Measurement and Data Aggregation in Small-n Social Scientific Research

Dirk Leuffen[1] ETH Zürich
Susumu Shikano[2] Universität Konstanz
Stefanie Walter[3] Universität Heidelberg

Paper prepared for the Symposium "Reassessing the Methodology of Process Tracing", Carl-von-Ossietzky-University Oldenburg, November 26th 2010.

Abstract
In this paper, we argue that measurement is a theoretically equivalent concept in small- and large-n research. Namely, both approaches share the same goals of validity, reliability and objectivity. At the same time, small- and large-n researchers often face different challenges in implementation. After discussing similarities and differences in the measurement process and presenting a collection of strategies for improving measurement quality in small-n research, we analyze the question of how best to aggregate data in small-n research. We introduce and theorize different aggregation strategies that are commonly used in triangulation. We then evaluate their performance using computer simulations. Our simulation results show that averaging different information sources, in general, outperforms other aggregation strategies. However, this is not the case whenever poorly informed sources are biased in a similar direction.

[1] Senior Researcher at the Center of Comparative and International Studies, ETH Zürich, Haldeneggsteig 4, 8092 Zürich, Switzerland. Email: Dirk.Leuffen@eup.gess.ethz.ch.
[2] Professor of Methodology at the Department of Politics and Public Administration, University of Konstanz, Universitätsstr. 10, Box 92, 78457 Konstanz, Germany. Email: susumu.shikano@uni-konstanz.de
[3] Junior Professor at the Institute for Political Science, University of Heidelberg, Bergheimer Str. 58, 69120 Heidelberg, Germany. Email: s.walter@uni-heidelberg.de.

Es erfüllte Humboldt stets mit Hochgefühl, wenn etwas gemessen wurde; diesmal war er trunken vor Enthusiasmus. Die Erregung liess ihn mehrere Nächte nicht schlafen. (Daniel Kehlmann, Measuring the World, p. 39)

1. Introduction*

Measurement is a key component of empirical research. Without measurement, empirical tests of theoretical arguments are impossible. Therefore in both small- and large-n research we need to identify and measure the empirical referents of our theoretical concepts. In process-tracing there is a particularly strong demand for fine-grained measurement, since the individual observations constituting the causal chain are supposed to be of great importance. In the last fifteen years we have witnessed an explosion of research on small-n research methodology.While much of this research has focused on how to increase the external and internal validity of small-n research designs, somewhat less attention has been paid to the issue of measurement (exceptions are e.g. Adcock and Collier 2001; Thies 2002; Geddes 2003; Goertz 2006). Therefore our paper seeks to advance the methodological debate on measurement in small-n social science research.

For the purpose of this article we define measurement broadly as "the process of making empirical observations in relation to a theoretical concept" (Collier et al. 2004: 295). Measurement thus provides the central linkage between a theory and its real-world implications. It includes any assignment of particular values or categories of the theoretical concept to empirical observations (Geddes 2003: 145). This definition is more generic than the classic definition of measurement as the rule-based assignment of numeric values to objects or events (Stevens 1946; Blalock 1982). It explicitly recognizes that the classifying or rank-ordering of empirical objects or events into verbally-described categories is theoretically analogous to assigning each object a numerical value on an interval or ratio-level scale.[4]

* We thank Thomas Jensen and Hillel David Soiffer for useful comments on a previous version of this paper.

[4] In contrast to our conceptualization of measurement, some authors argue that the process of classifying objects into different categories (or measurement on a nominal scale) is not measurement at all (Sartori 1975). Following Stevens (1946), however, we argue that nominal and ordinal scales are equally valid levels of measurement as interval and ratio scales.

There is a general agreement that good measurement strives to maximize validity, reliability and objectivity. Measurement error poses significant obstacles to the goal of drawing valid causal inferences. While the pitfalls of measurement error have been widely discussed (e.g. King et al. 1994: 155ff.; Bartels 2004), there is still limited research on *how* to minimize measurement error in qualitative social research. While most methodologists agree on the importance of triangulation, it is similarly not at all obvious how to aggregate the rich information typically collected by qualitative researchers. This paper therefore concentrates on the issue of aggregation of sources and data types in the context of triangulation. We first introduce and discuss a set of simple aggregation strategies. We then use computer simulations to test the performance of the different strategies. Our simulations highlight that calculating weighted averages is generally the most promising strategy. This finding holds under the assumption that different sources are not systematically biased in one direction. In case a researcher possesses multiple sources which seem to be biased in one direction, it is advisable to base one's measurement on the better informed sources. Thus the choice of aggregation strategy depends strongly on the informational assumptions a researcher formulates about the sources. This also means that the aggregation strategy can differ for different units of a causal argument. Depending on the sources and the data types, it can be advisable to shift the aggregation strategy from one observation to another – a strategy that would seem rather alien and ad-hoc for quantitative analysts striving for a stronger standardization.

The paper begins with an overview over insights from previous research on the issue of measurement quality and the challenge of collecting objective, reliable, and valid information in small-n research. The following section deals more narrowly with the issue of triangulation. Drawing on examples from the measurement of preferences, we discuss different aggregation strategies before using computer simulations to test the performance of these strategies. We conclude this paper with a discussion of our findings and their usefulness for small-n social scientific research.

2. Objectivity, Reliability, and Validity

As to measurement, empirical research shares a common goal: to produce precise and accurate measures of the theoretical concept in question. Precise and accurate measures are a prerequisite for informed statements about empirical regularities. Measurement error can pose serious obstacles to the goal of drawing valid inferences. To minimize measurement error and to maximize the quality of measurement and the inference drawn from empirical research, the literature distinguishes three general standards of good measurement: objectivity, reliability and validity. While there is a long debate on how to maximize these criteria in large-n research – for instance, test theory offers a rich arsenal in psychometrics (Lord et al. 1968) – it is less clear how these criteria can be met by small-n researchers. In the following we will therefore collect and present some suggestions for improving objectivity, reliability and validity in small-n research.

2.1 Improving Objectivity and Reliability of Qualitative Measures

Objectivity and reliability are key objectives of good measurement. Objectivity means that measurement results are independent of the individual researcher. That means if several persons assign the same score to a phenomenon, the objectivity of the measurement process seems high. One of the challenges of social science in this regard is the prevalence of concepts that are difficult or even impossible to measure in an objective manner, forcing researchers to rely on subjective assessments (Bollen and Paxton 1998). The criterion of reliability goes one step further and requires that applying the same procedure in the same way should always produce the same measure (King et al. 1994: 25). Objectivity and reliability are consequently closely linked to the issue of replicability. The more objective and reliable an indicator is, the easier it will be to replicate the scoring of each case, and vice versa. One practical implication of this mutual relationship is that researchers can increase the objectivity and reliability of their indicators by striving to maximize replicability.[5] In general, objectivity and reliability can be achieved more easily in quantitative research than in qualitative research. Clear, detailed, and standardized coding guidelines typically leave less room for individual

[5] This explains the strong emphasis textbooks have placed on the criterion of replicability (e.g. King et al. 1994; Geddes 2003).

discretion than verbal and non-standardized classification schemes. However, as we discuss below, there are a number of ways in which qualitative researchers can increase the replicability of their measures.

