Improving the Measurement Validity of Quantitative Empirical Assessments of Democracy. Recommendations for Future Comparative Research on the Quality of Democracy and Political Support

Inauguraldissertation
zur
Erlangung des Doktorgrades
der
Wirtschafts- und Sozialwissenschaftlichen Fakultät
der
Universität zu Köln

2020

vorgelegt
von

Wiebke Breustedt, M.A.

aus

Heidelberg

Referent: Prof. Dr. André Kaiser, Universität zu Köln
Korreferentin: Prof. Dr. Susanne Pickel, Universität Duisburg-Essen
Tag der Promotion: 24.02.2021

Für Ida und Sven

Danksagung

Die Promotionszeit war eine Phase des intensiven Lernens für mich, fachlich wie persönlich. Umso dankbarer bin ich den vielen Menschen, die mich in dieser Zeit begleitet und unterstützt haben. An erster Stelle danke ich meinen Betreuer:innen. Prof. Dr. André Kaiser (Universität zu Köln) danke ich für sein ausgesprochenes Vertrauen in meine Fähigkeiten und den richtigen Rat zur rechten Zeit. Prof. Dr. Susanne Pickel (Universität Duisburg-Essen) habe ich nicht nur mein Interesse an der empirischen Demokratieforschung und quantitativen Methoden zu verdanken. Sie ist für mich ein Vorbild für universitäre Lehre auf Augenhöhe und außerordentliches, professorales Engagement für Studierende und Promovierende.

Mein besonderer Dank gilt auch der Universität zu Köln für die Finanzierung meiner Promotion im Rahmen der Cologne Graduate School sowie des Promotionsabschlussprogramms.

Im Rahmen meiner kumulativen Dissertation habe ich mit einer ganzen Reihe kluger Köpfe zusammen gearbeitet. Ich danke Dominik Becker, Theresia Smolka, Toralf Stark und Christina Zuber für den intensiven Austausch und die fachlich-freundschaftlichen Diskussionen. Prof. Dr. Achim Goerres und Dr. Jutta Wergen (Universität Duisburg-Essen) möchte ich meinen Dank aussprechen für den informellen, fachlichen Austausch und hilfreiche Ratschläge zu Beginn meiner Dissertation.

Meinen Freundinnen Vanessa Dräger, Anna Ebert, Annkatrin Kaiser, Maria Schubert und Katja Staack danke ich für die vielen Gespräche und den festen Rückhalt, den sie mir gegeben haben. Herzlichen Dank insbesondere an Annkatrin und Katja für die Unterstützung in der Abschlussphase.

Im gesamten Promotionsprozess konnte ich mir der Ermutigung und Unterstützung durch meine Familie sicher sein. Ich danke meinen Eltern Bettina Haentjens und Jürgen Breustedt, meinen Schwestern Rebekka und Deborah, meinen Schwägern Kai-Uwe und Thomas sowie meinen Schwiegereltern Sieglinde und Hans-Bernd Wackerbeck. Es ist nicht selbstverständlich, eine so große Familie stets hinter sich zu wissen. Gerne hätte ich noch mit Dir – Bernie – auf die Promotion angestoßen.

Als mein Mann und ich uns kennenlernten, war ich nach eigener Einschätzung gerade dabei, meine Dissertation fertig zu schreiben. Ich danke Dir, Sven, für Deinen langen Atem, Deine ehrlichen Worte und Deine wertvolle Unterstützung. Nach Abgabe meiner

Dissertationsschrift wurde mir gegenüber des Öfteren Bewunderung dafür ausgesprochen, dass ich trotz der Geburt unserer Tochter Ida fertig promoviert habe. Nicht trotz, wende ich ein, sondern wegen. Diese Dissertationsschrift ist euch beiden gewidmet.

Duisburg, den 4. März 2021

Table of Contents

List of Tables

List of Figures

1 Introduction

After three decades of successful democratization since the end of the Cold War, scholars and practitioners are keen to maintain and improve democracy around the globe (Diamond, 2016, p. 76). Major crises such as the recent global COVID-19 pandemic or the multiyear eurozone crisis have amplified their ambitions. On the one hand, achieving this aim involves evaluating and reforming the quality of democracies' political institutions, processes, and policies at the macro level (Landman, 2012, p. 462; Lauth, 2011, pp. 59–60; Ringen, 2007, p. 1). On the other hand, it also requires assessing and enhancing citizens' political support at the individual level (Norris, 2011, p. 8; S. Pickel & G. Pickel, 2006, pp. 50–51). As Przeworski (2010, pp. xii-xiii) writes: "[H]aving followed liberalization, transition, and consolidation, we have discovered that there is something still to improve: democracy".

In order to evaluate the state of democratic political regimes and to develop informed actions to meet their challenges, fortunately, practitioners and researchers can draw on a number of concepts and corresponding quantitative empirical assessments in comparative political science. From among these, political support and the quality of democracy provide an encompassing picture of the state of democracy at the individual and macro level. Political support focuses on people's values and attitudes toward the political regime as a whole, its institutions, processes, outcomes, and incumbents (Easton, 1965, p. 157). It was introduced to political science in the 1960s as a means to study the individual-level prerequisites of democratic persistence in the aftermath of World War II. Since then, the study of political support has established itself as part of the standard repertoire of political culture research (Almond, 1990, p. 8; Almond & Verba, 1965, pp. 1, 3, 337; Easton, 1965, p. 158, 1975, p. 445). In addition, its individual dimensions have become objects of analysis in their own right in this field.[1] Analyses of political support and its attributes are usually based on Easton's multidimensional conceptualization or subsequent developments thereof using cross-national survey data (Almond, 1980, pp. 15–16; Fuchs, 2007, pp. 164–165; S. Pickel & G. Pickel, 2006, pp. 31, 78–79). The concept has been used, for example, to study disparities between

[1] For conceptualizations and analyses of political values see for example Bratton and Mattes (2001), Dalton (2000), Inglehart and Welzel (2005), Thomassen (1995), and Welzel (2013). For models and analyses of political trust see for example D. Braun (2013), Citrin (1974), Gabriel (2018), Göhler (2002), Hooghe (2011) and K. Newton (2008). For conceptualizations and evaluations of political performance see for example Fuchs (1998) and Roller (2005).

citizens' expectations and democratic regimes' performance (Dalton & Welzel, 2014; Norris, 2011; Pharr & Putnam, 2000). Recent applications include an analysis of the effect of COVID-19 lockdowns on political support (Bol, Giani, Blais, & Loewen, 2020).

The quality of democracy joined the canon of common concepts in comparative political science more recently. Researchers' interest in studying it empirically arose when the 'people's rule' established itself as "the only broadly legitimate form of government in the world" (Diamond, 2016, p. 76) in the early years of the new millennium (Altman & Pérez-Liñán, 2002, p. 85; Fuchs & Roller, 2008, p. 77; Roberts, 2010, pp. 4–5). As a result of these changing circumstances, a number of comparative political scientists sought ways to determine in which respects a 'deepening' of democracy was attainable (Diamond & Morlino, 2004a, p. 20; Levine & Molina, 2011a, p. 259; Ringen, 2007, pp. 1–2). Others searched for means to establish whether countries were facing a 'democratic rollback' (Erdmann & Kneuer, 2011, p. 9; Lauth, 2015, p. 5; Roberts, 2010, p. 3). In order to facilitate such studies, social scientists proposed a variety of concepts and corresponding quantitative empirical assessments of the quality of democracy (Altman & Pérez-Liñan, 2002, pp. 86–87; D. F. Campbell, Carayannis, & Scheherazade, 2015; Erdmann, 2011, pp. 23–25; Lauth & Kauff, 2012; Merkel et al., 2018a, 2018b; Schraad-Tischler & Seelkopf, 2018). These indices largely consist of expert judgments and official statistics. They allow researchers to rank and compare countries in terms of political regimes' procedural aspects, structural characteristics, the results of its political processes or a combination of any of these facets at the macro-level.[2]

The usefulness of quantitative empirical assessments of the quality of democracy and political support for academia and practice depends on their measurement validity, however. That is to say, it is contingent on the extent to which the measurement process results in data that reflect the concept of the 'quality of democracy' and 'political support' (Adcock & Collier, 2001, pp. 529, 530; J. Behnke, Baur, & N. Behnke, 2006, p. 119; Carmines & Zeller, 1979, p. 16). The magnitude of this match is determined by the goodness of their conceptualization as well as the validity of their operationalization, measurement, and aggregation. In comparative research, it also rests on the extent of cross-national measurement equivalence of the measurement process (Adcock & Collier,

[2] A complementary approach is to use qualitative assessments such as the democracy assessment framework developed by the International Institute for Democracy and Electoral Assistance (Beetham, Carvalho, & Weir, 2008a). Its goal is to assist in improving democracy based on case-specific assessments that use a comparable set of criteria (Beetham & Weir 2000, pp. 75–76; Landman, 2012, p. 458).

2001, pp. 534–536). The poorer this match, the greater the bias in the quantitative empirical assessments. At best, this leads to imprecise data, at worst, it generates erroneous numbers (Carmines & Zeller, 1979, pp. 14–15). Researchers who use such invalid data may make incorrect inferences about the state of the quality of democracy and political support, cross-country similarities and differences as well as their causes, and effects (Carmines & Zeller, 1979, p. 11; Döring & Bortz, 2016, p. 98). In addition, invalid data and research results based on these data could misinform the public and prompt practitioners to draw wrong conclusions about which actions to take to maintain and improve democracy. Thus, assessing and enhancing measurement validity is essential so that quantitative empirical assessments of the quality of democracy and political support provide accurate and meaningful information that helps to sustain and ameliorate democracies.[3] Accordingly, the dissertation's overarching research question is: How can researchers improve the measurement validity of quantitative empirical assessments of the quality of democracy and political support for comparative research?

Judging by the state of debate, in many respects, researchers do not agree on how to measure the two concepts in a valid manner across countries. The unresolved issues differ depending on the concept. As mentioned above, political support has remained a prominent topic on the research agenda in comparative political science for several decades. The debate concerning its valid measurement has matured – to date, it has addressed detailed aspects pertaining to its conceptualization, operationalization, measurement, and cross-national comparability (see for example Canache, Mondak, & Seligson, 2001; Dalton, 2004; Fuchs, 1989; Marien, 2017). By contrast, as the quality of democracy is a fairly recent topic in comparative democracy studies, the discussion regarding its valid measurement has just commenced. It addresses fundamental decisions regarding its conceptualization as well as all stages in the measurement process (see for example Kaina, 2008; Munck, 2016; Ringen, 2007).

The articles included in this dissertation advance the debates by addressing four key unresolved issues. The first article (S. Pickel, Stark, & Breustedt, 2015) evaluates the comparative validity of measurement instruments of the quality of democracy. The second article (S. Pickel, Breustedt, & Smolka, 2016) reflects on the relevance of

[3] This dissertation uses the term 'assessments' of democracy rather than 'measurements' of democracy. The latter term commonly refers to macro-level indices (S. Pickel & G. Pickel, 2006, pp. 157, 159, Footnote 100). Since this dissertation addresses measurement validity issues of measurement instruments at both the macro and individual level, the term 'assessment' is deemed more appropriate.

including citizen's perspective for the valid measurement of the quality of democracy. The third article (Breustedt, 2018) considers the cross-national equivalence of political trust. The fourth article (Becker, Breustedt, & Zuber 2018) studies the implications of different methods of analysis for the valid aggregation of citizens' support for democracy. The articles study these issues on the basis of a common theoretical foundation in the critical rationalist tradition (Albert, 1991, 2000; D. Miller, 1994, 2006; Popper, 1959/2005, 1962, 1972, 1974, 1979, 1985, 1994). This foundation comprises a specific understanding of measurement validity, a set of aspects and procedures to evaluate and improve it, as well as certain epistemological assumptions as to the extent to which measurement validity can be improved. In line with this foundation, all articles include theory-guided empirical analyses that are based on a comparative research design, use quantitative empirical data, and apply quantitative empirical methods. All articles proceed in the same fashion. In a first step, each article evaluates the validity of quantitative empirical assessments of democracy with regard to the issues raised above. Based on these validations, each develops recommendations on how to enhance the validity of the measurement process concerning the issue in question. Jointly, they thus help to improve the measurement validity of quantitative empirical assessments of democracy for comparative research.

The remaining part of the introduction to this dissertation is divided into three sections. Section 1.1 clarifies the theoretical foundation underlying the dissertation's articles. This clarification serves to explicate how they evaluate and improve measurement validity. In addition, it helps to delineate the contributions and limitations of the articles. Section 1.2 summarizes the main unresolved measurement validity issues with regard to the quality of democracy and political support and explains which of these issues are addressed by the dissertation's research articles. In addition, it gives a synopsis of each article and describes the recommendations they propose.[4] Section 1.3 highlights the overall contributions of the articles to comparative research on the quality of democracy and political support as well as measurement validation in general and acknowledges limitations. It concludes with future research opportunities regarding the measurement validity of quantitative empirical assessments of democracy for comparative research.

[4] Articles one, two, and four were co-authored; article three is a single-author publication. The authors of the co-authored articles jointly discussed and decided all aspects regarding the papers' content. At the same time, each author took primary responsibility for certain parts of each paper. The synopsis of each article in section 1.2 includes a brief summary of the contribution of the dissertation's author.

1.1 Measurement Validity

The subsequent sections outline the common theoretical foundation underlying the dissertation's articles by answering the following questions: What is measurement validity (section 1.1.1)? How can it be evaluated and improved (section 1.1.2)? And to what extent (section 1.1.3)? Different schools in the philosophy of science such as positivism, interpretivism, and critical rationalism propose diverging answers to these questions (Blaikie, 2007, p. 109). The answers in the subsequent sections are informed by a critical rationalist point of view in the Popperian tradition – one of the mainstream philosophical schools that informs quantitative empirical analyses in comparative political science (Döring & Bortz, 2016, p. 36; Mouritzen, 2011, p. 2208).[5]

1.1.1 Defining Measurement Validity

Measurement validity, as understood in this dissertation, is a particular kind of validity in two respects. First, it is particular in that it is associated with a certain meaning of validity in general. The meaning of the term 'validity' depends on the research field (P. Newton & Shaw, 2014, p. 3). For quantitative empirical analyses such as those performed by the dissertation's articles, validity has been defined as "the approximate truth of an inference. When we say something is valid, we make a judgment about the extent to which relevant evidence supports that inference as being true or correct" (Shadish, Cook, & D. T. Campbell, 2002, p. 34; see also Adcock & Collier, 2001, pp. 530–531; Döring & Bortz, 2016, p. 93).[6]

Second, measurement validity is particular as it refers to the validity of certain stages in the research process. In line with the deductive model of scientific inquiry (Popper, 1959/2005, pp. 9–10), the research process in theory-based, quantitative empirical studies such as the dissertation's articles consists of a number of stages and follows a linear

[5] Like any philosophy of social science, critical rationalism too has its variants (Furlong & Marsh, 2010, p. 189). The following sections refer to its core axioms. For an overview of critical rationalism, its variations, as well as alternative philosophies of science that predominate in the social sciences see Chalmers (2013), Blaikie (2007, pp. 109–205) as well as della Porta and Keating (2008).

[6] By contrast, in philosophy, for example, validity is the property of a deductive argument "where the conclusion in some sense (actually, hypothetically, etc.) follows from the premises *necessarily*" (Baggini & Fosl, 2010, p. 13; see also Popper, 1962, p. 243; Tomassi, 1999, p. 4). In jurisprudence, validity refers to "the norms in law and the acts executed in the name of the law" (Varga, 1999, p. 883) that possess legal power. In the social sciences, the meaning of validity is also disputed. For discussions of different conceptualizations see Markus and Borsboom (2013, pp. 196–220) as well as Shadish et al. (2002, pp. 475–478). For an extensive historical overview of the changes in meaning over time see P. Newton and Shaw (2014, pp. 27–181).

research logic (S. Pickel & G. Pickel, 2018, pp. 13, 21; Schmitter, 2008, p. 264; Schnell, Hill, & Esser, 2013, p. 4).[7] Scientific practice calls for these stages to comply with certain quality criteria of scientific inquiry – namely relevance, comprehensible documentation, as well as ethical strength and methodological strength (Döring & Bortz, 2016, pp. 85–92). The methodological strength of theory-based, quantitative empirical studies can be expressed in terms of their validity (Döring & Bortz, 2016, p. 93).[8] According to Cook and D. T. Campbell's classic distinction, validity in turn can be distinguished into four types – measurement validity (originally referred to as construct validity[9]) statistical conclusion, internal, and external validity (Cook & D. T. Campbell, 1979, pp. 37–39; Peters, 2013, p. 91; Shadish et al., 2002, pp. 37–39). Each of these types applies to different stages of the research process. Measurement validity refers to the measurement process, that is, the operationalization, measurement, and aggregation stage.[10, 11]

As the research stages build on one another, the extent of measurement validity is assessed with reference to the preceding stage in the research process – the conceptualization. As the oft-cited definition of measurement validity puts it, it refers to "*the degree to which [the measurement process] measures what it purports to measure*" (Ruch, 1924, p. 13, emphasis in original; see also Carmines & Zeller, 1979, p. 12; Pennings, Keman, & Kleinnijenhuis, 2006, p. 67). That is to say, in order to be valid, the measurement process should result in data that reflect the theoretical concept of interest. The extent of the overall measurement validity depends on the degree to which the stages

---

7 This is not to say, however, that the research process is unidirectional. While the stages build on one another in principle, they inform one another in practice. As Schmitter (2008) emphasizes these stages are "a schematic and idealized representation" (p. 264) of the research process.

8 Validity is one of several standard quality criteria of methodological strength in empirical social scientific research. Another key criterion commonly referred to is reliability (Jackman, 2008, pp. 121–125; B. Miller, 2007, pp. 131–136; S. Pickel & G. Pickel, 2018, pp. 46–48). Reliability is considered to be a necessary but not sufficient condition for validity (P. Newton & Shaw, 2014, p. 14).

9 The use of the term 'construct validity' in the social sciences varies. To avoid conceptual confusion, Adcock and Collier (2001, p. 537) therefore suggest using the term 'measurement validity' for validity issues related to the measurement process instead (see also P. Newton & Shaw, 2014, p. 5).

10 Researchers have specified additional subtypes of measurement validity. Adcock and Collier (2001, p. 530, Footnote 2) as well as P. Newton & Shaw (2014, p. 8) provide extensive lists. These subtypes are not used in this dissertation as they do not add additional insights to answering the research question.

11 The remaining three types refer to subsequent stages of the research process. Statistical conclusion validity and internal validity pertain to the analysis. Both describe the approximate truth of inferences regarding the relationship between empirical indicators. Statistical conclusion validity refers to inferences regarding the covariation of the indicators of interest. Can researchers' claim that the relationship is statistically significant and practically relevant (in terms of effect size) be supported by sufficient evidence? Internal validity concerns the approximate truth of inferences about the *causal* relationship between the indicators. Is there evidence against this claim, such as omitted variables or a bias in case selection? External validity pertains to the conclusion in the research process. It reflects the extent to which causal inferences can be generalized to a broader set of cases, time periods, or contexts (Cook & D. T. Campbell, 1979, pp. 37–39; Döring & Bortz, 2016, p. 97).

jointly coincide with the respective concept (Adcock & Collier, 2001, pp. 530–531; J. Behnke et al., 2006, p. 119; Perron & Gillespie, 2015, p. 35). In addition, in comparative research, the measurement process needs to be equivalent in all units of analysis in order for comparisons to be valid.

Combining these defining attributes, measurement validity is defined as follows. It refers to the degree to which researchers provide sufficient evidence to support their claims regarding the match between the theoretical concept in question on the one hand and the operationalization, measurement, and (if applicable) aggregation on the other hand (Adcock & Collier, 2001, p. 529; King, Keohane, & Verba, 1994, pp. 55–63; P. Newton & Shaw, 2014, p. 3; Perron & Gillespie, 2015, pp. 38–39). In comparative research, this includes evidence of measurement equivalence across the units of analysis. To clarify, measurement validity is *neither* an inherent property of the measurement process, *nor* of the measurement instrument that applies this process, *nor* of the results of the measurement process. It is a property of researchers' judgment about the approximate truth of their inferences from the data to the theoretical concept they are intended to measure (Shadish et al., 2002, p. 34). For the purpose of brevity, the phrases 'validity of the measurement process' and 'validity of the measurement instrument' are used nonetheless in this dissertation.

1.1.2 Evaluating and Improving Measurement Validity

Building on the understanding of measurement validity outlined above, the following principle guides the articles' efforts to evaluate and improve the measurement validity of quantitative empirical assessments of democracy: Measurement instruments' validity can be assessed and improved for comparative research by considering and refining their theoretical concept in the conceptualization stage, scrutinizing and enhancing their operationalization, measurement, and aggregation as well as assessing and establishing cross-national invariance of the measurement process. Each of these tasks involves several aspects. Sections 1.1.2.1 to 1.1.2.4 outline these aspects. They are summarized in Table 1.1, which guides all of the subsequent sections of this introduction. Section 1.1.2.5 describes common strategies to assess the correspondence between the concept and the measurement process. Section 1.1.2.6 addresses the particularities of comparative research in these respects. Each section consists of a brief overview and the interested reader is referred to the relevant literature.

1.1.2.1 Good Conceptualization

A measurement instrument's conceptualization is "the starting point for assessing the validity of data sets" (Herrera & Kapur, 2007, p. 367). Conceptualization is a "triangular operation" (Gerring, 1999, p. 358), whereby (i) a set of attributes that define a concept's meaning is aligned with its (ii) referents, that is, the phenomena or events that the concept refers to as well as (iii) a term that covers both (i) and (ii) (Sartori, 1984, pp. 22–28). In addition, since most concepts in the social sciences are complex, conceptualization usually involves specifying a concept's dimensions, which are subdivided into several levels. Conceptualization results in a theoretical model of a concept, the so-called systematized concept. Specifying the systematized concept is necessary as there is often a variety of meanings associated with a concept. These 'background concepts' are usually too broad and too vague for scientific research (Adcock & Collier, 2001, pp. 530–532; Goertz, 2006, p. 6; Wonka, 2007, pp. 66–67). The systematized concept, on the other hand, provides the conceptual template against which to appraise the validity of its measurement instrument.

A measurement instrument's systematized concept helps to evaluate its measurement validity if it fosters theoretical clarity and facilitates empirical applicability (Popper, 1979, pp. 27–29; Wonka, 2007, pp. 65, 75). That is to say, it should be informed by theory and should differentiate the systematized concept from 'neighboring concepts' since "[n]o research can be conceptualized *ex novo* without reference to what has been produced already on that and related topics" (Schmitter, 2008, p. 269; see also Collier, LaPorte, & Seawright, 2012, p. 222; Gerring, 1999, p. 365). In addition, the systematized concept should be formed with sufficient detail to permit researchers to apply, test, and criticize their theory of interest empirically (Popper, 1979, pp. 28, 36).

When evaluating the validity of a measurement instrument for comparative research, its systematized concept is only useful if it is not subject to 'conceptual stretching' (Collier & Mahon, 1993; Sartori, 1970, pp. 1034–1035). That is to say, those who develop the measurement instrument should ensure that the meaning of their concept is not distorted when it is applied across cases (Lauth, G. Pickel & S. Pickel, 2014, p. 357). Thus, the attributes and linguistic label of the concept in question should be equivalent in the events or phenomena they are applied to (Peters, 2013, pp. 92–94; van de Vijver &

Tanzer, 2004, pp. 124–125; van Deth, 2009, pp. 87–90).[12]

In order to determine whether a measurement instrument's systematized concept provides a useful template, validations should consider the following aspects (see column C in Table 1.1). The first set of aspects concerns the conceptual content, that is, the meaning, nature, and level of the concept as well as its level of origin. The presentation of the 'meaning' of a concept (C1) should include not only a description of the attributes that signify its presence but also an outline of the attributes that indicate its absence as well as the substantive content of the continuum between these two poles (Goertz, 2006, pp. 30–35). The description of the 'nature' of the concept should point out whether its meaning suggests a categorical or continuous differentiation (C2; Goertz, 2006, p. 34; Schnell et al., 2013, pp. 128–129). The 'level' of the concept "is the level at which it is hypothesized to be manifest in a given theoretical model" (Kozlowski & Klein, 2000, p. 27) (C4). By contrast, the concept's 'level of origin' (C3) refers to "where, when and how the construct forms and is manifest" (Kozlowski & Klein, 2000, p. 28). A measurement instrument's conceptualization should address both levels as the referents of the concept as carriers of its attributes are not necessarily identical with the entities from where the attributes emerge (Diekmann, 2013, pp. 122–123; Lazarsfeld & Menzel, 1961).

The second set of aspects pertains to the systematized concept's conceptual logic. Evaluations of a measurement instrument's validity should determine whether its documentation specifies the number of conceptual dimensions and sub-dimensions (C5). In addition, it should clarify the relationship between the dimensions (C6): Are they substitutable? Are they equally important for the concept? Furthermore, it should describe the relationship between the levels, that is, between the dimensions and their sub-dimensions: Are the sub-dimensions a cause or an effect of their higher-level dimensions (Goertz, 2006, pp. 44–58) (C7)?

Third, the documentation should address the systematized concept's range (C8). This aspect calls for a specification of the referents of the concept. It should include a description of the scope conditions that these referents have to fulfill in order for the concept to be applicable to them. These pertain to the temporal, spatial, or otherwise specified conditions as stipulated by theory that sufficiently identify them as referents and distinguish them from non-referents (Foschi, 1997, p. 537; Sartori, 1984, pp. 42–44).

[12] See Gerring (1999, 2001), Goertz (2006), Kozlowski and Klein (2000), as well as Sartori (1970, 1984) for detailed refinements of these criteria for conceptual goodness.

Table 1.1
Conceptualization and Corresponding Aspects in the Measurement Process

		Measurement process					
Conceptualization (C)		Operationalization (O)		Measurement (M)		Aggregation (A)	
Conceptual content							
C1	• meaning ➢ attributes that define the negative and positive pole ➢ substantive content of the continuum between the poles	O1	• coverage of content of dimensions at large • coverage of content of individual dimensions ➢ unambiguous ➢ no omission or inappropriate inclusion	M1	• method of data collection • instrument of data collection • data sources	A1	• extent of data aggregation
C2	• nature of the concept ➢ continuous, categorical	O2	• measurement scale & level • thresholds	M2	• method of data collection	A2	• method of aggregation • aggregation rules
C3	• level of origin	O3	• content of indicators	M3	• method of data collection • method of case selection • level of data collection • units of observation	A3	• level of aggregation • units of analysis
C4	• level of the concept	O4	• content of indicators	M4	• method of data collection • method of case selection • level of data collection • units of observation	A4	• level of aggregation • units of analysis
Conceptual logic							
C5	• number of dimensions and sub-dimensions	O5	• measurement model's dimensionality ➢ number of indicators	M5	---	A5	• method of aggregation
C6	• relationship between dimensions	O6	• content of indicators • measurement model ➢ substitutability ➢ weights	M6	---	A6	• aggregation rules • method of aggregation
C7	• relationship between dimensions and sub-dimensions	O7	• measurement model ➢ reflective ➢ formative	M7	---	A7	• aggregation rules • method of aggregation
Conceptual range							
C8	• range (scope conditions)	O8	---	M8	• method of case selection	A8	• units of analysis

Note. Own compilation. Sources: Goertz (2006), Kozlowski and Klein (2000); Munck and Verkuilen (2002).

Ideally, evaluations of a measurement instrument's validity conclude that its documentation outlines all of the aspects described above in its description of the systematized concept underlying the measurement instrument. If not, this indicates the need for improvement. After all, "precise explication lays the foundation for sound measurement" (Kozlowski & Klein, 2000, p. 26; see also Fuchs & Roller, 2008, p. 77).

1.1.2.2 Valid Operationalization

The next stage to address when evaluating measurement instrument's validity is the operationalization. The operationalization serves to translate the theoretical model of the concept of interest into a measurement model. Generally speaking, the measurement model should describe how researchers infer from the observable indicators to their concept of interest. The description of the measurement model should thus indicate how the concept as a latent, non-observable variable is linked with observable indicators. Multidimensional, multilevel concepts such as the quality of democracy and political support should be translated into measurement models with first- and second-order latent variables. The description should also clarify the links between the latent variable(s) and the indicators, that is, the 'rules of correspondence' that researchers assume to determine how the observable facts correspond with the latent variable(s) and how they can be combined to represent the overall concept (Döring & Bortz, 2016, pp. 228–229; Dreier, 1997, pp. 236–237; Jackman, 2008, p. 119).

