SYMPOSIUM

measurement and data aggregation in small-*n* social scientific research

dirk leuffen[a],*, susumu shikano[b] and stefanie walter[c]
[a]University of Konstanz, Universitätsstr. 10, Box 83, Konstanz 78457, Germany
[b]University of Konstanz, Universitätsstr. 10, Box 92, Konstanz 78457, Germany
[c]Institute for Political Science, University of Heidelberg, Bergheimer Str. 58, Heidelberg 69115, Germany

*Corresponding author.

Abstract

How should small-*n* researchers aggregate the information collected during their research in an effort to measure the relevant theoretical concepts with high levels of validity and reliability? This article specifically focuses on the method of triangulation, which is frequently used in process-tracing approaches. We introduce and theorise different aggregation strategies commonly used in triangulation, such as weighted and simple averages or 'the winner takes it all' strategy. We then evaluate their performance with regard to their proneness to measurement error using computer simulations. Our simulation results show that averaging different information sources, in general, outperforms other aggregation strategies. However, this is not the case if poorly informed sources are biased in a similar direction; in these situations the 'winner takes it all' strategy shows a superior performance.

Keywords measurement; data aggregation; triangulation; small-*n*; computer simulation

Measurement is a key component of empirical research. Without measurement, empirical tests of theoretical arguments are impossible. In process-tracing, there is a particularly strong demand for high-quality measurement, since correctly classifying the individual observations is of great importance. While the significance of accurate and precise measurement has been widely recognised, and even though the pitfalls of measurement error have been discussed at length (e.g., King *et al*, 1994: 155ff.; Bartels, 2004), there is still limited research on how to minimise measurement error in qualitative social

research. For example, even though most methodologists agree on the importance of triangulation in small-*n* research, it is not at all obvious how to aggregate the rich information typically collected by qualitative researchers (but see Wagner *et al*, 2010). Our article therefore investigates the merits of different aggregation strategies of sources and data types in the context of triangulation.

For this purpose, we first introduce and discuss a set of simple aggregation strategies frequently used when triangulating from multiple sources and then use computer simulations to test the relative performance of these different strategies. Our simulations highlight that calculating weighted averages from the information drawn from multiple sources is generally the most promising strategy. However, if a researcher has collected information from multiple sources, which are biased in one direction, it is advisable to rely only on the better informed source. As a consequence, 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 observations in a study. Depending on the quality of sources, 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 stronger standardisation.

MEASUREMENT QUALITY AND TRIANGULATION

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

'*… there is still limited research on* how *to minimise measurement error in qualitative social research'*.

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 recognises 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.[1]

Most empirical researchers – both working in the qualitative and the quantitative research tradition – share a common goal with regard to measurement: to produce precise and accurate measures of the theoretical concept in question. For the purpose of minimising measurement error and maximising the quality of inferences drawn from empirical research, two main standards of good measurement have been proposed: reliability and validity. With regard to small-*n* research, the literature suggests that triangulation is well suited to improving qualitative measurement in relation to these two criteria, because the small number of cases under study typically gives researchers more opportunities to carefully consider several data sources for each specific case (e.g., Flick, 2008).

Data triangulation means that multiple sources or data types 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 or data type separately, the researcher can increase the number of individual

Figure 1 The black box of measurement aggregation. S1 is Source 1, S2 is Source 2, and S3 is Source 3, *M* means measurement.

measures. Figure 1 illustrates the process of triangulation. For each observation, the researcher consults a variety of sources (in this example Sources S1, S2, and S3). She then aggregates these sources through some aggregation mechanism to arrive at a single measure *M* for each observation.