The problems associated with measures that exhibit low degrees of objectivity and reliability are the same in qualitative and quantitative research: the possibility of severe measurement error. This error can come in two forms: Systematic measurement error occurs when a researcher systematically biases the scoring of the individual cases in one direction. One example is confirmation bias, where the researcher systematically scores the different cases according to his or her expectations. Such a behavior threatens the internal validity of the research design, because it is no longer clear that the observed variation actually exists, or whether it is simply a result of this systematic measurement error.[6] Unsystematic measurement error arises when measurement is not very precise, so that the assigned scores are sometimes lower and sometimes higher than the true score (for ordinal or higher levels of measurement), or sometimes specify other categories than the true category in which the case belongs (for nominal indicators). In contrast to systematic measurement error, these errors occur in a random fashion, so that the measures are correct on average. But this poses considerable challenge for qualitative research, since the small number of cases typical for qualitative research means that the likelihood that measurement errors will cancel out is smaller than in large-n research designs. With few cases, unsystematic measurement can therefore lead to wrong causal inferences.[7]

How, then, can the objectivity and reliability of qualitative measurement be increased and measurement error reduced? One option is to increase the number of observations. With a larger number of observations, non-systematic measurement error is more likely to cancel out and hence is less likely to bias our results (King et al. 1994). However, two objections have been raised against this strategy. First, this route is frequently not feasible in qualitative research (see, for example, Brady and Collier 2004), either because are no more comparable units because the domain for the theoretical

[6] Note that when systematic measurement affects all units by the same constant amount, it biases descriptive, but not causal inference (King et al. 1994). However, systematic measurement error in the form of confirmation bias will affect causal inference.

[7] For a more detailed, though disputed, discussion of the effects of nonsystematic measurement error see King et al. (1994) and the reply in Brady et al. (2004).

concept is limited, so that increasing the number of observations would result in conceptual stretching (Sartori 1970; Collier and Mahon 1993; Leuffen 2007), or because increasing the number of observations comes at the price of not being able to measure an object of interest as intensely as necessary, which decreases measurement validity (as well as the price of additional resources and time). Increasing the number of observations also does not necessarily remedy the problem of systematic measurement error.

A second option is to increase measurement precision. The following four recommendations have been proposed to improve the objectivity and reliability of qualitative measures (King et al. 1994; Geddes 2003; Yin 2003):

(1) Use unambiguous, concrete, and complete classification criteria.

Detailed coding (or classification) schemes can be used to collect empirical information in an objective and reliable fashion (Geddes 2003). To maximize the usefulness of these coding schemes, the classification criteria employed should meet three standards: first, the coding guidelines should enable the researcher to classify each observation into one of the different categories covered by the theoretical concept. Second, the classification criteria should be as concrete as possible. Here, the difficulty is to construct the coding scheme in such a way that it is concrete enough to adequately capture the empirically observable attributes of the theoretical concept "on-the-ground" but simultaneously is flexible enough to be applied in a variety of contexts (Przeworski and Teune 1970; Locke and Thelen 1995; Munck 1998). Finally, the classification criteria should be complete in that they cover the entire possible value space. This implies that each case can be classified as belonging to one of the specified categories so that the coding scheme is collectively exhaustive.[8]

(2) Provide detailed documentation of the data collection process, all data sources, and the collected data.

Documentation allows others (including oneself after some months) to be able to recall, retrace and possibly even repeat each step in the research process.

[8] Geddes (2003: Appendix C, see also the discussion in chapter 4) presents an excellent example of a coding scheme that fulfills all of these criteria.

Importantly, this procedure disciplines the researcher to approach the measurement process in a diligent manner and to actively reflect on the quality of his or her inferences. To ensure replicability, a project's documentation should specify how the data was gathered, which sources were used and how they were selected, which categories were used and how the units of observation were classified. In short, the entire reasoning and practice of the data collection should be revealed in detail to the reader (King et al. 1994: 23). This is best achieved by preparing a detailed documentation file that can be provided to interested readers on request and by summarizing this information in a chapter or section of the main manuscript.

(3) Use multiple researchers to classify observations based on the same evidence base.

By employing multiple researchers to assign scores to individual observations based on the same evidence and using the same coding scheme, inter-subjective reliability or objectivity can be assessed and improved. Once more, more detailed coding schemes increase the chance that multiple researchers come to the same coding conclusions based on the available evidence.

(4) Use several sources and triangulate the information obtained from them.

Triangulation means that multiple sources (and potentially methods) such as primary and secondary sources, interview data, and participant observation, are used to measure the same concept for a single unit (King et al. 1995: 479-480). By evaluating the evidence provided by each source separately, the researcher can increase the number of measures. This can decrease the magnitude of unsystematic measurement error. Triangulation can also reduce bias and increase the objectivity of measurement, because potentially biased statements receive less weight in the final evaluation. Coding all available information and retaining this information in a well-organized documentation package additionally increases the replicability, reliability, and validity of the measurement process. As a result, multiple measures of the same variable for the same observation in as diverse forms as possible are usually better than a single measure. The major challenge with this research strategy lies in the question of how the information provided by

these different sources is to be aggregated. The third part of this paper explicitly addresses this question.

When these four recommendations are followed, the objectivity and reliability of qualitative measures is generally enhanced. Moreover, the first two recommendations are likely to significantly strengthen the replicability of qualitative research.

2.2 Measurement Validity

Measurement validity means that an indicator actually measures the theoretical concept it is supposed to measure; its scores "meaningfully capture the ideas contained in the corresponding concept" (Adcock and Collier 2001: 530). Achieving high levels of measurement validity is important, because a lack of valid measures makes it impossible to assess the internal validity of theoretical arguments: when measurement is not valid, an empirically observed relationship does not give us any information about the true relationship between the two theoretical concepts. Achieving accurate, or valid measurement is thus of central importance for any type of empirical research.

Measuring empirically what one has conceptualized theoretically requires attention to several issues. We focus on concept validity, or the question, whether the indicator actually measures what it is theoretically supposed to measure.[9] To achieve a high degree of concept validity, two issues need to be resolved. First, the researcher needs to make sure that all the different dimensions or attributes of the theoretical concept are measured and that these dimensions are aggregated in a manner that is consistent with the structure of the concept (for a detailed discussion of this point see Goertz 2006). For this task, a good concept specification on the theoretical level is essential. Second, the researcher needs to pay attention to the issue of context-dependence, or contextual specificity (Przeworski and Teune 1970; Locke and Thelen 1995; Munck 1998; Adcock and Collier 2001). Since contexts differ widely – in terms of the cultural, economic, or political environment – the same concept may take different forms in different contexts. The main challenge associated with contextual specificity is that a coding scheme needs to be constructed in such a way that it is concrete enough to

[9] See Adcock and Collier (2001) for an excellent treatment of other types of validation.

adequately capture the attributes of the theoretical concept "on-the-ground" and allows the researcher to produce comparable measures, but simultaneously is flexible enough to be applied in a variety of contexts.