A measurement instrument's validation should assess how well its operationalization is aligned with its systematized concept. The better this alignment, the greater the validity of its operationalization. This involves a number of aspects (see column O in Table 1.1). Regarding the conceptual content, it entails that the content of the indicators unambiguously reflects the meaning of the overall concept and its individual dimensions while neither omitting relevant nor including irrelevant aspects (O1; Adcock & Collier, 2001, p. 538; Messick, 1995, p. 742; van de Vijver & Tanzer, 2004, p. 124). The categorical or continuous nature of the concept should be considered in the choice of measurement scale. Accordingly, a measurement instrument's evaluation should determine whether the numerical values of the indicators – forming a measurement scale with nominal, ordinal, interval, or ratio level of measurement – correctly represent the empirical manifestations of the concept's attributes. In the case of categorical measurement scales, this includes thresholds that mark the transition from one category

to another (O2; Lauth, 2009, pp. 160–165). In addition, evaluations should establish how the level of the concept and its level of origin are taken into account by the indicators' content (O3 and O4; Kozlowski & Klein, 2000, pp. 37–38).

Another set of aspects to consider is the correspondence between the structure of a measurement instrument's measurement model and the conceptual logic of its systematized concept. As shown in Table 1.1 (O5), the number of conceptual dimensions should guide the number of dimensions, which affects the minimum number of indicators in the measurement model (Perron & Gillespie, 2015, p. 43). Furthermore, the relationship between the dimensions should be reflected in the content of the indicators (O6). In addition, the conceptual substitutability and relative importance of the concept's dimensions should be expressed in the rules of correspondence between the latent variable(s) and the indicators in the measurement model (O6). Can high values on one indicator substitute low values on another indicator? Do all indicators carry equal weight (Goertz, 2006, pp. 46–50)? Finally, regarding the conceptual relationship between the concept's dimensions and sub-dimensions, evaluations should also take care to discern whether the measurement model of the measurement instrument is reflective or formative (O7; Jarvis, Mackenzie, & Podsakoff, 2003, p. 201). That is to say, their rules of correspondence should outline whether the latent variable is a cause or effect of the indicators (Coltman, Devinney, Midgley, & Venaik, 2008, p. 1252; Döring & Bortz, 2016, pp. 229–230; Goertz, 2006, pp. 53–58). In conjunction, these aspects delineate what to consider when evaluating the validity of a measurement instrument's operationalization and developing recommendations to improve it.

1.1.2.3 Valid Measurement

The next stage to consider when validating a measurement instrument is the measurement stage. Measurement instruments such as those evaluated in the dissertation at hand use measurement to apply their operationalized concept empirically. According to measurement theory, measurement can be defined as the assignment of numbers to objects such that the numerical relational structure preserves the empirical relational structure of the attributes of those objects (J. Behnke et al., 2006, pp. 87–93; Diekmann, 2013, pp. 281–282; Krantz, Luce, Suppes, & Tversky, 1971/2007, p. 1).

The validity of the measurement stage depends on two criteria. First, it is contingent on how well it pays heed to the characteristics of the indicators specified in the

operationalization (Adcock & Collier, 2001, p. 531).[13] Second, it is affected by researchers' decisions on how to apply the indicators empirically.

Accordingly, evaluations of the validity of a measurement instrument's measurement stage should consider a number of aspects (see column M in Table 1.1). Regarding the concept's meaning, the method of data collection – such as surveys or expert judgments – should allow researchers to gather suitable information on the meaning of their concept (Baur & Blasius, 2014, p. 45). In addition, the chosen instrument of data collection – such as standardized questionnaires – should include sufficient indicators to cover the concept's content. Likewise, in secondary data analyses, researchers' data sources should furnish enough information in order for the indicators to provide adequate coverage (M1; Friedrichs, 1981, pp. 357–360; Rathke, 2007, p. 153). As for the nature of the concept, the indicators' measurement scale should be taken into account by the method of data collection (M2). Furthermore, the level of the concept and its level of origin should be given thought when choosing the method of data collection, the method of case selection, the level of data collection, and the units of observation (M3, M4; Kozlowski & Klein, 2000, pp. 36–37; Niedermayer & Widmaier, 1997, pp. 80–84). Finally, the method of case selection should be in line with the concept's range (M8).

Overall, evaluations of the validity of the measurement stage should determine to what extent these aspects are taken into account by the quantitative empirical assessment in question. If it does not pay heed to an aspect, this indicates the need for improvement of its measurement stage. As Ringen (2007) puts it: "[M]easurement is never about piling up data. It is about considering carefully what the relevant data are and then arranging those data with plan and economy" (p. 17).

[13] Strictly speaking, mere reference to the systematized concept and its operationalization alone does not suffice to ensure the measurement stage's validity. Instead, according to measurement theory, the representational adequacy of the scale must be proven empirically (Diekmann, 2013, pp. 282–284; Orth, 1974, pp. 21–23; Schnell et al., 2013, p. 130). Aside from methodological studies, such proof is seldom carried out in social scientific research, however. Instead, social scientific analyses are usually based on measurement by fiat, that is, measurement based on researchers' judgments rather than proven relationships (J. Behnke et al., 2006, pp. 97–101; Schnell et al., 2013, pp. 135, 138–139; Torgerson, 1958, pp. 21–25). Schedler (2012a, pp. 31–33) argues that such judgments are justified as long as they are not based on subjective arbitrariness but rather abide by certain methodological standards. If so, the validity of measurements can be taken as given as it is grounded in "informed and reasoned public argument" (Schedler, 2012a, p. 31; see also S. Pickel & G. Pickel, 2012, pp. 9–10).

1.1.2.4 Valid Aggregation

Usually, quantitative empirical assessments of democracy provide aggregated data. Thus, the final stage to consider when evaluating their measurement validity is their aggregation stage. The goal of the aggregation stage is to capture the dimensions of their measurement models' latent variable in a single number (Gehring & Weins, 2009, p. 18; Pennings et al., 2006, p. 86). More specifically, aggregation refers to the mathematical combination of measurements of the indicators from the units of observation at the level of data collection to measurements on the units of analysis at the level of analysis. If the level of data collection is lower than the level of analysis, aggregation is a means to combine the data observed at a lower level (such as individual-level data) to a higher level (such as data on regions or countries) (S. Pickel & G. Pickel, 2018, p. 85). If the level of data collection and the level of analysis coincide, aggregation can be used to reduce the data provided by the observable indicators on the different dimensions to a single measure of the latent variable (Müller & S. Pickel, 2007, p. 527; Peters, 1998, p. 71).

There are various aggregation methods available. Common aggregation techniques include multiplication, taking the sum, computing the average or percentages, as well as related forms that weigh the indicators' values on the basis of conceptual considerations (Diekmann, 2013, p. 121; Gehring & Weins, 2009, p. 18; B. Miller, 2007, p. 139). More complex methods are factor analysis and principal components analysis, which determine the weight of each indicator statistically (Krishnakumar & Nagar, 2008, p. 482; Nardo, Saisana, Saltelli, & Tarantola, 2005, p. 12). For concepts measured with categorical indicators in formative measurement models, another approach to aggregation is to construct descriptive typologies. These are obtained by cross-tabulating two or more indicators with two or more categories (Collier, Laporte, & Seawright, 2008, p. 153; Lauth, 2009, p. 154). Each of these aggregation techniques has its own advantages and disadvantages in terms of adherence to the systematized concept and statistical refinement (Collier et al., 2008, pp. 165–166; Goertz, 2008, pp. 95–127; Nardo et al., 2005, pp. 74–85; Saisana & Tarantola, 2002, pp. 9–11).

A measurement instrument's validity depends on the extent to which its aggregation is based on a valid operationalization and measurement and corresponds with the systematized concept in question. Evaluating this correspondence involves several aspects (see column A in Table 1.1). Regarding the concept's meaning, generally

speaking, the data should be aggregated to an extent that permits researchers to apply, test, and criticize their theory of interest (A1). Preserving the meaning of concepts in the aggregation stage is a 'balancing act' particularly for cross-national measurement instruments based on multidimensional, multilevel concepts (Munck & Verkuilen, 2002, pp. 22–23; S. Pickel & G. Pickel, 2012, pp. 2–3, 10; Weischer, 2015, p. 15). If a measurement instrument does not provide aggregated data on multidimensional concepts it may be difficult for researchers who use its data to discern patterns, establish relationships, and reach generalizing conclusions regarding the concept across cases. If the extent of aggregation is too great, however, systematic variation in the empirical manifestations of the concept's dimensions across cases may be obscured. This may cause researchers to reach invalid conclusions.

In order to reflect the concept's content, the aggregation should also be based on rules and methods that take the nature of the concept into account (A2). This entails maintaining the measurement scale of the concept's indicators (S. Pickel & G. Pickel, 2012, pp. 10–11). This is particularly important to keep in mind when the indicators have different measurement scales since not all methods of aggregation are suitable for all measurement scales (Schnell et al., 2013, pp. 135–136, 161–167). In case of typologies, researchers should carefully reflect how the aggregation rules account for the thresholds set for membership in the different categories (Lauth, 2009, pp. 163–165).

When aggregating data, a measurement instrument should also consider the level of the concept, its level of origin, as well as the conceptual range (A3, A4, and A8). The level of aggregation as well as the resulting units of analysis should match the level of the concept. If the level of the concept does not coincide with its level of origin, researchers who develop measurement instruments should bear this in mind when deciding on the manner and extent to which the data are aggregated. Relatedly, the units of analysis should correspond with the scope conditions that specify the concept's range.

Aside from these aspects pertaining to conceptual content and range, a measurement instrument's validation should determine whether it takes the conceptual logic of its systematized concept into account. This involves several aspects. The aggregation rules should reflect the relationship between the dimensions by accounting for the substitutability and weights of the indicators specified in the operationalization (A6). If needed, the method of aggregation should permit such computations (A5 and A6). In addition, the rules and methods of aggregation should match the reflective or formative

structure of the measurement model (A7; Goertz, 2006, pp. 39–58). Together, these steps describe what to consider when evaluating how well measurement instruments aggregate their data in order to reflect their concept of interest in a valid manner. If a measurement instrument fails to reflect on any of these aspects, this indicates that it may require improvement.

1.1.2.5 Validation Strategies

Using the aspects outlined above as a guide, the dissertation's articles validate the measurement process of current quantitative empirical assessments of the quality of democracy and political support. Validation pertains to procedures that help to assess the extent to which the measurement process of a measurement instrument results in data that reflect the systematized concept it is intended to measure (Adcock & Collier, 2001, p. 530; Perron & Gillespie, 2015, p. 39). Originally developed for the field of psychometrics in the 1950s and 1960s, these procedures are typically distinguished into three different strategies: content, criterion, and construct validation (Adcock & Collier, 2001, pp. 536–537; American Psychological Association, 1954; American Psychological Association et al. 1966; Carmines & Zeller, 1979, pp. 17–27; Schnell et al., 2013, pp. 145).[14]

Validation strategies share several common features. First, the evidence they provide for researchers' claim to validity is based on empirical analyses or derived from logical arguments. Second, failure to provide such evidence indicates that certain aspects of the measurement process require improvement. Third, as each strategy provides different kinds of evidence for validity, no strategy is sufficient by itself to establish researchers' claim to measurement validity (Adcock & Collier, 2001, p. 530; P. Newton & Shaw, 2014, pp. 8, 22–23; Perron & Gillespie, 2015, p. 39; Rupp & Pant, 2007, pp. 1032–1033).

The validation strategies differ in so far as they serve to validate different aspects of the correspondence between the systematized concept and its operationalization, measurement, and aggregation. The dissertation's articles use content validation and construct validation because of the validity issues they address. Content validation refers

[14] In the social science literature, these strategies are sometimes referred to as *types* of validity (Schnell et al., 2013, p. 145). In line with the unified approach to validity in the psychometric literature (Messick, 1995; P. Newton & Shaw, 2014), Adcock and Collier (2001, pp. 536–537) convincingly argue that they should not be regarded as types of validity in their own right but rather as types of validation that provide different kinds of evidence for validity.

to procedures that analyze how well the operationalization reflects the conceptual content of the systematized concept (J. Behnke et al., 2006, p. 120; Carmines & Zeller, 1979, p. 20).[15] These procedures usually involve qualitative expert judgments. There are no quantitative guidelines regarding the extent to which a measurement instruments' data should coincide with the systematized concept (Carmines & Zeller, 1979, p. 22; Kimberlin & Winterstein, 2008, p. 2279; Litwin, 1995, p. 35). Instead, experts use theoretical reasoning and argumentation to validate the content of the operationalization.[16]

Construct validation comprises procedures that appraise the extent to which the measure of interest performs in line with theoretical expectations about the systematized concept that is being measured (Carmines & Zeller, 1979, p. 27; Cronbach & Meehl, 1955, pp. 282–283). These theoretical expectations are derived from the so-called 'nomological net' of the systematized concept. The nomological net consists of the theoretical model of the systematized concept, its measurement model as well as its relationship with other concepts. As such, it comprises a latent, non-observable variable, related latent variables and the variables' observable indicators. In addition, it contains the relationships between these elements, that is, theoretical 'laws' concerning the relationship between the latent variables, 'rules of correspondence' pertaining to the connection between the latent variables and their observable indicators, as well as empirical hypotheses regarding the correlation between the observable indicators (Cronbach & Meehl, 1955, pp. 290, 294; Hartig, Frey, & Jude, 2008, pp. 145–146).

Construct validation consists of a series of (usually quantitative) tests of these empirical hypotheses, several of which are applied in the dissertation's articles. One kind of test assesses whether the hypothesized relationship between the systematized concept of interest and another latent variable holds empirically (Cronbach & Meehl, 1955, p. 283; Schnell et al., 2013, p. 147). If the hypothesized relationship is positive and "if the

[15] In its original sense, this type of validation was thought to reflect how well the selected items "provide an adequate and representative sample of all the items that might measure the construct of interest" (Kimberlin & Winterstein, 2008, p. 2279; see also Cronbach & Meehl, 1955, p. 282). This definition is based on the idea that it is possible to specify the content of a systematized concept in full and to draw a sample from this content. A lot of the time, this is not feasible in the social sciences (Carmines & Zeller, 1979, pp. 21–22). What is more, this approach does not coincide with the critical rationalist viewpoint that it is impossible to establish the 'true' meaning of a concept (see section 1.1.3). This dissertation therefore uses a moderated definition that is frequently found in the social scientific literature.

[16] B. Miller (2007, pp. 132–133) adds that such a quantitative criterion is not sensible. According to him, it is logically impossible to test the correspondence between an indicator and a concept empirically as the concept is a latent, non-observable variable that requires observable indicators in order to be measurable (see also section 1.1.2.2).

correlation is positive and substantial, then *one piece of evidence* has been adduced to support (...) construct validity" (Carmines & Zeller, 1979, p. 23).

Another kind of test addresses the operationalization of the systematized concept of interest. Since operationalizations serve to enable a measurement of a specific systematized concept, ideally, different indicators of the same systematized concept should be strongly related. Such tests are also referred to as 'convergent validation'. They are usually undertaken by correlating the data provided by different measurement instruments assumed to measure the same concept. Conversely, operationalizations of the systematized concept of interest should be clearly distinguishable empirically from operationalizations of similar but distinct latent variables. This is also referred to as 'discriminant validation'. It is commonly tested by relating the measurement instrument of interest to measurement instruments that are supposed to measure the other latent variables (Litwin, 1995, pp. 43–44; Schnell et al., 2013, pp. 147–148). Multitrait-multimethod-matrices allow researchers to perform convergent and discriminant validation simultaneously (J. Behnke et al., 2006, pp. 122–123; D. T. Campbell & Fiske, 1959).

Still another kind of construct validation test refers to empirical hypotheses about the systematized concept itself. Such hypotheses pertain to the dimensional structure of the systematized concept or the relationship between the items used to measure a certain dimension. Confirmatory factor analysis (CFA) and item response theory (IRT) models are common methods applied to test such hypotheses (Cronbach & Meehl, 1955, pp. 287–288; Hartig et al., 2008, pp. 153–154; Moosbrugger & Kelava, 2008, p. 14).[17]

Applying these validation strategies in the dissertation's articles not only serves to assess the extent of measurement validity of the measurement instrument in question, it also provides suggestions on how to improve it. Results in line with the hypotheses serve as pieces of evidence that support the claim that the measurement instrument of interest

[17] Criterion validation is not applied in this dissertation because of a lack of prerequisite criterion variables. These variables are necessary because criterion validation serves to assess the degree to which a measurement instrument correctly estimates or predicts the values of a defined criterion variable (Rupp & Pant, 2007, p. 1033). The criterion variable is measured with a different instrument, which, ideally, is accepted as the 'gold standard' in the scientific community (Kimberlin & Winterstein, 2008, p. 2279; Litwin, 1995, p. 37). Standard methods of criterion validation include regression and correlation (J. Behnke et al., 2006, p. 120; Rupp & Pant, 2007, p. 1033). If the data provided by a measurement instrument are related to the criterion variable in the expected manner, this is interpreted as evidence for the claim to validity of that instrument. The dissertation's articles do not apply this type of validation since such criterion variables are difficult to come by in the social sciences in general and in the dissertation's research field in particular (Carmines & Zeller, 1979, p. 19; Schnell et al., 2013, p. 146).

measures its systematized concept in a valid manner. Results that falsify the hypotheses undermine this claim and indicate the need for improvement (Cronbach & Meehl, 1955, pp. 290, 294–295; Hartig et al., 2008, p. 146; Kimberlin & Winterstein, 2008, p. 2279).

1.1.2.6 Measurement Validity in Comparative Research

In comparative analyses such as the dissertation's articles, the final key aspect to evaluating and improving a measurement instrument is to assess and enhance the comparability of its measurement. The issue at stake is that the measurement process has to yield measures of similar conceptual attributes among similar referents in different countries in order for the data to be comparable (Horn & McArdle, 1992, p. 117; van Deth, 2009, pp. 84–85). This is challenging: On the one hand, the operationalization, measurement, and aggregation of quantitative empirical assessments have to be sufficiently similar so as to facilitate generalizing statements; on the other hand, the measurement process has to allow for country-specific features so as to measure the concept in a valid manner in a particular national context (Adcock & Collier, 2001, pp. 529–530, 534–535; Bachleitner, Weichbold, Aschauer, & Pausch, 2014, p. 66; van Deth, 2013, p. xiv; Westle, 2005, p. 157). Thus, measurement instruments have to strike a balance between the "Scylla of losing national or cultural validity and the Charibdis [*sic*] of endangering cross-cultural or cross-national comparability" (van Deth, 2009, p. 85) in order for researchers to be able to draw valid inferences about the similarities and differences of their concept of interest across countries.

Provided that the concept is able to 'travel' across units of analysis, the solution proposed to resolving this dilemma is to apply an equivalent (or, more technically, invariant) measurement process across cases (M. Braun, 2006, pp. 17–18; S. Pickel & G. Pickel, 2018, p. 93; Przeworski & Teune, 1966, 1970, pp. 106–110). The measurement process is invariant if the chosen means are equally effective in numerically representing the relevant aspects of the concept of interest – that is, the aspects of the phenomena or events that the researcher seeks to compare (van Deth, 2009, p. 86; Westle, 2005, pp. 151–152).[18]

Countries' contextual specificity may complicate establishing the invariance, and thus the comparability, of the measurement process for comparative analyses, however (Adcock

[18] See Bachleitner et al. (2014) for an extensive presentation of the types of equivalence needed to carry out valid comparisons as well as a critical discussion of the assumptions behind this procedure.

& Collier, 2001, p. 534). According to van de Vijver and Tanzer (2004), three types of 'nuisance factors' may affect the invariance of the operationalization, measurement, and aggregation across countries: construct bias, method bias, and item bias. Construct bias occurs if the same items are used to measure the concept in cross-national analyses even though its attributes differ across countries. In this case, the operationalization does not represent the conceptual content in a valid manner because the measurement model omits relevant or includes irrelevant aspects. Method bias pertains to the incomparability of samples, problems regarding the instrument of data collection, as well as administration problems. Item bias occurs when items "have different psychological meanings across cultures" (van de Vijver & Tanzer, 2004, p. 121).

Altogether, if the measurement process is biased, differences in the data do not reflect differences in the empirical manifestations of the concept of interest across countries. Instead, one or several of the contextual specificities listed above systematically distorts the results of the measurement process (van de Vijver & Leung, 2011, p. 18). If this is the case, researchers who use the data provided by the measurement instruments run the risk of drawing incorrect inferences about the similarities and differences of the manifestations of their concept of interest across countries.

In order to avoid such inferential errors, evaluations of measurement validity for comparative research should test the measurement invariance of the measurement instrument in question (Cole & Maxwell, 1985, pp. 389–390). A widely applied method is multiple group confirmatory factor analysis (MGCFA). Alternative methods include IRT models and latent class analysis (Davidov, Meulemann, Cieciuch, Schmidt, & Billiet, 2014; Kankaraš, Vermunt, & Moors, 2011; Millsap, 2011). Tests applying these methods show whether the measurement instrument in question "yield[s] measures of the same attributes" (Horn & McArdle, 1992, p. 117). In other words, they establish in which respects and to what extent the data are comparable across countries.

Similar to the validation strategies outlined in the previous section, in the dissertation's articles, tests for measurement invariance not only serve to evaluate to what degree the measurement instrument of interest measures its systematized concept in a valid manner across countries, they also indicate which aspects require improvement. Model fit evaluations that signify measurement invariance support the claim that the measurement instrument provides a valid numeric representation of the underlying systematized concept across countries. Model fit evaluations that indicate non-invariance undermine

this claim and point to possible aspects that lack equivalence, therefore requiring improvement (see Brown, 2006, pp. 103–211, 236–319 for an extensive presentation).

To sum up, the dissertation's articles evaluate and improve the measurement validity of quantitative empirical assessments of democracy for comparative research by examining potential sources of non-invariance, assessing specific aspects regarding the match between the operationalization, measurement, and aggregation and the systematized concept of interest and making suggestion on how to enhance deficient aspects.

1.1.3 Limitations to Improving Measurement Validity

One caveat applies to the articles' aim to improve the measurement validity of quantitative empirical assessments of democracy, however: Measurement validity cannot be perfected. Three epistemological limitations curb the articles' efforts in this respect. First, according to critical rationalist assumptions, researchers cannot determine the exact 'essence' of a concept for its own sake. Such an attempt would lead to an infinite regress of convoluted definitions of definitions to establish their 'true' meaning (Adcock & Collier, 2001, p. 532; Popper, 1972 pp. 18–21, 1974, p. 337, 1979, pp. 20–37). When developing suggestions on how to improve the measurement validity of quantitative empirical assessments of democracy, it is thus unproductive to criticize their systematized concepts in terms of their meaning.

Second, when seeking to measure concepts empirically, it is impossible to establish a perfect match between the concept of interest and its observable indicators. This is referred to as the 'problem of correspondence'. "Any attempt to define universal names [or theoretical terms, WB] with the help of individual names [or observational terms, WB] is bound to fail" (Popper, 1959/2005, p. 45). Instead, as noted above, researchers have to clarify which 'rules of correspondence' they apply to justify how and why the observational terms reflect the theoretical terms (see section 1.1.2.2). These rules are based on auxiliary theories regarding the relationship between the two and are thus subject to refutation (Blalock, 1968; Costner, 1969; Schnell et al., 2013, pp. 68–74). Thus, the articles cannot provide concluding recommendations in this respect. "The best that can be achieved are collectively agreed-upon approximate matches" (Elder, 2005, p. 560).

This is related to a third issue, the 'problem of the empirical basis' (Popper, 1959/2005, pp. 74–94). Accordingly, measurements are not "*records or protocols of immediate*

observation, or perceptions" (Popper, 1959/2005, p. 78; emphasis in original), they are *statements* about these observations. Even though these statements describe the state or extent of a concept under certain spatio-temporal conditions, they always include so-called 'universals' (Popper, 1959/2005, p. 76). These universals "denote (...) structural or relational or 'dispositional' properties of things which are 'abstract'" (Popper, 1985, p. 109). As such, all universals incorporate theories. On the one hand, the universals' theories can be tested in order to corroborate the truth of the measurements. On the other hand, these tests can never be exhaustive, because all statements about the results of these test would themselves include universals, thus leading to an infinite regress.[19] Thus, it is epistemologically impossible for the articles to determine whether observational statements perfectly describe reality. Instead, it is up to the scientific community to decide whether the claim to measurement validity is based on sufficient evidence (Popper, 1959/2005, p. 92).

Given these three limitations, according to critical rationalism, the recommendations proposed by the dissertation's articles will not perfect the measurement validity of current quantitative empirical assessments of democracy. As Popper (1994) writes: "We cannot reasonably aim at certainty. Once we realize that human knowledge is fallible, we realize also that we can *never* be *completely certain* that we have not made a mistake" (p. 4; emphasis in original). Still, in line with the dissertation's theoretical foundation outlined in this section, the articles can enhance measurement instruments by logically scrutinizing and empirically testing the validity of their measurement process as well as its invariance for comparative research and developing recommendations on how to improve deficient aspects. "Since we can never know anything for sure, it is simply not worth searching for certainty; but it is well worth searching for truth; and we do this chiefly by searching for mistakes, so that we can correct them" (Popper, 1994, p. 4).

[19] The inability to verify observational statements is further aggravated by critical rationalism's assumption that "the customary distinction between '*observational terms'* (or '*non-theoretical terms'*) and *theoretical terms* is mistaken, since all terms are theoretical to some degree, though some are more theoretical than others" (Popper, 1972, p. 119; emphasis in original).

1.2 Issues and Recommendations

Researchers dispute the valid measurement of the quality of democracy and political support in a number of respects. Using Table 1.1 from the previous section as a guide, first, the following sections give a brief account of the main points of dispute regarding the two concepts and how they can be measured in a valid manner (sections 1.2.1 and 1.2.2). Second, they provide an overview of current measurement instruments for each concept (sections 1.2.1.1, 1.2.2.1, and 1.2.2.3). Third, with reference to these reviews, the sections summarize two key unresolved measurement validity issues for each concept. With regard to each issue, the respective sections outline how the dissertation's research articles examine the validity of current measurement instruments and what they recommend to improve them (sections 1.2.1.2, 1.2.1.3, 1.2.2.2, and 1.2.2.4).

1.2.1 The Quality of Democracy

Regarding the valid measurement of the quality of democracy, numerous questions are unresolved. This is largely because researchers seeking to assess it empirically are faced with a double challenge: "a normative one (finding the correct standards for assessing the functioning of a democracy) and an empirical one (determining how democracies actually work and the degree to which they live up to these standards)" (Roberts, 2010, p. 22).

First and foremost, the undecided points concern its conceptualization, that is, the conceptual template for determining the validity of the empirical measurements (see column C in Table 1.1). While the concept is generally considered to encompass multiple dimensions, researchers differ on the meaning as well as the relationship between the conceptual dimensions (C1, C6).[20] In terms of meaning, strictly procedural, contextualized procedural, and expansive substantive approaches differ with regard to the aspect of the democratic political system they refer to and the normative standards they apply when specifying the set of qualities that make up 'good' quality of democracy (Altman & Pérez-Liñan, 2002, pp. 86–87; Munck, 2016, pp. 4, 16; Roberts, 2010, p. 26). Regarding the relationship between the dimensions, researchers differ in the way they

[20] The concepts vary from two-dimensional (Ringen, 2007) to eight-dimensional (Diamond & Morlino, 2004a, 2005) (C5 in Table 1.1). Many authors opt for a three-dimensional conceptualization (Altman & Pérez-Liñan, 2002; Bühlmann, Merkel, & Wessels, 2008; Lauth, 2004, 2015; Munck, 2016; Roberts, 2010). Researchers then specify these dimensions further in terms of sub-dimensions. Here, the variety is even more pronounced.

conceive them to interact. Some authors focus on trade-offs (Plattner, 2004, pp. 107–108). Others emphasize the complementary nature of the dimensions (Munck, 2016, p. 20). Many authors take on an intermediate position (Beetham & Weir, 2000, pp. 79–80; Bühlmann, Merkel, & Wessels, 2008, pp. 14–15; Diamond & Morlino, 2004a, pp. 21, 29; Lauth, 2011b, pp. 65–67, 2015, p. 10; Roberts, 2010, pp. 41–44).[21]

What is more, there is conceptual disagreement on whether it is sensible to study the quality of democracy in non-democracies (C8). Several researchers apply it to all kinds of political regimes (Beetham et al., 2008a, pp. 251–252; Lauth, 2015, p. 16; Munck, 2016, pp. 9–10). Many others consider the concept of the quality of democracy to be applicable only to democracies (Altman & Pérez-Linan, 2002, pp. 86–87; Bühlmann, Merkel, Müller, & Wessels, 2012, p. 520; Diamond & Morlino, 2004a, p. 21; Levine & Molina, 2011b, p. 2; Roberts, 2010, p. 25).

