correspond to existing knowledge? First, we suggest conducting a traditional convergent validity analysis, visually and statistically comparing the new measure to extant ones. Second, we recommend statistically examining the extent to which observable aspects of the data generating process predict systematic divergence between new and extant measures. Finally, we recommend examining the convergence between the measure and original or existing cases studies.

Basic Quantitative Convergent Validity. A traditional convergent validity test aims to assess whether various measures appear, on aggregate, to tap into the same concept. However, an aggregate convergent validity assessment can also be used to examine a measure’s comparative advantage: When using a measure for the first time, what are its strengths and limitations compared to existing measures? What is gained by using this measure instead of others?

Since the measures most comparable to the V-Dem Corruption Index, WGI and CPI, explicitly discourage comparisons over time, we assess aggregate convergent validity on a year-by-year basis. As Figure 3 and Figure 4 show, V-Dem and extant measures agree about which countries are more corrupt. Both pooled correlation coefficients are around 0.90: clear evidence of convergent validity. Nonetheless, there are differences in how V-Dem compares to WGI versus CPI. The deviations from WGI are more uniformly distributed over the range of the V-Dem Corruption Index, whereas the V-Dem Corruption Index is systematically lower than CPI for countries with a moderate level of corruption, and systematically higher for countries with extreme levels of corruption.

Two measures can be highly correlated at the aggregate level but systematically differ from one another in important ways. We therefore see standard quantitative convergent validity assessments as only the first step in a more comprehensive convergent validity analysis.

Statistical Analysis of Measure Convergence. Explaining areas that lack convergence is as, or more, important as demonstrating strong correlations (Adcock and Collier 2001;

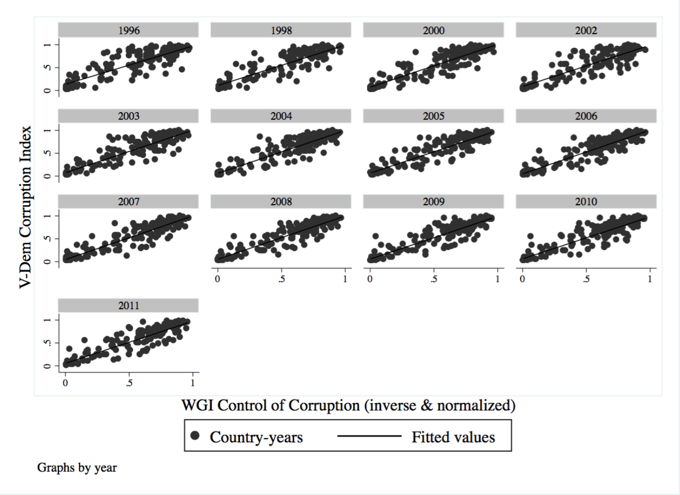

Figure 3. *Comparing the V-Dem Corruption Index and WGI Control of Corruption Index*

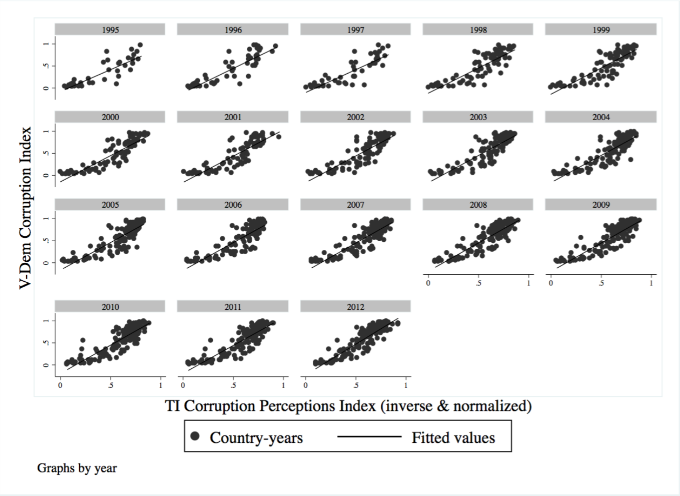

Figure 4. *Comparing the V-Dem Corruption Index and TI Corruption Perceptions Index*