Measurement validity is often considered a particular strength of qualitative research (George and Bennett 2005). This is because of the small number of cases under study qualitative researchers typically have more opportunities to consider carefully all the attributes of the concept for each specific case – think, for example, of the potential of open interviews to dig deeply into specific meanings. Moreover, triangulation is easier when the number of cases is relatively small, so that it is possible to use different sources to measure the same concept in a variety of ways. In addition to increasing the objectivity and reliability of qualitative measures, it can also help to offset systematic measurement error because of its strength of identifying certain biases (such as false memories). Triangulation therefore is also the most important tool for improving (convergent) validity in small-n research (Yin 2003; Brewer and Hunter 1989; Adcock and Collier 2001: 540). Once more, however, the crucial question is *how* the information provided by different sources should be aggregated in the process of triangulation.

3. Triangulation

Originally, triangulation is a geodetic technique for locating points in a space. For instance, in the past cartographers made use of angles and their geometric characteristics for fixing distant places on a map. In the social sciences triangulation is sometimes understood as an application of different methods on a single issue of interest (cf. Tarrow 2004: 178). We here use the term more narrowly; for us in the context of measurement triangulation means that a researcher uses multiple sources or data types to measure the same concept for a single unit. Such data triangulation is often considered key in improving (convergent) validity and minimizing bias (Yin 2003; Brewer and Hunter 1989; Adcock and Collier 2001: 540; Marks 2007).

To illustrate the use of and difficulties associated with triangulation, consider the following example: in her study of societal preferences on exchange-rate levels, Walter (2008) triangulates information from four data types. Table 1 provides an excerpt of her measurement of sectoral vulnerability to a depreciated exchange rate for the export sector

in Hong Kong during the Asian Financial Crisis of 1997/8. It shows that in order to measure the extent of this vulnerability, Walter has interviewed a government official, used secondary information from the literature and a local newspaper, as well as primary information from a government document. Since the export sector faced competitiveness problems because of a relatively appreciated exchange rate, the two secondary sources report a 'low' level of vulnerability to a depreciation of the exchange rate, whereas the level of vulnerability reported by the two primary sources is classified as 'low to intermediate.' Based on these assessments, Walter coded the sector's vulnerability as "low."

Source	**Statement**	**Coding**
Expert interview	"Exporters were affected by the devaluations of the other Asian countries, but Hong Kong's exports did not do particularly badly compared to all the other Asian countries. HK had other advantages that made the suffering less severe, such as many re-exports to China." (Interview 9)	Low to intermediate
Secondary literature	"Hong Kong's export-oriented services sectors (e.g., finance and tourism) had become uncompetitive with the fall in the currencies of its major customers and competitors in the region. Domestic industry and services also suffered from severe import competition." (Lim 1999: 104)	Low
Newspaper	"The part of the economy that is trade-related is dominated by re-exports from goods manufactured in the mainland, which devalued its currency two years ago and was un-likely to do so again soon, economists said. [...] "Along with China, Hong Kong's major exports are consumer products or textiles and garments, none of which are major exports by Southeast Asia," Kwok Kwok-chuen, chief economist at Standard Chartered Bank, said." (South China Morning Post (Hong Kong) Nov 23, 1997)	Low
Archival record	"Notwithstanding the sharp currency depreciation in the region, the price competitiveness of Hong Kong's exports was preserved to a considerable extent by the continued downward adjustment in export prices, moderating domestic inflation, and upgrading of productive efficiency and product quality." (Government of the Hong Kong SAR 1998: 5)	Low to Intermediate
OVERALL CODING		LOW

Table 1: Vulnerability of Hong Kong's export sector to relative price changes.

In this example, the level of vulnerability of Hong Kong's export sector to a depreciated currency was relatively easy to determine since all sources roughly agreed that Hong Kong's export sector is not very vulnerable to relative price changes. Walter

accordingly classified the sector's vulnerability as 'low'. But how should she have proceeded, had the sources contradicted one another? Imagine the interviewed expert had reported that a depreciation would have had extremely damaging effects on the export sector's viability. Should she have ignored this information, given that the other sources agreed? Or should she have changed her final coding for Hong Kong's export sector to "intermediate"? This example shows that while most authors agree that triangulation can reduce measurement errors, it is less clear *how* triangulation works in practice.[10] What do researchers do when they triangulate data from different sources and different data types? The aggregation of data in qualitative research is often done intuitively, such as using the information provided by the source judged to be most 'trustworthy'. While following one's intuition might lead to valid results, it does not necessarily yield this outcome, and the reliability and objectivity of one's measurement can be limited if the procedure is not reported.

To open up the black box of measurement aggregation (figure 1), we in the following systematize five different strategies of data aggregation. We later use computer simulations to test the performance and examine the effects of these different aggregation mechanisms.[11]

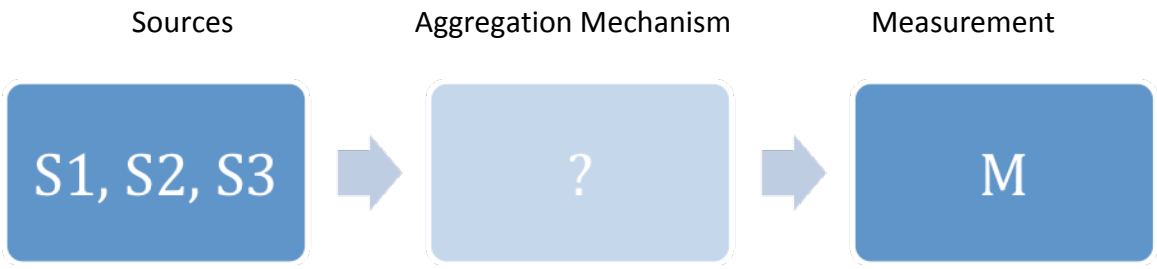

Figure 1:The black box of measurement aggregation. S1 is source 1, S2 is source 2 and S3 is source 3, M means Measurement.

[10] Gary Goertz and William Dixon cover the issue of aggregation in the context of forming a dyadic-level concept from individual-level attributes (cf. Goertz 2006: 129-155). From this we take step back by again focusing on individual-level attributes.
[11] Plümper et al. (2009) similarly use simulations in order to compare different case selection strategies.

3.1 Aggregation Strategies

We examine five simple aggregation strategies that can be used to aggregate information unearthed from different sources. To illustrate these approaches, we always assume that the researcher is triangulating from three different sources. Our examples are drawn from research measuring preferences, since preferences are important concepts for most political science analyses, despite being unobservable (Frieden 1999).

a) Random Selection

Random selection is our baseline strategy. Imagine we have information from three different sources. The researcher randomly just picks any of these three sources to derive his measurement. Of course, this is a poor strategy and there is essentially no triangulation. We therefore expect all other mechanisms to outperform random selection.

b) Arithmetic mean

Our next strategy consists of calculating the arithmetic mean of the values proposed by the individual sources. The arithmetic mean treats all sources equally, giving each source the same weight. When assuming a constant and unbiased quality of sources, the arithmetic mean should approach the 'true' value of our object with a growing number of sources. When a researcher has no information about the quality of her sources or data types, the simple average can therefore help to reduce measurement error. Figure 2 illustrates the arithmetic mean strategy for three sources that either score '0' or '1'.[12] Scenarios I and IV yield unambiguous results since all sources agree on either '0' or '1'. In scenarios II and III, however, intermediary scores of .3 and .7 result.

