International Interactions, 37:340–361, 2011
Copyright © Taylor & Francis Group, LLC
ISSN: 0305-0629 print/1547-7444 online
DOI: 10.1080/03050629.2011.596028

SPECIAL DATA FEATURE

Events Data as Bismarck's Sausages? Intercoder Reliability, Coders' Selection, and Data Quality

ANDREA RUGGERI
University of Amsterdam

THEODORA-ISMENE GIZELIS
University of Essex and CSCW, PRIO

HAN DORUSSEN
University of Essex

Precise measurement is difficult but essential in the generation of high-quality data, and it is therefore remarkable that often so little attention is paid to intercoder reliability. It is commonly recognized that poor validity leads to systematic errors and biased inference. In contrast, low reliability is generally assumed to be a lesser concern, leading only to random errors and inefficiency. We evaluate the intercoder reliability of our recently collected data on governance events in UN peacekeeping and show how poor coding and low intercoder reliability can produce systematic errors and even biased inference. We also show how intercoder reliability checks are useful to improve data quality. Continuous testing for intercoder reliability ex post enables researchers to create better data and ultimately improves the quality of their analyses.

A previous version of this paper was presented at the ISA annual conference in New York, March 2009 and at the Folke Bernadotte UN PKO working group in Columbia University, New York, October 30–31, 2009. We acknowledge financial support from the Folke Bernadotte Akademin and the ESF/ESRC (RES-062-23-0259). We thank the participants at the Folke Bernadotte UN PKO working group for their comments. We also thank Kristian Skrede Gleditsch, María Belén González, Birger Heldt, Spyros Kosmidis, Nikolay Marinov, Will H. Moore, Vera Troeger, and Steffen Weiss and three anonymous reviewers of *International Interactions* for their comments. Replication data are available on the *International Interactions* dataverse page at http://dvn.iq.harvard.edu/dvn/dv/internationalinteractions.

Address correspondence to Andrea Ruggeri, Department of Political Science, University of Amsterdam, Amsterdam, The Netherlands. E-mail:A.Ruggeri@uva.nl

KEYWORDS coders, events data, measurement, peacekeeping, reliability

> *"Data, like sausages, cease to inspire respect in proportion as we know how they are made."*
>
> Freely adapted from Otto von Bismarck

Accurate measurement remains a thorny issue in the social sciences. Any empirical test obviously requires using relevant indicators and data of the highest possible quality. Data quality thus encompasses the need for a clearly defined conceptualization of the theoretical framework, measurement validity—i.e., linking theoretical concepts to observed data—as well as accurate coding, or data reliability. Even though quantitative research has become increasingly important in the study of international relations, measurement issues still receive only limited attention. Our survey of articles published in five major international relations (IR) journals during the last 5 years reveals that less than one third of all articles using quantitative methods pays any attention to issues of measurement error or data reliability. It is even less common to tackle validity or reliability explicitly; this is only done by 6% of all quantitative articles.[1]

In this article we argue that intercoder reliability is crucial to both data quality and the development of efficient and unbiased estimators. In contrast to existing literature (see, for example, Carmines and Zeller 1979), we highlight that intercoder reliability is not just related to random errors but often involves systematic ones. Consequently, failure to take into account and improve intercoder reliability is bound to lead to inefficient and biased inferences. We highlight the importance of intercoder reliability for our data collection on governance events in United Nations (UN) peacekeeping operations and give particular attention to the problems associated with coding events data.[2]

Arguably coders who perform poorly do not solely make random coding errors but often exclude cases in a systematic way. They tend to identify events that are easier to code and systematically disregard events that require more effort or sophistication in order to be properly coded. They will also be more likely to exclude or discount information which is harder to interpret and may thus fail to code the correct attributes of an event. Intercoder reliability tests can assist in identifying coders who behave accordingly. Consequently, examining intercoder reliability ex post can improve data quality and allows

[1] The findings of the survey are discussed in more detail in the concluding section. See also Table 7.

[2] In this article we pay special attention to events data; yet, the issue of intercoder reliability applies more broadly to research methods. Even qualitative data, like interviews, archival research or even participatory observation, require "coding," that is, the selection and interpretation of information (Kalyvas 2006:393–422). Although often probably unfeasible for practical reasons, in principle researchers could independently code the "raw" information in order to gauge the reliability of qualitative research.

researchers to create the best possible dataset. We apply this approach to our own data and show how we selected data collected by "better" coders.

The following section is a brief survey of important recent data collection efforts in IR, distinguishing three common practices to address intercoder reliability. Next, we elaborate the problem of poor intercoder reliability using our peacekeeping data. Then we analyze the impact of coder-selection for our models on cooperation and conflict with peacekeeping efforts to improve local governance. We conclude with a brief discussion of the practical implications of our research, presenting a checklist and steps for researchers who are involved in events data collection.

DENIAL, RESIGNATION, AND SELECTION: VARIOUS APPROACHES TO DATA RELIABILITY

Events data often require coding or "translating" textual information into numerical values (Franzosi 1994; Olzak 1992; Schrodt 2006). "Translation" is even more important for qualitative categories (for example, quality of governance) than quantitative figures (for example, number of deaths), since the coder needs to interpret a continuous latent concept into a limited set of categories. As Rothman (2007) argues, the more the coders are required to interpret, the more the data are susceptible to systematic errors. Any coding process has two essential components. First, it requires a set of rules that identify the dimensions of the phenomenon under consideration where scores are assigned to these dimensions. This information is normally contained in a codebook. The second component is the actual coding of the raw information.

Intercoder reliability requires that different coders give the same "scores" on all dimensions when provided with the same raw information. Imprecise coding rules undermine reliability because coders are uncertain about their task. Variation in the quality of coders is another challenge to intercoder reliability. Intercoder reliability would obviously be a moot issue if we knew which coders coded the raw information correctly. Reliability would either be perfect, or we could simply discard the incorrect data. However, even in the absence of a clear gold standard it may still be worthwhile to distinguish between "better" and "worse" coders and to emphasize the data collected by the former.