Bowman, Lehoucq, and Mahoney 2005). While one rarely has access to a "gold standard" against which to assess convergence, researchers can model systematic determinants of divergence. In Table 8, we extend the analysis of the effect of coder-level determinants to explain deviations from WGI, which has the broadest coverage.[13,14] We ask whether the composition of V-Dem coders per country and year, measured with average coder traits, affects the tendency for V-Dem to deviate from WGI. In other words, what can explain the absolute residuals in the year-by-year comparisons in Figure 3?

Consistent with the finding that women rate corruption differently than men, the gender composition coefficient is positive and statistically significant; the larger the share of female coders, the larger the absolute difference between V-Dem and WGI. This is not necessarily a sign of bias in the V-Dem Corruption Index, and could even be seen as a virtue. This is a topic worthy of further study.

Otherwise, there are few systematic patterns in coder composition. Notably, V-Dem coder disagreement is a statistically significant predictor of the absolute residual between V-Dem and WGI. Disagreement may be most common in hard-to-rate cases, a finding more indicative of stochastic error than systematic bias. On the other hand, WGI and V-Dem disagree less when V-Dem relies more heavily on PhD-holding coders. Insofar as PhD holders are the "correct" set of experts, this result may indicate that including other coders may systematically bias V-Dem.[15] Yet overall, the pattern is clear: there are few systematic predictors of the deviations between WGI and V-Dem Corruption Index.

Convergent Validity Testing with Case Studies. As Hawken and Gerardo L. Munck (2009) note, "Consensus is not necessarily indicative of accuracy and the high correlations by themselves do not establish validity." Even when a new measure converges, on aggregate, with existing measures, it is useful to unpack disagreement across measures and examine validity in the context of specific examples. Researchers can use case studies to scrutinize particularly salient examples of disagreement and examine how the information presented by quantitative measures corresponds to actual cases. The case studies are labor intensive, so it is important to select cases purposively to assess the measures in question. It is also preferable to perform the analysis "blind," meaning that one writes a description of

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

[14] These are the results for the V-Dem Corruption Index. We provide measure-disaggregated models in the online appendix.

[15] Of course, the bias may go in the other direction, in which case V-Dem might benefit from relying less on PhD holders.

TABLE 8 *Explaining Deviations from WGI Control of Corruption Index with Aggregate Coder Traits*

	DV: Absolute Residual
Share female coders	0.052*
	(0.025)
Average age of coders	-0.002
	(0.009)
Average age of coders^2	0.000
	(0.000)
Share of PhD coders	-0.084*
	(0.023)
Share of coders employed by government	-0.068
	(0.042)
Share of coders born in country	-0.009
	(0.028)
Share of coders residing in country	0.010
	(0.027)
Average free market support	0.006
	(0.010)
Average electoral democracy support	0.001
	(0.015)
Average liberal democracy support	-0.005
	(0.013)
Mean coder discrimination (beta)	0.004
	(0.004)
Coder disagreement	0.345*
	(0.043)
No of coders	-0.008*
	(0.002)
R-squared	0.099
No. Countries	164
No. Observations	54235

Note: Entries are regression coefficients, with standard errors, clustered on countries, in parentheses. The dependent variable is the absolute residuals from regressing each V-Dem measure on WGI, including year-fixed effects. Individual-level coder traits are included. Measure-fixed effects are omitted from the table. $^* p < 0.05$

corruption in a country before seeing the measures.