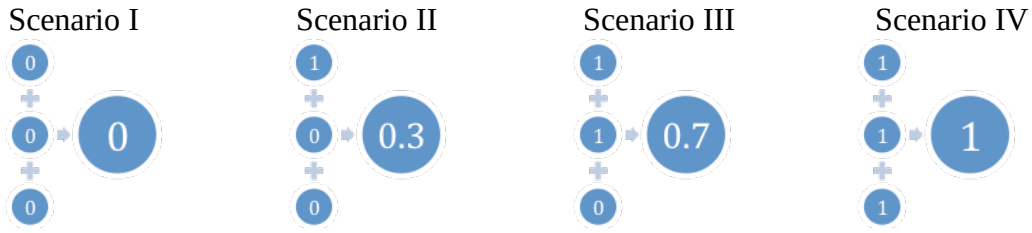

Figure 2:Arithmetic Mean Strategy: (A + B + C) / 3

[12] Note that we here assume that the mean has a substantive meaning; an application of this method to nominal scales is, of course, problematic.

A political science example for this simple strategy is Ray (1999: 288): In his expert survey on party orientations to European integration, the author averaged the evaluations of a minimum of five expert opinions to produce estimates of political parties' positions on the issue of European unification.[13]

c) Majority Strategy

According to the majority strategy, the researcher selects the most frequently recorded value, i.e. the mode, as the "true" value associated with an observation. This strategy assumes that agreement between different sources is a good indicator for correct measurement, but a requirement is that the sources are independent from one another. This means that the researcher can, for example, exclude the possibility that a newspaper has received its information from the same expert who was interviewed by the researcher. In our example in figure 3, if two out of three values agree, they determine the measurement, leading to one of the two possible outcomes '0' or '1'. As such, if restricted to a maximum of three sources, a researcher could stop collecting additional data, if one source is backed by an additional one. Note that if the sample is bimodal, multimodal or does not have a mode, this strategy cannot be applied in a straightforward manner.

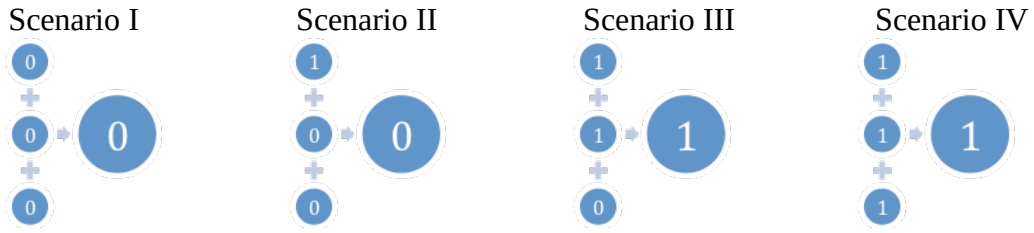

Figure 3:Majority strategy

d) Weighted average

In the weighted average strategy, the researcher possesses additional information on the quality of different sources. This information is, however, not perfect. Defining the

[13] An exclusion of responses qualified as deviant by the author, highlighted that (Ray 1999)'s expert positions were not biased by outliers. In addition, Ray compared his estimates to other sources such as Eurobarometer surveys and data provided by the Comparative Party Manifesto project.

quality of sources often is an arduous task, and can be achieved in an *ex ante* and *ex post* fashion. *Ex ante* rankings classify the quality of different types of sources according to previous knowledge or experience. For example, it has been argued that primary data from governmental archives should be preferred to interview data when measuring governmental preferences (Moravcsik 1998: 80ff.), because the memory of interview partners can be fragmentary and they might have strategic incentives to misrepresent their past preferences ex post. When interview data from various experts is available, the researcher can classify the different interviewees as to the access they had to the information concerned, but also their incentives to misrepresent the situation. For instance, when collecting data on French executive decision-making, it might not be advisable to talk to a President who currently faces an election campaign, and one might instead prefer interviewing a retired bureaucrat who had previously worked in the Elysée. Similarly, asking a random mayor of the President's party might not make sense when interested in decisions taken at the highest executive level. *Ex post* ranking of different sources can be based on the researcher's evaluation of the quality of different sources. When conducting interviews, one usually gets a good understanding of the respondent's engagement, his memory and the consistency of his explanations. Using this information, the researcher can qualitatively classify each interview in terms of its informational quality.

When different sources or data types are ranked according to their informational quality, a weighted aggregation strategy can be used. By including additional information about the quality of sources, this strategy is more complicated than simply using the arithmetic mean; however, the resulting measurement can be more nuanced.

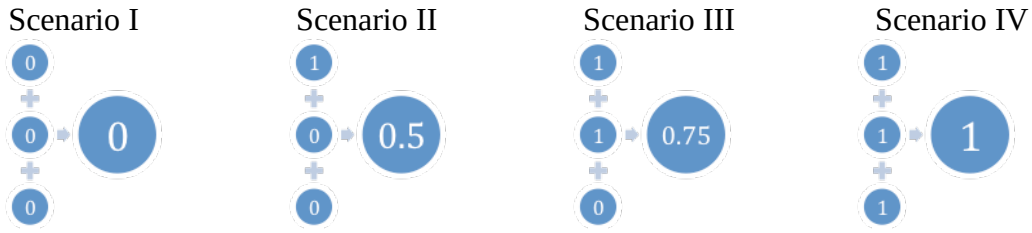

*Figure 4:Weighted average ((2*I) + II + III) / 4*

In our example in figure 4, the researcher considers source A as two-times more trustworthy than sources B and C. When comparing scenarios II and III from the arithmetic mean to the weighted average we accordingly find that the results are stronger driven by the source classified as being most trustworthy.

e) Winner takes it all

While in the weighted average scenario the final measurement always points towards the most reliable source, in the winner takes it all scenario the most reliable source wins and the other sources are not taken into account. In this strategy, the researcher thus again disposes of some information about the quality of each source. In our example in figure 5, source I is once more considered the most reliable source. As a consequence, the other two sources are not taken into account.

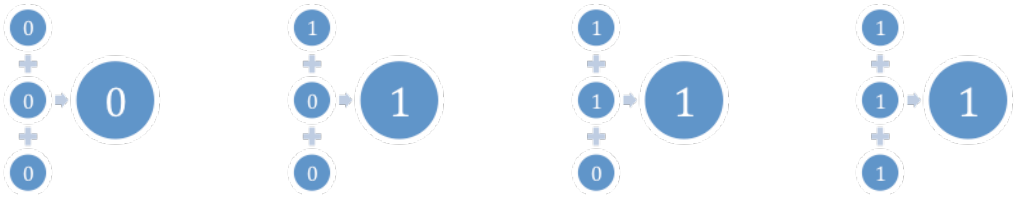

Figure 5: Winner takes it all

Thomson et al. (2006) provide an application of this strategy. To measure governmental preferences in EU decision-making, they only use the information provided by the interviewee with the best 'quality of the argumentation' (Thomson et al. 2006: 347). While this strategy assures consistency, it also causes waste since the information provided by other interviewees is not taken into consideration (or is only taken into consideration in order to evaluate and rank the different interviews). The winner takes it all strategy thus seems reasonable when the researcher is very sure about the superiority of one source. However, with less certainty, this strategy might not be preferable to the weighted average.