The above discussion raises a number of challenges with regards to measurement as well as the search for practical solutions. First, can tests for intercoder reliability be used to improve data quality, for example by weeding out low quality coders? Second, what are the implications for empirical research if there is variation in the quality of coding and coders? The study of measurement errors suggests that random errors can be tolerated at the cost of higher imprecision in the model estimates. Widespread variation between coders remains obviously problematic, but is also quite common in studies employing more subjective concepts, such as governance and state capacity that require interpretation. Rothman (2007) and Mitchell and Rothman (2006) highlight how little attention has been paid to coder reliability in data collection. A likely reason for the relative neglect of reliability is that it remains unclear what (or who) causes unreliable coding and how serious the problem is for empirical analysis.[3]

Validity and reliability are basic properties of empirical measurement and have different implications for measurement error. Measurement validity concerns the link between empirical reality and concepts, or as Bollen (1989:184) suggests, validity is "concerned with whether a variable measures what it is supposed to measure." The closer the operational definition of a variable is to its conceptual definition, the higher the measurement validity of a variable. Most researchers who are involved in data collection are rightly concerned with measurement validity; for example, a large literature has developed on how to define democracy and the characteristics of a democracy (Gleditsch and Ward 1997; Jaggers and Gurr 1995; Treier and Jackman 2008).

Reliability refers to the consistency of measurements, or whether the tools to measure concepts used in research projects can be replicated by different coders (Carmines and Zeller 1979). The theory of reliability suggests that for measurements to be reliable, the variance around the true value of a concept must be as low as possible (low random error and no systematic bias). In this sense, reliability is related to measurement validity. However, intercoder agreement does not guarantee validity or that the data correctly measure the underlying theoretical concept (see Dorussen, Blavoukos, and Lenz 2005). Both criteria are essential for measures to be precise, but are too often treated as two separate problems in data collection (Singer 1984).

Figure 1 represents our understanding of how the error distribution relates to different levels of reliability. In the two top figures (a), reliability is conceptualized as stochastic measurement errors. Perfect reliability, where coders always agree on the measurement, would be represented as the center of a bull's-eye. The closer the hits are to the center, the more reliable is the measurement. The literature suggests that low reliability increases primarily the stochastic component of the measurement error. Consequently, the distribution of the errors will still be still normal but with fatter tails (see Rothman 2007). However, if low levels of intercoder reliability are related to certain values of an item, the measurement error would not be only stochastic but systematic as well. For instance, in the two bottom figures (b) the hits

[3] Once measurement bias has been identified, it is straightforward to evaluate the impact of low validity on the statistical analysis and even to correct for any bias. However, it is far from straightforward to determine whether unreliable data have significant impact on the statistical analysis or how to use information on reliability to improve data quality. Baugh (2003) has developed a method that allows the adjustment of coefficients using the estimated data reliability.

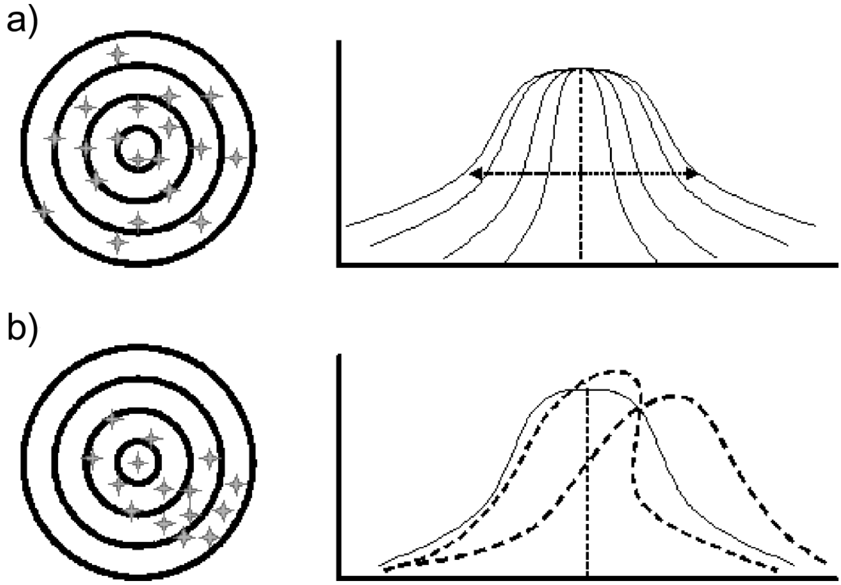

FIGURE 1 Reliability: Stochastic (a) and systematic (b) errors.

on the bull's-eye do not have a white noise distribution but a biased one leading to a skewed distribution of the measurement error.

In the social sciences it is often quite hard to reach high levels of reliability. The problem is not only limited construct validity, but also human-induced errors. Some of the human error is still clearly in the remit of the principal investigators; for example, the coding rules may be poorly defined or there may have been insufficient coder training. Yet, one has to acknowledge that there is also simply variation in the capabilities of the coders. Even precise coding rules will confuse some coders, and the best possible instructions will occasionally be wasted on them. Coders may also have preconceived notions that bias their coding decisions; for example, in our experience, some of the worst coders were highly knowledgeable on a particular issue area and overwhelmingly chose events that were linked to their area of interest.

In reviewing existing studies, we identify three common approaches to intercoder reliability, in short: denial, resignation, and selection. We argue that the latter approach may well be optimal whenever the quality of coders is the main cause for reliability problems.

Denial

A brief survey of major dataset with events data on conflict still shows that many researchers do not explicitly address intercoder reliability. The codebook (Version 4, 2008) for the *UCDP/PRIO Armed Conflict* dataset extensively discusses issues of measurement validity, for example, definition of civil conflict, intrastate and internationalized intrastate conflict. There is also some discussion of missing data. However, there is no reference to explicit reliability checks and intercoder consistency, although they require two independent sources to code a conflict. Similarly, the *ACLED* (Armed Conflict Location Events Data) project (Raleigh and Hegre 2005) does not make any references to intercoder reliability. The related *PKOLED* (Peacekeeping Location Events Data) project (Dorussen 2007) also relies on only one coder per UN mission. The decision to rely on a single coder often stems from resource and time constraints. A further reason may be that the coding (mainly place and time coordinates) are deemed fairly unambiguous. However, this ignores the important and potentially problematic issue of event selection.

Even when the importance of intercoder reliability is recognized, this does not mean that the issue is confronted systematically; for example, the codebook of the *Minorities at Risk Project* (*MAR version 8*) mentions only occasional reliability checks. There is no report of either consistent and systematic intercoder reliability checks, or a systematic screening of indicators for internal consistency. Instead efforts to minimize discrepancies between coders focus on extensive and rigorous training of coders and revisions of the coding rules by senior staff. Most of these procedures are informal, although they appear to have led to changes in the coding rules and procedures.