Measures derived from triangulating different sources are usually expected to have a lower risk of both unsystematic and systematic measurement error, because potentially incorrectly measured or biased information receives relatively less weight in the final evaluation. For this reason, the use of multiple measures of the same variable for the same observation, in as diverse forms as possible, are usually argued to be better than a single measure. In addition, triangulation can help the researcher to identify certain biases (such as false memories), which might lead to systematic measurement error.[2] Triangulation therefore serves as a helpful tool for improving reliability and validity in small-*n* research (Yin, 2003; Brewer and Hunter, 1989; Adcock and Collier, 2001: 540; Marks, 2007).

The major challenge for the research strategy of triangulation is the question of *how* the information provided by these different sources is to be aggregated. What strategy should researchers follow when they triangulate data from different sources and different data types? Oftentimes the aggregation of different pieces of information is done intuitively in qualitative research, such as relying predominantly on the information provided by the sources judged to be most

'trustworthy' by the researcher. While following one's intuition may lead to valid results, it does not necessarily yield this outcome. Moreover, the reliability and objectivity of measurement can be limited if the aggregation procedure is not reported. To illuminate the question *how* to triangulate, this is *how* to aggregate data from different sources, the remainder of this article analyses the strengths and weaknesses of five different possible aggregation strategies.

FIVE DIFFERENT AGGREGATION STRATEGIES

To open up the black box of measurement aggregation, we examine five different 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, each of which suggests a certain value for the specific observation under study. Our examples are drawn from research measuring preferences, since preferences are unobservable but important concepts for most political science analyses (Frieden, 1999), but the logic discussed below equally applies to other concepts as well.

STRATEGY 1: RANDOM SELECTION

In the random selection strategy, the researcher randomly just picks any of these three sources to derive his

measurement. Of course, this is a poor strategy and triangulation effectively does not occur in this strategy. The strategy thus serves as a baseline strategy and we expect all other mechanisms to outperform random selection.

STRATEGY 2: ARITHMETIC MEAN (UNWEIGHTED AVERAGE)

The second strategy consists of calculating the arithmetic mean of the values proposed by each individual source. 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'. Scenarios I and IV yields unambiguous results, since all sources agree that the observation should be rated either '0' or '1'. In Scenarios II and III, however, the strategy

leads to intermediary scores of 0.3 and 0.7, respectively. 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.[3]

STRATEGY 3: MAJORITY STRATEGY

In the majority strategy, the researcher selects the most frequently recorded value, that is, 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, for example, that the researcher can 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 final value for this observation, leading to one of the two possible
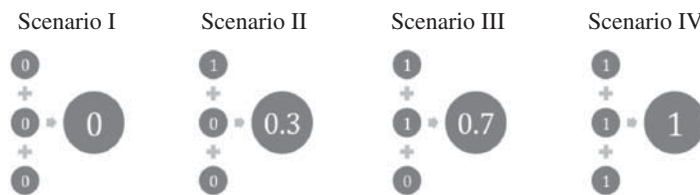

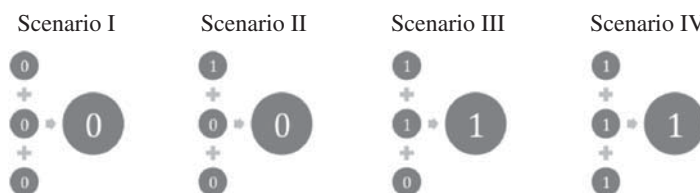**Figure 2** Arithmetic mean strategy: $(A + B + C)/3$.

Figure 3 Majority strategy.

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 straight-forward manner.

STRATEGY 4: 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 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 pre-ferred 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 not only classify the different interviewees as to the access they had to the information concerned, but also with regard to their incentives to misrepresent the situation. *Ex post* rankings 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 ex-planations. 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 informa-tional quality, a weighted aggregation strategy can be used. This means that the researcher attaches a greater weight to those sources judged to be of a higher quality. In our example in Figure 4, the researcher considers Source A as twice as trustworthy than Sources B and C. When comparing the outcomes of Sce-narios II and III, in which the sources disagree about the true value of the observation, of the weighted average strategy to the arithmetic mean strategy, we accordingly find that the aggregated value is influenced more strongly by Source A, which was classified as being most trustworthy. By including additional information about the quality of sources, this strategy is more complicated than simply using the arithmetic mean; how-ever, the resulting measurement can be more nuanced.