Overall, we see that as expected, all strategies agree in the two unambiguous scenarios I and IV, in which all sources unanimously point in one direction. However, they disagree in the more likely scenarios II and III, where there is considerable disagreement amongst sources about the true value of the observation under study. Which

strategy promises the most accurate measurements? To answer this question, the next section analyzes these aggregation strategies more systematically with the help of computer simulations. For this purpose, we determine their quality under different informational assumptions as well as number of sources used. For instance, we will determine at which level of trustworthiness the weighted average should be preferred over the simple average or the winner takes it all. By using simulations, we will thus make use of a method in order to test a method. But before turning to the simulations we will again illustrate the importance of such aggregation decisions by drawing on an example from our own empirical research.

3.2 Different Aggregation Strategies: An Example

In order to illustrate the practical implications of the aggregation strategies we will illustrate them by referring to a data collection on preferences of European Union member states. We explicitly selected an example where the information provided by the different sources differed to a great extent in order to highlight the importance of being clear about the aggregation mechanisms. From January to March 2009, one of us (Leuffen) collected data on preferences in EU decision-making. The data collection was closely modeled on the example of the Decision-making in the European Union dataset (Thomson et al. 2006). After having selected a sample of suitable European Commission legislative proposals, experts from member state representations in Brussels and of the European Commission were identified and contacted. For each proposal two to seven interviews were conducted. The experts were asked to classify the positions and saliences of all member states, the Commission, the European Parliament as well as the reversion point and the outcome for controversial issues on a scale ranging from 0 to 100. Table 2 shows information collected in three interviews with experts from three different member states (South, North and East) on one particular issue of COM (2007) 372 ("Proposal for a Council Regulation on the common organization of the market in wine and amending certain regulations"). The issue was about grubbing up of vineyards. While the Commission had initially envisaged a grubbing up of 400.000 ha of vineyards in the European Union the proposal cut this to 200.000 ha. Without this new piece of legislation, there would be no grubbing up. Therefore the reversion point is the status quo

with a grubbing up of 0 ha. In this case the position 100 refers to a grubbing up of 200.000 ha, the position 0 codes an opposition against the grubbing up of vineyards. The other positions refer to intermediary levels of grubbing up. For the sake of illustration, our table shows the positions on the scale from 0 to 100 reported by three anonymous respondents' for a particularly contested subgroup of member states, namely France, Germany, Italy, Spain, and the UK.

	Expert I	Expert II	Expert III	Mean	Mode	Weighted Average	Winner Takes it All
France	0	40	100	46	.	35	0
Germany	50	70	.	60	.	57	50
Italy	15	80	.	47.50	.	37	15
Spain	80	90	100	90	.	88	80
UK	100	50	25	58	.	69	100

Table 2: Preferences on grubbing up of vineyards; Source: Leuffen.

We see that the experts differ greatly in their assessment of member state positions on this issue. The four last rows highlight the resulting values from our different aggregation strategies. After having conducted the interviews, respondent number one was classified as the best informed source based on his argumentation and engagement – he later updated his interview information in an email after having again consulted his dossiers. For the calculation of the weighted average, we here decided that expert one should be weighted two times as strong as experts two and three. We see that the strategies – in this case of a particular high disagreement – produce some quite diverse solutions. Especially the winner-takes-it-all and the arithmetic mean differ substantively, for instance, in the case of France. The mode does not produce a result since there is no disagreement. This case can be considered particularly difficult since there is such great disagreement between the respondents. But all the more it highlights the importance of being clear about the aggregation mechanism. Following the advice given by (Thomson et al. 2006) in this case Leuffen opted for the winner-takes-it-all strategy, using only the evaluations made by expert I.

3.3 Simulating Different Aggregation Strategies

In recent years, computer simulations have been increasingly utilized in various fields of political science for different purposes. Some use this technique to derive implications from complex models which cannot be solved analytically (e.g. Laver, 2005). Others utilize it to evaluate and/or present the performance of statistical models (e.g. King, 2000). Despite different purposes, computer simulations share some common advantages. In particular, they allow researchers to create hypothetical situations that seem suitable for testing the internal validity of their theories. In fact, they allow us to design theoretically interesting scenarios that would be very hard to realize by collecting empirical data. Computer simulations can thus help researchers to explore more deeply and systematically the logics and consequences of their theoretical ideas. In this article, we use computer simulations to test the performance of the different aggregation strategies introduced above. The problem with empirical data in this context is that we often lack information about the true values political objects hold. Therefore it is hard to carefully evaluate the performance of different strategies. In the simulations, on the other hand, we can assume specific values for our objects of interest and then play around with and test different measurements and aggregation strategies.

In the following we thus assume that we know the true value of an object. We can then generate multiple sets of possible measures based on different number of experts. These experts differ in the information level that they possess. For each generated set of measures, we can then evaluate the performance of the aggregation rules and compare their performance in different settings.

More concretely, our computer simulations are set up as follows: We a-priori define a uni-dimensional continuous scale for a concept whose true value is assumed to be 50 without loss of generality. We further assume that every expert is uncertain about this true value to some degree. That is, every expert recognizes the true value with certain cognition error. For the errors, we assume a normal distribution whose expected value equals zero. That is, the probability of small errors is higher than that of large errors. We differentiate this probability of errors between experts with different information levels. This is visualized in Figure 6. The normal distribution of the left panel represents a better

informed expert which gives information near the true value (50) with quite high probability. In contrast, worse informed experts represented by the mid and right panel are more likely to give larger errors. This degree of making errors can be controlled in the simulation model by setting different levels of dispersion for the normal distribution. Technically, this is reached by varying the standard deviation.

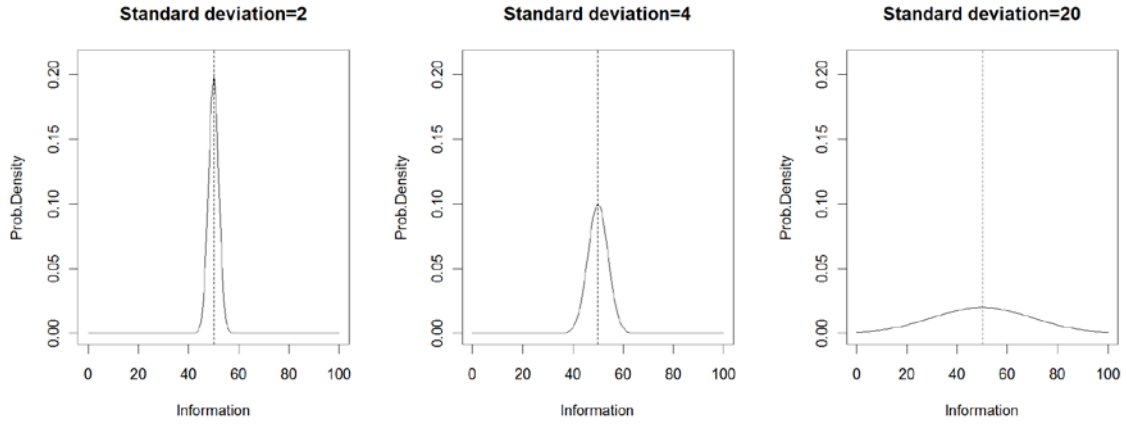

Figure 6: Sources with different information levels

Once we define the number of experts holding different information levels, we randomly draw information from the corresponding normal distributions independently for each expert.[14] After rounding the drawn number to an integer, the gained information is then aggregated by different rules discussed above. After we repeat this random draw and aggregation for 1000 times we can obtain 1000 measures for each aggregation rule. These results are, in turn, evaluated in terms of the 'true' value. For this purpose, we utilize the mean absolute error (MAE) defined as follows:

$$MAE_j = 1/1000 \sum_i | x_{ij} - 50 |,$$

where x_{ij} denotes the *i* th measure using aggregation rule *j*. In words, we take the average of absolute errors over 1000 simulated measures for each aggregation rule. As a matter of fact, a better measure should show a smaller MAE and a model with a perfect fit has a MAE of zero (cf. also Achen 2006).