Resignation

A number of studies explicitly report on reliability, but do not discuss potential implications. The *ICOW (Issue of Correlates of War)* project employed a limited number of coders, while the director of the project together with one primary research assistant determined the final coding scheme. In case of discrepancies, the director of the research project was ultimately responsible for assigning the values for cases identified as miscoded (Hensel 1998:12–13). Regarding the *Expert Survey on Power Distributions among Ethnic Groups*, Cederman et al. (2006) claim that reliability checks among expert interviewers have been conducted, but there is no reference to specific details.

Tests for intercoder reliability often rely on random checks (cf., the *MAR* and *ICOW* projects) by senior academic staff or experts to identify possible problems and discrepancies between coders. Random checks help to assess reliability, but it can be a daunting task to recode previously collected information. Limited monetary resources and time constraints can seriously derail a project if new coding of finished cases is required. The common practice to have senior researchers alter data entries can hardly be justified on the basis of random checks, since there is no guarantee that spot-checks conclusively identify all major problems. Random checks may allow researchers to find estimates for data reliability which can be used to adjust coefficients in analyses of data (see Baugh 2003). However, we have found no examples of studies in international relations in which these methods were actually applied.

Alternatively, the *Kansas Event Data System Project* (Schrodt 2006) uses automated events coding (latest development is TABARI) to avoid the problems of manual data collection. Although automated coding obviously minimizes human error, it is not without problems of its own. Automated coding is less appropriate when the coding is complex, relies more on interpretation, the level of subjectivity is very high, or the structure of the text is not consistent across the various sources (Gerner et al. 2002).

Selection

A third approach is to select the best possible information. For instance, for the *Behavioral Correlates of War* (*BCOW*) project, Leng and Singer (1988:165) report high intercoder reliability for pairs of coders. They comment, however, that the selection of events and identification of actors and targets had taken place at an earlier stage. They report lower intercoder reliability scores for more complex coding categories. Beardsley and Schmidt (2007) also use coder pairs who simultaneously and independently code all cases. The two coding reports are subsequently reconciled. In cases of discrepancy between the two coders, the two project directors took an executive decision. Beardsley (personal correspondence) noted that any discrepancies between the coders were generally minor, possibly because the coded data required relatively little interpretation.

We explore a related approach using intercoder reliability tests to select "good" coders in order to ensure the highest possible data quality. We propose a multistage process which depends on the ex post detection of "good" and "bad" coders, and a subsequent evaluation of the impact, if any, on the data analysis. The following section extensively discusses the process that we followed to improve reliability. We also explore two issues relevant to intercoder reliability: do coders identify the same events, and do they code an event the same way?

UNITED NATIONS PEACEKEEPING AND LOCAL GOVERNANCE EVENTS DATA

The objective of the *UN Peacekeeping and Local Governance* project is to systematically collect information on the state- and peace-building policies implemented as part of UN peacekeeping operations. The data collection involves events-coding of the reports of the UN Secretary General, and thus provides the UN's evaluation of its priorities as well as its assessment of these policies. The eventual goal of the project is to code governance events in all UN peacekeeping missions in civil wars since the end of the Cold War.

Governance events are defined as time- and place-specific interactions between stakeholders in peacekeeping, that is, the UN, its peacekeepers, and other external actors as well as government, rebel, and local authorities, with a direct effect on the provision of public goods and services. Our primary interest is to code the response to UN peacekeeping activities distinguishing between conflictual and cooperative responses. One of our assumptions is that these response variables are not mutually exclusive, since arguably events can provoke simultaneously conflict and cooperation. Conflict and cooperation are coded as attributes of a particular governance event. For example, the kidnapping of aid workers is not considered a governance event, but a conflictual response to the governance event of providing humanitarian aid.[4]

The models analyzed below use a small number of independent variables controlling for the actors involved in the event, the policy content and the role of the UN peacekeepers. The data distinguish between government, rebel, and (quasi-) independent local authorities. Further types of participants are external authorities and societal actors. The UN peacekeepers are coded among the external authorities. The set of dummy variables *Authorities Involved — Government*, *Rebel*, and *Local*, are thus neither mutually exclusive nor complete, where events that involve only the UN or the UN and other external authorities (like IGOs, NGOs, and the third-party government) are the baseline category.

The coding of policy content is based on Ratner's (1996:41) classification on the breadth of second-generation (or new) peacekeeping operations. The policy content of an event is classified as strengthening the center (*Policy: Strengthen Center*) if the event involves policies of military matters, law and order, governmental administration, economic reconstruction and external relations. The impact of these policies is contrasted with the impact of policies dealing with elections, democratization, human rights, national reconciliation, refugee care, and humanitarian relief.

Ratner (1996:41) identifies the following roles for peacekeepers: (a) monitoring without a mandate to intervene, (b) supervision defined as oversight over situations with a mandate to request changes in the behavior of actors, but not to order those actors directly to correct their behavior,

[4] Conflict and cooperation do not refer to the general background of the governance event, but must be in direct response to an event. A deteriorating security situation is likely to influence governance, but conflict is only coded if the governance actors or public goods are explicitly targeted. The baseline categories are when no conflict (cooperation) is recorded.

(c) control or direct line authority, and (d) conduct which involves the authority to perform certain tasks directly, with or without the assistance of local authorities and notwithstanding their views on those matters. As a fifth category, we added education defined as providing technical assistance and public information. Control and conduct are coded as replacing regulatory capacity (*Capacity: Replace*). In contrast, education, monitoring, and supervision are assumed to strengthen regulatory capacity (*Capacity: Strengthen*).[5]

From the outset we were aware that the coders were interpreting politically motivated documents, the UN reports of the Secretary General on peacekeeping, and that they were asked to identify theoretically abstract attributes of often complex and confusing events. Therefore, we were immediately concerned with the quality of coding.

INTERCODER RELIABILITY AND THE SELECTION OF CODERS

To assess intercoder reliability, we first analyzed whether coders identified the same governance events. Next, we evaluated whether they coded the same events identically. Initially there was large variation of coder quality, but intercoder reliability improved as the project progressed. The additional information on the (assigned) coder quality demonstrates that poor coder quality does not only result in white noise, but also leads to systematic error.

The project progressed in three stages. In the first stage, the project managers coded a small number of reports in order to clarify the conceptual and operation definitions of the instruments. As a result the coding procedures and the codebook were updated. Subsequently, one person coded the missions of Sierra Leone and Liberia as pilot projects. The coder and project managers discussed any questions and concerns, which resulted in the final version of the codebook.