With this in mind, we selected four countries for cases studies to evaluate V-Dem. We chose Georgia and Zambia, from their points of independence to the present, because V-Dem measures for these countries differ significantly from those produced by other corruption measurement projects, specifically WGI and CPI. We find that V-Dem measures, relative to other corruption measures, more closely mirror detailed descriptions from published accounts of corruption in the countries. We also selected historical Spain and the U.S. to check the quality of the V-Dem Corruption Index going back in time. We examine both countries from 1900 and stop with 1988 for Spain and 1955 for the U.S. to capture periods of dramatic change. In this case, we do not compare the V-Dem measures of corruption with other corruption measures because there are no other corruption measures with this level of historical coverage. We find that the case analysis of Spain and the U.S. validate the V-Dem Corruption Index, and thus increase our confidence in its quality. We also examine the individual low-level V-Dem corruption measures against the U.S. case, once again finding close alignment. This demonstrates the value of providing disaggregated measures along with the high-level, over-arching measure.

To develop the case studies, a research assistant used scholarly articles, books, and intergovernmental and nongovernmental reports to describe the extent and nature of corruption generally and, where possible, in each branch of government and the public sector. The reports he used included thick descriptions from the World Bank but not the data sources that include corruption measures – WGI and BEEPS. Importantly, the research assistant did not view the quantitative corruption measures from either V-Dem or other datasets prior to, or while writing, the case studies.

In presenting V-Dem measures for the four countries, each of the country graphs includes only the portion of the scale where a country's corruption scores fall. Figure 5 illuminates the absolute values of corruption in the countries to discourage readers from exaggerating the meaning of changes in corruption in a country. Due to space constraints, we present Zambia and the U.S. here and Georgia and Spain in the online appendix.

For Zambia, the contrast among the measures is substantial, as Figure 6 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. Yet, the V-Dem measure matches published accounts of corruption in that country more closely than other corruption measures (Chikulo 2000; Van Donge 2009; Mbao 2011; Szeftel 2000). During Zambia's First and Second Republic, from independence in 1964 until 1990, corruption was pervasive in the country, according to published accounts. The relatively high score on the V-Dem scale reflects this. As the economy worsened in the early 1970s, civil servants increasingly turned to theft of state resources to augment their salaries; the V-Dem measure captures this increase. Since then growth in corruption can mainly be attributed to the informal practices of government elites. In the first years of the Third Republic, government officials used the privatization

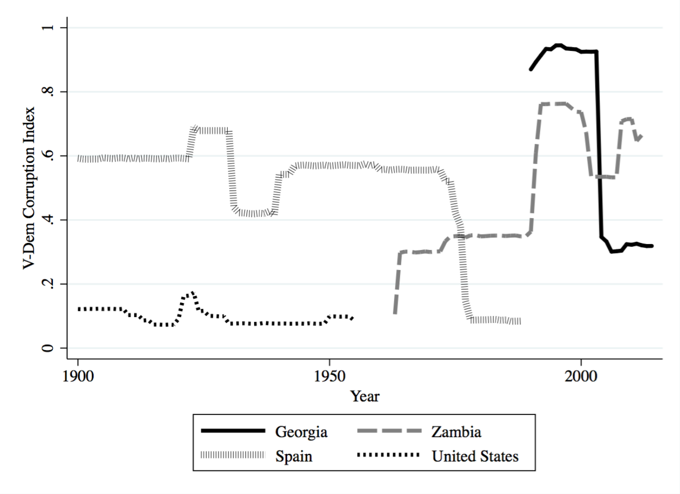

Figure 5. V-Dem's Corruption Measure for Georgia, Spain, the United States, and Zambia

campaign to enrich themselves, according to published reports. Thick descriptions do not mention the small dip in the late 1990s that the V-Dem measure depicts (as does WGI, but not CPI). Otherwise, the publications and V-Dem measure move in lockstep for this era. The published accounts allude to a decline in corruption with the 2001 exit of President Frederick Chiluba and other officials who were implicated in theft of state resources. Corruption in the country then began to increase in 2008 with the election of new presidents then and also in 2012, according to those accounts. The V-Dem measure mirrors this pattern, except for showing a small drop in 2011, which the publications do not mention (but the other measures depict).