[14] We relax this independence assumption later in the last scenario.

We first analyze a scenario in which all experts are relatively well informed. More specifically, we assign a standard deviation, (sd) of 2 for the better informed experts and and sd=4 for the other experts. We begin our simulations with two experts (one well informed expert and another worse informed expert) and thereafter increase the number of worse or better informed experts. Note that the second variation allows us to obtain a measurement using "winner takes it all" for two experts since there are multiple winners for three or more experts. Figure 7 shows the mean absolute errors (MAE) of different aggregation rules for an increasing number of experts. In the left panel we increased the number of the worse informed experts. We can clearly see that the MAE for both averaging rules, i.e. simple and weighted average, decrease over an increasing number of experts. It is reasonable that the difference among them is only marginal since all experts are almost equally well informed and weighting the best expert plays no significant role. In this simulation, the performance of "winner takes it all" is worse than both averaging rules and also the majority rule. While the "winner takes it all" strategy directly suffers from the cognition errors of the better informed expert, this does not hold for the other rules. Both averaging rules can cancel out the unsystematic cognition errors of multiple experts. The "majority" rule also suffers less from individual cognition errors. Since we draw the cognition of experts from a multivariate normal distribution centered on the true value, multiple experts more easily agree near true value rather than distant from the true value. As expected, the last strategy, "random choice", has the worst performance. This is no surprise since this rule suffers directly from the cognition errors of the randomly selected expert. The second panel of figure 7 shows the effect of increasing the number of better informed experts. Again we see that the two averaging strategies align rather closely. In fact, the performance of the averaging strategies of panel 1 is mirrored rather closely by panel 2. The reason for this is that the different sources again cancel each other out in the averaging process. However, this is not the case for the majority and random selection rule. While increasing better informed experts improves the measurement, increasing worse informed experts brings at best no improvement.[15]

[15] Note also that the individual steps on our X-axis can also be read as changing the ratio of less to better informed respondents.

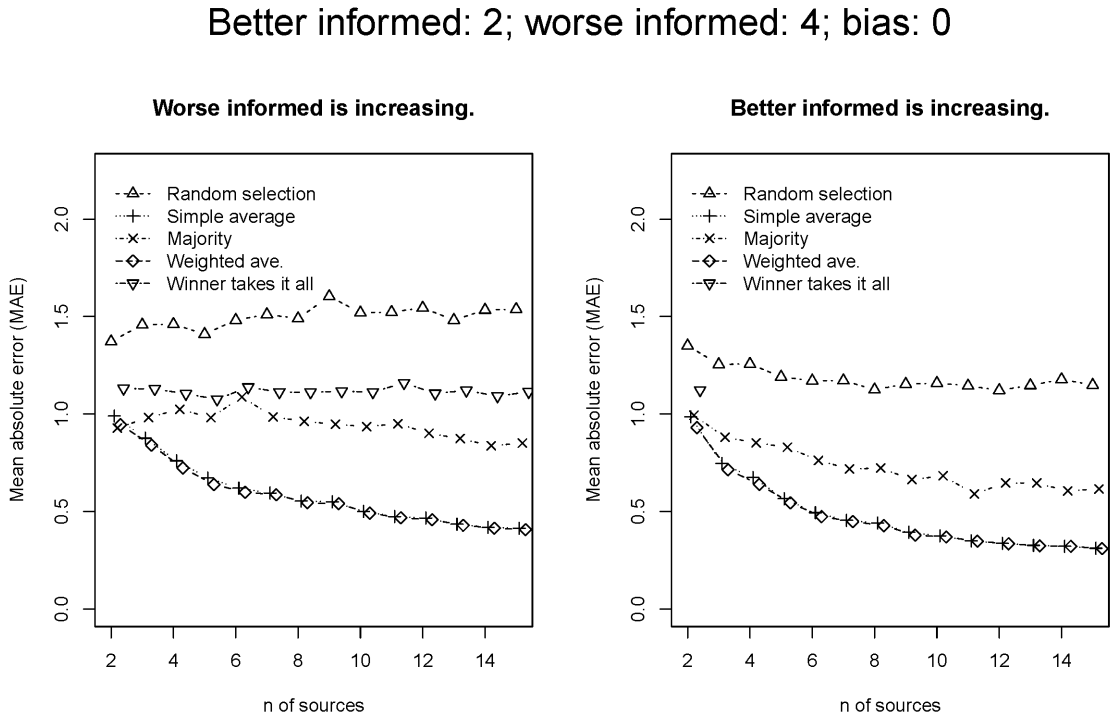

Figure 7: Scenario 1

In our next simulation, we keep the better informed expert with sd=2 and lower the information level of the other expert onto sd=20. That is, the better informed expert has 10-times better information about the true value than the others. Figure 8 gives us some different results from figure 7. First, while the level of MAE of "winner takes it all" remains constant the other aggregation rules have a worse performance. This is reasonable since all aggregation rules except for "winner takes it all" suffer from the less informed experts. Second, there is significant difference between both averaging rules. If only a limited number of experts is available the simple average is much worse than the weighted average. This is also the case in comparison with "winner takes it all" in the first variation increasing the number of the worse informed. If one has more than 10 experts available the simple average can outperform "winner takes it all". If one further increases the number of experts, the MAE of both averaging rules converges. Third, the majority rule seems to have a satisfactory performance with a small number of experts (2 to 4 experts). Behind this performance, however, there are also a large number of simulation runs in which the majority rule can achieve no aggregation since experts do not agree on a single value. This is more likely if only a small number of experts is

available. Fourth, both panels in Figure 8 clearly differ from one another. Not surprisingly the general performance of the strategies in panel 2 is higher.