STRATEGY 5: WINNER TAKES IT ALL

Like the weighted average strategy, the 'winner takes it all' strategy relies on information about the quality of each data source. In contrast to the Strategy 4, however, in the 'winner takes it all'
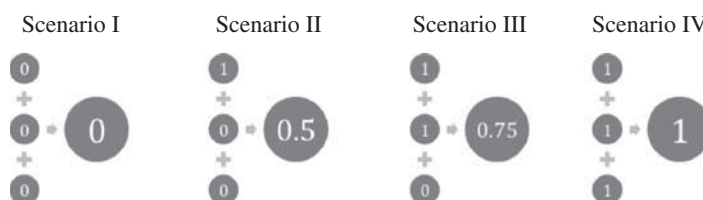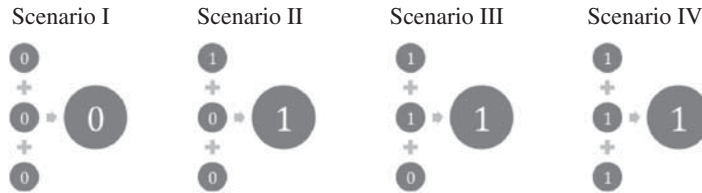

Scenario I	Scenario II	Scenario III	Scenario IV
0 + 0 + 0 → 0	1 + 0 + 0 → 0.5	1 + 1 + 0 → 0.75	1 + 1 + 1 → 1

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

Figure 5 Winner takes it all.

strategy only the most reliable source is considered, whereas the other sources are not taken into account. In our example in Figure 5, Source A is once again considered as the most reliable source. As a consequence, the information provided by the other two sources is discarded. 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).

Thomson *et al* (2006) provide an example for this strategy. To measure governmental preferences in EU decision making, they interview experts about the salience and position of governments regarding different European policy proposals. The experts are asked to identify the government positions on a scale ranging from 0 and 100, and to include all policy options that were on the table. The salience also ranges from 0, meaning no salience, to 100, meaning a very high salience. They then finally only use the information provided by the interviewee who displayed the best and most consistent 'quality of the argumentation' (Thomson *et al*, 2006: 347).

SIMULATING THE QUALITY OF DIFFERENT AGGREGATION STRATEGIES

Overall, we see that as expected, all five aggregation 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 among sources about the true value of the observation under study. Which strategy provides the most accurate measurements? To answer this question, we analyse the performance and effects of these aggregation strategies more systematically with the help of computer simulations.[4]

Computer simulations allow researchers to design hypothetical, but theoretically interesting scenarios that would be very hard to realise by collecting empirical data. In the context of triangulation strategies, for example, the problem is that we often lack information about the true values of individual observations, which makes it difficult to carefully evaluate the performance of different aggregation strategies. In computer simulations, in contrast, we can assume specific true values for each observation and can generate multiple sets of possible measures based on different numbers of sources and different assumptions about the accuracy of information that each source possesses. For each generated set of measures, we can then evaluate and compare the performance of each aggregation strategy. The computer simulations will therefore allow us to explore more deeply and systematically the logic and consequences of the different aggregation strategies discussed above.

SIMULATION SETUP

To compare the strengths and weaknesses of these strategies, 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 source is uncertain about this true value to some degree. That is, every source recognises the true value with certain cognition error, which is normally distributed with an expected value of zero. This means that the probability of small errors is higher than that of large errors and that the sources are not biased. The distribution of cognition errors depends on the information level of a source (see Figure 6): the better informed a source is, the narrower this distribution, signified by a small standard deviation, and the higher the probability that the source will report information near to the true value of the observation (in our case the value of 50). In contrast, worse informed sources (represented by large standard deviations, see the mid and right panel of Figure 6) are more likely to report values that deviate strongly from the true value. By setting different levels of dispersion for the normal distribution by varying the standard deviation, we can control the level of information of each source in the simulation model.