Second, the unresolved issues concern the validity of the measurement process itself. Existing measurement instruments have been criticized with regard to a variety of aspects regarding their choice of operationalization, measurement, and aggregation (columns O, M, and A in Table 1.1). Among other things, researchers' criticism pertains to the indicators' content, the choice of data, units of observation, or the appropriate aggregation procedure (Jäckle & Bauschke, 2009, 2010; Jäckle, Wagschal, & Bauschke, 2012, 2013; Kaina, 2008; Merkel, Tanneberg, & Bühlmann, 2013; Müller & S. Pickel, 2008; Ringen, 2007, pp. 14–19; Smolka, 2019, pp. 202–229). Altogether, this variety of critical reflections indicates that the question of how to assess the quality of democracy in a valid manner is highly contested.

In light of this heterogeneity of viewpoints, it is challenging to determine how the measurement validity of quantitative empirical assessments of the quality of democracy can be improved. As stated in section 1.1.2, the dissertation's articles are based on the assumption that measurement validity can be enhanced for comparative research by developing a good systematized concept in the conceptualization stage, enhancing the

[21] They maintain that the dimensions do tend to develop similarly but are not complementary to an extent that the quality of democracy becomes an "all or nothing affair" (Landman, 2012, p. 461), whereby its decline within one dimension unquestionably entails its deterioration within the other dimensions. At the same time, they do not regard the trade-offs between the dimensions to be so great as to prevent their simultaneous realization (Beetham & Weir, 2000, p. 80). In addition, a number of authors agree that "there is no unique model of good democracy" (Morlino, 2004a, p. 29). Instead, they consider it to be inherent to democracies that the balance between the quality dimensions is "an ongoing political and civil process" (Bühlmann, Merkel, & Wessels, 2008, p. 15; see also Diamond & Morlino, 2005, xxxix; Lauth, 2015, p. 10). Overall, many authors concur that it is neither realistic nor recommendable to expect the quality of democracy to maximize simultaneously across all dimensions.

operationalization, measurement, and aggregation and establishing cross-national invariance of the measurement process. The scientific community evidently does not agree on the core attributes of the quality of democracy let alone the basics of its measurement, however. Thus, the question is: Which measurement validity issues should be addressed by the dissertation's articles without yielding to conceptual partiality or engaging in methodological nitpicking – and how should current measurement instruments be improved accordingly?

The following section (1.2.1.1) provides an overview of current measurement instruments of the quality of democracy. It serves to point out the similarities and differences between the indices in order to highlight unresolved issues regarding its valid measurement. This is necessary in order to derive and justify the answers of the dissertation's first and second article to this question in the subsequent two sections (1.2.1.2. and 1.2.1.3).

1.2.1.1 Current Approaches to Operationalizing, Measuring, and Aggregating Data on the Quality of Democracy

On the basis of Table 1.1 of the dissertation's theoretical foundation, the following section briefly presents publicly available, cross-national quantitative measurement instruments of the quality of democracy: the Democracy Ranking (DR) (D. F. Campbell, 2008; D. F. Campbell et al., 2015), the Democracy Barometer (DB) (Merkel et al,. 2018a, 2018b), the democracy dimension of the Sustainable Governance Indicators (SGI) (Schraad-Tischler & Seelkopf, 2018), and the Democracy Matrix (DM) (Lauth, 2004, 2015; Lauth & Schlenkrich, 2018). The overview starts out with a comparison of their systematized concepts, followed by a presentation of the similarities and differences regarding their approach to operationalizing, measuring, and aggregating data on the quality of democracy.

As summarized in Table 1.2 below, the measurement instruments are based on systematized concepts that are similar in several respects (see row C). In terms of conceptual content, all of the conceptualizations are based on a procedural understanding of the quality of democracy (C1). Three of the indices go beyond a strictly procedural definition, however, as they include contextual aspects such as corruption prevention and levels of discrimination (Bühlmann, Merkel, Müller, & Wessels, 2012, pp. 520–521; Schraad-Tischler & Seelkopf, 2018, p. 3) or substantive attributes such as environmental

sustainability (D. F. Campbell, 2008, pp. 30–37). The Democracy Matrix strikes a compromise by focusing on strictly procedural attributes while acknowledging that contextual factors constitute necessary conditions for the promotion of the quality of democracy (Lauth & Schlenkrich, 2019a, p. 15). All of the conceptualizations refer to 'degrees' or 'levels' of quality, thus conceiving the nature of the concept (C2) to be continuous (Bühlmann, Merkel, Müller, et al., 2008, p. 118; Bühlmann, Merkel, Müller, & Wessels, 2012, p. 521; D. F. Campbell, 2008, pp. 36–37; Lauth, 2015, p. 7, Lauth & Schlenkrich, 2019a, p. 10; Schraad-Tischler & Seelkopf, 2018, p. 8). None of the conceptualizations explicate the level of the concept (C4) and its level of origin (C3). All conceptualizations imply, however, that the concept originates at both the individual and macro level (C3). To illustrate, Bühlmann, Merkel, Müller and Wessels (2012, pp. 521–522; see also Merkel et al., 2018a, pp. 32–33) consider inclusive participation a key aspect of the quality of democracy. According to them, 'participation' is indicative of a democracy's quality not only in terms of the rules that regulate its political participation but also in the sense of individuals' equal and effective participation (see also D. F. Campbell, 2008, pp. 36–37; Lauth & Schlenkrich, 2019b, pp. 2–16; SGI, 2018a, pp. 38–48). In addition, all conceptualizations suggest that the concept of the quality of democracy is located at the macro level (C4) (Bühlmann, Merkel, Müller, & Wessels, 2012, pp. 520–521; D. F. Campbell & Barth, 2009, p. 214; Lauth & Schlenkrich, 2019a, p. 2; Schraad-Tischler & Seelkopf, 2018, p. 2.).

As for the conceptual logic, all of the systematized concepts consist of multiple dimensions and levels. As indicated in Table 1.2, the number of dimensions and sub-dimensions (C5) differs, ranging from three dimensions and six attributes in the Global Democracy Ranking (D. F. Campbell, 2008, pp. 30–37) to 15 matrix fields, 27 components and 12 subcomponents in the Democracy Matrix (Lauth & Schlenkrich, 2019a, pp. 5–9, 2019b). Regarding the relationship between the dimensions (C6), while differing in the details, the systematized concepts generally consider it to be both complementary and conflicting (Bühlmann, Merkel, Müller, & Wessels, 2012, pp. 521–522; D. F. Campbell, 2008, pp. 32–33; Lauth & Schlenkrich, 2019a, pp. 10–13). Especially the authors of the Democracy Barometer and the Democracy Matrix go to great lengths to explicate the intricacies of the dimensions' interlacements (Lauth, 2016;

Lauth & Schlenkrich, 2018).[22] Concerning the relationship between the dimensions and sub-dimensions (C7), the systematized concepts of the Sustainable Governance Indicators and the Democracy Barometer lack a description. The Democracy Matrix conceptualizes the relationship between the sub-dimensions for the dimensions in logical terms of necessity and sufficiency (Lauth & Schlenkrich, 2019b). The Global Democracy Ranking takes on a more sociological perspective by arguing that the sub-dimensions reflect different sub-systems of society (D. F. Campbell, 2008, p. 33).

As for the conceptual range (C8), the systematized concepts differ in that some consider the quality of democracy to be applicable to all countries whereas others restrict the scope conditions to a smaller set of political regimes. According to the conceptualization of the Sustainable Governance Indicators and the Democracy Barometer, the quality of democracy is explicitly limited to democracies (Bühlmann, Merkel, Müller, & Wessels, 2012, p. 520; SGI, 2018a, p. 10); the Global Democracy Ranking refers to free and partly free countries (D. F. Campbell & G. Pölzlbauer, 2008, pp. 4–5); the Democracy Matrix considers the quality of democracy to be applicable to all political regimes when seeking to create 'quality profiles'. It applies the concept to democracies in order to summarize their differences in terms of so-called 'democracy profiles' (Lauth & Schlenkrich, 2019a, pp. 15–16).

Whereas the systematized concepts bear certain similarities, their empirical application in the measurement process differs considerably (see row O in Table 1.2). Regarding the operationalization, all measurement instruments document how they cover the conceptual content (O1) (D. F. Campbell, 2008, pp. 38–41; Merkel et al., 2018b; Schraad-Tischler & Seelkopf, 2018, pp. 8–9). The Democracy matrix is exceptional in that it not only operationalizes the quality of democracy in terms of degrees (the 'core measurement') but also covers the content by explicitly operationalizing the trade-offs between the dimensions (the 'trade-off measurement') as well as context factors (the 'context measurement') affecting the quality of democracy (Lauth & Schlenkrich, 2019a, pp. 10–15).

[22] See also Kaiser, Lehnert, Miller, and Sieberer (2002) on a detailed conceptualization and analysis of the trade-off between inclusion of preferences and responsibility of government in representative democracies as a measure of democratic quality of institutional regimes.

Table 1.2

Comparison of Measurement Instruments of the Quality of Democracy

	SGI	DB	GDR	DM
Conceptualization (C)				
Conceptual content				
• meaning (C1)	• contextualized procedural	• contextualized procedural	• substantive procedural	• strictly procedural ➢ recognition of impact of contextual factors
• nature of the concept (C2)	• continuous	• continuous	• continuous	• continuous
• level of origin (C3)	• individual & macro level	• individual & macro level	• individual & macro level	• individual & macro level
• level of the concept (C4)	• macro level	• macro level	• macro level	• macro level
Conceptual logic				
• number of dimensions and sub-dimensions (C5)	• 4 dimensions, 15 sub-dimensions	• 3 democratic principles, 9 democratic functions, 18 components, 53 subcomponents	• 3 dimensions, 6 attributes	• 15 matrix fields; 27 components; 12 subcomponents
• relationship between dimensions (C6)	• n.a.	• equal importance, interdependence, partial compensation, no substitution ’optimise interdependence between equality and freedom by means of control	• complementary & conflicting	• complementary & conflicting
• relationship between dimensions and sub-dimensions (C7)	• n.a.	• n.a.	• complementary (reflect sub-systems)	• necessary and (jointly) sufficient relationship
Conceptual range				
• range (scope conditions) (C8)	• democracies	• democracies	• (partly) free countries	• political regimes & democratic regimes

Table 1.2 (continued)

	SGI	DB	GDR	DM
Operationalization (O)				
• coverage of content (O1)	• democracy & the rule of law	• rules in law & use	• quality of politics & quality of society	• core measurement, context factors, trade-offs between dimensions
• measurement scale & level (O2)	• ordinal scale: 1-10	• different measurement scales depending on the indicator	• interval scale: 1-100	• interval scale ➢ core measurement: 0-1 ➢ trade-off: 0.75-1 ➢ context: 0.75-1
• thresholds (O2)	• n.a.	• n.a.	• n.a.	• 0.5/0.75 threshold on the core measurement scale to distinguish autocracies/deficient dem./working dem.
• content of indicators (O3, O4, O6)	• exclusively based on global property indicators	• 87 global property indicators	• 22 global property indicators	• exclusively based on global property indicators
• number of indicators (O5)	• 15 indicators	• 105 indicators	• 42 indicators	• 86 quality-measuring, 15 trade-off, 9 context indicators
• measurement model (O6, O7)	• n.a.	• reflective	• n.a.	• n.a.

Table 1.2 (continued)

	SGI	DB	GDR	DM
Measurement (M)				
• instrument of data collection/data sources (M1)	• questionnaire-based qualitative assessments	• secondary data	• secondary data	• secondary data (V-Dem dataset)
• method of data collection (M1, M2, M3, M4)	• expert judgments	• collection of existing data including observational data, reports, archival material etc.; expert judgments; survey data	• differs depending on the data source	• expert judgments
• level of data collection (M3, M4)	• macro level	• individual & macro level	• individual & macro level	• macro level
• units of observation (M3, M4)	• pol. parties, citizens, media, government etc.	• institutions, individuals,, government, pol. parties etc.	• individuals, socio-economic units, institutions etc.	• individuals, institutions, pol. parties, social groups etc.
• method of case selection (M3, M4, M8)	• EU + OECD states	• democracy	• based on Freedom House assessments	• all countries (data availability provided)

Table 1.2 (continued)

	SGI	DB	GDR	DM
Aggregation (A)				
• extent of data aggregation (A1)	• composite index; dimensional indices	• overall index, principles, functions, components & sub-components	• overall index	• core measurement, context measurement, trade-off measurement ➢ total value, dimensional & institutional indices ➢ core & context measurement ➢ core & trade-off measurement
• level of aggregation (A3, A4)	• country level	• country level	• country level	• political regime level
• units of analysis (A3, A4, A8)	• countries	• countries	• countries	• countries, institutions (depending on the level of aggregation)
• method of aggregation (A2, A5, A6, A7)	• average	• average, aggregation formula	• weighted average per attribute	• methods of aggregation ➢ cumulative distributive function + min-max transformation ➢ average or weighted average ➢ multiplication + root extraction ➢ multiplication ➢ min. or max. value
• aggregation rules (A2, A6, A7)	• equal weighting of indicators and dimensions	• stepwise aggregation by concept level ➢ standardization to an interval scale from 0-100 based on empirical min./max. values ➢ average of indicators & subcomponents	• attributes weighted differently ➢ politics (50%) ➢ gender, economy, knowledge, health, environment (10% each)	• different aggregation rules depending on the component/sub-component

Note. Own compilation; n.a.: no information available.

As for the reflection of the nature of the concept in the operationalization, the measurement scales and levels (O2) of the measurement instruments differ substantially. The Sustainable Governance Indicators uses an ordinal scale that ranges between one and 10 (SGI, 2018a, pp. 38–48). The Democracy Barometer's scales differ depending on the indicator (Merkel et al., 2018b). The Global Democracy Ranking and the Democracy Matrix both use an interval scale that ranges from one to 100 in the former case and from zero to one in the latter case (D. F. Cambell & G. Pölzlbauer, 2008, p. 5; Lauth & Schlenkrich, 2019c, p. 2). The only measurement instrument to apply thresholds (O2) is the Democracy Matrix. The authors use two thresholds on their core measurement scale to distinguish autocracies, deficient democracies and working democracies (Lauth & Schlenkrich, 2019c, p. 5).

The measurement instruments are similar with regard to the manner in which the content of their indicators reflects the level of origin and the level of the concept (O3, O4). As shown in Table 1.2, in all indices, the content of the majority of indicators focuses on 'global properties' of democracies. These indicators describe the quality of the principles, structures, and outcomes of democratic institutions in terms of properties that are "not based on information about the properties of individual members" (Lazarsfeld & Menzel, 1961, p. 503). In comparison, indicators that measure individual-level states, behavior, beliefs, attitudes, or values are seldom used.

Concerning the translation of the conceptual logic in operational terms, the indices differ considerably with regard to the number of indicators they use (O5). For example, whereas the Sustainable Governance Indicators are based on only 15 indicators, the Democracy Barometer uses 105 indicators to operationalize its systematized concept of the quality of democracy (Merkel et al., 2018a, p. 3; SGI, 2018a, pp. 38–48). Other than that, only one measurement instrument explicitly reflects on its underlying measurement model (O6, O7). In line with a reflective measurement model, the Democracy Barometer deduces its indicators from the democratic principles and functions specified in its systematized concept of the quality of democracy (Bühlmann, Merkel, Müller, et al., 2008, p. 118).[23]

Regarding the measurement stage in the measurement process, the indices differ for the most part (see row M in Table 1.2). Regarding the instrument of data collection and the

[23] The other measurement instruments proceed similarly, suggesting that their assessments are also based on a reflective measurement model.

data sources they use (M1), the Democracy Barometer, the Global Democracy Ranking and the Democracy Matric rely on secondary data sources whereas the Sustainable Governance indicators conducts its own, questionnaire-based qualitative assessments. In terms of the method of data collection (M1, M2, M3, M4), expert judgments and official statistics predominate (D. F. Campbell & Barth, 2009, p. 216; Lauth & Schlenkrich, 2019d, p. 2; Merkel et al., 2018b, p. 10; SGI, 2018a, pp. 15–16). As for the level of data collection, the indices are constructed using only macro-level data (Lauth & Schlenkrich, 2019, p. 2; SGI, 2018a, pp. 38–48) or individual-level as well as macro-level data, depending on their data basis (D. F. Campbell & G. Pölzlbauer, 2008, pp. 7-8; Merkel et al., 2018b) (M3, M4). The Democracy Barometer is unique in that it is the only measurement instrument that includes survey data (Merkel et al., 2018b, p. 10). The units of observation (M3, M4) underlying the data vary greatly. Among many others, they include citizens, the media, government, districts, political parties, and social groups (D. F. Campbell & G. Pölzlbauer, 2008, pp. 19–26; Lauth & Schlenkrich 2019b; Merkel et al., 2018b; SGI, 2018a, pp. 38–48). The method of case selection also differs (M3, M4, M8). The Democracy Barometer, for example, seeks to include all countries, provided that the data are available (Lauth & Schlenkrich, 2019e, p. 1). In contrast, the Sustainable Governance Indicators explicitly limit their assessments to EU and OECD states (Schraad-Tischler & Seelkopf, 2018, p. 2).

The measurement instruments differ most profoundly with regard to the aggregation (see row A in Table 1.2). Regarding the extent of data aggregation (A1), all measurement instruments provide overall indices, dimensional aggregations as well as the disaggregated data. In doing so, the Democracy Barometer and the Democracy Matrix explicitly theorize the different levels of aggregation (Bühlmann, Merkel, Müller, et al., 2008, pp. 117–120; Lauth & Schlenkrich, 2019c). For instance, the Democracy Matrix not only offers a 'total value index' of its core measurement of the quality of democracy. It also provides an institutional as well as a dimensional index. Whereas the former assesses the extent to which the five core democratic institutions function well, the latter reflects the degree to which the three core democratic dimensions (freedom, democracy, and control) are developed within a given country (Lauth & Schlenkrich, 2019c, pp. 3–4). Independent of the extent of data aggregation, all indices aggregate the data to the same level (A3, A4): the country level. Likewise, they all intend countries to be the primary units of analysis (A3, A4, A8) (D. F. Campbell & Barth, 2009, p. 217; Lauth &

Schlenkrich, 2019a, p. 2; Merkel et al., 2018a, pp. 5–7; SGI, 2018b, pp. 62, 68). The Democracy Matrix is an exception in that its institutional indices also permit analyses that focus on political institutions within a given country (Lauth & Schlenkrich, 2019c, pp. 3–4). Notwithstanding these similarities, the indices differ profoundly in terms of the methods of aggregation (A2, A5, A6, A7) as well as the aggregation rules (A2, A6, A7) they apply. The methods vary from simple averages (D. F. Campbell & Barth, 2009, p. 217; Schraad-Tischler & Seelkopf, 2018, pp. 17–18; SGI, 2018a, p. 14) to complex aggregation formulas (Lauth & Schlenkrich, 2019c; Merkel et al., 2018a, p. 10). Some indices apply the same rules to all dimensions (Schraad-Tischler & Seelkopf, 2018, pp. 17–18; SGI, 2018a, p. 14) whereas others change them depending on the dimension and sub-dimension (Bühlmann, Merkel, Müller, & Wessels, 2012, p. 532; D. F. Campbell, 2008, p. 34; D. F. Campbell & Barth, 2009, p. 217; D. F. Campbell & G. Pölzlbauer, 2008, p. 6; Lauth & Schlenkrich, 2019c; Merkel et al., 2018a, pp. 8–11).

In sum, the differences between the measurement instruments of the quality of democracy undoubtedly outweigh the similarities. In terms of conceptualization, the indices share common ground in that they all include procedural elements in their systematized concepts. All indices conceive of the quality of democracy as a multidimensional concept with complementary and conflicting dimensions and agree on the nature, level of origin, and level of the concept. The differences between the indices are clearly evident with regard to the choices they make in the measurement process, however. Aside from the level of aggregation and countries as units of analysis, each index takes a different approach to operationalizing, measuring, and aggregating their assessments of the quality of democracy. Altogether, the summary of the state of research on the systematized concepts and measurement instruments shows that the approaches to conceptualizing and measuring the quality of democracy differ with regard to the majority of aspects included in the dissertation's theoretical foundation (see section 1.1.2).

1.2.1.2 The Need for an Overview

The heterogeneity of conceptualizations and measurement instruments of the quality of democracy is clearly evident. It is not surprising, then, that previous reviews pertain to individual measurement instruments and provide recommendations on how to enhance specific aspects of the measurement process in light of the respective systematized

concept. For example, Kaina (2008, p. 522) criticizes the Democracy Barometer in terms of the match between its systematized concept and its operationalization. According to her, the index suffers from a discrepancy between the conceptual nature of the quality of democracy and the chosen measurement scale (see C2 and O2 in Table 1.2). The authors of the Democracy Barometer consider a democracy to reach the best level of quality when freedom, equality, and control are combined in a way that allows them to be fulfilled in an optimal manner (Bühlmann, Merkel, Müller, et al., 2008, p. 119). Kaina (2008, p. 522) argues that the term 'optimal' requires a dichotomous measurement scale - the functions are either combined optimally or they are not. The index measures the quality of democracy on a continuous scale from one to ten, however. Concerning the Sustainable Governance Indicators, Jäckle and Bauschke (2009, pp. 368–370), find fault with the operationalization of the quality of democracy with regard to the coverage of the content of its dimensions (see O1 in Table 1.2). They argue that the indicators do not allow for a sufficient distinction between measurements of the quality of democracy and its neighboring concept, the degree of democracy.

Previous efforts to improving the measurement validity of the quality of democracy have their merits but also their shortcomings. They benefit individual measurement instruments insofar as they acknowledge the conceptual differences within the research community. In addition, these recommendations help to improve measurement instruments' validity in specific respects. The downside is that such isolated validations add little benefit to empirical research on the quality of democracy as a whole. Instead, they run the risk of fragmenting the field similar to democratization research and its 'democracies with adjectives' (Collier & Levitsky, 1997, p. 431). What is more, they do not always take into account the measurement process as a whole. As noted earlier, the operationalization, measurement, and aggregation work in conjunction, however. Consequently, the extent of a measurement instrument's overall measurement validity depends on the degree to which each of these stages coincides with the respective concept. By concentrating on individual aspects at certain stages, previous evaluations do not consider these interdependencies.

Thus, the debate on the measurement validity of the quality of democracy would benefit from a more comprehensive review. Article 1 'Assessing the quality of quality measures of democracy. A theoretical framework and its empirical application' (S. Pickel, Stark, & Breustedt, 2015) addresses this issue by answering the following research question: How

valid are current measurement instruments of the quality of democracy and how should they be improved?[24]

According to the dissertation's theoretical foundation outlined in section 1.1.2 (see Table 1.1), there are many aspects to consider at each stage. The initial part of the dissertation's first article develops a systematic framework that provides evaluation criteria for each of these aspects. The framework comprises a standardized set of 20 criteria and corresponding coding rules that pertain to the quality of the systematized concept, the operationalization, measurement, and aggregation. The criteria are based on methodological standards that can be applied to all indices independent of their conceptual foundation. It includes criteria that serve to evaluate the goodness of the conceptualization as well as the validity of indices' operationalization, measurement, and aggregation.[25] In order to avoid partiality when evaluating the conceptual template for the measurement process, the criteria do not address researchers' conceptual choices regarding the meaning, nature, level of origin, or level of the concept (see C1 to C4 in Table 1.1). The coding rules specify when a measurement instrument fulfills, partly fulfills or does not fulfill the 20 criteria. These judgments can then be used to construct a tripartite quality index that allows researchers to compare the number of positive, intermediate, and negative evaluations per measurement instrument (Müller & S. Pickel, 2007, p. 518). Altogether, the framework supplies the means for an overall, comparative evaluation of the quality of quality measures of democracy.

The framework incorporates several validation strategies to collect different types of evidence for validity. From among the classic validation strategies, it applies content validation by means of qualitative expert judgments. These are intended to evaluate the degree to which the content of the measurement instruments' indicators represents the attributes of their systematized concepts of the quality of democracy (Adcock & Collier, 2001, p. 537; Carmines & Zeller, 1979, pp. 20–22). In addition, the framework goes beyond the traditional approach to applying this validation procedure in two respects.

[24] In article one, the author of the dissertation at hand was mainly responsible for explicating the theoretical criteria for the empirical analysis (see section 2.2 'The quality assessment criteria').

[25] The framework also includes criteria concerning the reliability and replicability of the measurement process. It addresses reliability because it affects the validity of measurements (Munck & Verkuilen, 2002, p. 18; P. Newton & Shaw, 2014, p. 14). In addition, it contains criteria regarding the replicability of the measurement process as "claims about either validity or reliability hinge upon the replicability of measures. Yet because issues of measurement are inescapably subjective, involving a variety of judgments rather than firmly objective criteria, it is absolutely vital that the community of scholars retain the ability to scrutinize and challenge the choices that shape the generation of data" (Munck & Verkuilen, 2002, pp. 18–19).

First, it not only addresses the match between the indicators and their conceptual foundation, it also includes criteria to evaluate the validity of the measurement and aggregation stage of each index as well. Validating these stages is important as invalid measurement or aggregation offset an otherwise valid operationalization. Second, the framework provides the means to quantify the extent to which the measurement instruments reflect their systematized concepts, thereby permitting a comparison of their validity.

The second part of the article applies this systematic framework to three indices suitable for large N comparative analysis, namely the Democracy Barometer, the Global Democracy Ranking and the Sustainable Governance Indicators. Drawing on publicly available documents and data for each index, the article develops logical arguments and provides empirical evidence regarding the match between the systematized concept and its operationalization, measurement, and aggregation.

Using this evidence[26], the second part of the article concludes by comparing the number of times each index fulfills, partly fulfills, or does not fulfill the 20 quality criteria. The Democracy Barometer fares best as it fulfills 13 out of 20 criteria, partly fulfills three criteria and fails to fulfill four criteria. The Sustainable Governance Indicators come in second. They fulfill 11 out of 20 criteria, partly fulfill five criteria and do not fulfill four criteria. The Global Democracy Ranking ranks third as it fulfills only eight criteria, partly fulfills two criteria and fails to fulfill 10 criteria. Overall, the comparison shows that their quality varies substantially.

On the basis of the indices' comparison, the third part of the article makes three recommendations to improve the measurement validity of current quantitative empirical assessments of the quality of democracy. First, it points out that the validity of the indicators used to measure the quality of democracy should be enhanced. The match between the indicators' content and the content of the conceptual attributes they are intended to measure, the empirical applicability of the indicators, as well as redundancies and conflations among the indicators require improvement in all indices (O1 in Table 1.2). Second, the article notes that while some indices' conceptual range is limited to full democracies, they nonetheless provide empirical data on electoral democracies ('conceptual range' C8 and 'method of case selection' M3, M4, M8 in Table 1.2). It cautions that researchers should limit data collection to their index's conceptual

[26] as well as an evaluation of the measurement instruments' reliability and replicability

boundaries in order to avoid conceptual stretching. Third and relatedly, the article proposes to limit the assessment of the quality of democracy to full democracies in the future. It argues that measurements of the quality of democracy in electoral democracies are of little substantive value compared to measurements of the degree of democracy as electoral democracies' quality is worse by definition.

1.2.1.3 The Call for Citizens' Perspective

Another way to discern how the measurement validity of quantitative empirical assessments of the quality of democracy could be improved is to look for common measurement validity issues among the measurement instruments. This approach helps to avoid methodological nitpicking. One such issue concerns the match between the level of origin of the quality of democracy as specified in the indices' systematized concepts on the one hand and its operationalization and measurement on the other hand (see C3, O3, and M3 in Table 1.2). As described in section 1.2.1.1, all of the indices suggest that the quality of democracy originates at both the individual and macro level. As outlined in the dissertation's theoretical foundation (see Table 1.1), the level of origin should be reflected in the content of the indices' indicators. In addition, in terms of measurement, it should be considered when choosing the units of observation in order for the indices to provide valid assessments. Even though current indices include indicators and units of observation that pertain to the individual and macro level (O3 and M3 in Table 1.2), they rarely include citizens' assessments of the quality of democracy. This may lower the indices' measurement validity, however.

Researchers have provided methodological and theoretical arguments as to why the neglect of the citizens' perspective may lead to a bias in current quantitative empirical assessments. Ringen (2007) argues that "[s]ystems have *potential* but the *value* contained in that potential is manifested [...] in the lives of persons" (p. 17; emphasis in original). Methodologically, he therefore concludes that the assessment of the quality of democracy requires 'double bookkeeping', that is, the inclusion of observations of both the system as well as individual citizens (Ringen, 2007, pp. 18–19). More specifically, the observations at the individual level should contain citizens' *evaluations* of the quality of democracy since "people are themselves the best judge of their well-being" (Ringen, 2007, p. 9; see also Beetham et al., 2008a, p. 19 for a similar argument). Researchers working in the tradition of political culture theory also advocate including the citizens' perspective.