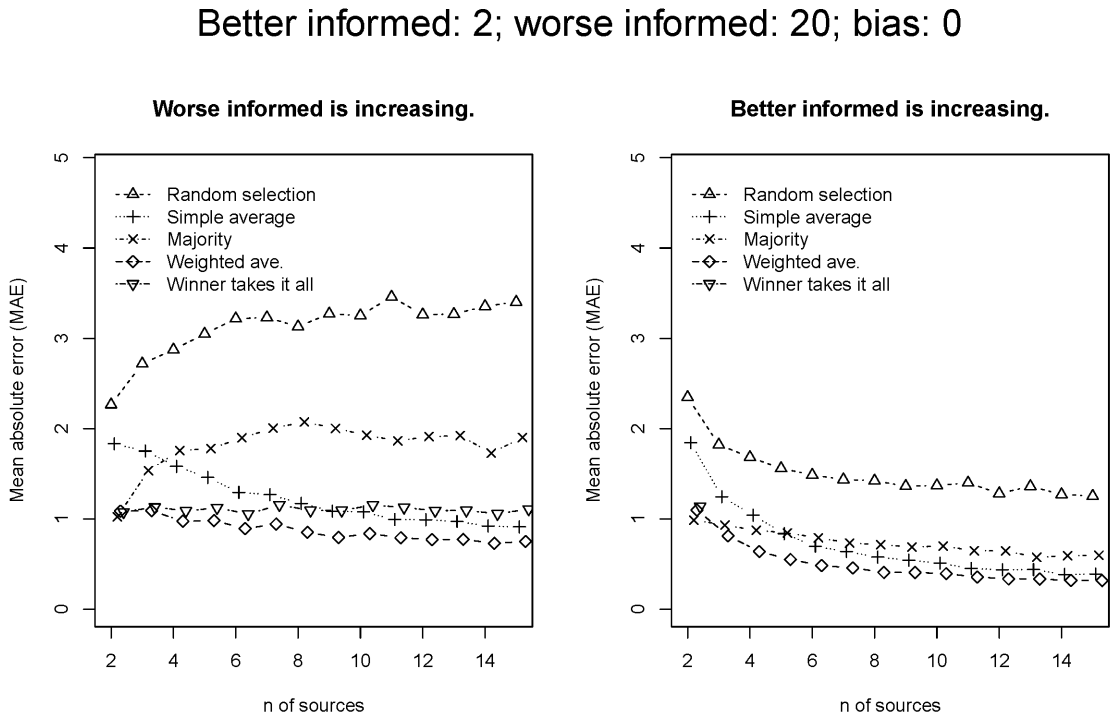

Figure 8: Scenario 2

From both scenarios above, we learn that averaging rules in general outperform the other rules. Only if the better expert is much better informed than the others and a limited number of experts is available the "winner takes it all" rule can outperform the simple average. But also in this case the weighted average is the best rule with the smallest MAE. This implicates that a certain number of less informed experts can outperform a small number of much better informed expert. This is, however, based on a strong assumption of the simulation scenarios above: the less informed experts' cognition errors are unsystematic and independent from each other. As introduced above, we draw expert cognitions from different normal distributions independently from each other. However, it might be more realistic that less informed experts can suffer from drawing their information from the same biased source; and thus their error is systematic. Therefore, we relax the independence assumption in our next simulation scenario.

To make the cognitions of less informed experts dependent from each other, we introduce a covariance of cognition errors among less informed experts. More

technically, we draw the cognition of experts from a multivariate normal distribution the non-diagonal elements of variance-covariance matrix of which have a value corresponding to a correlation of 0.5. That is, if a less informed expert overestimates in reporting the true value other experts also tend to make overestimated reports. We, however, assume that the better informed experts do not suffer from the biased information source. More technically, the co-variances between the better informed and the others is 0. Figure 9 presents our simulation results with the bias introduced above keeping other parameters the same as in the second scenario. The right panel is identical with that of the second scenario since the added experts who are better informed do not suffer from the bias. In contrast, the left panel shows different results as compared to the second scenario above.

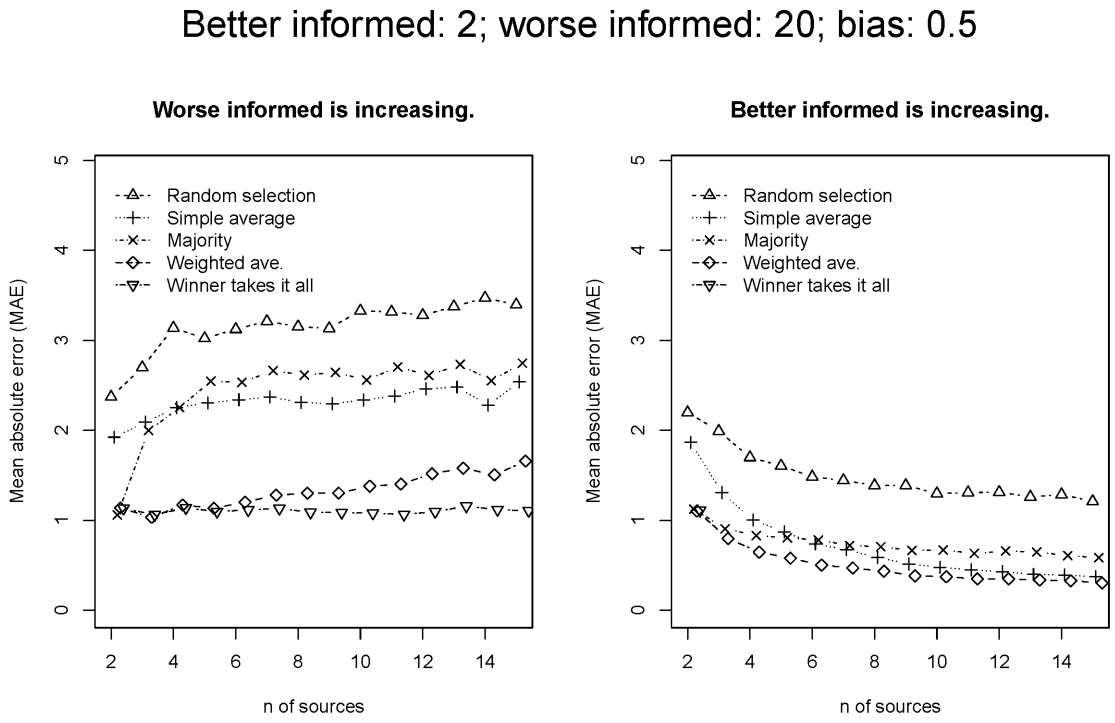

Figure 9: Scenario 3

First, the "winner takes it all" strategy shows the best performance. We can clearly see that the important feature of the best informed expert is its independence from the biased sources. Second, the increasing number of experts does not improve the other aggregation rules. The opposite is true. If one adds experts who possess the same biased information

of the already interviewed experts, one's effort is not only in vain, but also makes the final results worse. This tendency can be found most clearly in the weighted average since this rule considers only the individual uncertainty level of experts and not its covariance.

For us, these simple findings again highlight the importance of carefully selecting one's sources. One should attempt to collect information from different types of sources or actors to reduce the risks of systematic bias. For instance, when collecting information on the European policy positions of the French split-executive, Leuffen (2009) therefore always contacted supporters of the two main political camps in order to avoid partisan biases.

Our simulations should incite researchers to carefully consider their aggregation strategies. Their choice of strategy should depend on the number of sources to which they get access and their respective trustworthiness. In process-tracing, this should differ for different objects under investigation. Based on our analysis we therefore find that there are good reasons for changing the aggregation strategies in one study. This, as well as the definition of trustworthiness of sources, can be considered a qualitative choice researchers can take in small-n designs (and this constitutes a difference to more standardized procedures of large-n research). Based on our analysis we find that the straightjacket of standardized procedures might come at the price of reducing validity. Therefore, because of their in-depth case expertise, small-n researchers should be trusted to take such qualitative choices, but at the same time, for the sake of replicability, they should also be encouraged to be explicit about their decisions and actively discuss the tradeoffs behind their choices.