Both the V-Dem Corruption Index for the U.S. and its individual measures match the details provided by published accounts of individual cases, increasing our confidence in the V-Dem measures going back in time and demonstrating the utility of providing disaggregated measures of corruption in addition to a high-level measure (Benson, Maaranen, and Heslop 1978; Grossman 2003; Menes 2003; Reeves 2000; Woodiwiss 1988). At the turn of the century, U.S. government bureaucrats stole state resources and exchanged state services for personal material gain. However, the Progressive Movement of the early 1900's discouraged and lessened this corruption. The V-Dem Corruption Index depicts this decrease in corruption, as Figure 5 shows. Corruption increased in 1921 with the administration of Warren Harding, fueled by Prohibition-era bribes from liquor smugglers, and declined upon his death in 1923. The V-Dem Corruption Index approximates this account well. The measure shows a small increase in 1920 but then,

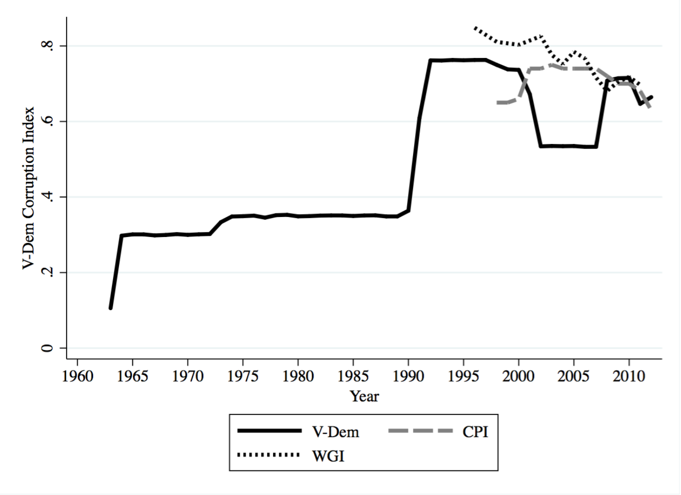

Figure 6. *Corruption in Zambia*

like the thick description, a significant increase in 1921 and a dramatic decrease in 1924.

The value of the individual V-Dem measures become especially apparent with the Harding administration. The measures diverge, reflecting the published accounts. It is evident in Figure 7 that most of the increase is attributable to executive and public sector bribery, and then embezzlement. This period is not characterized by a dramatic increase in legislative corruption, as is clear from the published reports.[16] Legislative corruption, such as the awarding of military contracts in exchange for bribes, was central to corruption during World War II and sustained it during this period. With the end of the war and prosecutions for the schemes, these opportunities subsided. The V-Dem legislative corruption measures captures the dip in corruption at the end of the war in 1945. The individual V-Dem measures also match the published accounts of increased corruption, by bureaucrats in numerous agencies, during the Truman administration. The V-Dem measure shows, in Figure 5, that corruption increased during the Truman administration (1945 to 1953); corruption levels jump in 1950 and drop in 1955. Individual V-Dem measures 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 7. Overall, the V-Dem measures present a picture similar to thick descriptions of

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

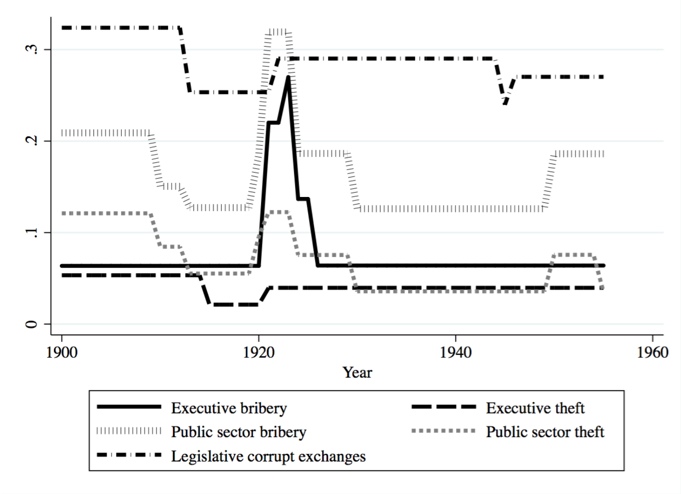

Figure 7. Disaggregated Corruption in the USA

corruption in the U.S. historically.