In the second stage, events were selected and coded independently by pairs of research assistants.[6] The training procedure was standardized for all coders. They were introduced to the project and the codebook was explained step-by-step. Reports previously coded by the project managers were used as examples in order to explain how to use the codebook in relation with the text. Finally, they were asked to code independently two reports previously coded by the project managers. In the review, every

[5] For more information on the other variables please refer to the codebook downloadable from http://privatewww.essex.ac.uk/~hdorus/.

[6] In the second stage, not all reports were double coded; Angola was double coded for the 93% of the reports, Burundi for 28%, Central African Republic 100%, and Democratic Republic of Congo 24%.

TABLE 1 UN Peacekeeping Missions and Coding Strategy

Country	Time-Span	Stage 1 Single-Coded	Stage 2 Double-Coded	Stage 3 Events preselected; Double-Coded
Liberia	1993–1997	X		
	2003–2005	X		
Sierra Leone	1997–2005	X		
Angola	1991–1999		x	
Burundi	2004–2005		x	
Central African Republic	1998–2000		x	
D. R. Congo	2000–2005		x	
Ivory Coast	2004–2005			x
Mozambique	1992–1994			x
Namibia	1989–1990			x
Somalia	1992–1995			x

coding discrepancy was discussed in order to get a more consistent coding strategy.

In the third stage, based on the discussions between the project directors and the coders, it was decided to assign one coder to preselect events and next have two different coders code these events independently. The coders used in the third stage were a subsection of those used in the second stage. The project managers felt strongly that the average quality of the coders improved from stage two to stage three. The contracts of some, less motivated, coders were discontinued, while the remaining coders had obviously gained more experience. Table 1 summarizes how the various missions were coded.

Identifying Events

The selection of events emerged as a potentially serious issue in early discussions between the coders and the project managers.[7] The coders indicated that it was often difficult to identify events, because the reports referred to policies in general terms lacking specific information on their spatial and/or temporal domain. These discussions led to a change of the coding practice with events being selected independently from their coding. It remains, however, interesting to assess the seriousness of the problem were coders able to identify the same governance events. The comparison is relevant for four countries (Angola, Burundi, the Democratic Republic of Congo, and the Central African Republic), where events were not preselected and coded by two persons.

[7] As one of the reviewers pointed out, event selection may well be the major difference between high frequency data, such as TABARI, BCOW or the UN Peacekeeping events data, and country-year data, like the UCDP-PRIO data.

TABLE 2 Intercoder Reliability Regarding Identification of Events

Country	Total events	Unique events	Double identified	Reports	Reports Double-Coded
Liberia	590			38	
Sierra Leone	1246			42	
Angola	1638	1303	41%	56	52 (93%)
Burundi	195	170	28%	7	2 (28%)
Central African Rep.	360	309	40%	10	10 (100%)
D.R. Congo	1264	1143	18%	25	6 (24%)
Ivory Coast	225	129	85%	7	7 (100%)
Mozambique	239	134	88%	12	12 (100%)
Namibia	93	53	86%	4	4 (100%)
Somalia	680	338	88%	15	15 (100%)

In Table 2, the second column presents the number of unique events, where an event is unique if it was identified by either both coders or only one. The third column gives the percentage of the *matching rate*, indicating how many unique events were spotted by two coders out of the total number of unique events. The matching rate thus gives the level of agreement between coders in selecting governance events. It is clear that coders found it difficult to distinguish the same governance events. At best, they agreed on 41% percent of the events. In the worst case, the matching rate was a mere 18%. Since not all reports were double-coded, the poor matching rate is worrying. Closer inspection of the data revealed that most of the variance could be explained by the proclivity of some coders to disaggregate information in terms of geographic location and actors involved. Coders largely identified the same events, but some ended up with one event with multiple actors over a broadly specified geographical area, whereas others identified several events with unique combination of actors interacting at specific locations.

A further concern was that the selection/nonselection of events was systematic instead of random. Some coders were only identifying events that had specific attributes. For example, they preferred events where conflict was more evident; they left out events where conflict was less obvious, thereby generating a selection process that introduced systematic bias in favor of high conflict events and at the expense of reporting either low conflict or cooperative events.

In the third stage one coder preselected the governance events, and the remaining two coders were provided with UN reports in which the governance events were highlighted. However, they were allowed to contest the preselection and to add events. As to be expected, the matching rate is much higher in these cases. It varies from 85% for Ivory Coast to 88% for Mozambique and Somalia. Table 3 reports how many events were preselected and the numbers of events that the other two coders finally selected.

TABLE 3 Identification of Events; Number of Events Pre-selected, Coder 1 and Coder 2

	Preselected	Coder 1	Difference Pre–Coder 1	Coder 2	Difference Pre–Coder 2
Ivory Coast	95	106	+11%	121	+27%
Mozambique	117	119	+2%	122	+4%
Namibia	43	47	+9%	48	+12%
Somalia	266	257	−4%	313	+18%

It is clear that the coders were not just coding the preselected events, but that they were also evaluating whether the preselection of events was correct. The final number of events was on average 10% more than the initial number of the preselected events.

Intercoder Agreement in Events Coding

Even if coders agree on the identification of events, they may still disagree on the coding or interpretation of events. The literature suggests numerous ways to calculate intercoder reliability (Neuendorf 2002; Rothman 2007). The two most commonly used indicators are Cohen's Kappa and Percent Agreement. The Percent Agreement indicator simply calculates the number of times coders agree as a percentage of the number of times they could possibly agree. Cohen's Kappa corrects Percent Agreement for chance agreement, calculating the so-called beyond-chance agreement. Generally, this means that Cohen's Kappa indicators will be lower than Percent Agreement (Sim and Wright 2005:258). Cohen's Kappa is defined as:

$$k = \frac{P_o - P_c}{1 - P_c}$$

where *Po* is the proportion of observed agreements and *Pc* is the proportion of agreements expected by chance. The Kappa can range from –1 to 1, with 1 signaling perfect agreement and 0 indicating agreement no better than chance (Liebetrau 1983). In practice a negative Kappa, indicating observed *dis*agreement greater than expected by chance, is rare. Achievement of perfect agreement is equally difficult and often impractical given finite resource and time constraints (Hruschka et al. 2004:313).[8] Table 4 reports intercoder reliability for the conflict scale, our main dependent variable. Again, the difference in reliability between the countries coded in the second and third stage is striking.