After defining the number of sources and their respective information levels, we randomly draw information from each source's normal distribution, round the value given by the source to an integer, and then aggregate the thus collected information by each of the different aggregation strategies discussed above. We repeat the random draws and aggregate the information 1,000 times, so that we obtain 1,000 measures for each aggregation rule. These results are then evaluated in relation to the 'true' value (i.e., 50) for the observation under study. For this purpose, we calculate the absolute error for each of the 1,000 simulated measures for each aggregation rule and use this information to compute the mean absolute error (MAE), which is defined as follows:

$$\text{MAE}_j = (1/1{,}000) \sum_i |x_{ij} - 50|$$

where x_{ij} denotes the i th measure using aggregation rule j. Since the MAE tells us how strongly the aggregated measure deviates on average from the true value, smaller MAE values denote more accurate measurement. An aggregation strategy that perfectly predicts the true value would have an MAE of zero (Achen, 2006).

SIMULATION RESULTS

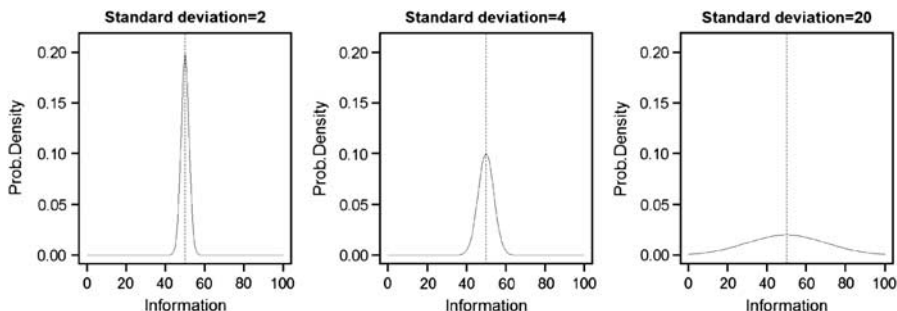We analyse two scenarios. In both scenarios, sources differ significantly

Figure 6 Sources with different information levels.

in their level of information. In the first scenario, we assume that the cognition errors of each source are unsystematic and independent from each other; in technical terms, we draw expert cognitions from different normal distributions independently from each other. However, less informed experts often draw their information from the same biased source so that their cognition errors are systematically related. One example is tendentious reasoning by partisan interviewees that may introduce a systematic bias if an analyst focuses on sources from one partisan camp.[5] In the second scenario, we therefore relax the independence assumption and allow the information to be systematically biased.

Scenario 1

In the first scenario, we assume that the better informed source has information about the true value of the observation that is three times more accurate than the information provided by the less well-informed sources, reflected by a standard deviation of 2 for the well-informed sources, and a standard deviation of

6 for the less well-informed sources. Figure 7 shows the mean average errors for different aggregation strategies. The left-hand panel shows the results for the different aggregation strategies when the proportion of worse informed sources relative to well-informed sources is increasing, whereas the right-hand panel shows the results for the opposite case (increasing number of well-informed sources).[6]