According to Lauth (2011a, pp. 72–73), from among the set of indicators developed in this field, individual-level evaluations of the political system could contribute to analyses of the quality of democracy. These evaluations cover key aspects of current systematized concepts of the quality of democracy, namely citizens' evaluations of the institutional accountability procedures as well as evaluations of the policy performance of the institutions. Thus, the measurement validity of current assessments of the quality of democracy could be improved on both methodological and theoretical grounds by including the citizens' perspective.

At the same time, those who support this point of view draw attention to two issues that should be considered prior to using citizens' assessments to improve current measurement instruments. Regarding the conceptual template for the evaluations, they note that people's understanding of democracy may differ within and across countries. This poses a challenge for comparisons between current indices – which draw on a certain understanding themselves – and citizens' assessments – which are not necessarily based on the same concept of the quality of democracy. Concerning the evaluations, researchers also point out that citizens generally may not have all the information needed to be able to make overall assessments of the quality of democracy. Still, they conclude that citizens' evaluations constitute a valuable contribution as they help to identify issues of concern to the public (Lauth, 2011a, pp. 73–74; Ringen, 2011, pp. 17, 31).

Researchers' appeal and the uncertainties they address indicate that the debate on the measurement validity of the quality of democracy would benefit from a clarification of the relationship between citizens' and indices' systematized concepts on the one hand and their evaluations on the other hand. Article 2 'Measuring the quality of democracy: Why include the citizens' perspective?' (S. Pickel, Breustedt, & Smolka, 2016) attends to this research gap by answering the following research question: Do citizens' understandings and evaluations of the quality of democracy coincide with the concepts and assessments of the quality of democracy by existing indices or do they provide a complementary perspective?[27]

Applying the principles of construct validation (see section 1.1.2.5), the dissertation's second article tests two hypotheses within the nomological net of the 'quality of democracy' concept. The first hypothesis assumes a correspondence between indices'

[27] In article two, the author's key task was to convey the results in the empirical analysis (see section 3.5.2 'Evaluation'), to formulate the line of argument in the introduction and conclusion and to describe the method in the research design section (section 3.4).

and citizens' systematized concept; the second hypothesis expects a convergence between their empirical evaluations. The article tests these hypotheses by comparing the Democracy Barometer's operationalization and measurement with citizens' understandings and assessments of the quality of democracy in 20 European established democracies on the basis of European Social Survey data (European Social Survey, 2012a).

In a first step, the article analyzes whether citizens' understanding of democracy corresponds to the systematized concept underlying the Democracy Barometer. Using ESS survey items that cover the meaning of the Democracy Barometer's systematized concept, it performs a principle component analysis for each country. The results show that, generally speaking, people's understanding of a 'good' democracy is similar to the Democracy Barometer's systematic concept. Items covering the idea of representation, competition, freedom, vertical accountability, and the rule of law load strongest on the first principal component in the majority of countries. Beyond this shared core understanding, it appears that countries' economic circumstances give rise to country-specific associations with the quality of democracy.

In a second step, the article determines how citizens evaluate the quality of their democracy compared to the Democracy Barometer. Paying heed to researchers' concerns described above, this step takes into account only those evaluative items that reflect citizens' understanding of democracy in their respective country. On the basis of these 'real-life' understandings of democracy, the article adds up the rates of approval for the respective evaluative items and standardizes them by dividing the sum by the number of items used. Using this empirical evidence, the article compares countries' rankings according to citizens' evaluations of the quality of democracy with the assessments by the Democracy Barometer. At the time point of analysis in 2012, the rank of eight out of 20 countries was same or only differed by one.[28] By contrast, in five countries, the quality of their democracy fared far better in the eyes of the citizens than according to the Democracy Barometer.[29] In seven countries, the opposite was the case.[30]

On the basis of this comparison, the article concludes that citizens' assessments provide a valuable complementary perspective to measuring the quality of democracy that should be taken into account in future research. The empirical comparison shows a

[28] Sweden, Denmark, Norway, Netherlands, Iceland, Germany, Poland, and Spain.
[29] Finland, Czech Republic, Cyprus, France, and Hungary.
[30] Belgium, Ireland, Italy, Switzerland, the United Kingdom, Slovenia, Portugal.

correspondence between citizens' understandings and evaluations on the one hand and the Democracy Barometer on the other hand. This assuages the above-mentioned concerns about citizens' ability to make judgments about the quality of democracy. At the same time, the differences between citizens' and the index's systematized concepts indicate that citizens add additional dimensions to the core meaning, depending on their country. In addition, the discrepancies in the evaluations testify to the fact that citizens perceive the quality of democracy somewhat differently compared to experts judgments and official statistics.

Thus, the article recommends to incorporate citizens' perspective in future empirical studies of the quality of democracy to improve their measurement validity. This could be accomplished in two ways. First, existing indices could be expanded to include further aspects of the quality of democracy (O1 in Table 1.2). Second, researchers could consider citizens' perspective by incorporating appropriate survey data in their empirical analyses (O3 'content of indicators' and M3 'units of observation' in Table 1.2).

1.2.2 Political Support

As for the second concept of interest in this dissertation – political support – the scientific community appears to have resolved many issues regarding its conceptualization and valid measurement. In over half a century of research, social scientists have applied the concept and its individual dimensions in many empirical studies to describe and explain democratic transitions, reversals, consolidation, as well as the effects of a critical citizenry on the democratic process and persistence of democracies (G. Pickel, 2010). In the course of these empirical applications, researchers have identified a number of weaknesses in its original conceptualization and measurement and have developed suggestions on how to improve them. The dissertation's articles can thus build on a broad range of efforts to improve the measurement validity of its quantitative empirical assessments. The following section summarizes the state of debate on its conceptualization and valid measurement to highlight what remains to be addressed.

Regarding the conceptual foundation for its empirical assessment, several researchers have substantiated the content and structure of Easton's (1965, 1975) original multidimensional conceptualization of political support (see for example Dalton, 2004; Fuchs, 2002, 2007; Norris, 1999, 2011, 2017; S. Pickel, 2013, pp. 162–164; Tables A1 to A3 in Appendix A). Regarding the meaning of the concept, the attributes remain largely

unchanged. Researchers still conceive of political support to encompass people's support of political objects at different levels of abstraction (C1 in Table 1.1).[31] All systematized concepts thus continue to consider political support to be a multidimensional concept (see C5 in Table 1.1; Dalton, 2004, pp. 7, 23; Fuchs, 2002, pp. 36–37, 2007, pp. 164–166; Norris, 1999, p. 9, 2011, pp. 21–23, 2017, p. 23; S. Pickel, 2013, pp. 163–164). Judging by their choice of operationalization, they likewise generally concur with the idea of its continuous (as opposed to categorical) nature (see C2 in Table 1.1; Dalton, 2004, pp. 24–45; Easton, 1965, p. 163; Fuchs, 2007, pp. 167–169; Norris, 1999, pp. 10–12; S. Pickel, 2013, p. 166). In terms of the level of origin and level of the concept (see C3 and C4 in Table 1.1) all authors subscribe to the original conceptualization of political support as a concept that originates at the individual level and can be analyzed at the individual or macro level (Dalton, 2004, pp. 5, 14; Easton, 1965, pp. 158, 166, 168–169; Fuchs, 2007, pp. 165–166, 173; Norris, 1999, pp. 9, 26). As for the range (see C8 in Table 1.1), according to Easton's (1965) original conceptualization, political support can be applied to any political system. Recent conceptualizations usually limit the scope of application to democratizing and democratic political regimes, however.[32] Last but not least, among the authors, Fuchs (2002, pp. 37–38) is the only one to explicitly theorize the relationship between the dimensions in his model in terms of a causal relationship (see C6 in Table 1.1).[33] Overall, these specifications have rendered the concept more easily accessible to empirical research. This provides a better conceptual foundation in order to operationalize, measure, and aggregate the data in a valid manner.

As a result of these efforts, at least in the Eastonian tradition, the scientific community is approaching common ground regarding the conceptualization of political support. There is one qualification to this overall assessment, however. Over time, each dimension of the original concept has advanced as a field of research in its own right. Notwithstanding the

[31] Recent conceptualizations expand the concept's dimensions by differentiating between individuals' support for democratic regime principles and values on the one hand and individuals' assessment of democratic regimes' institutions in practice. This allows researchers to grasp people's assessments of democracy in a more refined manner (Dalton, 1999, p. 59, 2004, pp. 6–7; Fuchs, 2002, pp. 37–39, 2007, p. 166; Norris, 1999, p. 9, 2011, pp. 23–25; S. Pickel, 2013, pp. 163–164). With regard to the number of dimensions and, relatedly, the meaning associated with each dimension (see C1 and C5 in Table 1.1), Fuchs' and S. Pickel's models differ from Dalton's and Norris' model as they do not include support for the nation state. Fuchs (2002, pp. 43–45, 2007, pp. 169–170) justifies this by arguing that Easton (1965, p. 185) described it as a prerequisite for the persistence of political systems as such.

[32] See Breustedt and Stark (2015) as well as S. Pickel and Stark (2010) for an application of the concept of political support to authoritarian regimes.

[33] None of the authors address the relationship between the dimensions, sub-dimensions, and indicators (see C7 in Table 1.1).

commonalities concerning the overall concept described above, researchers have developed different conceptualizations of its individual dimensions. This has resulted in a diversification of systematized concepts. With regard to people's support for regime principles, Fuchs (2002, pp. 40–43) goes beyond the original conceptualization by distinguishing between support for libertarian, liberal, republican, and socialist models of democracy; according to Inglehart and Welzel (2005, pp. 247–248), a broader set of civic values, which they refer to as 'self-expression values', is required in order to sufficiently describe the values that make democracy work. Concerning citizens' evaluations of regime performance, Roller (2005, pp. 19–73) develops a multidimensional model to evaluate the effectiveness of liberal democracies. As for people's political trust, researchers have proposed three distinct systematized concepts. Some consider it to be a single-dimensional concept (see for example Hooghe, 2011, p. 274; Marien, 2011a, pp. 16–17); others distinguish between trust in political authorities and political institutions (see for example Dalton, 2004, p. 24; Denters, Gabriel, & Torcal, 2007, p. 68; Norris, 2011, pp. 23–31); still others separate trust in representative and implementing institutions conceptually (see for example Gabriel, 1999, pp. 206–207; G. Pickel & Walz, 1995, p. 146; Rothstein & Stolle, 2003, pp. 193–194). Thus, researchers' overall agreement regarding the conceptualization of political support does not apply to the same extent to its individual dimensions.

As for the improvements of its measurement process, social scientists have devoted particular attention to validating the operationalization and measurement of political support. Regarding its operationalization, researchers have discussed and tested the match between the conceptual meaning and common indicators for several dimensions of the concept. These studies pertain to 'support for regime principles' (Ariely & Davidov, 2011), 'trust in regime institutions' (Brunner & Walz, 2000; Citrin, 1974; Feldman, 1983; A. Miller, 1974a, 1974b) and 'evaluations of regime performance' (Canache et al., 2001; Linde & Ekman, 2003).

The validity of its measurement has been examined in several respects. Researchers have scrutinized people's ability to empirically discriminate between the dimensions of political support as a response to critics who argued that this overtaxes survey respondents (see for example Allenspach, 2012; Booth & Seligson, 2009; Caballero, 2009; Fuchs, Gabriel, & Völkl, 2002; Klingemann, 1999). Furthermore, scholars have discussed whether the measurements of different dimensions of political support can be

compared across countries. Several studies have addressed the comparability of the dimension of 'support for the nation-state' (Gabriel, 1998; S. Pickel. 2010) and 'support for regime principles' (Mishler & Rose, 2001a; Schedler & Sarsfield, 2007). Recently, scholars have tested the cross-national measurement invariance of several measurement models used to assess people's political trust (see for example Coromina & Davidov, 2013; Marien, 2017; I. Schneider, 2017). Overall, social scientists have questioned the validity of the measurement process in a number of respects in order to improve the operationalization and measurement of political support in general and for comparative research in particular.

The state of research outlined above indicates that the debate on the measurement validity of quantitative empirical assessments of political support would benefit from analyses that address two key issues. First, researchers have attended to many aspects with regard to the measurement validity of political trust as a key attribute of political support. Concerning its measurement invariance, their analyses are inconclusive, however. Second, whereas social scientists have devoted substantial attention to validating the operationalization and measurement of political support, they have not tested the validity of the standard approach to aggregation in cross-national studies, that is, the construction of index scores. Thus, the question is: How can the measurement validity of political trust and the aggregation of political support be improved for comparative research?

In order to further demarcate the issues in question and to illustrate what requires improvement, section 1.2.2.1 describes the items available to operationalize and measure political trust; section 1.2.2.3 gives a succinct account of current approaches to aggregating political support data. On the basis of the respective reviews, sections 1.2.2.2 and 1.2.2.4 summarize the articles' efforts to evaluating and improving the measurement validity of political support.

1.2.2.1 Current Approaches to Operationalizing and Measuring Political Trust

In the tradition of political culture research, political trust is usually measured by means of survey data (Almond & Verba, 1965, pp. 40–44; S. Pickel & G. Pickel, 2006, p. 31). On the basis of Table 1.1, the following chapter briefly outlines the characteristics of the respective items included in the most recent wave of publicly available cross-national surveys: the Afrobarometer, the Asian Barometer, the Eurobarometer, the European

Values Study, the European Social Survey, the International Social Survey Program, the Latinobarómetro and the World Values Survey (S. Pickel & G. Pickel, 2006, pp. 33–36). This helps to understand why researchers still disagree on how to measure political trust in a valid manner across countries.

As indicated in Table 1.3 below, the survey projects usually do not explicitly refer to a specific systematized concept (see row C). Instead, they list political trust under the topics included in their questionnaires such as public attitudes on democracy and governance (Afrobarometer, 2017a, p. i) or trust in institutions (Asian Barometer, 2020a, n.a.). The European Social Survey is exceptional in this respect. The survey project website provides a detailed description of the questionnaire design process, including statements by academic specialists. With regard to political trust, Thomassen (2001, pp. 181–185) makes explicit reference to Easton's (1965) concept of political support as summarized above.

Concerning the operationalization, the survey projects take a similar approach, notwithstanding certain variations (see row O in Table 1.3). As for the content of the indicators (O3, O4, O6), all questionnaires include items pertaining to political objects. The number of political objects referred to – and thus the number of items included (O5) – varies, however, ranging from five in the European Social Survey to ten in the Asian Barometer. The question wording is very similar. Regarding the coverage of content (O1), all questionnaires include items that address trust in several political objects. Items covering trust in parliament, the police, political parties[34] and the courts of law are included in in all surveys. Six questionnaires ask about respondents' trust in the army, five inquire about their trust in government. Four of the seven survey projects include items on trust in the national electoral commission, civil service and local government council. Less frequent items include trust in the president or prime minister, politicians, the health care system and the social security system. As for the measurement scale and level (O2), people's political trust is usually gauged on a four-point ordinal scale. The labels differ slightly, though (see also Breustedt 2015, 8). The Eurobarometer and the European Social Survey take a different approach – while the former allows respondents to choose between two options, the latter asks them to their level of trust on a scale from zero to 10. Finally, mirroring the lack of explication of a systematized concept, none of

[34] The Afrobarometer differentiates between trust in the ruling party and opposition political parties, however.

the survey projects make reference to any measurement model of political trust (O6, O7) (Afrobarometer, 2017b, p. 14; Asian Barometer, 2014-2016, p. 2; Eurobarometer, 2019, n. a.; European Social Survey, 2018, p. 8; European Values Study, 2018, pp. 12–13; Latinobarómetro, 2018a, p. 1; World Values Survey, 2012, pp. 8–9).

As for the measurement stage (see row M in Table 1.3), the survey projects' approach varies somewhat. All survey projects use a standardized questionnaire as the instrument of data collection (M1) to collect data at the individual level (M3, M4). Likewise, they all carry out face-to-face interviews as the method of data collection (M1, M2, M3, M4).[35] The surveys use probability sampling as the method of case selection, the sample design differs somewhat depending on the survey, though (M3, M4, M8). Also, the designated units of observation (M3, M4) vary. While some survey projects only include citizens of a certain age in their samples, others are less restrictive in that they include the resident population (Afrobarometer, 2020a, 2020b; Asian Barometer, 2020b; Eurobarometer, 2020a, 2020b; European Social Survey, 2020a, 2020b; European Values Study, 2020a, 2020b; Latinobarómetro, 2018b; World Values Survey, 2020).

In sum, whereas early research on political trust applied at least nine different item scales (Citrin & Muste, 1999, p. 469) today, at least the large cross-national public opinion surveys share a number of common features regarding their approach to measuring political trust. A standard scale of items for comparative research on political trust has yet to be developed, however.

[35] In 2017, the European Values Study also allowed alternative methods including interviews via post and the web.

Table 1.3
Comparison of Measurement Instruments of Political Trust

	Afrobarometer	Asian Barometer	Eurobarometer	European Social Survey	European Values Study	Latino-barometro	World Values Survey
Conceptualization (C)							
• conceptual content, logic, and range (C1 to C8)	• public attitudes on democracy and governance	• trust in institutions	• n. a.	• political support	• n. a.	• n. a.	• n. a.
Operationalization (O)							
• coverage of content (O1)	• pol. institutions and authorities ➢ president ➢ parliament/ national assembly ➢ national el. commission ➢ local gov. council ➢ ruling party ➢ opposition pol. parties ➢ police ➢ army ➢ courts of law	• pol. institutions and authorities ➢ president ➢ courts ➢ nat. gov. ➢ pol. parties ➢ parliament ➢ civil service ➢ military ➢ police ➢ local gov. ➢ election commission	• pol. institutions and authorities ➢ public administration ➢ regional or local public authorities ➢ government ➢ parliament ➢ pol. parties ➢ legal system ➢ police ➢ army	• pol. institutions and authorities ➢ parliament ➢ legal system ➢ police ➢ politicians ➢ pol. parties	• pol. institutions and authorities ➢ armed forces ➢ police ➢ parliament ➢ civil service ➢ social security system ➢ health care system ➢ justice system ➢ pol. parties ➢ government	• pol. institutions and authorities ➢ armed forces ➢ police ➢ parliament ➢ government ➢ legal system ➢ pol. parties ➢ election commission	• pol. institutions and authorities ➢ armed forces ➢ police ➢ courts ➢ government ➢ pol. parties ➢ parliament ➢ civil service
• measurement scale & level (O2)	• ordinal (0-3)	• ordinal (1-4)	• nominal (1-2)	• ordinal (0-10)	• ordinal (1-4)	• ordinal (1-4)	• ordinal (1-4)
• content of indicators (O3, O4, O6)	• pol. objects	• pol. objects	• pol. objects	• pol. objects	• pol. objects	• pol. objects	• pol. objects

Table 1.3 (continued)

	Afrobarometer	Asian Barometer	Euro-barometer	European Social Survey	European Values Study	Latino-barometro	World Values Survey
Operationalization (O)							
• number of dimensions/ indicators (O5) • measurement model (O6, O7)	• 9 • n.a.	• 10 • n.a.	• 8 • n.a.	• 5 • n.a.	• 9 • n.a.	• 7 • n.a.	• 7 • n.a.
Measurement (M)							
• instrument of data collection (M1)	• standardized questionnaire	• standardized questionnaire	• standardized questionnaire	• standardized questionnaire	• standardized questionnaire	• standardized questionnaire	• standardized questionnaire
• method of data collection (M1, M2, M3, M4)	• face to face interviews	• face to face interviews	• face to face interviews	• face to face interviews	• face to face/web/post interviews	• face-to-face interviews	• face-to-face interviews
• method of case selection (M3, M4, M8)	• Nationally representative, random, clustered, stratified, multistage area probability sample (selection of units at different stages varies by country)	• random national probability samples (country-specific variations)	• Multistage random probability sampling	• Multistage probability sampling with national adaptations	• representative single stage or multistage random probability sampling	• multistage random sampling, in most countries combined with quotas sampling for sex and age (education) in the 3rd stage;	• full probability/combination of probability and stratified sampling
• level of data collection (M3, M4)	• individual level	• individual level	• individual level • EU citizens aged 15+	• individual level	• individual level • resident population aged op 18+	• individual level	• individual level
• units of observation (M3, M4)	• citizens aged 18+ (voting age)	• adult, voting-age population		• population aged 15+		• national pop.	• resident pop. aged op 18+

Note. Own compilation. n.a.: no information available.

1.2.2.2 The Question of Cross-National Comparability

As outlined in the previous section, ample international survey data is available for cross-national research on political trust (Cautrès, 2011; Heath, Martin, Spreckelsen, 2009; Norris, 2009). This "profusion of data" (Cautrès, 2011, p. 505) has given rise to the question of how political trust can be measured in a comparable manner across countries in order to facilitate generalizing statements. Researchers have compared the dimensionality of the measurement model of political trust in a number of exploratory, cross-national studies (see for example D. Braun, 2013; Hooghe & Kern, 2015; Lu, 2014; Marien, 2011b). Others have tested to what extent specific measurement models of political trust are invariant across countries (see for example André, 2014; Ariely, 2015; Schaap & Scheepers, 2014). Both groups of researchers have thus sought to provide insights into the valid cross-national measurement of political trust.

Despite these efforts, the question of the appropriate measurement model of political trust for comparative research remains unresolved. First of all, this uncertainty concerns the choice of survey items. To date, there is no agreement which political objects political trust refers to. Accordingly, as noted in the previous section, international survey projects have not implemented a common set of political trust items (O1 in Table 1.3). Likewise, the set of political institutions and authorities included in the operationalizations varies in studies who use this survey data to explore political trust' dimensionality and test its measurement invariance. Some use a small set of items such as trust in parliament, politicians, and political parties (Marien, 2011a). Others include a broader set of institutions such as government, parliament, political parties, civil service, the legal system, and the police (D. Braun, 2013; Fuchs et al., 2002; K. Newton & Zmerli, 2011). Still others go beyond the state level and include trust in regional and local government institutions (I. Schneider, 2017) or EU parliament and the UN (André, 2014). Thus, previous studies provide no guidance as to the appropriate set of items for cross-national research on political trust.

Second, previous analyses addressing the measurement invariance of political trust are inconclusive. They support measurement models whose number of dimensions differs (O5 in Table 1.3). While some indicate that political trust is a single-dimensional concept (André, 2014; Ariely, 2015; Coromina & Davidov, 2013; Marien, 2011a), others provide evidence that a two-dimensional model fits the data best (Schaap & Scheepers, 2014; I. Schneider, 2017). What is more, depending on the items and countries included in the

analysis, the levels of measurement invariance reached by these models vary, suggesting different degrees of comparability across countries (see section 1.1.2.6). It thus remains unresolved whether political trust contains the same number of dimensions across countries, if so, which political objects these dimensions comprise and to what extent its measurements can be compared in a valid manner.

Third, the range of countries included in previous studies testing the measurement invariance of political trust has been limited. Most analyses have focused on Europe (see for example André, 2014; Marien, 2011a, 2017; Schaap & Scheepers, 2014). I. Schneider's (2017) study also includes former Soviet countries. The results of these studies cannot unquestioningly be extended to democracies in Asia, Africa, and Latin America, however. As Przeworski and Teune (1970) caution: "Inferences leading to measurement statements must be validated in each social system" (p. 106).

The measurement validity of political trust could be improved for comparative research by testing its measurement invariance while taking into account previous studies' contradictory results and limitations. Article 3 'Testing the measurement invariance of political trust across the globe. A multiple group confirmatory factor analysis' (Breustedt, 2018) fills this research gap by answering the following research question: To what extent can the measurement invariance of political trust be established across the globe and if so, on the basis of which measurement model?

The dissertation's third article uses construct validation and measurement invariance testing (see section 1.1.2.5). The state of debate outlined above indicates that researchers favor different nomological nets of political trust. They have divergent theoretical assumptions about the dimensionality of the systematized concept and the respective measurement models as well as varying notions about the set of items to use to measure political trust (O1 and O5 in Table 1.3). In line with critical rationalism's assumption that researchers cannot establish the 'true' meaning of a concept through theoretical reasoning (see section 1.1.3), the article uses construct validation to test which of the alternative models suggested in the literature fit the data best. In addition, it applies this validation strategy to determine how the observable indicators relate to one another and whether all of them relate equally well to the latent variables representing the conceptual dimensions of political trust in the measurement model. Furthermore, the article uses measurement invariance testing in order to establish to what extent political trust can be compared across countries (see section 1.1.2.6). The article tests the respective hypotheses by

conducting country-specific confirmatory factor analyses as well as multigroup confirmatory factor analyses using data on a set of six items from the World Values Survey (wave 6, 2010-2014) for 32 electoral and liberal democracies.

In doing so, it addresses previous studies' limitations and contradictory results in three respects. First, it draws on a large set of items to measure political trust. Second, it considers three common systematized concepts of political trust currently being discussed in the literature. Third, it extends the analysis by including non-Western democracies.

The article proceeds as follows. On the basis of a review of the three most common systematized concepts of political trust, the article first seeks to determine which of the respective measurement models is supported by the data. In order to do so, it tests the three measurement models separately in each country by means of confirmatory factor analysis. The results show that a two-dimensional model of trust in implementing and representative political institutions fits the data best. This model is therefore chosen as the so-called 'baseline model', that is, the model that is tested for measurement invariance in the subsequent analytical step.

Having established the baseline model, the article then performs several multigroup confirmatory factor analyses. The first multigroup confirmatory factor analysis serves to determine to what extent the baseline model reaches measurement invariance in the 32 democracies included in the study. The results indicate that the measurement model is not comparable as such across all countries. On the basis of the results of the country-specific confirmatory factor analyses and multigroup confirmatory factor analysis, the article therefore goes on to adapt the baseline model and tests the altered versions in subsequent multigroup confirmatory factor analyses. Different levels of measurement invariance are reached when excluding trust in civil service, adding a correlated error term between the item 'trust in parliament' and 'trust in political parties' and limiting the set of countries.[36] The revised model reached configural invariance in 19 democracies and was fully invariant in three post-communist countries.[37] This provides reason to infer that people have the same overall concept in mind when responding to political trust items. Apparently, however, only the survey respondents in the post-communist countries used

[36] Taking account of the ordinal measurement scale of the items, the article only tests two levels of measurement invariance, namely configural and full invariance (Davidov, Datler, Schmidt, & Schwartz, 2011, pp. 160–161).

[37] Poland, Romania, and Slovenia.

the response scale in the same manner when indicating whether they have a great deal, quite a lot, not very much or no trust at all.

On the basis of these findings, the article makes three recommendations to improve the measurement validity of quantitative empirical assessments of political trust for future comparative research. First, the results make a strong case for a two-dimensional model of political trust (O5 in Table 1.3). Researchers should take this into account when seeking to enhance the systematized concept of political trust or aiming to determine its levels, causes, and effects. Second, the analysis underlines that social scientists should test the measurement invariance of their measurement model of political trust prior to conducting comparative studies (row M in Table 1.3). Since its comparability depends on the set of countries included in the study, the measurement model should be adapted accordingly. Third, future analyses should be mindful of the fact that the item 'trust in civil service' seems to measure the concept of political trust less well than the other items (O1 in Table 1.3).

1.2.2.3 Current Approaches to Aggregating Political Support Data

As outlined above, researchers have scrutinized the validity of the operationalization and measurement of political support for comparative research in a number of respects. The validity of the aggregation stage has received considerable less attention, however. Based on Table 1.1, this section provides a brief overview of current approaches to aggregating political support data. Individuals are the most common units of analysis in comparative studies of the levels, causes, and effects of political support (Fuchs, 2007, p. 172). The use of countries as units of analysis is less common and is largely restricted to studies that address the levels of political support as well as its effects on countries' degree of democracy or the persistence of democracies (A3, A4, A8; for the former, see for example Cho, 2014; Martini & Quaranta, 2020; for the latter, see for example Fails & Pierce, 2010; Norris, 2011). Thus, depending on the research question, the level of aggregation is either the individual or the country level (A3, A4). When studying political support, researchers aggregate the data on its individual dimensions to a different extent (A1). They frequently choose to summarize the data on single items at the country level (see for example Erlingsson, Linde, & Öhrvall, 2016; Raschke & Westle, 2018). Another common approach is to create separate indices for the individual dimensions of political support on the basis of several survey items (see for example

Chu, Welsh, & Chang, 2013; Citrin, Levy, & Wright, 2014). It appears to be uncommon for researchers to calculate a single overall measure of political support.[38] Frequently applied methods of aggregation include summing up the scores of the individual items (see for example Catterberg & Moreno, 2006; Hutchison & Johnson, 2011), taking their average (see for example Anderson & Singer, 2008; Chang & Chu, 2006; Wong, Wan, Hsiao, 2011) or using the proportion of those who agree with a certain statement (see for example Inglehart & Welzel, 2005) (A2, A5, A6, A7). In doing so, the most common aggregation rule is to weight the item responses equally (A2, A6, A7). Overall, current approaches to aggregating data on political support vary depending on the phenomenon of interest to the researcher.