[8] Hruschka et al. (2004) suggest fairly stringent cutoffs of Kappa $\geq$0.80 or 0.90, while Landis and Koch (1977:165) suggest a range of intercoder reliability based on Kappa, distinguishing between poor (< 0.0),

TABLE 4 Intercoder Reliability, Conflict Scale

Country	Percent agreement	Expected agreement	Kappa	Std. Error	Z
Angola	72.94%	46.92%	.22	.07	3.18***
Burundi	Insufficient observations				
Central African Rep.	16.67%	30.56%	−.21	.24	−.81
D.R. Congo	50.00%	35.89%	.18	.14	1.27
Ivory Coast	81.05%	68.99%	.39	.07	5.01***
Mozambique	90.48%	80.02%	.52	.08	6.51***
Namibia	91.11%	78.53%	.63	.13	4.79***
Somalia	94.12%	90.02%	.41	.05	6.91***

*$p < .05$, **$p <$ 01, ***$p < .001$.

The agreement rate for the group of countries coded in the second stage ranges from 16.67% to 72.94% whereas agreement for the third stage ranges from 81% to 94%. Furthermore, a *t* test comparing the mean of the Kappa statistics of the two groups suggests that they differ at the 5% significance level. This supports our impression that the quality of the coders improved from the second stage of coding to the third.

THE IMPACT OF VARIATION IN THE QUALITY OF CODING

In the second and third stage of the project, every coder could be compared directly with at least one other one, since both coded a number of identical reports. Noting significant variation in the quality of the coders employed in the project, we selected coders on three criteria: commitment, understanding, and sophistication. We selected for each pair the best coder based on: (a) the number of events coded for each report, (b) the number of cases with information on the conflict/cooperation scale relative to the total number of cases coded, and (c) the ability to differentiate between the conflict and cooperation scales. The criteria are admittedly subjective, but in our experience the first criterion sets apart committed and focused coders. The second criterion identifies coders that were meticulous, and the final criterion singles out coders that had a more precise and sophisticated understanding of the main dependent variables. The coders were neither aware that their coding quality would be assessed, nor were they informed about our criteria for distinguishing "good" from "bad" coders.

One of the objectives of the data collection is to examine the response of local stakeholders to the activities of the UN peacekeepers (see Dorussen and Gizelis 2008). Tables 5 and 6 replicate the basic model to evaluate the

slight (0.00–0.20), fair (0.21–0.40), moderate (0.41–0.60), substantial (0.61–0.80), and almost perfect (0.81–1.00).

TABLE 5 Conflict Response to UN Peacekeeping, Coder Selection

	(1)	(2)	(3)	(4)	(5)	(6)
				"Good"		
				+ Sierra		
		Only	Only	Leone	2nd	3rd
Cases/Events	All	"Bad"	"Good"	/Liberia	Stage	Stage
Government involv.	−0.12	−0.12	−0.05	−0.17	−0.30	0.16
	(0.08)	(0.08)	(0.06)	(0.06)**	(0.10)**	(0.18)
Rebel involv.	0.45	0.45	0.67	0.88	0.31	0.52
	(0.10)**	(0.10)**	(0.07)**	(0.06)**	(0.12)**	(0.24)*
Local involv.	0.28	0.28	−0.12	−0.05	0.36	–[a]
	(0.16)	(0.16)	(0.19)	(0.18)	(0.18)*	
Capacity Strengthen	0.47	0.47	0.45	0.44	0.47	0.36
	(0.08)**	(0.08)**	(0.06)**	(0.06)**	(0.10)**	(0.17)*
Policy Strengthen Center	0.12	0.12	−0.35	−0.51	0.09	−0.35
	(0.10)	(0.10)	(0.08)**	(0.07)**	(0.11)	(0.29)
GovInv × "Good"	0.07				0.18	0.03
	(0.10)				(0.12)	(0.25)
RebInv × "Good"	0.22				0.28	0.28
	(0.13)				(0.14)*	(0.32)
LocInv × "Good"	−0.40				−0.61	–[a]
	(0.24)				(0.26)*	
CapStre × "Good"	−0.02				−0.03	0.08
	(0.10)				(0.12)	(0.25)
PolCen × "Good"	−0.47				−0.43	−0.55
	(0.12)**				(0.14)**	(0.43)
"Good" Coder	0.10				−0.04	−0.06
	(0.07)				(0.08)	(0.15)
Constant	−1.06	−1.06	−0.96	−1.11	−0.76	−1.50
	(0.05)**	(0.05)**	(0.04)**	(0.04)**	(0.07)**	(0.10)**
Observations	3952	1604	2348	3690	2787	1139

Standard errors in parentheses; *significant at 5%; **significant at 1%.
[a]Variable not included in submodel.

impact of using information from "good" and "bad" coders respectively. The analysis further distinguishes between events coded in the second and third stage of the project.

In the conflict model (Table 5), most results are robust for the quality of the coders. Comparing Models 2 and 3 (or 4) shows that the only exception is *Policy: Strengthen Center*, where using exclusively data from "good" coders leads to a significant and negative effect of building governance capacity of central authorities on conflict. The differences between "good" and "bad" coders are much larger when cooperative responses are analyzed (models 8–10 in Table 6). Firstly, the impact of government involvement is not robust. The effect is nonsignificant when "good" coders are used and Sierra Leone and Liberia are excluded from the analysis. Including the latter or using "bad" coders suggests, however, that government involvement may lead to more cooperation. The effects for rebel and local involvement are robust. Secondly, the selection of coders affects the results for policy content