The results of the simulations in Figure 7 suggest a number of conclusions. First, not surprisingly, the mean average error in the aggregated measure tends to be smaller in the latter scenario, in which more well-informed sources are consulted. Second, the different aggregation rules differ in the average quality of measures, which they produce. As expected, the baseline strategy of random selection performs worst. Third, the quality of the average measure initially increases when more sources are consulted, but then quickly stabilises. Collecting information from more sources hence, is an endeavour of decreasing returns. Finally, the weighted average aggregation strategy outperforms all
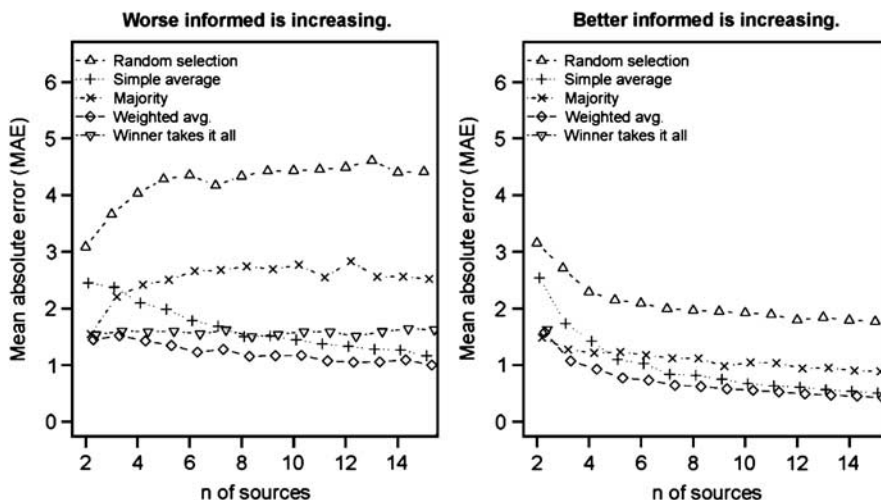

Figure 7 Scenario 1.
Notes: Better informed: 2; worse informed: 6; bias: 0.

other strategies in all simulations. The simple average strategy, in contrast, performs much worse than the weighted average strategy, especially when the number of sources is relatively small. Importantly, the weighted average strategy also performs better than the 'winner takes it all' strategy. This suggests that the collective knowledge of significantly less informed sources can still lead to better measurement than reliance on a single well-informed source.

Scenario 2

In the second scenario, the less informed sources draw their information from related sources; this means that their information is biased in the same direction. Examples for such instances include interviews with journalists who all draw their information from the same press releases and meetings, or party members who all share a similar partisan bias in their opinions. To model the fact that their assessments are no longer independent from each other, we introduce a covariance of cognition errors among less informed experts. More technically, we draw the information of multiple sources simultaneously from a multivariate normal distribution, which is based on a variance–covariance matrix whose non-diagonal elements correspond to a correlation of 0.5. That is, if a less informed source overestimates the value of an observation, other sources also tend to report overestimated values. Because of their superior access to information, we further assume that the better informed sources do not share the same information source, so that their estimates are still independent from each other.[7] Figure 8 presents the simulated mean absolute errors for the five different aggregation strategies using the same informational assumptions as in Scenario 1 with the exception that the information provided by the less informed sources is not independent. Note that the right-hand panel is identical with that of Scenario 1 since adding more well-informed sources does not introduce bias.

The results show that compared with Scenario 1, the 'winner takes it all' strategy now shows the best performance when the proportion of well informed to less well-informed sources exceeds the
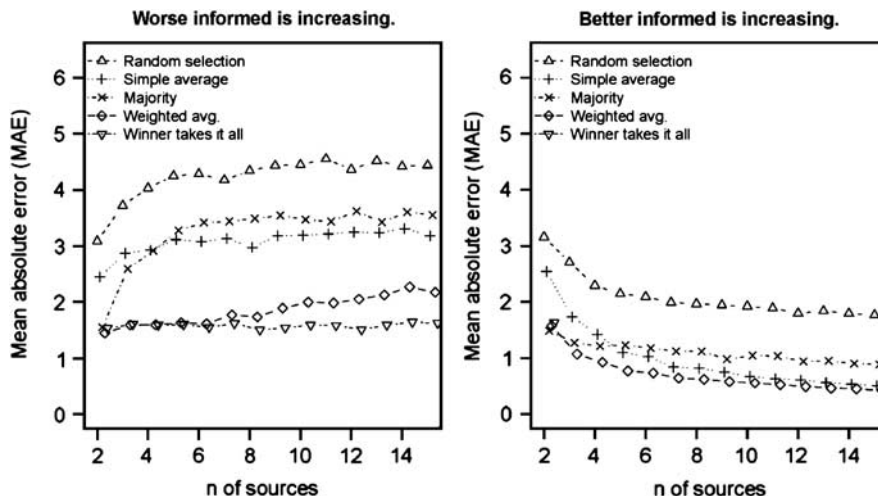