Although only four cases, the analysis of Georgia, Zambia, Spain and the U.S. that we present here and in the online appendix boosts our confidence in the V-Dem measures. V-Dem outperforms the alternative measures, which do not capture all the trends revealed in the thick descriptions of Georgia and Zambia. Like thick descriptions, V-Dem measures capture an increase or decrease within a short time period, such as a year or two. The cases of Spain (supplemental appendix) and the U.S. increase our confidence in the V-Dem data generating process to gather historical information and translate it into valid measures of corruption. The U.S case also demonstrates the value of individual V-Dem corruption measures. Overall, the convergent validity assessment using case studies suggests that V-Dem measures correspond closely to existing knowledge.

Discussion

Greater attentiveness to measurement quality can improve social science research. Rather than providing abstract advice or suggestions relevant only to *creating* a measure, this paper describes and demonstrates an approach to assess the strengths and limitations of *existing* measures. Specifically, our approach helps reveal systematic and random measurement error in order to judge the validity and reliability of measures. We advocate for three components in the measurement assessment approach, each incorporating multiple tools:

1) assessing content validity by using face validity and factor analysis tools; 2) evaluating the validity and reliability of the data generating process; and 3) assessing convergent validity with case studies and comparisons across coders and measures.

In a world of limited data, it is often tempting to conduct validation tests, mention they have been done in a footnote of a paper, and then say no more about it. The literature on validation has provided scant guidance about what to do with the findings of a validation exercise and how they might affect substantive research. Yet, validation exercises provide rich information about how strengths and limitations of the measure might affect the findings of substantive research, or more specifically, the conditions under which substantive conclusions might be more or less robust. We therefore now provide five examples of how the findings of our validation exercise might be incorporated by researchers using the V Dem corruption measures.

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

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

Third, also from the data generating process assessment we learned that V-Dem coder disagreement for a country-year observation is inversely related to the level of freedom of expression and that it is greatest for country-years with a moderate level of corruption. This in turn means there will be more uncertainty in V-Dem Corruption Index estimates for countries with low freedom of expression or with a moderate amount of corruption. This uncertainty has the potential to diminish the robustness of results when testing theories pertaining to middle-corruption countries or less free societies. Researchers are encouraged to use the information the V-Dem project provides to estimate and mitigate the impact of this coder uncertainty. In addition to the point estimates for each country-year observation, the V-Dem dataset includes the confidence intervals surrounding the point estimates. These can be incorporated into robustness checks to ascertain how sensitive findings are to variations in estimates within the confidence intervals.

Fourth, the data generating process assessment highlighted the relative value of using V-Dem measures for time-series, cross-sectional research on corruption. The consistency of the V-Dem coding procedures and aggregation procedures across all years will enable researchers to use the V-Dem Corruption Index to examine corruption dynamics over time. For other high-level corruption measures, the data sources and aggregation procedures change over time. Similarly, V-Dem's use of a sophisticated measurement model, bridge coders, and anchoring vignettes facilitates cross-country comparison. The extensive temporal and geographic coverage of the measures also enables time-series, cross-sectional research.

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

These five points highlight how researchers might begin to think about mitigating concerns and utilizing strengths in working with the V-Dem corruption measures. More generally, this article offered and illustrated a complementary, flexible, practical, and methodologically rigorous approach to measurement assessment so that researchers can more completely understand the strengths and limitations of existing measures. With such tools in hand, social scientists can become more attentive to the quality of the measures they use.

References

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

Benson, George C.S., Steven A. Maaranen, and Alan Heslop. 1978. *Political Corruption in America.* Lexington: D.C. Heath / Company.