TABLE 6 Cooperation Response to UN Peacekeeping, Coder Selection

	(7)	(8)	(9)	(10)	(11)	(12)
				"Good"		
				+ Sierra		
		Only	Only	Leone	2nd	3rd
Cases/Events	All	"Bad"	"Good"	/Liberia	Stage	Stage
Government involv.	0.32	0.32	0.05	0.22	0.34	0.10
	(0.07)**	(0.07)**	(0.06)	(0.05)**	(0.09)**	(0.13)
Rebel involv.	0.28	0.28	−0.23	−0.53	0.12	0.59
	(0.10)**	(0.10)**	(0.07)**	(0.06)**	(0.11)	(0.22)**
Local involv.	.83	0.83	0.48	0.37	.98	0.56
	(0.16)**	(0.16)**	(0.20)*	(0.19)*	(0.21)**	(0.26)*
Capacity Strengthen	0.19	0.19	−0.16	−0.20	0.16	0.16
	(0.07)*	(0.07)*	(0.06)**	(0.05)**	(0.09)	(0.13)
Policy Strengthen Center	0.04	0.04	0.22	0.38	0.10	−0.30
	(0.09)	(0.09)	(0.07)**	(0.06)**	(0.10)	(0.18)
GovInv × "Good"	−0.27				−0.25	0.16
	(0.09)**				(0.11)*	(0.25)
RebInv × "Good"	−0.51				−0.36	0.21
	(0.12)**				(0.14)**	(0.40)
LocInv × "Good"	−0.36				−0.32	—[a]
	(0.25)				(0.29)	
CapStre × "Good"	−0.35				−0.19	−1.34
	(0.10)**				(0.12)	(0.22)**
PolCen × "Good"	0.17				0.21	0.15
	(0.11)				(0.13)	(0.28)
"Good" Coder	.94				0.55	2.20
	(0.06)**				(0.08)**	(0.15)**
Constant	−0.28	−0.28	.66	.86	−0.14	−0.41
	(0.04)**	(0.04)**	(0.04)**	(0.04)**	(0.06)*	(0.07)**
Observations	3952	1604	2348	3690	2787	1164

Standard errors in parentheses; *significant at 5%; **significant at 1%.
[a]Variable not included in submodel.

and the role of the UN. Only "good" coders find that building the governance capacity of central authorities is associated with more cooperation. Finally, "bad" coders find that more indirect UN involvement—education, monitoring, and supervision—leads to *more* cooperation, while "good" coders find that it leads to *less* cooperation.[9]

It is interesting to compare the second and third stages of the project. Comparing models 5 and 6 (in Table 5) shows quite clearly that, in the second stage of the project, the practices of "good" and "bad" coders significantly affect the analysis of conflict: three of the five interaction variables with the quality of the coders are significant. However, using data collected

[9] The probit models of Tables 5 and 6 largely replicate these findings in Dorussen and Gizelis (2008), at least for the "good" coders. This is noteworthy since Dorussen and Gizelis (2008) rely exclusive on data collected in stage 2, use a more extensive model, disaggregate conflict/cooperation levels and estimate the effects on conflict and cooperation simultaneously.

in the third stage of the coding process there is no longer a significant difference. This finding supports our intuition that the overall quality of the coders improved between the second and third stage—some of the "bad" coders in the third stage were actually identified as "good" coders in the second; they were simply teamed up with even "better" coders in the third stage. The same finding is, however, less obvious for the cooperative response models. Comparing models 11 and 12 (in Table 6), "good" coders still differ significantly from "bad" coders in both the second and third stage.

Perhaps somewhat unsurprisingly, these findings suggest that the quality of coders matters. However, so far we have no clear indication that the "good" coders are indeed better than the 'bad' coders. Figures 2 and 3 provide some evidence supporting the classification of coders.

Models 5 and 6 (Table 5) for conflict, and models 11 and 12 (Table 6) for cooperation are used to estimate the predicted probabilities of conflict

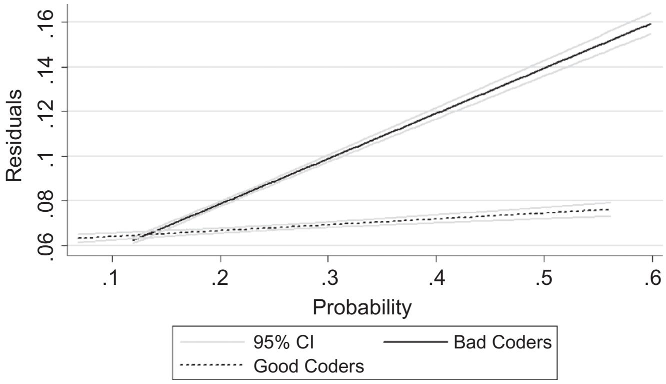

FIGURE 2 Estimated probability of conflict against standard errors of residuals, comparing "good" and "bad" coders (estimates based on Models 2 and 3 in Table 5).

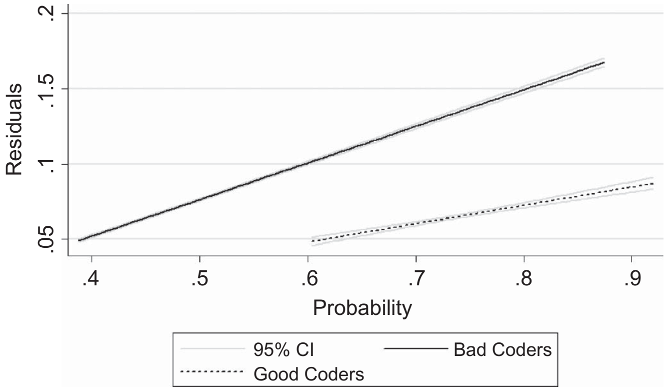

FIGURE 3 Estimated probability of cooperation against standard errors of residuals, comparing "good" and "bad" coders (estimates based on Models 8 and 9 in Table 6).

and cooperation respectively. Next, these estimates are plotted against the standard errors of predicted probability of conflict/cooperation. Arguably, in the case of "good" coders, the residuals of the predictions should have less variance (the predictions are more precise) and remain constant irrespective of how likely cooperation (conflict) is. The latter would indicate that the precision of the prediction is unrelated to its content. In contrast, the residuals of the predictions of "bad" coders are not only expected to have a larger variance, but also to be conditional on the likelihood of predicting conflict/cooperation.

The idea that the predictions of "good" coders are more precise fits our understanding of data reliability. The suggestion that the precision is unrelated to the estimated probability of conflict/cooperation is perhaps less obvious. Intuitively, the argument is similar to that of heteroskedasticity. A pattern in the standard errors may reflect the exclusion of a significant influence on the variance (here, excluding coder quality). Our argument is not simply that "good" coders recognize both (lack of) cooperation and conflict, but rather that they code all information relating to events involving conflict (/cooperation) as precise as events without conflict (/cooperation).