1.2.2.4 The Issue of Aggregation

Researchers have advanced several reasons as to why aggregated scores based on predominant aggregation procedures such as the sum score or the mean may not be valid. These relate to conceptual considerations in Easton's political systems theory. Here, Easton (1965) noted that political support at the macro level is "a function not only of actions or intensities of feelings, pro or con, but of the number of members who hold these feelings" (p. 165) as well as their political relevance and effectiveness (Easton, 1965, pp. 154, 165–166; see also Verba, 1980, p. 404). First, social scientists have criticized that, in current studies of political support, citizens have an equal chance to be included in the sample and their responses receive equal weight in the aggregation process. They have questioned the conceptual appropriateness of this approach as the levels of political support of certain groups of citizens may actually be more relevant than others' (Scheuch 1968, pp. 188–189, 197; Verba, 1980, p. 404). Second, researchers have criticized that current aggregation procedures do not sufficiently account for within-country heterogeneity of political support. They argue that certain groups of citizens – such as the political elite – may differ systematically in their kind and levels of political support (Kaase, 1983, p. 161; S. Pickel & G. Pickel, 2006, p. 44; Reisinger, 1995, pp. 333, 336–337).[39]

[38] This is not surprising as the concept encompasses a variety of attitudes and values whose consequences for the political system are assumed to vary. Changes in people's support for the political authorities are thought to affect their prospects of reelection; a decline in support of the institutional structure of a democratic regime is said to challenge the persistence of the type of democracy in a given country; decreasing levels of commitment to democratic values are assumed to threaten the persistence of democracy as such (Easton, 1975, p. 437; Fuchs, 2007, p. 166).

[39] In addition, some researchers reject the aggregation of individual-level data on political support to the

Paying heed to these points of criticism would require several adaptations in the measurement process. In the conceptualization stage it would necessitate a more detailed specification of the population to which the systematized concept applies (see C3 in Table 1.1). In the measurement and aggregation stage, it would entail an adaption of the units of observation, an adjustment of the method of case selection in terms of sampling as well as a reconsideration of the level of aggregation (C3, M3, and A3 in Table 1.1). In spite of this, the validity of the predominant aggregation procedures has not been evaluated empirically and little has been suggested on how to improve them (Fuchs, 2007, p. 173). Easton (1965) himself concludes that the sampling and weighting scheme he proposes "is a large order, one that would require considerable ingenuity to execute adequately" (p. 169).[40]

Given these challenges, it is not surprising, therefore, that these issues have not been resolved so far in terms of changes in the measurement process. Instead, proponents of cross-national research on political support have resorted to argumentation. They have suggested that the above-mentioned concerns about current aggregation procedures can be sidestepped by adding a premise to political support's nomological net. Accordingly, current aggregation procedures based on data collected through representative random sampling are acceptable for studies of political support in democratic political regimes because here, the value of the individual "is legally codified by the constitutions (...) and is expressed in the equal weight of each citizen in the political system" (Fuchs, 2007, p. 173; see also Kaase, 1983, p. 155).[41]

Since researchers' efforts to improve the aggregation stage have been limited to these conceptual considerations, the debate on the measurement validity of quantitative

---

macro level outright. These critics argue that, by aggregating the data, researchers commit an individualistic fallacy because "the collectivity is assumed to have the properties of the individuals that comprise it" (Peters, 1998, p. 44; see also Pye, 1972, pp. 291–292). Proponents of the aggregation procedure have countered that this criticism is unwarranted because it is based on a false conception of the individualistic fallacy (Inglehart & Welzel, 2005, pp. 231–233; S. Pickel & G. Pickel, 2006, p. 41).

[40] "In summary, (...) [i]f we are to speak inteIligibly about some minimal support, even though we cannot identify what this minimum may be except to recognize its presence, it is essential to know what kinds of variables would need to be taken into account to obtain an ordinal measure of support. We have seen that we would have to balance the number of members supporting and opposing a system, their power position, the intensity of their feelings in these respects, their capacity to express these feelings in action, and their readiness to do so under the circumstances. Finally, any summation of support would also weigh simultaneous positive and negative feelings and actions entertained by a member or group in order to ascertain the net consequences with respect to the level of support being extended to each of the political objects" (Easton, 1965, p. 169).

[41] Fuchs (2007, pp. 173–174) points out that this premise may not be applicable to more communitarian societies outside the Western realm, however. In Asian democracies, for example, the importance of the community and higher-ranking individuals may have to be taken into account.

empirical assessments of political support would benefit from an empirical validation of predominant aggregation procedures using cross-national survey data. Article 4 'Surpassing simple aggregation. Advanced strategies for analyzing contextual-level outcomes in multilevel models' undertakes this task by answering the following research question: How do predominant aggregation procedures perform compared to more advanced analytical strategies when studying the relationship between citizens' support for democratic values and democracies' persistence?[42]

The dissertation's fourth article compares a common aggregation procedure – the group means approach – with two advanced strategies to handling individual-level data in multilevel analyses with a dependent variable at the macro level – multilevel structural equation modeling, and the two-step approach. In order to do so, it tests a conceptually well-established hypothesis within the nomological net of political support: the assumption that citizens' support for democratic values has a positive impact on democracies' persistence (Easton, 1975, p. 445; Fuchs, 2007, p. 166; S. Pickel & G. Pickel, 2006, pp. 80–81). The analysis draws on data from the World Values Survey and the Quality of Government project for 98 countries between 1946 and 2014.

To begin with, the article outlines how well, in principle, each of the three strategies translates theoretical multilevel models with a dependent variable at the macro level into empirical measurement models. That is to say, it describes how each method handles data on macro- and individual-level determinants of macro-level dependent variables and how it relates them to one another. With regard to the group means approach, it points out that the reliability of the arithmetic mean of political support at the country level is affected by the variance between individuals' level of political support as well as the sample size (Lüdtke et al., 2008, p. 207). The greater the individual-level variance and the smaller the sample, the greater the expected bias of the regression estimates describing the relationship between political support and democracies' persistence at the macro level (Croon & van Veldhoven, 2007, pp. 49–50, pp. 52-54; Lüdtke et al., 2008, pp. 207–209). This in turn affects the validity of researchers' causal inferences based on the estimates.

In comparison, the estimates provided by multilevel structural equation modeling and the two-step approach are more valid as they utilize the information provided by the survey

[42] In article four, the author contributed large parts of the theoretical background of the paper (see section 5.3.1 'Substantive application. A multilevel explanation of the persistence of democracy'), a comparison of the advantages and disadvantages of the analytical strategies used (see section 5.2.2 on the 'Three analytical strategies'), the introduction as well as the summary of the period of analysis and data in the research design section (section 5.3.2).

data in a different manner. The former takes account of the heterogeneity within each group by treating political support as a macro-level latent variable (Lüdtke et al., 2008, pp. 209–210). The latter avoids validity issues related to aggregation by splitting the analysis in two consecutive steps. In a first step, the explanatory variable of interest (political support) is regressed on all other independent variables at the macro and individual level; in a second step, the estimated macro-level variance of the explanatory variable of interest (net of the effect of all other independent variables) is regressed on the dependent variable of interest (democratic persistence) at the macro level (Griffin, 1997).

The article continues by studying the relationship between citizens' support for democratic values and democracies' persistence empirically. It runs several models including and excluding additional control variables on the basis of the three methods of aggregation in combination with event history modeling.[43] The results show that the strategies do indeed perform differently. When applying the group means approach, support for democratic values had no significant effect on democracies' persistence. By contrast, the effect was significant in several models when using multilevel structural equation modeling and the two-step approach. In line with Easton's hypothesis, support for democratic values was negatively associated with democracies' breakdown. In addition, the confidence intervals of the estimates were smaller when applying the two more advanced strategies, indicating more precise estimates.

The article concludes that the group means approach performs less well compared to more advanced analytical methods. The results corroborate earlier simulation studies that indicated that researchers run the risk of committing type-two errors when applying the group means approach (Lüdtke et al., 2008). With regard to substantive applications, they may thus fail to reject a false null hypothesis, thereby drawing wrong inferences such as conclusions about the relationship between citizens' political support and democratic persistence.

Even though the article focuses on the validity of causal inferences, the comparison can also be used to draw conclusions how well the rules of aggregation inherent in the group means approach take into account the units of analysis (A3 in Table 1.1). In line with the

[43] Besides support for democratic values (World Values Survey Association, 2015) and countries' POLITY score (Marshall, Gurr, & Jaggers, 2015), the analysis included several control variables, namely economic development, ethnic heterogeneity at the macro level and citizens' age, highest educational level attained as well as subjective social class at the individual level (Teorell et al., 2016).

principles of construct validation (see section 1.1.2.5), one way to validate an aggregation strategy is to compare the results of empirical analyses that test the same, well-established hypothesis but use different methods of aggregation. If the analysis using the aggregation strategy in question yields results that do *not* support the hypothesis and analyses using alternative methods of aggregation that are assumed to be more valid provide results that corroborate the hypothesis, this suggests that the aggregation strategy under discussion is indeed less valid (Carmines & Zeller, 1979, p. 27).[44] This proved to be the case in the article's analysis. The results of the article therefore advise researchers to reconsider applying the group means approach as an aggregation procedure in multilevel studies of political support and should opt for alternatives such as multilevel structural equation modeling or the two-step approach instead.

1.3 Contributions and Limitations of the Articles

As demonstrated by the dissertation at hand, the measurement validity of quantitative empirical assessments of democracy is in need of improvement. Unless these shortcomings are addressed, researchers' inferences about democracies' state of affairs, cross-country similarities and differences, as well as their origins and consequences may be incorrect (Carmines & Zeller, 1979, p. 11; Döring & Bortz, 2016, p. 98). This in turn may misinform the public, politicians, and civil society, causing them to draw wrong conclusions about the actions needed to maintain and improve democracy.

The four articles included in the dissertation thus sought to answer the following research question: How can researchers improve the measurement validity of quantitative empirical assessments of the quality of democracy and political support for comparative research? Based on a common theoretical foundation in the critical rationalist tradition, each article developed several recommendations. While the articles do not provide indisputable, final solutions to the validity problems, they help to advance the field by attending to several key issues. The following section serves to summarize these contributions and limitations as well as ensuing research questions.

The first and second article address fundamental issues regarding the measurement

[44] Provided that the concept is operationalized and measured in a valid manner and that the analysis is not affected by other potential sources of bias (see section 1.1.2).

validity of quantitative empirical assessments of the quality of democracy. The dissertation's first article provides a comprehensive, comparative review of existing measurement instruments' quality by developing and applying a standardized framework that takes into account all stages of the measurement process. The framework accommodates both the diverse challenges involved at the individual stages as well as the heterogeneous conceptualizations of the quality of democracy that characterize the current state of research. The review pinpoints and compares strengths and weaknesses of current measurement instruments. It provides logical arguments supported by empirical evidence on the state of measurement validity of the quality of democracy.

The dissertation's second article consists of a systematic, cross-national comparison of survey-based assessments of the quality of democracy with measurements predominantly based on expert judgments and official statistics. This comparison sheds light on their similarities and differences. It shows that subjective assessments of the quality of democracy provide a complementary perspective to macro-level measurements. This insight adds empirical facts to an as yet theoretical line of reasoning as to why citizens' perspective should be incorporated in empirical analyses in order to improve their operationalization and measurement of the quality of democracy.

Building on previous efforts, the third and fourth article focus on refining specific aspects regarding the measurement validity of quantitative empirical assessments of individual dimensions of political support. The dissertation's third article revisits the question of cross-national measurement validity of political trust by testing the comparability of three competing measurement models in a systematic, global analysis. Its results support researchers who conceptualize political trust in terms of a two-dimensional model. At the same time, the article cautions that the measurement invariance of the measurement model and its items should always be established prior to conducting comparative analyses. This particularly applies to the item 'trust in civil service'. The article therefore provides empirical evidence on how to enhance the operationalization and measurement of political trust for comparative research.

The dissertation's fourth article follows up on criticism of the current approach to aggregating political support data to the country level by comparing it with two advanced analytical strategies. It demonstrates why researchers should reconsider current standard aggregation procedures in multilevel studies of political support and why they should opt for alternatives such as multilevel structural equation modeling or the two-step approach

instead. The analysis thus contributes empirical insights on how to improve the aggregation of political support data.

Each of the four articles thus contributes to answering the research question by addressing a specific issue and providing instructive suggestions on how to improve the stages of the measurement process concerned. Table 1.4 summarizes these recommendations. Jointly, they help to improve the measurement validity of quantitative empirical assessments of democracy for future comparative research.

The dissertation not only provides recommendations on how to enhance the measurement validity of quantitative empirical assessments of democracy – it also contributes to advancing validation efforts in comparative social science at large. First, the dissertation demonstrates different approaches to dealing with competing systematized concepts when assessing measurement validity – be it by applying a systematized framework that is applicable independent of the systematized concept (article 1), by taking into account different systematized concepts (article 3) or variations of a systematized concept (article 2). Second, the dissertation illustrates how future efforts to validate the comparability of cross-national survey data can accommodate to countries' contextual specificities in the operationalization stage (see section 1.1.2.6): While the original selection of items for the analysis should be based on a systematized concept's theoretical model that is applied to all cases, the set of items can be adapted for each country based on the country-specific model fit of the measurement model (articles 2 and 3). Third, the dissertation provides guidance for a more comprehensive validation of the aggregation stage. As outlined in the dissertation's theoretical foundation (see section 1.1.2), the aggregation stage should take into account all aspects of the systematized concept (see A1 to A8 in Table 1.1). Beyond these aspects, ideally, it should also accommodate for the characteristics of the multivariate model that is to be tested (article 4). Thus, applying the articles' insights and recommendations helps to improve the measurement validity of quantitative empirical assessments of democracy for comparative research as well as future validation efforts of similarly challenging concepts.

Table 1.4
Recommendations on How to Improve Quantitative Empirical Assessments of Democracy for Comparative Research

Assessment of democracy concept		Recommendations		
		Operationalization	Measurement	Aggregation
Quality of democracy	Article 1[a]	O1: coverage of content of individual dimensions • mind match between indicators' content and conceptual attributes • ensure empirical applicability • avoid redundancies and conflation	M3, M4, M8: method of case selection • observe match with conceptual range • limit assessments of the quality of democracy to full democracies	
		include citizens' perspective in future empirical studies of the quality of democracy		
	Article 2[b]	O1: coverage of content of individual dimensions • include additional aspects deemed relevant by citizens O3: content of indicators • citizens' assessments	M3: units of observation • use survey data more extensively to assess the quality of democracy	
Political support	Article 3[c]	O1: coverage of content of dimensions at large • trust in civil service measures political trust less well than other items O5: measurement model's dimensionality • political trust is two-dimensional (trust in implementing and representative institutions)	measurement invariance should be tested and measurement models adjusted depending on the cases selected	
	Article 4[d]			use of alternative methods of analysis to avoid inferential validity issues resulting from the use of aggregation methods such as the group means approach

Note. [a] S. Pickel, Stark, & Breustedt, 2015; [b] S. Pickel, Breustedt, & Smolka, 2016; [c] Breustedt, 2018; [d] Becker, Breustedt, & Zuber, 2018.

The articles' recommendations do not constitute ultimate solutions to the measurement validity issues addressed in this dissertation, however. Researchers may question the conceptual foundation or the research design of the validation efforts. This may challenge the articles' conclusions regarding the extent of measurement validity of current measurement instruments of the quality of democracy and political support and the resulting recommendations on how to improve them. First and foremost, such reservations may relate to the conceptualizations. According to the dissertation's theoretical foundation, assessments of measurement instruments' validity require a conceptual template. Because of the variety of systematized concepts of the quality of democracy and political support, there is no common theoretical foundation against which to evaluate whether measurement instruments' indicators reflect the concept in a valid manner, however. Researchers who favor systematized concepts other than those applied in the articles may thus have reservations about their inferences. While the articles sought to consider the most prominent systematized concepts in the field, the inability of researchers to establish the 'true' meaning of a concept precludes a final judgement as to the validity of current measurement instruments of the quality of democracy and political support in this respect (see section 1.1.3).

In addition, researchers may hesitate to implement the recommendations for reasons relating to the articles' research designs. The generalizability of the conclusions may be questioned because the analyses (like all analyses) were limited in terms of the chosen time period, cases, and the data available (Lucas, 2003, p. 239). Critics may object that the validation framework failed to include certain aspects that they consider to be relevant (article 1) (Bevir & Kedar, 2008; Gerring, 1999); they may criticize the use of survey data on the grounds that (non-expert) individuals are incapable of assessing democracies effectively (articles 2 and 3) (Hardin, 2000, pp. 32–35; Roberts, 2010, pp. 7–8; Skaaning, 2018, p. 111); or they may find fault with the chosen methods of analysis (articles 3 and 4) (Axelrod, 1997; Welzel & Inglehart, 2016). These points of criticism certainly deserve consideration. Because of the 'problem of correspondence' and the 'problem of the empirical basis' (see section 1.1.3), there is no ultimate research design to validate measurement instruments of the quality of democracy and political support, however.

Researchers' doubts regarding the research design may also concern the chosen validation strategies. The articles relied heavily on variants of construct validation as one

of the main procedures in social scientific research that serves to assess the extent to which a measurement instrument reflects the systematized concept it is intended to measure (see section 1.1.2.5). First, critics rightly note that construct validation requires a well-established nomological net, in other words, a tried and tested theoretical foundation (Anastasi, 1986, p. 6; Hartig et al., 2008, pp. 146–148). Second, as Cronbach and Meehl (1955) emphasized in their groundbreaking article on construct validation: "Unless substantially the same nomological net is accepted by the several users of the construct, public validation is impossible" (p. 291). Most social scientific concepts, including the quality of democracy and political support, fall short of these two prerequisites, however. Third, and most crucially, the results of construct validation procedures allow researchers substantial leeway to interpret negative evidence in terms of their implications for the concept's nomological net. A straightforward conclusion is that negative test results indicate that the measurement instrument needs improvement as it does not measure the systematized concept in a valid manner in general, or is not equivalent across countries. Another interpretation of the results is that the empirical hypothesis that was tested is incorrect, suggesting that this part of the nomological net should be changed. Still another explanation for negative test results is that the method used to test the empirical hypotheses is inaccurate or inappropriate. A final conclusion is that the measurement process applied to measure another latent variable that was included in the hypothesis that was tested is invalid (Carmines & Zeller, 1979, pp. 24–25; Cronbach & Meehl, 1955, p. 295). This array of possible interpretations shows that "construct validation is not only continuous (a matter of degree, not a categorical distinction between valid and invalid) but continual (a perpetual, self-refining process)" (Westen & Rosenthal, 2003, p. 609).

Leaving these limitations aside, the dissertation indicates several potential tasks for future research. First, comparative politics' age-old question of how to attain comparability in the face of heterogeneity remains pertinent. Adaptations at the aggregation stage of current measurement instruments contain considerable (hitherto unused) potential in this respect. Future research should consider which adaptations might allow for country-specific variations of what is considered to be the optimal balance between the different dimensions of the quality of democracy. Adaptations at the aggregation stage could also be a helpful means to take into account local, regional, or group-specific patterns of political support when studying its causes and effects.

Second, democracy researchers should debate how different kinds of data can be

combined sensibly. Each data type has its advantages and disadvantages when assessing democracies (Skaaning, 2018). The question is whether, and if so, when it makes sense to combine objective assessments based on administrative data and expert judgments, as well as subjective assessments based on survey data into an index – especially if the assessments that result from these different kinds of data do not coincide.

Third, comparative social scientists could draw lessons from the dissertation's articles for teaching comparative politics in higher education. As emphasized by the dissertation's foundation and reiterated in each of the four articles, valid measurement is founded on good conceptualization (see section 1.1.2.1). Conceptualization and the consideration of its many elements in the measurement process should receive greater attention in students' methodological training. This would add substance to the curriculum on quantitative empirical methods of data collection and analysis. Furthermore, by adding additional focus on teaching the ability to critically reflect on measurement and its validity, comparative social scientists could help students to distinguish the manipulative use of data in so-called 'fake news' from trustworthy facts in the media. Enhancing students' data literacy in this respect would empower them to reach their own, well-founded conclusions on the state of democracy in times of crises and beyond.

1.4 Appendix A

Table A1
Conceptualization and Operationalization of Political Support by Norris (2011) and S. Pickel (2013)

		Norris (2011)	S. Pickel (2013)	
Type of support	Levels of support	Survey measures and operational indicators	Attitudinal level	Operationalization
diffuse	support for the nation-state	feelings of national pride, such as in national achievements in the arts, sports, or the economy, feelings of national identity, and willingness to fight for country	--	--
	support for regime-principles	adherence to democratic values and principles, such as the importance of democracy, respect for human rights, separation of religious and state authorities, and rejection of autocratic principles	legitimacy	appropriateness of democracy for own society
	evaluations of regime performance	judgments about the workings of the regime, including satisfaction with the democratic performance of governments, and approval of decision-making processes, public policies, and policy outcomes within each nation-state	system support	internal efficacy
	confidence in regime institutions	confidence and trust in public sector institutions at national, regional, and local levels within each nation-state, including the legislature, executive, and civil service, the judiciary and courts, the security forces, and political parties	trust	trust in government
specific	approval of incumbent office-holders	approval of specific incumbents including popular support of individual presidents and prime ministers, ministers, opposition party leaders, and elected representatives	performance evaluations	satisfaction with government

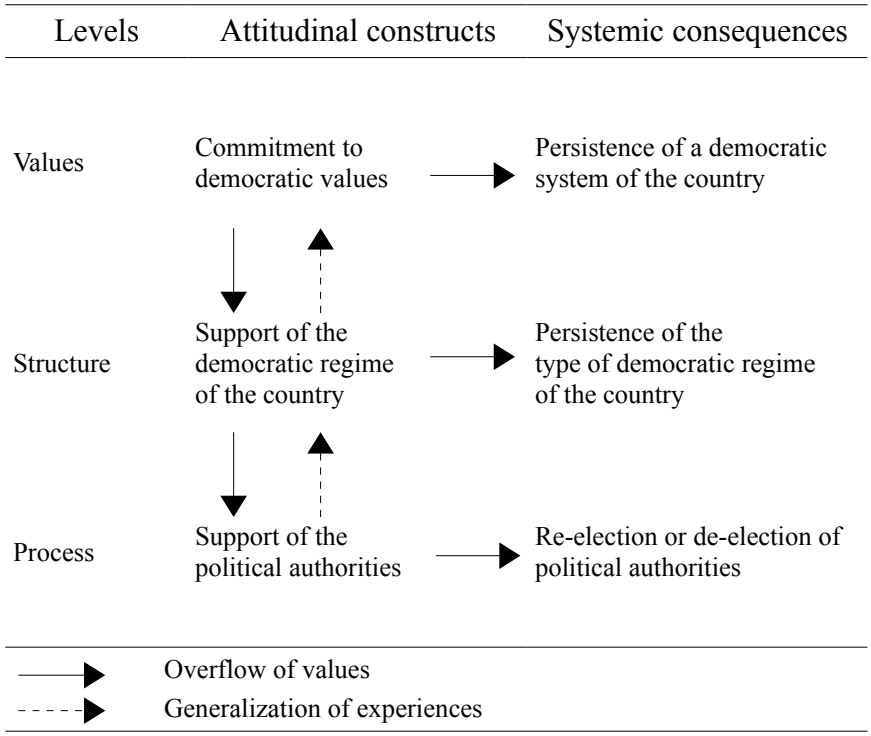
Note. Own compilation based on Norris (2011, p. 44); S. Pickel (2013, pp. 163–166; own translation).

Table A2
Conceptualization and Operationalization of Political Support by Dalton (2004)

Level of analysis	Evaluations (specific)	Affective orientations (diffuse)
	Operationalization	
community	best nation to live	national pride, sense of national identity
regime: principles	democracy best form of government	democratic values
regime: norms and procedures	evaluations of rights, satisfaction with democratic process	political rights, system norms, participatory norms
regime: political institutions	performance judgments, output expectations,	trust institutions, support party government
authorities	candidate evaluations, voting support,	trust politicians in general, identify with party

Note. Dalton (2004, pp. 23–24).

Table A3
Conceptualization of Political Support by Fuchs (2007)



Note. Fuchs (2007, p. 166).

2 Article 1 "Assessing the Quality of Quality Measures of Democracy: A Theoretical Framework and its Empirical Application"[45]

Abstract

Over recent decades, comparative political scientists have developed new measures at a rate of knots that evaluate the quality of democratic regimes. These indices have been broadly applied to assess the quality of democracy cross-nationally and to test the generalizability of theories regarding its causes and effects. However, the validity of these inferences is jeopardized by the fact that the quality of democracy is an abstract and contested concept. In order to address this eventuality, researchers constructing indices measuring the quality of democracy as well as researchers applying these indices should critically examine the quality of the indices. Owing to the absence of a standardized framework that is both suitable for the evaluation of contested concepts and that includes explicit coding rules so as to be directly applicable, this article seeks to fill this gap. The application of our framework is demonstrated by an evaluation of the Sustainable Governance Indicators, the Global Democracy Ranking and the Democracy Barometer. As indicated by our evaluation, the framework is a practical tool that helps to assess the conceptual foundation, validity, reliability, and replicability of indices. In addition, it can be used to study the quality of indices in a comparable manner.

Keywords: quality of democracy; validity; reliability; replicability; index; conceptualization

[45] Published as: Pickel, S., Stark, T., & Breustedt, Wiebke (2015). Assessing the quality of quality measures of democracy: A theoretical framework and its empirical application. *European Political Science 14*(4), 496–520. doi:10.1057/eps.2015.61

New measures of the quality of democracy in comparative political science surpass commonly applied indices such as Freedom House and Polity IV in that they allow researchers to study the nuances of the quality of democratic regimes. Indices such as the Sustainable Governance Indicators (SGI), the Democracy Barometer (DB) and the Global Democracy Ranking (GDR) have been applied broadly to assess the quality of democracy cross-nationally and to test the generalizability of theories regarding its causes and effects. However, the validity of these inferences is jeopardized by the fact that the quality of democracy is an abstract and contested concept. In addition, it is not directly observable. Consequently, its definition and measurement are susceptible to bias (Hopkin, 2010, p. 299; Jackman, 2008, pp. 121–122).[46]

In order to address this eventuality, we recommend that researchers constructing indices measuring the quality of democracy as well as researchers applying these indices should critically examine the quality of the indices. Such a critical examination should meet three criteria: first, it should be based on a common framework in order to be able to make comparable judgements; second, it should encompass all of the steps involved in applying abstract theoretical concepts to empirical phenomena, that is, conceptualization, operationalization, measurement, and aggregation; third, and consequently, the framework should focus on methodological criteria that can be applied to all indices while allowing for sufficient flexibility given different conceptualizations of the quality of democracy, namely validity, reliability, and replicability as well as the transparency and adequacy of the concept (Müller & S. Pickel, 2007, p. 517; Munck & Verkuilen, 2002, p. 8).

Owing to the absence of a standardized framework that is both suitable for the evaluation of contested concepts such as the quality of democracy and that include explicit coding rules, this article seeks to fill this gap. Our framework combines several influential and related efforts in the field: the standard for measurement validity as developed by Adcock and Collier (2001), the framework for the assessment of data measuring democracy provided by Munck and Verkuilen (2002) and the operationalization of Munck and Verkuilen's (2002) framework by Müller and S. Pickel (2007). We complement these

[46] We define bias as a 'systematic and culpable error' (Hammersley & Gomm, 1997, para. 4.13). While bias can affect the quality of research at all stages of the research design (for an overview see Indrayan, n.d.), given the task at hand, we focus on how to prevent biases arising from the perspective of the researcher of the concept as well as the measurement of the concept (Gujarati, 2004, pp. 65–67; Moses, 2011, pp. 798–801).

standards with Gerring's (1999, 2001) criteria for good conceptualization as well as established procedures concerning data aggregation (OECD, 2008).

Given the fact that we have to agree to differ regarding the meaning of quality of democracy, the article starts out by summarising current conceptualizations of the quality of democracy. We then go on to present the quality assessment criteria included in the framework for evaluating the quality of measures of the quality of democracy. The application of the framework is demonstrated by evaluating the SGI, the GDR, and the DB. The article concludes by considering the implications of the evaluation for cross-national comparisons and by providing advice on how to ensure the quality of quality measures of democracy.