Bollen, Kenneth. 1993. "Liberal Democracy: Validity and Method Factors in Cross-National Measures." *American Journal of Political Science* 37 (4): 1207–1230.

Bollen, Kenneth A. 1986. "Political Rights and Political Liberties in Nations: An Evaluation of Human Rights Measures, 1950 to 1984." *Human Rights Quarterly* 8 (4): 567–591.

Bollen, Kenneth A, and Pamela Paxton. 2000. "Subjective Measures of Liberal Democracy." *Comparative Political Studies* 33 (1): 58–86.

Bowman, Kirk, Fabrice Lehoucq, and James Mahoney. 2005. "Measuring Political Democracy Case Expertise, Data Adequacy, and Central America." *Comparative Political Studies* 38 (8): 939–970.

Chikulo, Bornwell C. 2000. "Corruption and Accumulation in Zambia." In *Corruption and Development in Africa: Lessons from Country Case Studies,* edited by K. R. Hope, Sr. and B. Chikulo, 161–182. London, UK: Palgrave Macmillan UK.

Christiane, Arndt, and Charles Oman. 2006. *Uses and Abuses of Governance Indicators.* Development Centre Studies, OECD Publishing.

Collier, David, Judy LaPorte, and Jason Seawright. 2012. "Putting Typologies to Work: Concept Formation, Measurement, and Analytic Rigor." *Political Research Quarterly* 65 (1): 217–232.

Coppedge, Michael, John Gerring, David Altman, Michael Bernhard, Steven Fish, Allen Hicken, Matthew Kroenig, et al. 2011. "Conceptualizing and Measuring Democracy: A New Approach." *Perspectives on Politics* 9 (2): 247–267.

Coppedge, Michael, John Gerring, Staffan I. Lindberg, Daniel Pemstein, Svend-Erik Skaaning, Jan Teorell, Eitan Tzelgov, et al. 2015. "Varieties of Democracy Methodology v4." *Varieties of Democracy Project: Project Documentation Paper Series.*

Dahlström, Carl, Victor Lapuente, and Jan Teorell. 2012. "Public Administration Around the World." In *Good Government. The Relevance of Political Science,* edited by S. Holmberg and B. Rothstein, 40–67. Cheltenham, UK: Edward Elgar.

Donchev, Dilyan, and Gergely Ujhelyi. 2014. "What Do Corruption Indices Measure?" *Economics & Politics* 26 (2): 309–331.

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

Galtung, Fredrik. 2006. "Measuring the Immeasurable: Boundaries and Functions of (Macro) Corruption Indices." In *Measuring Corruption,* edited by F. Galtung and C. Sampford, 101–130. Aldershot, UK: Ashgate.

Gerring, John. 2012. *Social Science Methodology: A Unified Framework.* 2nd ed. Cambridge, UK: Cambridge University Press.

Gingerich, Daniel W. 2013. "Governance Indicators and the Level of Analysis Problem: Empirical Findings from South America." *British Journal of Political Science* 43 (3): 505–540.

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

Hawken, Angela, and Gerardo L. Munck. 2009. "Do You Know Your Data? Measurement Validity in Corruption Research." *Working Paper, School of Public Policy, Pepperdine University.*

Hawken, Angela, and Gerardo L Munck. 2009. "Measuring Corruption: A Critical Assessment and a Proposal." In *Perspectives on Corruption and Human Development,* edited by A. K. Rajivan and R. Gampat, 1:71–106. New Delhi, India: Macmillan India for UNDP.

Herrera, Yoshiko M, and Devesh Kapur. 2007. "Improving Data Quality: Actors, Incentives, and Capabilities." *Political Analysis* 15 (4): 365–386.

Huckfeldt, Robert, and John Sprague. 1993. "Citizens, Contexts, and Politics." In *Political Science: The State of the Discipline II,* edited by A. W. Finifter, 281–303. Washington, DC: American Political Science Association.