Figure 2 clearly shows that with respect to the prediction of conflict, "good" coders are clearly different from "bad" coders in both respects. The estimates of the likelihood of conflict are not only much more precise (smaller residual variance), but the residuals also seem independent of the estimated probability of conflict. In contrast, the estimates of "bad" coders become much less precise if the estimated probability of conflict increases. The findings for cooperation in Figure 3 are much less clear-cut. As previously, the residuals are still less precise for "bad" coders. However, in this case the standard errors for predicting cooperation are conditional on the estimated probability of cooperation case for both "good" and "bad" coders. To us, these findings strongly suggest that the "good" coders were indeed better, but also that the coding of cooperative responses is more difficult and contentious than the coding of conflict responses.[10] Conflicts are more likely to grab the attention of the UN reporters and are particularly salient for the primary audience (that is, diplomats) of the reports.

DISCUSSION

Based on our experience, and strengthened by the analyses presented above, we hold that variation in the quality of coders is an important reason

[10] Though we provide only graphic tests for heteroskedasticity, we have also consistent findings using heteroskedasticity probit models (Alvarez and Brehm 1995) where we use the "good versus bad" variable to model the variance of the residuals. Indeed, we find that using "good" coders shrinks the residuals' variance.

for imprecise measurement. This by no means denies that there are several other concerns in coding events data, such as vaguely defined coding rules or an ill-conceived conceptual framework. However, whereas these latter problems are widely recognized, the problem of low quality coders has received little systematic attention and seems to be mainly bemoaned privately. Regardless of the precision of coding rules and without denying the importance of training, some variation in the quality of coders remains inevitable. Most disturbingly is that intercoder variability is not just random variation around the true value of the measurement, but that it can introduce systematic bias in the data collection process, as our peacekeeping data illustrate. Poor quality coding, as indicated by low reliability, is not only less precise but can also induce systematic error. This finding definitely warrants further research, and is especially relevant to IR scholars who regularly rely for their research on human-coded events data based on textual sources.

To assess the broader relevance of this conclusion, we surveyed the articles published by seven key journals in IR and political science from 2005 to 2010. As shown in Table 7, more than half of the articles published in the selected IR journals make use of quantitative methods.[11] In leading IR journals, like the *Journal of Conflict Resolution* (*JCR*) and *International Organization* even 60% of articles use quantitative methods. In this respect, IR compares well with political science more generally; in comparison, 49% of the articles in the *American Political Science Review* and 80% of the articles in the *American Journal of Political Science* use quantitative analysis. The articles published in IR journal pay, however, less attention to measurement issues. On average 30% of the articles that use quantitative methods in IR journal discuss measurement error with *JCR* as the positive exception. Moreover, only 6% of the IR articles address validity or reliability explicitly. Measurement error, and particularly validity and reliability are addressed more frequently in articles in general political science journals. In the *American Political Science Review*, 40% of the articles address measurement error, while 28% and 18% are explicitly concerned with validity and reliability respectively.

IR scholars thus regularly disregard reliability issues, or at least do not feel obliged to discuss the reliability of their data. At the same time as we have shown above, their data are commonly based on highly subjective analysis of text sources. The relative lack of reports on reliability should be a serious concern for the IR research community. In IR concepts are often controversial and difficult to define, such as interstate cooperation,

[11] We classify an article as quantitative if it includes at least one regression table. We excluded in our category "quantitative" articles that introduce new datasets, if they did not satisfy our criterion of at least one regression table. Moreover, we excluded formal theory articles that did not have any empirics.

TABLE 7 Survey- Measurement Error, Validity and Reliability

Journal	No. of Articles	Quantitative	%	Measurement error	%	Validity	%	Reliability	%
American Political Science Review	213	105	49	42	40	29	28	19	18
American Journal of Political Science	321	262	82	94	36	32	12	26	10
International Organization	151	91	60	27	30	3	3	5	5
International Studies Quarterly	211	116	55	28	24	7	6	8	7
International Interactions	97	53	55	13	25	3	6	4	8
Journal of Peace Research	224	115	51	30	26	9	8	7	6
Journal of Conflict Resolution	207	127	61	55	43	12	9	8	6
IR Journals Average			57		30		6		6

intrastate conflict, state capacity, democracy, human rights abuse and governance. Further, IR researchers frequently rely for their information on textual sources, e.g., newspapers, media, and speeches.

Even though there are significant attempts to use machine coding to create events data (see Gerner et al. 2002; King and Lowe 2003), most of the IR literature still relies on human coding. Two of the most widely used datasets used in conflict studies, Uppsala Conflict Data and the Relative Power Data UCLA/ETH, are collected by human coders. Even more frequently, the IR literature does not rely on original datasets but on secondary data analysis. The most widely used data all employed human coding, such as the Polity IV project, COW data, Uppsala/PRIO data, Goldstein Conflict-Cooperation Scale for the WEIS events data, Minority at Risk, and Political Terror Scale, to list a few examples. The finding, at least in our data, that low reliability is not only related to stochastic errors but also to systemic, should alert IR scholars to the potential biases introduced in their analyses and trigger further research.

“Denial” and “resignation” appear to be the most common responses to issues of intercoder reliability in conflict studies using events data. It is still commonplace to ignore or not report intercoder reliability, or to treat it as a minor issue resolved along the way. As an alternative approach, we suggest “selection,” where researchers use information on coders’ performance to improve their data quality. Selection should be an attractive alternative to discarding all data if ex post intercoder reliability is shown to be poor. Discarding all data would mean that good as well as poor data get thrown away. Although ideally one might prefer to start all over, this is often not a realistic option given time and money constraints. A reasonable alternative is to select the “good” (or at least “better”) coders and to rely exclusively on their information. Obviously, it is important to avoid possible bias because coders become overly concerned with meeting the criteria set by the principal investigators (instead of simply accurately applying the coding rules). It is also important to inspect whether the selection of coders has an impact on one’s findings, and whether evidence can be found to corroborate the categorization of “good” and “bad” coders.

We want to stress that our findings and suggestions are not unique to the content of our data—the UN Secretarial General reports on peacekeeping—but derive from the *nature* of events data. The selection of the events is often a thorny issue, and the concerns raised in this paper are relevant for all events data. The criteria we develop to select bad/good coders are transferable to other projects collecting events data from textual sources. The core of our argument is not unique to our peacekeeping data, but indicates a larger problem with events data. Table 8 lists a few suggestions that researchers can follow to address intercoder reliability. Sharpening the measurement instrument (codebook) is obviously crucial. It is worthwhile to take sufficient time for completing a few pilot studies before devising the final codebook. Extensive training of coders will definitely improve the quality of any data collected. In addition, we suggest that it is crucial to carefully select coders and to regularly assess the quality of their work. For obvious reasons, it is important that the coders are unaware of the specific selected criteria to allow for meaningful evaluation. The criteria should also be relevant for the research project and ideally as objective as possible.