2.1 Current Conceptualizations of the Quality of Democracy

Researchers analysing the quality of democracy primarily refer to Dahl's (1970) criteria of democracy as outlined in his concept of polyarchy as the key components of democracy (Gastil, 1988; Marshall, Gurr, & Jaggers, 2014; Vanhanen, 1997). To conceptualize the quality of democracy, a number of researchers (e. g., Altman & Pérez-Liñán, 2002; Beetham, Carvalho, Landman, & Weir, 2008; Bühlmann, Merkel, Müller, & Wessels, 2012; Bühlmann, Merkel, Müller, Giebler, & Wessels, 2012; Diamond & Morlino, 2004b; Lauth, 2004, 2015; Levine & Molina, 2011c; Ringen, 2007;) have expanded Dahl's key components.

The resulting approaches vary in their conceptual complexity. They can be assigned to a continuum ranging from approaches that include merely procedural characteristics to concepts that comprise both procedural and substantive characteristics of the quality of democracy (Kneuer, 2011, pp. 134–135; Munck, 2016, pp. 3–13; Stoiber, 2011, pp. 173–184). The procedural approaches (e. g., Bühlmann, Merkel, Müller, & Wessels, 2012; Bühlmann, Merkel, Müller, Giebler, et al., 2012; Lauth, 2004, 2015) solely refer to democratic procedures and institutions that generate collectively binding decisions. In contrast, more extensive concepts of the quality of democracy (e. g., Diamond & Morlino, 2004b; Morlino, 2011) include the output or outcome of the political process (Fuchs & Roller, 2008, p. 91). In the following, the aforementioned conceptual approaches to the study of the quality of democracy will be exemplified briefly.

Lauth (2004, 2015) develops a procedural concept of the quality of democracy. According to him, a high quality democracy requires that political liberty, political equality, and control of political power are fulfilled.[47] On the basis of Lauth (2004), Bühlmann, Merkel, Müller, and Wessels' (2012) concept of the quality of democracy is by far the most extensive of all. Underlying the DB index, it distinguishes between freedom, equality, and control as the key dimensions of the quality of democracy. Individual liberties, rule of law, the public sphere, competition, mutual constraints of constitutional powers, governmental capability, transparency, participation, and representation further specify these dimensions. The level of the quality of democracy depends on the extent to which political regimes fulfill these nine functions (Bühlmann, Merkel, Müller, & Wessels, 2012, p. 523).

Diamond and Morlino (2004b) develop an extensive concept of the quality of democracy that includes both procedural and substantive characteristics. They argue that in order to study the quality of democracy, certain minimal standards of democracy have to be fulfilled: universal, adult suffrage; recurring, free, competitive, and fair elections; at least two (effectively) competing political parties as well as alternative sources of information. They emphasize that it makes little sense to assess the scope to which hybrid regimes, defective democracies, or electoral autocracies fulfill the criteria of 'good democracy' (Diamond & Morlino, 2004b, p. 3). Once political systems meet the minimal democratic standards, the extent of political and civil freedom, popular sovereignty and political equality "as well as broader standards of good governance (such as transparency, legality, and responsible rule)" (Diamond & Morlino, 2004b, p. 3) can be studied in detail.

Further developing this previous work, Morlino (2011) presents an enhanced theoretical conceptualization, which he refers to as 'quality democracy', that is, what constitutes a 'good democracy' (Morlino, 2011, p. 195). He promotes this concept as an analytic tool that "cover[s] the main empirical aspects that are consistent with at least all the main existing and important normative conceptions of democracy" (Morlino, 2011, p. 193), namely the "rule of law, electoral accountability, institutional accountability, participation […], competition, freedom, equality/solidarity, and responsiveness" (Morlino, 2011, p. 44). He then relates these eight criteria to a tripartite notion of quality derived from the industrial and marketing sectors, that is, quality in terms of procedure, content, and

[47] Lauth has developed an index that serves to measure the quality of political regimes, the so-called 'Combined Index of Democracy' (CID). Given the intention of this index, it goes beyond the scope of this article (Lauth & Kauff, 2012).

results (Morlino, 2011, pp. 194–195; see also Diamond & Morlino, 2004b, p. 4; Morlino, 2004, p. 11). The procedural dimension of the quality of democracy includes the rule of law, participation, competition as well as electoral and inter-institutional accountability. The substantive dimension of the quality of democracy comprises respect for civil and political freedoms and increasing political equality. The results dimension concerns the responsiveness of the political representatives, that is, the correspondence of public policies with citizens' demands and preferences (Morlino, 2011, p. 196; see also Diamond & Morlino, 2004b, p. 5).[48]

Overall, while researchers agree to a certain extent on what constitutes the core meaning of democracy, they apparently differ with regard to the concept of the quality of democracy. "Conceptualizations of the quality of democracy are still far from providing a well-founded and widely accepted basis for identifying a distinct subject matter" (Munck, 2016, p. 2). This leaves us with two conclusions. First, because of this apparent lack of agreement, the quality of quality measures of democracy has to be assessed in terms of the respective conceptualization of the quality of democracy underlying the individual index. Second, and qualifying this statement to some extent, the concept of quality of democracy cannot be defined arbitrarily. Given the conceptualizations outlined above, it should include the idea that the concept of the quality of democracy can only be applied to countries that meet procedural criteria to a predetermined extent.

2.2 The Quality Assessment Criteria

Our framework specifies common standards of evaluation within all of the relevant steps in the research process that precede and characterize "the assignment of numerals to objects or events according to rules" (Stevens, 1946, p. 677), namely the conceptualization, operationalization, and measurement (Mueller, 2004, pp. 161–162). Since the measures of the quality of democracy are usually aggregated into indices, we also consider the aggregation process. The quality criteria in the framework are specified as follows: transparent and adequate conceptualization of the quality of democracy; valid, reliable, and replicable operationalization and measurement of the concept and

[48] For a detailed explanation of the relationship between these dimensions see Morlino (2011, pp. 197–211) and Diamond and Morlino (2004b, pp. 7–29).

valid, reliable, and replicable aggregation of the data. Each of these criteria is specified further, yielding 20 quality criteria overall.

In each of the following sections, we define each quality criterion (QC) and explain how it can be used to assess the quality measures of democracy. We operationalize each of the 20 criteria, thereby delineating explicit coding rules for the evaluation of the indices (see Table 2.1). These coding rules are a revised version of Müller and S. Pickel's (2007) operationalizations.

The assessments according to the 20 criteria can be used to construct an index that permits a comparable evaluation of the indices measuring the quality of democracy (Müller & S. Pickel, 2007, p. 518). In order to be able to construct a comparable index while taking heed of the particular characteristics of each index measuring the quality of democracy, the quality criteria are operationalized using a three-point ordinal scale. The coding rules therefore provide a guideline for a qualitative assessment according to three categories: do the indices fulfill (+), partly fulfill (0) or do not fulfill (-) the respective criterion.[49] On the basis that assumption that each of the 20 quality criteria is equally important, the results of the assessment of each index are added up for each of the three categories. This shows how many of the quality criteria each index fulfills, partly fulfills or does not fulfill. This information is used to construct a tripartite quality index that allows researchers to compare the number of positive (+), intermediate (0) and negative evaluations (-) per index at a glance (Müller & S. Pickel, 2007, p. 518).

2.2.1 Conceptualization

Generally, in the social sciences, the meaning of concepts is susceptible to interpretation and change (Gerring, 1999, p. 359, 2001) that affects the validity of the measurement. The "specification of the meaning of the concept […] affects the entire process of data generation, given that it provides the anchor for all subsequent decisions" (Munck & Verkuilen, 2002, p. 7). Valid measurement of concepts therefore depends on the transparent and adequate conceptualization of the concept that is to be measured.

[49] Since some criteria are dichotomous by definition, they had to be scaled accordingly. Consequently, the dichotomous criteria receive a larger weight in the evaluation. This slightly biases the positive or negative tendency of the quality assessment (Müller & S. Pickel, 2007, p. 518). However, dichotomising all of the criteria would lead to an overly harsh evaluation of those criteria where the indices fulfill the standards to some extent.

Table 2.1
Quality Assessment Criteria: Coding Rules

Conceptualization				
Quality criteria	Coding rules			
	Indicator	POS (+)	INTERMED (0)	NEG (-)
1.1 Transparency	1.1.1 Transparency	Detailed visual or text-based outline of the characteristics of the quality of democracy at different levels of abstraction (‘concept tree’)	Insufficiently detailed visual or text-based outline of the characteristics of the quality of democracy at different levels of abstraction (‘concept tree’)	No visual or text-based outline of the characteristics of the quality of democracy at different levels of abstraction (‘concept tree’)
1.2 Concept specification	1.2.1 Differentiation	Discussion of the difference between the concept of quality of democracy and related concepts	Difference between the concept of quality of democracy and related concepts mentioned	No mention of the difference between quality of democracy and related concepts
	1.2.2 Range of political systems	Application of the concept of the quality of democracy is restricted to democracies	[---]	Application of the concept of the quality of democracy is not restricted to democracies
	1.2.3 Range of institutional settings	Attributes applicable in a comparable manner in different democratic institutional settings	At least one but less than half of the attributes not applicable in a comparable manner in different democratic institutional settings	Half or more of the attributes not applicable in a comparable manner in different democratic institutional settings
	1.2.4 Parsimony	Definition includes a sufficient number of attributes to define it and distinguish it from related concepts and does not include attributes that are irrelevant in this respect	Inclusion of one irrelevant attribute	Inclusion of more than one irrelevant attribute
1.3 Conceptual logic	1.3.1 Coherence	The dimensions of the concept are functionally or logically related	One of the dimensions of the concept is not functionally or logically related to the others	More than one dimension of the concept is not functionally or logically related to the others
	1.3.2 Conflation	All of the less abstract characteristics are properly assigned to the respective attribute at a higher level of abstraction that they are intended to specify	One of the less abstract characteristics is improperly assigned to an attribute at a higher level of abstraction	More than one of the less abstract characteristics is improperly assigned to an attribute at a higher level of abstraction
	1.3.3 Redundancy	The scope of all of the characteristics of the quality of democracy that are intended to specify an attribute at a higher level of abstraction does not overlap	The scope of one of the characteristics of the quality of democracy that is intended to specify an attribute at a higher level of abstraction overlaps	The scope of more than one of the characteristics of the quality of democracy that are intended to specify attributes at a higher level of abstraction overlaps

Table 2.1 (continued)

Operationalization & Measurement				
Quality criteria	Coding rules			
	Indicator	POS (+)	INTERMED (0)	NEG (-)
2.1 Validity	2.1.1 Selection of valid indicators	(1) The indicators cover all of the attributes of the concept (2) The indicators actually measure what they are intended to measure (3) The indicators are applicable (is the indicator actually an indicator?) (4) There are no redundancies or conflation (see 1.3.2, 1.3.3) (5) The indicators are equivalent cross-nationally	(1) The indicators cover all of the attributes of the concept AND (2) The indicators actually measure what they are intended to measure AND (3) the applicability of the indicators is problematic AND (4) there are some redundancies or conflation (see 1.3.2, 1.3.3) AND (5) The indicators are equivalent cross-nationally	(1) The indicators *do not* cover all of the attributes of the concept AND/OR (2) The indicators *do not* measure what they are intended to measure AND/OR (5) The indicators are *not* equivalent cross-nationally (negative evaluation independent of the quality of the indicators regarding (3) and (4))
	2.1.2 Range of sources of information	(1) multiple sources (2) sources from both state- and non-state actors (3) Information from sources with varying national or cultural background	(2) sources only from state- or non-state actors OR (3) Information from sources with a single national or cultural background (negative evaluation independent of the number of sources (1))	(2) sources only from state- or non-state actors AND (3) Information from sources with a single national or cultural background (negative evaluation independent of the number of sources (1))
	2.1.3 Theory-based justification of the measurement level	The relevance of the choice of measurement level is reflected upon and the choice is justified based on theoretical considerations	[---]	No justification of the measurement level
2.2 Reliability	2.2.1 Intercoder reliability	Assessment of intercoder reliability	[---]	No assessment of intercoder reliabilty
2.3 Replicability	2.3.1 Publication of coding rules	(1) Coding rules are published (2) Coding rules include a list of all indicators, their measurement level and sufficient information for the unambiguous replication of the coding process	(1) Coding rules are published (2) Coding rules do *not* include sufficient information for the unambiguous replication of the coding process	(1) Coding rules are not published (non-fulfillment of (2) follows from the non-fulfillment of (1))
	2.3.2 Publication of sources of information	Publication of the sources of information	[---]	No publication of the sources of information
	2.3.3 Publication of disaggregated data	Disaggregated data of all indicators at all time points are publicly available to the scientific community free of charge	Some data available free of charge	Disaggregated data are not publicly available

Table 2.1 (continued)

Aggregation				
Quality criteria	Coding rules			
	Indicator	POS (+)	INTERMED (0)	NEG (-)
3.1 Validity	3.1.1 Appropriate level of aggregation	The level of aggregation strikes a sensible balance between the comparability of the index (high level of aggregation) and its detailedness (low level of aggregation of the attributes)	[---]	The level of aggregation emphasizes either the comparability or tries to maintain a high level of detail
	3.1.2 Theory-based justification of the rules of aggregation	The relevance of the choice of the aggregation rule is reflected upon and the choice is justified based on theoretical considerations	The choice of the aggregation rule is only partly justified based on theoretical considerations (e.g. in the case of multilevel aggregation procedures, some steps are not justified)	No justification of the choice of the aggregation rule
	3.1.3 Theory-based justification of weights	Theory-based justification of the weighting process (also concerns the lack of weighting)	Logical and practical justification No theory-based justification	No justification of the weighting process
3.2 Reliability	3.2.1 Robustness tests	Robustness of the index is tested (effect of including/excluding an indicator on the results, the choice of weights etc.)	[---]	No robustness tests
3.3 Replicability	3.3.1 Publication of aggregation rules	(1) Rules of aggregation are published (2) Rules of aggregation are sufficiently detailed for the unambiguous replication of the aggregation process	(1) Rules of aggregation are published (2) Rules of aggregation are *not* sufficiently detailed for the unambiguous replication of the aggregation process	(1) Rules of aggregation are not published (non-fulfillment of (2) follows from the non-fulfillment of (1))

In the interest of scientific progress, researchers have to make their definitions of concepts transparent (QC 1.1, see Table 2.1). ‘Opaque concepts’ (Schedler, 2012b, p. 253) prevent other researchers from positioning themselves regarding the conceptualization. They may also result in a concept bias, that is, a “systematic error [in the conceptualization] that the researcher should have been able to recognize and minimize, as judged either by the researcher him or herself (in retrospect) or by others” (Hammersley & Gomm, 1997, para. 4.13). Transparent conceptualization is best achieved by developing a ‘concept tree’ (Munck & Verkuilen, 2002, p. 13). Researchers should define the dimensions of the concept of the quality of democracy, that is, their attributes, and spell out in detail the characteristics that define these dimensions, that is, the components of these attributes, at different levels of abstraction (QC 1.1.1, see Table 2.1).

Adequate conceptualization requires a point of reference. Since concepts can take on different meanings, adequacy cannot refer to the extent to which the conceptualization reflects the ‘true’ meaning of the concept. The question is whether we can ever hope to ever find the ‘true’ meaning of social scientific concepts or whether the meaning of a concept is necessarily always restricted to the range of possible meanings agreed upon within the scientific community. But in the case of the meaning of ‘quality of democracy’, even the current scientific community has not come to an agreement regarding the core meaning of the term. The only exception seems to be that researchers implicitly or explicitly conceptualize the quality of democracy to comprise the elements of procedural democracy and then some. Thus, there is no common normative standard of comparison. It is therefore futile to evaluate the indices in terms of their conceptual validity (Seawright & Collier, 2014, p. 115). Consequently, adequacy has to refer to the ‘systematized concept’, that is, “the specific formulation of a concept adopted by a particular researcher or group of researchers” (Adcock & Collier, 2001, p. 530), as opposed to the ‘background concept’ “which encompasses the constellation of potentially diverse meanings associated with a given concept” (Adcock & Collier, 2001, p. 530).

Even though there are no absolute criteria that determine the adequacy of the *meaning* of the systematized concept, it can nevertheless be evaluated in terms of methodological standards as to how well the term (the linguistic label describing the concept) is aligned with the phenomena it defines (the referents) and the attributes that define these

phenomena (the meanings) (Gerring, 2001, pp. 39–40; Sartori, 1984, pp. 22–23)[50] – the specification of the concept (QC 1.2, see Table 2.1) and the conceptual logic (QC 1.3, see Table 2.1) (Munck & Verkuilen, 2002, p. 8). Concept specification refers to the identification of the attributes of the concept and is evaluated in terms of the differentiation, contextual range, and parsimony of a concept (Gerring, 1999, pp. 371–373, pp. 375–379, 2001, pp. 54–58; Munck & Verkuilen, 2002, pp. 7–12). Conceptual logic concerns the horizontal and vertical logical relationship between these attributes and by level of abstraction. It can be assessed by considering the extent of coherence, conflation, and redundancy in the definition (Gerring, 1999, pp. 373–375; Müller & S. Pickel, 2007, p. 529; Munck & Verkuilen, 2002, p. 8).

In order to usefully contribute to the accumulation of knowledge regarding the specification of social scientific concepts, concepts should be based on a definition with sufficient external differentiation. "A concept's differentiation derives from the clarity of its borders within a field of similar terms. A poorly bounded concept has definitional borders which overlap neighbouring concepts" (Gerring, 1999, pp. 375–376). With regard to the concept of the quality of democracy, constructors of indices measuring the quality of democracy should specify and justify how the concept differs from related concepts such as democracy, democratisation, and good governance.[51] This is an important prerequisite for the operationalization of the concept. The quality criterion of differentiation (QC 1.2.1, see Table 2.1) is therefore assessed in terms of whether the researchers outline the attributes that define and distinguish the quality of democracy from other similar concepts (Gerring, 1999, pp. 377–378).

The adequacy of the conceptualization of the quality of democracy also depends on its contextual range. Contextual range refers to the meaningful application of concepts in different cultural and spatial contexts (Gerring, 2001, pp. 54–56). Following the principle of differentiation, the study of the concept of the quality of democracy should be limited in range to democratic political systems (QC 1.2.2, see Table 2.1). Since the concept of the quality of democracy is used in order to compare the units of analysis, the attributes

[50] We must concede that all evaluations are influenced by individual presuppositions, knowledge, and limitations. We suggest to make a virtue out of necessity, plead for transparency, and propose to leave it to the scientific community to make a judgment.

[51] Gerring (1999, pp. 370–384, 2001, pp. 50–60) includes additional criteria that we do not take into account because they pertain to the usefulness of the concept as such and how it is received by the scientific community. In our case, we take this as given.

of the concept should also have an equivalent meaning across the range of different democratic institutional settings (QC 1.2.3, see Table 2.1).

Adequately conceptualized concepts are also parsimonious (Munck & Verkuilen, 2002, p. 9) (QC 1.2.4, see Table 2.1). Conceptualizations should not include too many attributes as this limits the usefulness of a concept. However, too few attributes prevent the researcher from being able to discriminate between cases (Hopkin, 2010, p. 300). The "matrix of potential meanings [commonly associated] with the background concept" (Adcock & Collier, 2001, p. 532) provides a general guideline. Since there is no agreement among researchers regarding the core meaning of the quality of democracy, it is difficult to say what constitutes a 'parsimonious' conceptualization. We therefore use the criterion of differentiation as the baseline. conceptualization of the quality of democracy should include a sufficient number of attributes to define its meaning and distinguish it from the concept of democracy and good governance while excluding characteristics that are irrelevant in this respect.

Once the attributes that constitute a concept have been specified, researchers should assure the conceptual logic of these attributes. The logical structure of the concept, the 'concept tree', should reflect the horizontal and vertical relationship between the attributes and by level of abstraction (Munck & Verkuilen, 2002, pp. 12–13). At the first level of abstraction, all of the attributes that are said to define a concept should be logically or functionally related, that is, horizontally, the dimensions of the concept should be internally coherent (Gerring, 1999, pp. 373–375) (QC 1.3.1, see Table 2.1). The subsequent levels of abstraction of the concept tree serve to further outline the attributes of the concept. Here, researchers should take care that the vertical relationship of the concept is properly specified. Less abstract characteristics have to be correctly assigned to the respective attribute at a higher level of abstraction that they are intended to specify (Munck & Verkuilen, 2002, p. 13). Researchers should thus avoid the conflation of attributes (QC 1.3.2, see Table 2.1).

Furthermore, researchers should avoid redundancy in the specification of the attributes (QC 1.3.3, see Table 2.1). "Attributes at the same level of abstraction should tap into mutually exclusive aspects of the attribute at the immediately superior level of abstraction" (Munck & Verkuilen, 2002, p. 13). The scope of all of the characteristics of the quality of democracy that are intended to specify an attribute at a higher level of abstraction should not overlap.

2.2.2 Operationalization and Measurement

After having conceptualized the quality of democracy transparently and adequately, the concept of interest has to be operationalized and measured in a valid (QC 2.1, see Table 2.1), reliable (QC 2.2, see Table 2.1), and replicable (QC 2.3, see Table 2.1) manner. Since "measurement processes build numerical bridges between abstract concepts and empirical realities" (Schedler, 2012a, p. 22), validity, reliability, and replicability (at least partly) ensure that the descriptive inferences from the observations to the unobserved phenomenon of the quality of democracy are transparent and correct (King et al., 1994, p. 55).

In the process of operationalization, it is important to ascertain that the indicators are valid (QC 2.1.1, see Table 2.1). First, they should accurately reflect the concept they have been selected to measure (Babbie, 2004, p. 143). Second, content validation and cross-national equivalence are central to the selection of valid indicators (Adcock & Collier, 2001, pp. 534–536, pp. 538–540). Content validation ensures that the indicators wholly operationalize the content of the concept (Jackman, 2008, p. 121). In this respect, the principles of redundancy and conflation outlined above also apply at the indicator level of the concept tree. Cross-national equivalence can comprise a number of aspects (van de Vijver, 1998, p. 47). With regard to macro-level concepts such as the quality of democracy, researchers should ensure the functional equivalence of the indicators across countries in order for them to be valid. The indicators do not have to be identical to be equivalent; instead "[a]n instrument is equivalent across systems to the extent that the *results* provided by the instrument reliably describe with (nearly) the same validity a particular phenomenon in different social systems" (Przeworski & Teune, 1970, p. 108). Since an indicator cannot validly measure a concept if it is not applicable in practice, we include the applicability of the indicator as an additional aspect that should be considered (Müller & S. Pickel, 2007, p. 526).

In order to maximize the validity of the measurement of the indicators, researchers should rely on multiple sources (Munck & Verkuilen, 2002, p. 16) (QC 2.1.2, see Table 2.1). While the fact that any source of information may be prone to random or systematic, data-induced measurement error (Bowman, Lehoucq, & Mahoney, 2005, p. 940), selecting multiple sources at least helps to reduce the amount of bias inherent in a single reference. In cross-national research, it is not only important to draw on multiple sources (Lauth, 2004, pp. 306–307). Researchers should use publications from varying

(state and non-state) actors with different national and cultural backgrounds, so as to avoid a Western state perspective, for example.

Since the measurement process requires the specification of the rules by which numerals are assigned to objects (Stevens, 1946, p. 677), the validity of the indicators also has to be evaluated with regard to the measurement level of the indicators (QC 2.1.3, see Table 2.1). The measurement level of choice should maximize "homogeneity within measurement classes with the minimum number of necessary distinctions" (Munck & Verkuilen, 2002, p. 17). This requires that the scale consist of a sufficiently large number of values so as to adequately grasp the details of the attribute it is intended to measure. The question of adequacy in turn can only be assessed in light of theoretical considerations.

The validity of the measurement is contingent on the reliable coding of the data (Juni, 2007, p. 834) (QC 2.2, see Table 2.1). When data are coded according to a coding scheme, it is essential that the raters interpret the coding instructions in the same manner. The precision of the coders' judgments of the sources should be tested for by applying inter-coder (or inter-rater) reliability tests (Jackman, 2008, pp. 122–124) (QC 2.2.1, see Table 2.1).[52]

Besides validity and reliability of the data, replicability of the indices (QC 2.3, see Table 2.1) is a key concern (Schedler, 2012b, p. 253). The publication of the research procedures is a key criterion of good social scientific research (King et al., 1994, p. 8). Since many researchers have neither the time nor the resources to develop their own indices of the quality of democracy, they usually assess the quality of democracy of political regimes or test the generalizability of theories related to the quality of democracy by means of secondary data analysis. They thereby depend on the quality of their colleagues' work. The coding rules, the sources of information, and the disaggregate data should therefore be easily available to the scientific community at all times and free of charge (Munck & Verkuilen, 2002, p. 19; Schedler, 2012b, p. 239). The coding rules should be detailed enough so as to allow an unambiguous replication of the coding process (QC 2.3.1-2.3.3, see Table 2.1) (Schedler, 2012b, pp. 253–255). This allows researchers to check the operationalization and measurement of the concepts involved and permits them to draw conclusions regarding the validity of the inferences they make.

[52] Munck and Verkuilen (2002, p. 18) caution not to take the results of reliability tests as indicative of the validity of a measure since indices can be highly reliable but invalid because of similar biases among the coders.

2.2.3 Aggregation

Since the quality of democracy is usually measured in terms of an index, the decisions regarding the aggregation of the indicators also have to be taken into account (Munck & Verkuilen, 2002, pp. 22–27; OECD, 2008, pp. 20–21). In order for the index to be a valid reflection of the concept it claims to measure (QC 3.1, see Table 2.1), the researchers have to justify the extent to which they aggregate the attributes (QC 3.1.1, see Table 2.1). A high level of aggregation reduces the amount of data, thereby facilitating the application, interpretation, and comparability of indices. However, it also reduces the validity of the measurement of the concept. A single index score may be overly simplistic as it may conceal systematic differences between the cases with regard to certain attributes (Munck & Verkuilen, 2002, p. 22; OECD, 2008, pp. 13–14). Researchers thus need to decide theoretically (depending on the level of generality they intend the measure to apply to) and test empirically whether the indicators can be aggregated into a single index or whether it is best to work with a multidimensional measure (Babbie, 2004, p. 154; OECD, 2008, p. 14). Furthermore, in order to be valid, the attributes of the concept have to be accumulated by means of aggregation procedures that take into account the horizontal and vertical relationships (QC 3.1.2, see Table 2.1). The weighting procedure should reflect the role of the attributes in defining the concept (QC 3.1.3) (Goertz, 2008, pp. 98–103; OECD, 2008, pp. 31–34).

As in the case of the measurement of the indicators, researchers should undertake reliability tests so as to assess the robustness of the composite indicator (QC 3.2., see Table 2.1). The process of operationalization, measurement, and aggregation involves uncertainty as it requires a number of choices on the part of the researcher. This includes, among others, the choice of indicators, the aggregation rules, and the weighting scheme. The question that follows is whether the values of a case fundamentally change when certain aspects are decided differently, for example, when an indicator is excluded. The quality of the index should therefore be assessed in terms of the robustness of its values to these choices (QC 3.2.1, see Table 2.1) (Munck & Verkuilen, 2002, p. 25; OECD, 2008, pp. 34–35).

Replicability is also important when it comes to the aggregation process (QC 3.3, see Table 2.1). The rules of aggregation should be published free of charge at all times and should be sufficiently detailed so as to allow an unambiguous replication of the aggregation process (QC 3.3.1, see Table 2.1). The aim of the following section is to

demonstrate the application of the quality assessment criteria outlined above to three current measures of the quality of democracy that cover the conceptual breadth of the debate on the quality of democracy.

2.3 Assessing the Quality of Quality Measures of Democracy

The SGI represents a comprehensive index measuring sustainable policy outcomes. The index was developed by the Bertelsmann Foundation with the aim to identify necessary reforms in the European Union (EU) and the Organisation for Economic Co-operation and Development (OECD). Given current challenges – economic globalization, demographic change and resource scarcity, to name a few – national governments are required to adapt their policies in order to provide sustainable solutions to lasting problems. According to the founders of the SGI, sustainable policies are closely related to good governance. Hence, the SGI furnishes a 'systematic, indicator-based comparison' (SGI, 2014a, p. 7) to monitor the sustainability of their policy performance.

The SGI ranks both the OECD countries and the EU member states in terms of their democracy performance, policy performance, and governance (Schraad-Tischler & Seelkopf, 2014, pp. 2–12).[53] According to the SGI, the quality of democracy is essential to sustainable governance and therefore constitutes the 'democracy' dimension of the SGI (Schraad-Tischler & Seelkopf, 2014, p. 8). In our assessment of the quality of the SGI, we therefore exclusively consider this dimension.