4. Conclusion

In this paper, we have taken stock of measurement in qualitative social research. After establishing that the fundamental logic of measurement is the same in quantitative and qualitative research and that both strands share the common criteria of validity, reliability and objectivity, we have explored how these goals can be met in qualitative social science research. Objectivity and reliability can be increased through clear and explicit documentations of the data collection process, concrete and unambiguous classification

criteria, inter-coder checks, and triangulation and source selection. Measurement validity can be strengthened through rigorous concept specification, attention to context, and the triangulation of different sources. Our main contribution here is that we have systematically introduced different strategies of aggregating sources in triangulation. Our simulations underline that generally more information should lead to better measurement results; however this only holds under the assumption that our sources are not systematically biased. Our simulations should encourage small-n researchers to work with (weighted) averages. In case of strong systematic biases they should opt for a "winner takes it all" strategy. Since the aggregation rule depends on informational criteria it can also vary for different objects of measurement. This seems a particularity of qualitative social science. All in all, we thus hope that our suggestions can contribute to solving some tensions between rigorous concept specification, replicable measurement processes and the explorative potential of case study research.

References

Adcock, Robert, and David Collier. 2001. "Measurement Vailidity: A Shared Standard for Qualitative and Quantitative Research." *American Political Science Review* 95 (3):529-45.

Arts, Bas, and Piet Verschuren. 1999. "Assessing Political Influence in Complex Decision-Making: An Instrument Based on Triangulation." *International Political Science Review* 20 (4):411-24.

Bartels, Larry. 2004. "Some Unfulfilled Promises of Quantitative Imperialism." In *Rethinking Social Inquiry. Diverse Tools, Shared Standards*, ed. H. E. Brady and D. Collier. Lanham: Rowman & Littlefield: 69-74.

Blalock, Hubert M. Jr. 1982. *Conceptualization and Measurement in the Social Sciences*. Beverly Hills: Sage.

Bollen, Kenneth, and Pamela Paxton. 1998. "Detection and Determinants of Bias in Subjective Measures." *American Sociological Review* 63 (3):465-78.

Brady, David, and David Collier, eds. 2004. *Rethinking Social Inquiry. Diverse Tools, Shared Standards*. Lanham: Rowman & Littlefield Publishers.

Brewer, John, and Albert Hunter. 1989. *Multimethod Research: A Synthesis of Styles*. Newbury Park, CA: Sage.

Collier, David, Henry E. Brady, and Jason Seawright. 2004. "Critiques, Responses, and Trade-Offs: Drawing Together the Debate." In *Rethinking Social Inquiry. Diverse Tools, Shared Standards*, ed. H. E. Brady and D. Collier. Lanham MD: Rowman & Littlefield Publishers: 195-228.

Collier, David, and James E. Mahon, Jr. 1993. "Conceptual "Stretching" Revisited: Adapting Categories in Comparative Analysis." *The American Political Science Review* 87 (4):845-55.

Frieden, Jeffrey A. 1999. "Actors and Preferences in International Relations." In *Strategic Choice and International Relations*, ed. D. A. Lake and R. Powell. Princeton: Princeton University Press: 39-76.

Geddes, Barbara. 2003. *Paradigms and Sand Castles. Theory Building and Research Design in Comparative Politics*. Ann Arbor: The University of Michigan Press.

George, Alexander L., and Andrew Bennett. 2005. *Case studies and theory development*. Cambridge: MIT Press.

Goertz, Gary. 2006. *Social science concepts. A user's guide*. Princeton, N.J.: Princeton University Press.

Government of the Hong Kong SAR. 1998. *First Quarter Economic Report 1998, May 1998*. Hong Kong: Economic Analysis Division, Financial Services Bureau.

King, Gary, Robert O. Keohane, and Sidney Verba. 1994. *Designing Social Inquiry. Scientific Inference in Qualitative Research*. Princeton: Princeton University Press.

King, Gary, Robert O. Keohane, and Sidney Verba. 1995. "Review: The Importance of Research Design in Political Science." *The American Political Science Review* 89 (2):475-81.

Leuffen, Dirk. 2007. "Case selection and selection bias in small-n research." In *Research Design in Political Science. How to practice what they preach*, ed. T. Gschwend and F. Schimmelfennig. Houndmills: Palgrave Macmillan: 145-60.

Lim, Linda Y. C. 1999. "Free Market Fancies: Hong Kong, Singapore, and the Asian Financial Crisis." In *The Politics of the Asian Economic Crisis*, ed. T. J. Pempel. Ithaca: Cornell University Press: 101-15.

Locke, Richard, and Kathleen Thelen. 1995. "Apples and Oranges Revisited: Contextualized Comparisons and teh Study of Comparative Labor Politics." *Politics & Society* 23 (3):337-67.

Lord, Frederic M., Melvin R. Novick, and with contributions ba Allan Birnbaum). 1968. *Statistical Theories of Mental Test Scores*. Reading: Addison-Wesley.

Mahoney, James, and Gary Goertz. 2006. "A Tale of Two Cultures: Contrasting Quantitative and Qualitative Research." *Political Analysis* 14:227-49.

Marks, Gary. 2007. "Introduction: Triangulation and the square-root law." *Electoral Studies* 26 (1):1-10.

Moravcsik, Andrew. 1997. "Taking Preferences Seriously: A Liberal Theory of International Politics." *International Organization* 51 (04):513-53.

Moravcsik, Andrew. 1998. *The choice for Europe. Social purpose and state power from Messina to Maastricht*. Ithaca, New York: Cornell University Press.

Munck, Gerardo L. 1998. "Canons of Research Design in Qualitative Analysis." *Studies in Comparative International Development* 33 (3):18-45.

Plümper, Thomas, Vera Tröger, and Eric Neumayer. 2009. *Case Selection and Causal Inference in Qualitative Research*. Unpublished manuscript, Toronto.

Przeworski, Adam, and Henry Teune. 1970. *The Logic of Comparative Social Inquiry*. New York: Wiley.

Ray, Leonard. 1999. "Measuring party orientations towards European integration: Results from an expert survey." *European Journal of Political Research* 36 (2):283-306.
Sartori, Giovanni. 1970. "Concept misformation in comparative politics." *American Political Science Review* 64 (4):1033-53.
Sartori, Giovanni. 1975. "The Tower of Babel." In *Tower of Babel: On the Definition and Analysis of Concepts in the Social Sciences*, ed. G. Sartori, F. Riggs and H. Teune. Interational Studies Association, Occasional Paper No.6: University of Pittsburgh.
Stevens, S.S. 1946. "On the Theory of Scales of Measurement." *Science* 103 (2684):677-80.
Thies, Cameron. 2002. "A Pragmatic Guide to Qualitative Historical Analysis in the Study of International Relations." *International Studies Perspectives* 3 (4):351-72.
Thomson, Robert, Frans N. Stokman, Christopher H. Achen, and Thomas König. 2006. *The European Union Decides*. Cambridge: Cambridge University Press.
Walter, Stefanie. 2008. "A New Approach for Determining Exchange-Rate Level Preferences." *International Organization* 62 (3):405-38.
Yin, Robert K. 2003. *Case Study Research. Design and Methods*. Thousand Oaks: Sage Publications.