Figure 8 Scenario 2.
Notes: Better informed: 2; worse informed: 6; bias: 0.5.

ratio of 1:4. Given the dependence of their information, adding sources, which possess the same biased information as sources that have already been consulted, decreases the performance of aggregation strategies that build on averaging.[8] This tendency is most evident in the weighted average strategy since it considers only the individual uncertainty level of experts and not its covariance. Note, however, that the superior performance of the 'winner takes it all' strategy rests on the assumption that this source's information is independent from those of the less well-informed sources. These findings highlight the importance of carefully selecting sources in qualitative research. In particular, the results suggest that researchers should attempt to collect information from different types of sources, or actors, in order to reduce the risk of systematic bias.

CONCLUSION

While most researchers agree that precise and accurate measurement is important for qualitative research strategies such as process tracing, it is less clear *how* such high-quality measurement can be achieved. It is particularly unclear how researchers should aggregate information collected from different sources when performing triangulation. To shed light on this question, this article has used computer simulations to evaluate the quality of five different aggregation strategies.

The results underline that researchers should carefully consider their aggregation strategies. In general, collecting more information and using all of this information, weighted by the quality of the source typically leads to better measurement results. However this only holds under the assumption that the sources are not systematically biased. This suggests that small-*n* researchers should

'... *researchers should attempt to collect information from different types of sources or actors in order to reduce the risk of systematic bias'.*

'... *small-n researchers ... should ... be encouraged to be explicit about their decisions and actively discuss the tradeoffs behind their choices'.*

work with (weighted) averages when they believe their sources to be independent, whereas they should opt for the 'winner takes it all' aggregation strategy when strong systematic biases can be suspected. Which aggregation rule is suited best in a particular context hence depends on the number of sources, their respective trustworthiness, and their level of independence. In process-tracing, this context can differ for different observations under investigation.

On the basis of our analysis, we therefore find that there can be good reasons for changing the aggregation strategies in one study, a finding which seems a particularity of qualitative social science. 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 trade-offs behind their choices. In doing so, the quality of their measures can be greatly enhanced.

Acknowledgements

Notes

1 We thus do not follow the argument that the process of classifying objects into different categories (or measurement on a nominal scale) does not constitute measurement (Sartori, 1975).

2 Not surprisingly, measurement validity is often considered a particular strength of qualitative research (George and Bennett, 2005).

3 To ensure that expert positions were not biased, the author recalculated his scores on a subsample by excluding those responses that were qualified as deviant by the author. This, however, did not substantively change his results. In addition, Ray compared his estimates with other sources such as Eurobarometer surveys and data provided by the Comparative Party Manifesto project.

4 Our approach is similar to that of Plümper *et al* (2009), who use computer simulations to compare different case selection strategies.

5 We thank an anonymous reviewer for making this point clear to us.

6 Note that we cannot use the 'winner takes it all' strategy in this scenario.

7 More technically, the co-variances between the better informed and the others is 0.

8 For similar results concerning the wisdom of crowd effect see, for example, Lorenz *et al* (2011).

References

Achen, C.H (2006) 'Evaluating Political Decision-Making Models', in R. Thomson, F.N. Stokman, C.H. Achen and T. König (eds.) *The European Union Decides*, Cambridge: Cambridge University Press, pp. 264–298.

Adcock, R. and Collier, D. (2001) 'Measurement validity: A shared standard for qualitative and quantitative research', *American Political Science Review* 95(3): 529–545.