Johnson, Valen E., and James H. Albert. 1999. *Ordinal Data Modeling.* New York: Springer.

Kaufmann, Daniel, and Aart Kraay. 2002. "Growth Without Governance." *World Bank Policy Research Working Paper,* no. 2928.

Knack, Stephen. 2007. "Measuring Corruption: A Critique of Indicators in Eastern Europe and Central Asia." *Journal of Public Policy* 27 (0=3): 255–291.

Lambsdorff, Johann Graf. 2007. "The Methodology of the Corruption Perceptions Index 2007." *Transparency International (TI) and the University of Passau.*

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

Martin, Andrew D, Kevin M Quinn, and Jong Hee Park. 2011. "Mcmcpack: Markov Chain Monte Carlo in R." *Foundation for Open Access Statistics.*

Martinez i Coma, Ferran, and Carolien van Ham. 2015. "Can Experts Judge Elections? Testing the Validity of Expert Judgments for Measuring Election Integrity." *European Journal of Political Research* 54 (2): 305–325.

Mbao, MLM. 2011. "Prevention and Combating of Corruption in Zambia." *Comparative and International Law Journal of Southern Africa* 44 (2): 255–274.

Menes, Rebecca. 2003. "Corruption in Cities: Graft and Politics in American Cities at the Turn of the Twentieth Century." *NBER Working Paper,* no. 9990.

Mudde, Cas, and Andreas Schedler. 2010. "Introduction: Rational Data Choice." *Political Research Quarterly* 63 (2): 410–416.

Munck, Gerardo L, and Jay Verkuilen. 2002. "Conceptualizing and Measuring Democracy: Evaluating Alternative Indices." *Comparative Political Studies* 35 (1): 5–34.

Pemstein, Daniel, Kyle Marquardt, Eitan Tzelgov, Yi-ting Wang, and Farhad Miri. 2015. "Latent Variable Models for the Varieties of Democracy Project." *Varieties of Democracy Institute Working Paper Series,* no. 21.

Pemstein, Daniel, Stephen A. Meserve, and James Melton. 2010. "Democratic Compromise: A Latent Variable Analysis of Ten Measures of Regime Type." *Political Analysis* 18 (4): 426–449.

Pemstein, Daniel, Eitan Tzelgov, and Yi-ting Wang. 2014. "Evaluating and Improving Item Response Theory Models for Cross-National Expert Surveys." *Varieties of Democracy Institute Working Paper Series,* no. 1.

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

Seawright, Jason, and David Collier. 2014. "Rival Strategies of Validation: Tools for Evaluating Measures of Democracy." *Comparative Political Studies* 47 (1): 111–138.

Steenbergen, Marco R, and Gary Marks. 2007. "Evaluating Expert Judgments." *European Journal of Political Research* 46 (3): 347–366.

Szeftel, Morris. 2000. "Eat With Us: Managing Corruption and Patronage Under Zambia's Three Republics, 1964-99." *Journal of Contemporary African Studies* 18 (2): 207–224.

Tanner, Martin A. 1993. *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions.* 2nd ed. New York, NY: Springer Verlag.

Thomas, Melissa A. 2010. "What Do the Worldwide Governance Indicators Measure?" *European Journal of Development Research* 22 (1): 31–54.

Treisman, Daniel. 2007. "What Have We Learned About the Causes of Corruption from Ten Years of Cross-National Empirical Research?" *Annual Review of Political Science* 10:211–244.

Van Donge, Jan Kees. 2009. "The Plundering of Zambian Resources by Frederick Chiluba and His Friends: A Case Study of the Interaction between National Politics and the International Drive Towards Good Governance." *African Affairs* 108 (430): 69–90.

Woodiwiss, Michael. 1988. *Crimes, Crusades, and Corruption: Prohibitions in the United States, 1900–1987.* Lanham, MD: Rowman & Littlefield Publishers.