It is our experience that the quality of coding vastly improves with an authoritative identification of events. While this approach remains subjective, it is at least external to the individual coders. The identification of events often presents the biggest hurdle to intercoder reliability, and appears to be a common concern to all events data projects based on text analysis. Finally, as we have argued, post hoc reliability tests can help to identify weak coders and improve the quality of the data and analyses that are ultimately reported.

TABLE 8 Tips on Data Events Coding

1) Spend sufficient time on creating the codebook (use pilot studies)
2) Coder training should be on sources already coded by project managers
3) First round coding: double coding
4) Run agreement and reliability tests on the previous stage
5) Select best coders (possible criteria: understanding, commitment, and sophistication)
6) Second round coding: one coder preselects events, two coders code
7) Run agreement and reliability tests on the second run of coders
8) Run final models using the best data available

REFERENCES

Alvarez, Michael, and John Brehm. (1995) American Ambivalence Towards Abortion Policy: Development of a Heteroskedastic Probit Model of Competing Values. *American Journal of Political Science* 39(4):1055–1082.

Beardsley, Kyle, and Holger Schmidt. (2007) UN Data Codebook. Unpublished manuscript, Emory University and George Washington University.

Baugh, Frank. (2003) Correcting Effect Sizes for Score Reliability. In *Score Reliability: Contemporary Thinking on Reliability Issues*, edited by Bruce Thompson. Thousand Oaks, CA: Sage.

Bollen, Kenneth. (1989) *Structural Equations with Latent Variables*. New York: Wiley.

Carmines, Edward, and Richard Zeller. (1979) *Reliability and Validity Assessment*. Beverly Hills: Sage.

Cederman, Lars-Erik, Luc Girardin, and Andreas Wimmer. (2006) Getting Ethnicity Right: An Expert Survey on Power Distributions among Ethnic Groups. Paper presented at the 102th Annual meeting of the American Political Science Association, Philadelphia, PA, 31 August–3 September.

Dorussen, Han. (2007) Introducing PKOLED: Peacekeeping Operations Locations and Event Dataset. Paper presented at the Annual GROW-Net Conference on Disaggregating the Study of Civil War and Transnational Violence, University of Essex, 24–25 November.

Dorussen, Han, Spyros Blavoukos, and Hartmut Lenz. (2005) Assessing the Reliability and Validity of Expert Interviews. *European Union Politics* 6(3): 315–338.

Dorussen, Han, and Theodora-Ismene Gizelis. (2008) Into the Lion's Den. The Local Reception of UN Peacekeeping. Paper presented at the 49th Annual ISA Convention, San Francisco, CA.

Franzosi, Roberto. (1994) From Words to Numbers. *Sociological Methodology* 24: 105–136.

Gerner, Deborah J., Philip A. Schrodt, Omur Yilmaz, and Rajaa Abu-Jabr. (2002) The Creation of CAMEO (Conflict and Mediation Event Observations): An Event Data Framework for a Post Cold War World. Presented at the annual meeting of the American Political Science Association, San Francisco, 29 August–1 September.

Gleditsch, Kristian Skrede, and Michael Ward. (1997) Double Take: A Reexamination of Democracy and Autocracy in Modern Polities. *Journal of Conflict Resolution* 41(3):361–383.

Hensel, Paul. (1998) Reliability and Validity Issues in the ICOW Project. Paper prepared for presentation at the 39th Annual Meeting of the International Studies Association, Minneapolis, 17–21 March.

Hruschka, Daniel, Deborah Schwartz, Daphne Cobb St. John, Erin Picone-Decaro, Richard A. Jenkins, and James W. Carey. (2004) Reliability in Coding Open-Ended Data. *Field Methods* 16(3):307–331.

Jaggers, Keith, and Ted R. Gurr. (1995) Tracking Democracy's Third Wave with the Polity III Data. *Journal of Peace Research* 32(4):469–482.

Kalyvas, Stathis. (2006) *The Logic of Violence in Civil War*. New York: Cambridge University Press.

King, Gary, and Will Lowe. (2003) An Automated Information Extraction Tool for International Conflict Data with Performance as Good as Human Coders. *International Organization* 57(3):617–642.

Landis, Richard, and Gary Koch. (1977) The Measurement of Observer Agreement for Categorical Data. *Biometrics* 33(1):159–174.

Leng, Russell J., and J. David Singer. (1988) Militarized Interstate Crises. The BCOW Typology and its Applications. *International Studies Quarterly* 32(2):155–173.

Liebetrau, Albert M. (1983) *Measures of Association*. Thousand Oaks, CA: Sage.

Mitchell, Ronald, and Steven Rothman. (2006) Creating Large-N datasets from Quality Information. Paper presented at the 102th Annual meeting of the American Political Science Association, Philadelphia, PA. 31 August – 3 September.

Neuendorf, Kimberly A. (2002) *The Content Analysis Guidebook*. Thousand Oaks: Sage Publications.

Olzak, Susan. (1992) *The Dynamics of Ethnic Competition and Conflict*. Stanford: Stanford University Press.

Raleigh, Clionadh, and Havard Hegre. (2005) Introducing ACLED: An Armed Conflict Location and Event Database. Paper presented to the conference, "Disaggregating the Study of Civil War and Transnational Violence," University of California Institute of Global Conflict and Cooperation, San Diego, CA, 7–8 March.

Ratner, Steven R. (1996) *The New UN Peacekeeping. Building Peace in Lands of Conflict after the Cold War*. New York: St Martin's Press.

Rothman, Steven. (2007) Understanding Data Quality through Reliability. *International Studies Review* 9(3):437–456.

Schrodt, Philip A. (2006) Twenty Years of the Kansas Event Data System Project. *The Political Methodologist* 14(1):2–8.

Sim, Julius, and Chris Wright. (2005) The Kappa Statistics in Reliability Studies. *Physical Therapy* 85(3):257–268.

Singer, J. David. (1984) Variables, Indicator, and Data. *Social Science History* 6(2):181–217.

Treier, Shawn, and Simon Jackman. (2008) Democracy as a Latent Variable. *American Journal of Political Science* 52(1):201–217.

Copyright of International Interactions is the property of Routledge and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.