The DB is another instrument that aims to measure the quality of democracy. It was developed by Wolfgang Merkel, Marc Bühlmann, Daniel Bochsler et al.[54] According to these authors, the index aims "to overcome the conceptual and methodological shortcomings of existing measures, in order to measure the subtle differences in the quality of democracy" (Democracy Barometer, 2014b). It is based on a middle-range definition of democracy that ranks between a liberal and a participatory understanding of democracy and comprises the principles of freedom, equality, and control. Nine democratic functions are required to fulfill these principles. These functions are divided further into components and subcomponents which are then operationalized with 105 indicators (Merkel et al., 2014a, pp. 13–35). Each subcomponent serves to assess both

[53] For detailed information on the SGI see Schraad-Tischler and Seelkopf (2014).
[54] All members of the research team are listed online (Democracy Barometer, 2014a).

the de jure institutional setting as well as the de facto effectiveness of the institutions. Unlike the SGI, the DB strives to avoid data based on qualitative expert judgements and uses quantitative indicators where possible.[55]

The GDR was developed by D. F. Campbell and Sükösd (2002, 2003) and D. F. Campbell (2008). The aim of the index is to provide a comparative ranking of the quality of all democracies[56] at certain time points. The authors argue that freedom, equality and performance are key dimensions of the quality of democracy. Accordingly, they determine countries' quality of democracy based on six attributes: politics (50 per cent), gender (10 per cent), economy (10 per cent), knowledge (10 per cent), health (10 per cent) and environment (10 per cent) (D. F. Campbell, Barth, P. Pölzlbauer, & G. Pölzlbauer, 2012; D. F. Campbell & G. Pölzlbauer, 2008, pp. 30–33; D. F. Campbell & Sükösd, 2002, pp. 5–6). These attributes are weighted differently (weighting factor in parentheses; see above) (D. F. Campbell & Sükösd, 2003). D. F. Campbell (2008, pp. 34–35) justifies the weights by arguing that the quality of democracies consists of the 'Quality of Politics + Quality of Society'. The political dimension is given the greatest weight since "without acknowledging the political system, it does not appear appropriate to talk about democracy" (D. F. Campbell, 2008, p. 34).[57]

The SGI and in particular the DB have received both positive and critical mentions in scholarly reviews. While Schmidt (2010), Müller and S. Pickel (2008) as well as Munck (2012) emphasize the positive aspects of the DB, Kaina (2008) and Lauth (2011a) have contributed critical remarks to the discussion. To date, the comments and rejoinders by Jäckle et al. (2012, 2013) and Merkel et al. (2013) in 'Comparative Governance and Politics' mark the climax of this debate. However, the arguments on both sides do not take into account key aspects of the measurement of the quality of democracy. As for the SGI, it has been evaluated critically by Czada (2010) and Nuscheler (2009) but these reviews focused on its theoretical foundation of good governance. This highlights the need to evaluate the indices anew based on our framework that accommodates the challenges inherent in measuring the quality of democracy.

The SGI generally provides valid, reliable, and replicable results with regard to the quality of democracy as a key component of the index. However, its conceptualization of

[55] For detailed information on the DB see Merkel et al. (2014a) and Merkel et al. (2014b).

[56] The country sample includes all countries that were considered 'free' or 'partly free' according to Freedom House throughout the respective two-year period of analysis.

[57] For detailed information on the GDR see D. F. Campbell et al. (2012) and D. F. Campbell and G. Pölzlbauer (2008).

the quality of democracy suffers from a lack of theoretical foundation. The description of the index briefly refers to the relevance of democratic theory and states that the SGI uses 'ideal representative democracy' (SGI, 2014b, p. 12) as its normative point of reference. Yet, the description of the SGI neither includes nor justifies its definition of democracy or the quality of democracy. However, the list of criteria of the quality of democracy provided shows that the quality of democracy as conceptualized by the SGI clearly goes beyond the procedural concept of democracy (SGI, 2014b, p. 13). The indicators used to operationalize the attributes are very complex, resulting in one of the more problematic aspects of the SGI. The attributes cover many different aspects of the phenomenon of interest and therefore at times refer to multiple dimensions of the quality of democracy. This complicates the task of evaluating their validity. The peer review process of assessing the expert judgements (inter-coder reliability) is very transparent and structured in an exemplary manner. While the aggregation rules are convincing from a methodological point of view, the SGI does not justify them theoretically.

External validation as well as robustness tests would help to enhance the quality of the index. The SGI includes two non-democratic countries: Mexico and Turkey. Given this fact, the evaluative statements pertaining to these countries must, by definition, differ from the assessments of the remaining countries. They refer to the quality of the political system rather than the quality of democracy.

The DB performs the best among the three indices in terms of the quality criteria outlined above. It is justified convincingly both in theoretical and methodological terms. The extensive operationalization of its middle-range concept of democracy constitutes both a strength and a weakness of the index. On the one hand, the DB outlines the underlying concept of the quality of democracy in detail and provides a wealth of sources of information. On the other hand, the authors themselves admit that the fact that they use 105 indicators to operationalize the concept renders it nearly impossible to avoid redundancies and conflations and to provide a thorough theoretical foundation up to the final leaves of the concept tree (Merkel et al., 2013, p. 78). Thus, the DB's range of indicators is too complex. The DB's documentation does not make any reference to inter-coder reliability tests. The index therefore cannot be assessed in this regard. Like the SGI, the DB also includes hybrid regimes, calling into question its aim of exclusively measuring the quality of democracy.

Table 2.2
Quality Assessment Criteria: Empirical Application

Conceptualization				
Quality criteria	Coding rules			
	Indicator	SGI	DB	GDR
1.1 Transparency	1.1.1 Transparency	+ · detailed explanation available online (www.sgi-network.org/2014/)	+ · detailed explanation available online (www.democracybarometer.org/concept_en.html)	+ · explanation available online (www.democracyranking.org/downloads/basic_concept_democracy_ranking_2008_letter.pdf)
1.2 Concept specification	1.2.1 Differentiation	- · The quality of democracy is defined as part of an overall measurement of Sustainable Governance · Includes only a very general reference to theory · its normative reference point is an ideal representative democracy'[a] · quality of democracy combined with political participation and legitimacy of the political system[b]	+ · theoretical reference to democratic theory results in a 'middle-range concept of democracy embracing liberal as well as participatory ideas'[c] · three core principles: freedom, equality and control	+ · extensive concept of the quality of democracy derived from theory àQuality of democracy = Quality of Politics + Quality of Society · Three core dimensions: freedom, equality and performance represent sub-systems of society · The subsystems include: the political system, gender equality, the economic system, knowledge- based information society, the health system and environmental sustainability[d]
	1.2.2 Range of political systems	- · includes Mexico and Turkey[e]	- · only democracies included in the blueprint sample · extended data set also includes hybrid regimes	- · 115 countries (2013) including autocracies (China, Egypt: 'virtual scores'[f]) and fragile states (Libya and Syria) · case selection based on FDH (free and partly free)
	1.2.3 Range of institutional settings	+[g]	+ · A few indicators do not seem equally applicable to parliamentary and presidential systems	+
	1.2.4 Parsimony	+	-	+

Table 2.2 (continued)

Quality criteria	Indicator	SGI	DB	GDR
		Conceptualization		
		Coding rules		
1.3 Conceptual logic	1.3.1 Coherence	**0** · No justification why access to information (D2) is assessed separately from other civil rights and political liberties (D3)[h]	**+** · Clearly comprehensible both theoretically and functionally up to the level of the components. Further specifications of the concept follow logically but not necessarily theoretically or functionally	**0** · Theory-based justification up to the level of the sub systems · Indicators for the final measurement not derived from theory
	1.3.2 Conflation	**+**	**0**	**-** · E.g. conflation in the dimensions measured with FDH data
	1.3.3 Redundancy	**+** · However, the assessment of media independence and the prevention of corruption overlap somewhat	**0** · Some indicators measure very similar aspects (e.g. votediff and seatdiff)	**-** · Redundancy among the indicators provided by FDH, Freedom of the Press and CPI
		Operationalization & Measurement		
2.1 Validity	2.1.1 Selection of valid indicators	**0** · 1) 0 (see 1.2) · 2) + · 3) 0 (multidimensional indicators: indicators are very complex because they assess too many aspects of the respective attribute) · 4) + · 5) +	**0** · 1) + · 2) 0 · 3) + (macro data) · 4) 0 · 5) +	**-** · 1) + · 2) - (neither the selection nor the number of indicators is quite comprehensible) · 3) 0 (non-political indicators: yes, political indicators: no (FDH/CPI) · 4) – · 5)+
	2.1.2 Range of sources of information	**+** · 1) + (country experts/ reviewers/ regional coordinators select the sources) · 2) + · 3) +	**+** · 1) + yes · 2) + · 3) 0	**+** · 1) + (many sources of information) · 2) + · 3) 0

Table 2.2 (continued)

Quality criteria	Indicator	SGI	DB	GDR
		Operationalization & Measurement		
		Coding rules		
2.1 Validity (continued)	2.1.3 Theory-based justification of the measurement level	- · numerical ten-point scale comprises a four-point scale (in terms of content) · assignment of values to empirical characteristics is not theory-based[i]	+ · theory-based reflection on possible scaling methods · empirical minimum and maximum values[j]	- · no justification of the scale (1-100)
2.2 Reliability	2.2.1 Intercoder reliability	+ · Five-stage reassessment process[k]	- · Not mentioned in the description of its methodology	- · Not mentioned in the publications
2.3 Replicability	2.3.1 Publication of coding rules	+ · 1) + · 2) +	+ · 1) + · 2) +	- · 1) - · 2) 0 (no information on transformation of the indicator scores)
	2.3.2 Publication of sources of information	+	+	+
	2.3.3 Publication of disaggregated data	**0** · the disaggregated data are only available for 2014, only country reports are available for 2009 and 2011	+	+
		Aggregation		
3.1 Validity	3.1.1 Appropriate level of aggregation	+ · aggregation of the four dimension to a single index for cross-country comparisons · Detailed information included in web charts published online [l]	+ · Justified based on theoretical considerations · Measurement scale not based on democratic theory but rather based on empirical results across thirty countries within eleven years (best practice)	+ · Includes scores for the sub-systems as well as the index as a whole · improvement ranking highlights changes over time

Table 2.2 (continued)

Aggregation				
Quality Criteria	Coding rules			
	Indicator	SGI	DB	GDR
3.1. Validity (continued)	3.1.2 Theory-based justification of the rules of aggregation	**0** · Justification in terms of measurement theory, no theory-based justification[m]	**+** · Aggregation rules based on theoretical and methodological justifications regarding the relationship between the different attributes of the concept at different levels[n]	**-** · Rules of aggregation mentioned · No theory-based or methodological justification
	3.1.3 Theory-based justification of weights	**0** · Practical but no theory-based justification	**+** · see 3.1.2	**-** · weighting of the sub-system scores is briefly elaborated · no theory-based justification of the weights · no methodological explanation of the weighting of the indicators
3.2 Reliability	3.2.1 Robustness tests	**-**	**-** (however, they perform a number of validation tests)[o]	**-**
3.3 Replicability	3.3.1 Publication of aggregation rules	**+** · 1) + · 2) + · transparent aggregation rules[p]	**+** · 1) + · 2) + · transparent aggregation rules	**0** · 1) + · 2) 0 (standardisation of the values is not entirely comprehensible, aggregation process is transparent in part
positive		**11**	**13**	**8**
intermediate		**5**	**3**	**2**
negative		**4**	**4**	**10**

Note. [a] SGI (2014b, p. 12); [b] SGI (2014a, p. 10); [c] Bühlmann, Merkel, Müller, & Wessels (2012, p. 520); [d] D. F. Campbell (2008, pp. 30–33); [e] If countries (such as Turkey and Mexico, both rated as partly free by FDH 2013) violate at least one crucial dimension of democracy, it is more than questionable whether they should be regarded as full democracies (Freedom House, 2014).; [f] D. F. Campbell, P. Pölzlbauer, Barth, & G. Pölzlbauer (2011, p. 4); [g] As the theoretical concept refers to representative democracies, referenda on the national political level are neglected as in the case of Switzerland. Nevertheless, Switzerland receives high scores on the index because of its national election process; [h] SGI (2014a, p. 11); [i] Schraad-Tischler and Seelkopf (2014, p. 13); [j] Merkel et al. (2014a, pp. 5–7); [k] Schraad-Tischler and Seelkopf (2014, pp. 13–14); [l] See for example SGI (2014c); [m] Schraad-Tischler and Seelkopf (2014, p. 18); [n] Bühlmann, Merkel, Müller, Giebler, et al. (2012, pp. 134–135); [o] Bühlmann, Merkel, Müller, Giebler, et al. (2012, pp. 139–144); [p] Schraad-Tischler and Seelkopf (2014, pp. 17–18).

The results of the GDR are less clear-cut. The constructors also provide a theoretically justified conceptualization of the quality of democracy in its own right based on O'Donnell (2004). The approach appears to be innovative and original. However, its main issue lies with the non-transparent, largely incomprehensible, and theoretically weak operationalization (see Table 2.2). It partly relies on data sources (Freedom House, Corruption Perception Index etc.) that are based on somewhat rudimentary theoretical foundations. In addition, the choice of indicators as well as the aggregation of the indicators (concept tree) is not thoroughly justified in methodological and theoretical terms. Its documentation does not include any statements regarding inter-coder reliability tests. As for the countries evaluated by the index, it includes quality assessments of states representing the whole range of political systems (e.g., China and Egypt as authoritarian regimes). Thus, the index somewhat resembles a measure of the quality of political systems rather than the quality of democracy. These issues provide reason for concern regarding the quality of the index. In addition, the GDR covers failed states such as Libya and Syria. Not only are these states non-democratic, they are also dysfunctional. This provides all the more reason to criticize the application of the concept of the quality of democracy.

2.4 Implications and Recommendations

What are the implications of these results for cross-national comparisons of the quality of democracy? Overall, the quality criterion 'range of political systems' (QC 1.2.2), the 'selection of valid indicators' (QC 2.1.1) and 'inter-coder reliability' (QC 2.2.1) seem to be the weak spots of the indices with regard to the conceptualization, operationalization and measurement of the quality of democracy.

Regarding the range of political systems, the constructors of some of the indices apply their instruments to countries that do not fulfill the prerequisite for assessing the quality of democracy, namely the fact that they are democratic political regimes. The indices are therefore applied to measure the quality of democracy of cases they are not calibrated to measure. The conclusions that follow from these evaluations are of little substantial value: since electoral democracies are not full democracies, it only follows that they fare worse in the evaluations of their quality of democracy, especially when the index has been carefully constructed to measure the sometimes subtle differences in quality in

established democracies. If applied to defective or electoral democracies, indices that measure the quality of democracy have to give lower scores to these political regimes compared with democracies if the index is to reflect democratic progress.

The selection of valid indicators is crucial for the measurement of these differences in quality. Such a measurement depends on the fact that the indicators measure what they are intended to measure (Przeworski & Teune, 1970, p. 108). A detailed theoretical concept tree is imperative in this respect in order to be able to judge the validity of the selected indicators.

Concerning inter-coder reliability, this is most likely the easiest issue to remedy. Constructors of indices should take care to involve several coders and publish the results of the reliability tests.

In conclusion, in order to analyse and ensure the quality of quality measures of democracy, we recommend that researchers should:

a) carefully study its theoretical foundation. Many indices, including some of the indices measuring the quality of democracy, rest on weak theoretical ground;

b) ensure that the attributes and indicators are combined into a coherent theoretical concept. Do the indicators measure the concept adequately and comparably? Are the sources upon which the measurements are based convincing in terms of their validity?

Only if these questions have been answered conclusively should researchers go on to consider the methodological details of the construction of their indices. Here, the image of the concept tree applies as well: only a sturdy trunk can sustain strong branches.

By applying our framework, researchers can test the quality of the conceptualization, operationalization, measurement, and aggregation of their index of interest in a systematic manner. The standards of assessment can be used both by researchers constructing as well as researchers applying indices for secondary analysis. This provides a clear picture of the strengths and weaknesses of the data regarding their conceptual foundation as well as their validity, reliability and replicability. These results can then be used to draw conclusions regarding the quality of descriptive and causal analyses based on these data. Overall, the quality criteria provided in this article help to enhance the validity of scientific inferences.

3 Article 2 "Measuring the Quality of Democracy: Why Include the Citizens' Perspective?"[58]

Abstract

New indices measuring the quality of democracy constitute a significant innovation in comparative political science. They might, however, provide a biased perspective because they largely focus on macro-level criteria. Thus, the question is whether the measurement of the quality of democracy can be improved by complementing the evaluations of these indices with assessments based on individual-level survey data. Using data from 20 established democracies in the European Social Survey 2012 and the Democracy Barometer, we compare the understandings and evaluations of the quality of democracy underlying these two measurement approaches. We demonstrate that while the results coincide to a certain extent, individual-level data provide an important complementary perspective that adds to the validity of the measurement of the quality of democracy.

Keywords: quality of democracy; understanding of democracy; meaning of democracy; evaluation of democracy; conceptualization of democracy

[58] Published as: Pickel, S., Breustedt, W., & Smolka, T. (2016). Measuring the quality of democracy: Why include the citizens' perspective? *International Political Science Review*, *37*(5), 645–655. doi:10.1177/019251 2116641179

3.1 Current Understandings and Evaluations of the Quality of Democracy: A Biased Perspective?

Indices used to measure the quality of democracy have come a long way. Initially, researchers frequently applied indices such as Polity IV and Freedom House's 'Freedom in the World' (Campbell, 2012). These measures were criticized, however, as they insufficiently grasp the nuances of the quality of democracy because of their underlying theoretical concept of democracy as well as the empirical measures (Bühlmann, Merkel, and Wessels, 2008, p. 4).

New indices such as the Democracy Barometer and the Sustainable Governance Indicators constitute a significant innovation in comparative political science. Their conceptualization and empirical measures are nuanced enough to permit researchers to empirically assess the quality of democracy (S. Pickel et al., 2015).

These indices can guide political decision-makers on how to improve the quality of their democracy "to achieve the broad and durable legitimacy that marks consolidation" (Diamond & Morlino, 2005, p. ix). In order to do so, each of these indices has its own set of quality criteria and applies these equally to all democracies. At the same time, their measurements largely focus on the "global properties" (Lazarsfeld & Menzel, 1961, p. 503) of democracies, that is, the quality of the principles, structures, and outcomes of democratic institutions.

The measures of these indices might result in a biased perspective, though. From an institutionalist viewpoint, assessing the quality of democracy in terms of the above-mentioned macro-level phenomena draws a comprehensive picture of what should be improved to achieve democratic legitimacy. According to the viewpoint of political culture research, though, these indices only provide a partial impression of the quality of democracy.

In line with the latter research tradition, citizens' individual-level evaluations of the quality of democracy are what matters for democratic legitimacy (Easton, 1975). These evaluations pertain to what Scharpf (2003) refers to as 'input legitimacy' and 'output legitimacy'. Input legitimacy derives from citizens' positive evaluation of the institutional accountability procedures. Output legitimacy concerns evaluations of the policy performance of the institutions (Scharpf, 2003).

This individual-level perspective also gives rise to the assumption that citizens might have varying quality criteria in mind when evaluating their democracy. This is corroborated by previous research on subjective understandings of democracy. While citizens generally share a liberal notion (Canache, 2012, p. 1133), there are also substantial within- and cross-country variations (Dalton, Shin, & Jou, 2007, pp. 146–147). Consequently, the political culture viewpoint differs from the institutionalist viewpoint on the quality of democracy in two respects. First, citizens' understanding of democracy can vary across countries, that is, unlike current indices they do not apply a common set of quality criteria when assessing the quality of democracy. Second, citizens' individual-level evaluations of the quality of democracy should be taken into account because they provide "above all a relatively accurate perception of *their own* needs and preferences" (Diamond & Morlino, 2005, p. xiii; emphasis in original) that is not necessarily reflected in the macro-level features and expert judgements that current indices largely rely on.

The differences between these two viewpoints evoke the following question: do the understandings and evaluations of the quality of democracies by citizens coincide with the concepts and assessments of the quality of democracies by existing indices or do they provide a complementary perspective? Most likely, citizens have a subjective perspective on the everyday workings of democracy that differs from and therefore complements expert judgements and macro-level data on the quality of democracy. Methodologically, considering both perspectives would therefore add to the validity of the measurement of the quality of democracy (Logan & Mattes, 2012, p. 471). Practically, by comparing existing indices with citizens' perspective, political scientists could better pinpoint the kinds of reforms political decision-makers should undertake in order to reduce public dissatisfaction, thereby enhancing the legitimacy of their democracy.

To date, the only data to permit a direct comparison of the macro- and individual-level perspectives are the Democracy Barometer (DB) and the 2012 round of the European Social Survey (ESS). The former – a macro-level index – was developed based on Bühlmann, Merkel, and Wessels' (2008) concept of the quality of democracy. When constructing the latter – individual-level survey items measuring the understanding and evaluation of the quality of democracy – Kriesi et al. (European Social Survey, 2013) also referred to Bühlmann, Merkel, and Wessels (2008) as well as Diamond and Morlino (2005; see also Morlino, 2009).

Our analysis builds on Logan and Mattes (2012) who address the match between individual-level and macro-level assessments of the quality of democracy. It differs in three important respects, though. First, the survey items they use were not originally designed to measure the quality of democracy unlike the ESS 2012 survey items. Second, they do not consider country-specific differences in understandings of democracy when measuring individual-level evaluations of the quality of democracy. Third, unlike us, they compare these individual-level assessments with macro-level measures of political transformation rather than with indices that were specifically designed to measure the quality of democracy.

In order to compare the macro-level and individual-level understandings and evaluations of the quality of democracy, we first present the two concepts of the quality of democracy. We then briefly describe the DB and the ESS survey items. We go on to test whether citizens' understanding of democracy in European established democracies is in line with the concept underlying the DB. Finally, we compare the evaluations of the quality of democracy by citizens with the assessment by the DB. We conclude with implications and suggestions of our research results for the study of the quality of democracy.

3.2 The Concept of the Quality of Democracy

Over the years, researchers have proposed a number of conceptualizations of the quality of democracy (for an overview, see Munck, 2016). The conceptualizations underlying the DB and the ESS survey items (Morlino, 2004b; see also Bühlmann, Merkel, & Wessels, 2008; Diamond & Morlino, 2005, as well as Morlino, 2011) are among those that have strongly influenced the current debate (Logan & Mattes, 2012, p. 470; Munck, 2016, p. 3).

Diamond and Morlino (2005, p. xi) and Morlino (2009, pp. 3–4) derive three different meanings of the term 'quality' from the industrial and marketing sector. They define quality of democracy in terms of its process, content, and results. They then identify eight different dimensions wherein the quality of democracy varies (Diamond & Morlino, 2005, p. xii). The five procedural dimensions include the rule of law, participation, competition, and accountability (vertical and horizontal) and the two substantive dimensions include respect for civic and political freedoms as well as equality. The

eighth dimension that addresses responsiveness links the procedural with the substantive dimensions and evaluates the government´s policy results in light of citizens' expectations.

Bühlmann, Merkel, and Wessels (2008) distinguish between three democratic core principles (political equality, freedom, and control of political power)[59] and these three principles are further divided into nine democratic functions and then into components and subcomponents of these functions. Thus, the quality of a democracy is measured in terms of the degree to which it fulfils the nine functions and their components (Bühlmann, Merkel, & Wessels, 2008, pp. 6, 27–29; see Table 3.1).

In summary, the two concepts are similar in so far as both contain the three basic principles of democracy: freedom; political equality; and control. At the same time, Bühlmann, Merkel, and Wessels (2008) on the one hand and Diamond and Morlino (2005) and Morlino (2009) on the other differ with respect to the substantive dimension. While the former framework does not include the outcome dimension, the latter one includes it as a separate dimension.

3.3 The Macro-Level and Individual-Level Measurement of the Quality of Democracy

The DB is a macro-level index of the quality of democracy. It was developed based on the theoretical concept by Bühlmann, Merkel, and Wessels (2008) outlined above. It consists of 105 indicators that reflect the lowest level of the concept. In line with Bühlmann, Merkel, and Wessels (2008), these indicators are aggregated into 53 subcomponents, 18 components, nine democratic functions, three democratic principles, and finally the overall quality score. The quality score for the blueprint countries ranges from 0 to 100[60] and the countries in the DB are ranked accordingly.[61]

[59] The categorisation of control as one of the three basic principle of democracy can be traced back to political theory. Control is regarded "as the instrument to influence the balance of equality and freedom and to guarantee them" (Bühlmann, Merkel, and Wessels, 2008, p. 13; see also Lauth, 2004; O`Donnell, 1998).

[60] 'Blueprint' countries are a set of countries that serve as the benchmark for all other countries. In the case of the DB established democracies selected on the basis of Polity IV and Freedom House represent the standard of comparison within the index (Bühlmann, Merkel, and Wessels, 2008, p. 5).

[61] For further information on the methodology see Merkel et al. (2014a).

Table 3.1
Quality of Democracy: Macro-Level Concepts and Individual-Level Items

Morlino (2009)	Bühlmann, Merkel, & Wessels (2008)	Items in module 'Europeans' understandings and evaluations of democracy' of the 2012 round of the European Social Survey (ESS)
• Rule of Law (*Procedural)*	• Rule of Law (*Freedom*)	Concept 1: Rule of Law • accessibility and equality of the judicial system (E10/E25)
• Electoral Accountability (Procedural)	• Vertical Accountability *(Control)* • Transparency (*Equality*)	Concept 2: Vertical accountability • retrospective accountability (E12/E26) • transparency: a) transparency of political decisions (E14/E28) b) availability of alternative sources of information (E6/E22)
• Inter-institutional Accountability (*Procedural*)	• Mutual constraints of constitutional powers (*Control)*	Concept 3: Horizontal accountability • "courts are able to stop the government…" (E11/n.a.)
• Participation (*Procedural*)	• Participation (*Equality*)	Concept 4: Participation • opportunities of effective participation (E9/n.a.) • forms of participation: a) referenda (E8/E24); b) deliberation (E2/E18)
• Competition (*Procedural*)	• Vertical Accountability: Competitiveness (*Control*)	Concept 5: Competition • elections free and fair (E1/E17) • differentiated offer (E3/E19) • viable opposition (E4/E20)
	• Representation (*Control*)	Concept 6: Representation • subjects of representation (E7/E23) • type of governmental coalition a) single party government (E42/E43) b) coalition government (E44/E45)
• Responsiveness (*Outcome*)	• Responsiveness (*Equality*)	Concept 7: Responsiveness • responsiveness to the citizens a) change planned policies (E37/E38) b) stick to planned policies (E39/E40) • responsiveness to other stakeholders (E16/E30)
• Freedom (*Substantive*)	• Individual liberty (*Freedom*)	Concept 8: Freedom • freedom of expression a) for all (E32/E33); b) not for extreme (E34/E35) • freedom of press (E5/E21)
• Equality (*Substantive*)		Concept 9: Equality • social equality (E15/E29) • welfare (E13/E27)

Note. Authors' compilation based on Bühlmann, Merkel, & Wessels (2008); European Social Survey (2013) and Morlino (2009). The former ESS item number refers to 'meaning' items, the latter to 'evaluation' items; n.a: there is no corresponding evaluation item.

Based on the macro-level conceptualizations of the quality of democracy by Bühlmann, Merkel, and Wessels (2008) and Diamond and Morlino (2005; see also Morlino, 2009), Kriesi et al. (European Social Survey, 2013, pp. 6–8) distinguish 10 different attributes of the quality of democracy.[62] They developed 45 corresponding 'meaning' and 'evaluation' survey items measuring the understanding and assessment of the quality of democracy for the rotating module of the 2012 round of the ESS. In closed-ended questions, citizens are asked to state (on a scale from 0 to 10) how important they think the stated aspect is to democracy in general and to what extent the statement applies in their country. Table 3.1 presents an overview of the operationalization of the macro-level concepts with individual-level survey items (European Social Survey, 2013, pp. 9–37).

As Table 3.1 shows, the items in the ESS 2012 cover all of the dimensions of the quality of democracy as specified by Diamond and Morlino (2005) and Morlino (2009). As for Bühlmann, Merkel, and Wessels (2008), the items reflect all but one function (governmental autonomy) of the control dimension of their conceptualization of the quality of democracy. Thus, the ESS 2012 survey items provide a suitable tool to compare the macro-level concept and assessments of the quality of democracy with the individual-level understandings and evaluations.