Bartels, L. (2004) 'Some Unfulfilled Promises of Quantitative Imperialism', in H.E. Brady and D. Collier (eds.) *Rethinking Social Inquiry: Diverse Tools, Shared Standards*, Lanham, MD: Rowman and Littlefield, pp. 69–74.

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

Brewer, J. and Hunter, A. (1989) *Multimethod Research: A Synthesis of Styles*, Newbury Park, CA: Sage.

Collier, D., Brady, H.E. and Seawright, J. (2004) 'Critiques, Responses, and Trade-Offs: Drawing Together the Debate', in H.E. Brady and D. Collier (eds.) *Rethinking Social Inquiry: Diverse Tools, Shared Standards*, Lanham, MD: Rowman and Littlefield, pp. 195–228.

Flick, U. (2008) *Managing Quality in Qualitative Research*, Thousand Oaks, CA: Sage.

Frieden, J.A. (1999) 'Actors and Preferences in International Relations', in D.A. Lake and R. Powell (eds.) *Strategic Choice and International Relations*, Princeton: Princeton University Press, pp. 39–76.

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

George, A.L. and Bennett, A. (2005) *Case Studies and Theory Development*, Cambridge: MIT Press.

King, G., Keohane, R.O. and Verba, S. (1994) *Designing Social Inquiry: Scientific Inference in Qualitative Research*, Princeton: Princeton University Press.

King, G., Keohane, R.O. and Verba, S. (1995) 'Review: The importance of research design in political science', *The American Political Science Review* 89(2): 475–481.

Lorenz, J., Rauhut, H., Schweitzer, F. and Helbing, D. (2011) 'How social influence can undermine the wisdom of crowds effect', *Proceedings of the National Academy of Sciences USA (PNAS)* 108(22): 9020–9025.

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

Moravcsik, A. (1998) *The Choice for Europe: Social Purpose and State Power from Messina to Maastricht*, Ithaca, NY: Cornell University Press.

Plümper, T., Tröger, V. and Neumayer, E. (2009) 'Case selection and causal inference in qualitative research', Unpublished manuscript, Toronto.

Ray, L. (1999) 'Measuring party orientations towards European integration: Results from an expert survey', *European Journal of Political Research* 36(2): 283–306.

Sartori, G. (1975) 'The Tower of Babel', in G. Sartori, F. Riggs and H. Teune (eds.) *Tower of Babel: On the Definition and Analysis of Concepts in the Social Sciences*, International Studies Association, Occasional Paper No.6, Pittsburgh, PA: University of Pittsburgh.

Stevens, S.S. (1946) 'On the theory of scales of measurement', *Science* 103(2684): 677–680.

Thomson, R., Stokman, F.N., Achen, C.H. and König, T. (2006) *The European Union Decides*, Cambridge: Cambridge University Press.

Wagner, S., Rau, C. and Lindemann, E. (2010) 'Multiple informant methodology: A critical review and recommendations', *Sociological Methods & Research* 38(4): 582–618.

Yin, R.K. (2003) *Case Study Research: Design and Methods*, Thousand Oaks, CA: Sage Publications.

About the Authors

Dirk Leuffen is a Professor of Political Science and International Relations at the Department of Politics and Public Administration, University of Konstanz. His research interests are on EU politics, on the linkage between international relations, and domestic politics and on small-*n* research methods.

Susumu Shikano is a Professor of Political Methodology at the Department of Politics and Public Administration, University of Konstanz. He is also affiliated with the Hanse Institute for Advanced Study in Delmenhorst and the Mannheim Centre for European Social Research at the University of Mannheim. His research interests lie in electoral politics, coalition formation, spatial models of politics, and simulation techniques in political science.

Stefanie Walter is a Junior Professor for International and Comparative Political Economy at the Institute for Political Science, University of Heidelberg. In her research she focuses on the politics of exchange-rate, monetary and fiscal policymaking, globalisation, distributional conflict, and economic voting.