3.4 Case Selection and Method of Analysis

In our analysis, we study 20 established democracies included in the ESS 2012, namely Belgium (BE), Cyprus (CY)[63], Czech Republic (CZ), Denmark (DK), Finland (FI), France (FR), Germany (DE), Hungary (HU), Iceland (IS), Ireland (IE), Italy (IT), Netherlands (NL), Norway (NO), Poland (PL), Portugal (PT), Slovenia (SI), Spain (ES), Sweden (SE), Switzerland (CH), and the United Kingdom (GB). These countries were selected as they are part of the set of 'blueprint' countries in the DB (Merkel et al., 2014a, p. 6).[64]

[62] Concept 10 'Support for Democracy' includes an item that asks for an assessment of how democratic the country is and an item that reflects an overall assessment of the importance of living in a democracy. Since these items do not pertain to a specific dimension of quality of democracy they are excluded from the analysis.

[63] Questionnaires in Cyprus were administered in Greek (European Social Survey, 2012b, p. 42).

[64] The samples in the ESS 2012 are selected according to random probability methods and are representative of the population aged 15 and over. In order to correct for possible sampling errors and non-response bias, we weight all data using the post-stratification weight provided by the ESS (European Social Survey, 2014, pp. 1–3).

Using the ESS 2012 'meaning' items, we perform a principal component analysis for each country to determine whether citizens' understandings of democracy match the principles of the quality of democracy underlying the macro-level index 'Democracy Barometer'. If the individual-level conceptualizations perfectly align with these principles, all of the items in the analysis should load on a single principal component in all of the countries.[65]

We deviate from Bühlmann, Merkel and Wessels' (2008) concept of the quality of democracy that underlies the DB in two respects, though. First, as mentioned earlier, the ESS 2012 items do not reflect governmental autonomy. Second, we take into account the items measuring substantive aspects of the quality of democracy (social equality and welfare) because they are an important part of 'output legitimacy'.

3.5 Comparing Macro-Level and Individual-Level Understandings and Evaluations of the Quality of Democracy in European Established Democracies

3.5.1 Understanding

The results of the principal components analyses are presented in Table 3.2. They show that individuals in the established democracies included in the ESS share a number of associations with democracy. In addition, the results imply that their idea of what constitutes a 'good' democracy is similar to the macro-level concept.

Item E7 ("the rights of minority groups are protected") primarily loads on the first principal component in all but three countries. This indicates that the subjects of representation are a key concern in individuals' understanding of democracy. Item E1 ("national elections are free and fair") loads strongest on the first principal component in 16 out of 20 countries. Item E4 ("opposition parties are free to criticize the government") loads strongest on the first principal component in 14 out of 20 countries and most

[65] We consider principal component analysis to be the appropriate technique for the following reasons. We assume that, empirically and at the individual level, the understanding of the quality of democracy is a composite measure of the meanings citizens associate with democracy. Thus we assume that the construct of democracy can be described as a formative model. The measurement in principal component analysis is based on this measurement model. If the citizens in the countries of analysis deem all of the theoretical principles of the quality of democracy addressed in the survey items as relevant, then their responses to the survey items can be "combined to form weighted linear composites" (Edwards, 2011, p. 370) of a single, principal component.

Table 3.2
Items with Primary Factor Loadings on the First Principal Component

Concepts	Items	Countries																			# of primary factor loadings
		BE	CY	CZ	DK	FI	FR	DE	HU	IS	IE	IT	NL	NO	PL	PT	SI	ES	SE	CH	GB	
Rule of Law	E10					x		x	x	x	x		x		x	x	x	1	x	x	x	**13**
Vertical Accountability	E12					x		x			x				x	x	x	x		x	x	9
	E14							x			x					x	x	x			x	6
	E6	x		x	x	x	x	x	x	1		x	x	x	x			x	x	1		**15**
Horizontal Accountability	E11							1	x		x				1	x		x		x	x	8
Participation	E9			x			x	x	x						x		x	x			x	8
	E8			x					x		x				x		x	x			x	7
	E2	x	x				x		x			x	x									6
Competition	E1	x	x	x	x	x	x	x	x	x		x	x	x	x			x	x	x		**16**
	E3	x	x	x			x		x			x	x	x								8
	E4	1	1	x	1	x	1	x	1	x		1	1	1					1	x		**14**
Representation	E7	x		1	x	x	x	x	x	x	x	x	x	x	x	x		x	x	x		**17**
	E44																					0
Responsiveness	E37																					0
	E16																					0
Freedom	E32													x					x			2
	E5	x	x	x	x	1	x	x	x	x		x	x	x					x	x		**14**
Equality	E15										x					x	x	x			x	5
	E13										1					1	1	x			1	5

Note. Extraction method: principal component analysis; rotation method: oblimin with Kaiser Normalization; maximum number of iterations: 100; pairwise deletion; weight: pspwght; x = item primarily loads on the first principal component; 1 = item with the highest factor loading on the first principal component in the respective country; E42 and E44, E37 and E39 as well as E32 and E34 could not be included simultaneously in the analysis as agreement with one indicates disagreement with the other; therefore, one item per item pair was excluded from the analysis; European Social Survey data (2012a).

frequently has the highest primary factor loading on the first principal component across all countries. Both items measure the importance of competition. Items E5 ("the media are free to criticize the government"), E6 ("the media provide citizens with reliable information to judge the government") and E10 ("the courts treat everyone the same") exhibit the highest primary factor loading on the first principal component in at least 13 out of 20 countries.

In summary, more than half of the ESS dimensions of the quality of democracy in Table 3.1 – namely, representation, competition, freedom, vertical accountability, and rule of law – are covered by individuals' associations with the concept of democracy. In most dimensions, however, only one out of several aspects is addressed by the above-mentioned items. In each of the established democracies, additional items augment the understanding of democracy beyond this general pattern.

Unlike suggested by Niedermayer (2009), there is no conceptual divide between individuals' understanding of democracy in Eastern and Western European countries that could arise because of differences in individuals' political socialization. Instead, the differences could be attributed to the societies' varying economic circumstances. In many of the countries that faced an economic crisis in 2012, individuals associate welfare and social equality with the concept of democracy. Here, economic expectations are often part of citizens' understanding of democracy.

Apparently, in these cases, individuals tend to combine expectations regarding a certain input, that is, procedural understandings, with expectations that pertain to a certain output, that is, substantive aspects related to social policies (see Table 3.1). This is in line with research in the tradition of political culture outlined in the introduction, which assumes that both input and output of a political system are relevant when it comes to the assessment of the quality of democracy. At the same time, responsiveness (E37 and E16; Morlino's (2009) 'outcome' dimension), does not find expression in individuals' primary associations with the understanding of democracy.

In conclusion, the comparison of citizens' understandings of democracy with the concept underlying the macro-level quality index 'Democracy Barometer' shows that individual-level data provide a complementary perspective. First, the concept of the quality of democracy underlying the macro-level index does not apply perfectly at the individual level in the European established democracies in our analysis. Second, despite the fact that we can make out a certain core understanding, the overall pattern of understandings

of democracy varies across countries. This variance in terms of meanings and the number of items is in spite of the fact that we only considered the first principal component, assuming it to reflect the predominant understanding of democracy. These results, in turn, have implications for the comparison of citizens' evaluations of the quality of democracy with the assessments by the DB.

3.5.2 Evaluation

The variance in understandings of democracy should be taken into account when researchers intend to use the ESS evaluation items as an individual-level measure of the quality of democracy. Respondents will most likely provide answers to all of the evaluative questions. Yet, unless the criteria they evaluate reflect what they expect of a 'good' democracy, their evaluations of these criteria may be meaningless. Merely combining their evaluative responses into an index and then comparing them across countries might bias the results.

We therefore suggest that researchers use those items that prove to be meaningful to the citizens in the respective countries as shown in the principal component analyses above. This evaluative measure reflects 'real-life' understandings of democracy more closely than if we selected the evaluation items based on a theoretical concept alone. Consequently, the number of items is not determined *a priori*, that is, our approach is data-driven rather than theory-driven in this respect.

As for the comparability of the measures, we would argue that the country-specific understandings of democracy outlined in Table 3.2 can be considered to be equivalent in the sense that they all bear a family resemblance. "The commonalities are quite evident, even though there may be no trait that all family members, as family members, have in common" (Collier & Mahon, 1993, p. 847).

In order to calculate the individual-level evaluations of the quality of democracy for each of the 20 European established democracies, we proceed as follows. We use those evaluation items that correspond to the meaning items that constitute the understanding of democracy in a given country, that is, which primarily load on the first principal component for each country (see Table 3.1 and Table 3.2).[66] In order to quantify the overall quality assessment, we add up the rates of approval for each evaluation item. The rates of approval are the valid percentages of the response frequencies of category 6 to

[66] Note, though, that there are no corresponding evaluation items for items E9 and E11.

category 10 on a scale from 0 to 10. We then standardize the measure by dividing it by the number of items used.

Table 3.3 shows the results of the comparison between citizens' evaluations of the quality of democracy based on the ESS and the assessments of the quality of democracy by the DB. The following should be noted. Even though both the DB and the aggregated individual-level measures range from 0 to 100 (see notes in Table 3.3), they are only comparable in terms of country ranks and not in terms of absolute numbers because of differences in the way the scales were constructed.

Table 3.3
Comparison of the Evaluation of the Quality of Democracy, based on European Social Survey (ESS) Data and the Democracy Barometer

Ranking	Country	ESS Evaluations	Country	Democracy Barometer Evaluations
1	Sweden	86.60	Denmark	73.69
2	Denmark	86.16	Sweden	69.99
3	Finland	85.83	Switzerland	69.31
4	Norway	85.59	Norway	67.95
5	Netherlands	81.44	Finland	67.94
6	Switzerland	80.24	Netherlands	65.85
7	Cyprus	74.46	Iceland	65.03
8	Iceland	73.13	Belgium	64.55
9	Germany	72.89	Germany	62.75
10	Belgium	72.30	Slovenia	59.32
11	France	64.06	United Kingdom	58.10
12	Czech Republic	62.29	Ireland	56.29
13	Hungary	58.58	Portugal	55.22
14	Ireland	58.29	Cyprus	55.21
15	United Kingdom	51.02	Italy	54.04
16	Poland	51.00	Poland	53.92
17	Italy	49.06	Czech Republic	53.69
18	Slovenia	34.57	Spain	50.87
19	Spain	34.40	Hungary	50.34
20	Portugal	19.42	France	49.87

Note. Authors' own compilation based on the European Social Survey (2012a) and Merkel et al. (2014c); ESS evaluations reflect the rates of approval of the country-specific evaluation items (valid percentages of the response frequencies of category 6 to category 10 on a scale from 0 to 10, standardized by the number of country-specific evaluation items); the scale of the indicators of the democracy barometer 'blueprint' countries ranges from 0 to 100. These values are then aggregated in a multistep procedure (Merkel et al., 2014a, p. 10).

What is notable in Table 3.3 is that eight out of twenty countries rank the same or the rank only differs by one, namely Sweden, Denmark, Norway, Netherlands, Iceland, Germany, Poland, and Spain. The rank of several countries differs substantially, though. In Finland and especially in the Czech Republic, Cyprus, France, and Hungary citizens evaluate the quality of their democracy far more positively than the macro-level index. In Belgium, Ireland, Italy, Switzerland, the United Kingdom, and particularly in Slovenia and Portugal, the opposite is the case.

The comparison of the individual-level and macro-level evaluations of the quality of democracy suggests two things. First, although individual-level measures of the quality of democracy may not perfectly reflect the entire breadth of meaning of the macro-level concept, individual-level and macro-level measures can nevertheless lead to similar assessments. Second, in more than half of the cases, the evaluations differed. This shows that individual-level evaluations of the quality of democracy do indeed provide a complementary perspective. In particular, the results show that the inclusion of social equality and welfare items, which affect 'output-legitimacy', makes a substantial difference. This is particularly evident in the case of Spain and Portugal, where the evaluations of the quality of democracy in these respects are far lower compared to the other items.

3.6 The Citizens' Perspective: Implications and Suggestions

The understandings of the quality of democracy by citizens in European established democracies and the concept of democracy underlying the macro-level index 'Democracy Barometer' coincide in certain respects and differ in others. Both relate 'democracy' to the principles of freedom, equality, and control. The empirical analyses show, however, that citizens differ in their understandings of how these principles should be enacted. Beyond that, in a number of countries, citizens associate quality of democracy with 'output' performance such as social equality and welfare. Based on our results, future evaluations of the quality of democracy by means of macro-level indicators should consider including these aspects in their measurement.

We come to a similar conclusion with regard to the evaluation of the quality of democracy by the citizens and the DB. On the one hand, the country rankings coincide in almost half of the countries. On the other hand, in more than half of the countries the

rankings differ substantially. These results indicate that citizens provide a complementary perspective with regard to the measurement of the quality of democracy.

It should be noted, though, that our approach to studying citizens' evaluations of the quality of democracy emphasizes that it is important to take varying understandings of democracy across countries into consideration. An alternative approach would be to use the same individual-level evaluation items for all countries according to a common theoretical standard, thereby mirroring the approach of macro-level indices of the quality of democracy. Future research on the citizens' perspective of the quality of democracy would benefit from discussing the pros and cons of these approaches theoretically and testing the differences empirically.

Either way, researchers should bear the individual level in mind when measuring the quality of democracy. The results of our analysis show that citizens' 'subjective' understandings and evaluations provide a meaningful complementary perspective to 'objective' measures of the quality of democracy. Considering both perspectives therefore adds to the validity of the measurement and provides 'bottom-up' insights on what needs to be improved to enhance the legitimacy of democracy in the given countries.

4 Article 3 “Testing the Measurement Invariance of Political Trust across the Globe: A Multiple Group Confirmatory Factor Analysis”[67]

Abstract

Today, comparative social scientists have ample survey data to test the generalizability of theories related to political trust. Unless its measurement invariance has been established, they run the risk of drawing invalid conclusions though. Based on different sets of items and dimensional models, previous studies have yielded diverging results regarding the measurement invariance of political trust in Europe and former Soviet countries. Using a set of six items and contrasting three competing dimensional models, this study tests the measurement invariance of political trust across the globe in 32 electoral and liberal democracies. It uses multiple group confirmatory factor analysis and draws on data from the World Values Survey (wave 6, 2010-2014). Configural invariance of a revised two-dimensional model of trust in implementing and representative political institutions was established in 19 democracies when excluding trust in civil service. Full invariance of this model was established in three post-communist countries in eastern and southeastern Europe. The results corroborate that the measurement invariance of political trust must not be assumed. Conceptually, they provide reason to infer that, by and large, people in democracies have a two-dimensional construct of political trust. Methodologically, they manifest that trust in civil service is an ambiguous item, which is not as meaningfully related to the construct of political trust as other items.

Keywords: measurement equivalence; measurement invariance; multiple group confirmatory factor analysis; political trust; trust in political institutions

[67] Published as: Breustedt, W. (2018). Testing the measurement invariance of political trust across the globe. A multiple group confirmatory factor analysis. *Methods, Data, Analyses*, *12*(1), 7–46. doi:10.12758/mda.2017.06

4.1 Introduction

Today more than ever, comparative social scientists can test the generalizability of theories pertaining to the changes, sources, and consequences of political trust thanks to the growing availability of cross-national survey data (D. Braun, 2013; Zmerli & van der Meer, 2017). This is a decisive, but not a conclusive step forward. Unless the comparability of political trust measures has been established, inferences about the generalizability of political trust theories across the globe may be invalid (Davidov et al., 2014, pp. 56–57).

The issue of comparability results from the fact that people's political trust is a construct. As such, it is a latent property of individuals that cannot be measured directly (Jackman, 2008, p. 119). Cross-national researchers therefore have to rely on observed measures such as survey items pertaining to trust in different political objects. According to the 'response process model' (Tourangeau, Rips, & Rasinski, 2000, p. 166), answers to these items allow inferences about people's underlying construct of political trust. Based on this assumption, studies commonly use political trust items to create additive or averaged index scores (see for example Catterberg & Moreno, 2006; Chang & Chu, 2006).

While indices are a common and convenient measurement instrument, the index scores are not necessarily comparable across countries and over time. A key to valid comparisons is to establish the invariance of the measurement instrument. "The general question of invariance of measurement is one of whether or not, under different conditions of observing and studying phenomena, measurement operations yield measures of the same attributes" (Horn & Mcardle, 1992, p. 117). Various forms of bias may systematically distort the invariance of measures (van de Vijver & Tanzer, 2004). For example, asking about people's trust in a political institution such as civil service may be biased because civil service's responsibilities and tasks differ across countries. Or, owing to the translation of the response scales, the difference between 'a great deal of trust' as opposed to 'quite a lot of trust' may not be judged in the same way by respondents from different countries, thereby biasing their responses.

Because of these potential biases, it is essential to test the measurement invariance of the political trust items beforehand. The goal is to determine whether and to what extent the proposed measurement model matches the observed structure of the data, thereby supporting the assumption that political trust can be measured across countries by a

common set of items using the same number of latent factors (Milfont & Fischer, 2010, p. 112). If measurement invariance is not tested beforehand, comparisons of observed differences in means may not reflect actual differences in people's average level of political trust and regression coefficients may suggest false relationships. In addition, true country-specific or temporal differences may be obscured (Chen, 2008, p. 1005). Either way, using political trust indices without testing for measurement invariance may lead to invalid conclusions regarding the changes, sources, and consequences of political trust (Ariely & Davidov, 2012, p. 364; Vandenberg & Lance, 2000, p. 9).

The lack of a common measurement model of political trust complicates such a test. First, there is no common set of political trust items and second, there is no agreement on the dimensionality of political trust.[68] This is best exemplified by previous cross-country exploratory studies (see Table 4.1). They reach different conclusions regarding the dimensionality of political trust depending on the estimation method and specifications, the design (pooled or country-specific), and the items used. This lack of consensus hampers valid comparisons.

Recently, several researchers tested the measurement invariance of political trust in European and former Soviet countries by means of multiple group confirmatory factor analysis. This method provides a stringent test because every element of the measurement model (not just the number of factors) is specified beforehand and the model outputs allow researchers to discern the reasons for invariance in detail (Brown, 2006, pp. 49–50). The studies tested and supported different dimensional models of political trust. Whereas some show that it is a single-dimensional construct, others provide evidence that a two-dimensional model of political trust in representative and implementing institutions reaches different levels of measurement invariance, depending on the countries of analysis and the chosen items (see Table 4.2).

Given these diverging measures and results, the question of the appropriate measurement model of political trust remains subject to debate. In addition, previous measurement invariance tests of political trust have focused on European and former Soviet countries, neglecting Asia, Africa, and Latin America. The purpose of this article is to determine: To what extent can the measurement invariance of political trust be established across the globe and if so, based on which measurement model?

[68] The issue of comparability is further exacerbated by the fact that there is no uniform wording and response scale for political trust items.

Table 4.1
Previous Cross-Country Exploratory Analyses of the Dimensionality of Political Trust

Author(s)	Survey	Time point/ period	Countries	Method	Trust in							
					govern-ment	parlia-ment	politi-cians	pol. parties	civil service	courts/ legal system	the police	the army
Hooghe & Kern (2015)	ESS	2002-2010	30 (Europe)	principal component analysis (pooled across countries and time)								
Rogge & Kittel (2014)	ESS	2008, 2010	29 (Europe)	principal component analysis (separately for each time point; pooled across countries)								
Lu (2014)	Asia Barometer Survey WVS	2006-2008 2005-2007	5 (Brazil, Russia, India, China, South Africa)	principal component analysis ((1) separately for each country but pooled over time; (2) pooled across countries and time)								
D. Braun (2013) (1)	EVS/WVS	1994-1999 2005-2007 2008	14 southern, eastern and southeastern European countries	principal component analysis (pooled by country groups; separately for each time point)								
Zmerli (2013) (2)	ESS	2002, 2004, 2006	European countries (# varies by year)	principle component analysis (separately for each time point and country); specifies no. of components								

Table 4.1 (continued)

Author(s)	Survey	Time point/ period	Countries	Method	Trust in							
					govern-ment	parlia-ment	politi-cians	pol. parties	civil service	courts/ legal system	the police	the army
Marien (2011b)	ESS	2006, 2008	23 (Europe)	principal component analysis (pooled across countries and time)								
K. Newton & Zmerli (2011)	WVS	2005-2007	22 democracies	principal component analysis (pooled across countries)								
Oskarsson (2010)	ESS	2002, 2004	23 (Europe)	principal component analysis (pooled across countries)								
Rose & Mishler (2010)	NEB	1993-2004	14 post-Communist countries	principal component analysis (pooled across countries and time)								
Slomczynski & Janicka (2009)	ESS	2006	23 (Europe)	factor analysis (pooled across countries and separately per country)								
Listhaug & Ringdal (2008)	ESS	2004	24 (Europe)	factor analysis (separately per country); specify no. of factors								
Zmerli & K. Newton (2008) (3)	ESS US CID	2002-2003 2006	23 (Europe), US	principal component analysis (separately for each country)								

Table 4.1 (continued)

Author(s)	Survey	Time point/ period	Countries	Method	Trust in							
					govern-ment	parlia-ment	politi-cians	pol. parties	civil service	courts/ legal system	the police	the army
Denters et al. (2007)	CID	1999-2002	13 (Europe)	factor analysis (pooled across countries); specify no. of factors								
Zmerli, K. Newton, & Montero (2007) (4)	CID	1999-2002	13 (Europe)	principal component analysis (separately for each country)								
Lühiste (2006)	NBB	2001	3 (Latvia, Lithuania, Estonia)	principal component analysis (pooled across countries)								
Zmerli (2004) (3)	ESS	2002	21 (Europe)	principal component analysis (separately for each country)								
Fuchs et al. (2002) (5)	WVS	1995-1997	6 (Europe), US	factor analysis (pooled across countries); specify no. of factors								

Note. The analyses also include item measuring trust in (1) the press and unions; (2) European Parliament; (3) European Parliament and the UN; (4) municipal board; (5) 14 additional items measuring political support; the shading of the cells indicates the dimensional structure found in the analyses. Own compilation.

Table 4.2
Previous Multiple Group Confirmatory Factor Analyses of Political Trust

Author(s)	Survey	Time point/ period	Number of countries	Level of measurement invariance reached	Trust in										
					politicians (1)	political parties (2)	government (3)	parliament (4)	courts (5)	police (6)	regional government (7)	local government (8)	army (9)	EU parliament (10)	UN (11)
Coromina & Davidov (2013)	ESS	2002-2008	7 3	partial metric; partial scalar											
Marien (2011a)	ESS	2004 2006 2008	23 21 22	partial metric	ϕ_{21}	ϕ_{21}			ϕ_{65}	ϕ_{65}					
Marien (2017)	ESS	2012	23	partial scalar	ϕ_{21}	ϕ_{21}			ϕ_{65}	ϕ_{65}					
Schaap & Scheepers (2014)	ESS	2010	19	scalar[a]	ϕ_{21}	ϕ_{21}									
Ariely (2015)	EVS	2008	32	metric											
I. Schneider (2017)	LITS II	2010	23	metric											
I. Schneider (2017)	LITS II	2010	21	partial scalar											
I. Schneider (2017)	LITS II	2010	29	partial scalar						ϕ_{96}			ϕ_{96}		
I. Schneider (2017)	LITS II	2010	35	partial metric											
André (2014)	ESS	2008	22	partial scalar	ϕ_{21}	ϕ_{21}			ϕ_{65}	ϕ_{65}				ϕ_{1110}	ϕ_{1110}

Note. The shading of the cells shows the factor structure of political trust in the model tested by the author(s). ϕ indicates an error covariance between the respective items. For example, ϕ_{96} indicates an error covariance between trust in the army and trust in the police. [a] the analysis focused on the invariance of trust in the police. Own compilation.

The study extends previous analyses in several ways. First, it tests the measurement invariance of political trust on a global scale in 32 electoral and liberal democracies. Second, the analysis provides a detailed debate and conclusion regarding the dimensionality of the construct of political trust. Third, it discusses the suitability of the available items for cross-national comparisons in detail. Overall, the article's conclusions and recommendations can be used to inform future cross-national studies of political trust.

Since "any equivalence procedure can only be implemented successfully if an unambiguous specification of the concept is available" (van Deth, 2013, p. XXI), the article begins by defining political trust and by outlining three competing dimensional models of political trust. The subsequent section describes the research design and the three alternative measurement models of political trust that follow from the dimensional models. In the analysis section, the measurement invariance test of political trust is presented. The article concludes by outlining the implications of the findings and recommendations for the comparative study of political trust.

4.2 Competing Dimensional Models of Political Trust

Political trust can be defined as people's positive anticipatory expectation that, despite uncertainty, the conduct of the political trustee in question will be in line with their normative expectations (A. Miller & Listhaug, 1990, p. 358; Möllering, 2006, p. 356).[69] Researchers generally agree that trust in different political trustees such as parliament, the judiciary, and government can be distinguished theoretically (Levi & Stoker, 2000, p. 497). They disagree on the empirical dimensionality of citizens' construct of political trust, though, resulting in three competing dimensional models.

The first dimensional model proposes a distinction between trust in political authorities and trust in political institutions. Building on Easton's (1975) classic model of political support, several researchers advocate that the two are related but separate dimensions of political trust (Dalton, 2004, pp. 5–7, 24; Denters et al., 2007, p. 68; Norris, 2011,

[69] To date, there is no commonly accepted definition of political trust. Some conceptualize it as a kind of supportive behavior (Fisher, van Heerde, & Tucker, 2010, p. 162) whereas others regard it as an attitude (A. Miller & Listhaug, 1990, p. 358). Relatedly, the elements of the definitions of political trust that they stipulate do not coincide. Furthermore, some researchers state that the term 'trust' can 'travel' to political institutions without over-stretching its conceptual core (Fuchs et al., 2002, p. 430). Others maintain that 'trust' in political institutions should be referred to as 'confidence' (Hardin, 2000, p. 31).

pp. 23–31). First and foremost, they assume that people perceive abstract and specific trustees separately: Abstract political institutions are characterized by rules that define relationships among political roles, thereby prescribing and constraining the interactions of political actors in general over time; specific political incumbents enact and interpret these roles within a particular period of time (March & Olsen, 1989, pp. 23–24). Second and consequently, while people may not trust the current political incumbents, they do not necessarily doubt that the conduct of the political institution in question will be in line with their normative expectations once the incumbents are no longer in office. At the same time, the two dimensions are related because incumbents affect the perception of the institutions. Proponents of this dimensional model assert that the distinction should be maintained all the same because it may yield more valid insights on the changes, sources, and consequences of political trust (Dalton, 2004, p. 7; Norris, 2011, pp. 43–46).

According to the second dimensional model, the distinction between trust in representative and implementing political institutions is more plausible. Several researchers assume that citizens' political trust has two dimensions because people broadly categorize the responsibilities and characteristics of the work of political institutions into two groups. On the one hand, representative political institutions such as political parties, government, and parliament serve to make collectively binding decisions. By and large, their work is characterized by political controversies and competition. On the other hand, implementing political institutions such as the courts and police are responsible for maintaining order and implementing the law. On the whole, political partisanship is less prominent in their daily work (Gabriel, 1999, pp. 205–207; G. Pickel & Walz, 1995, p. 146; Rothstein & Stolle, 2003, pp. 193–195). Within this group of researchers, there is disagreement regarding the attribution of trust in civil services, though. According to some, it is affected by people's overall trust in implementing political institutions as civil services serve to enact government policies (Gabriel, 1999, pp. 206–207). According to others, civil service officials may be perceived as agents of government precisely because they implement its laws, thereby politicizing the perception of the trustee (Rothstein & Stolle, 2008, pp. 444–445). This in turn may cause people to attribute it to their overall trust in representative political institutions. Leaving aside these differences, proponents of this two-dimensional model generally argue that trust in representative and implementing political institutions is

related because the latter act on the basis of laws that were drafted and adopted by the former (Fuchs et al., 2002, p. 439).

Still others have proposed a third, single-dimensional model of political trust. Some state that it especially applies to citizens in newly established democracies who have not had sufficient experience to distinguish between representative and implementing political institutions (Mishler & Rose, 1994, p. 25). Others maintain that this model also holds in established democracies. This may be because individuals learn to trust at an early age and generalize this socialization experience to the political realm. People's generalized trust attitude is assumed to 'spill up' to political institutions (Mishler & Rose, 2001b, p. 34). Another line of argument suggests that political trust is "a comprehensive assessment of the political culture that is prevalent within a political system" (Hooghe 2011, p. 275). As a system characteristic, political culture is assumed to impact political actors and institutions alike. As a result, people evaluate political objects and form political trust 'en bloc'. Therefore people are expected to trust political trustees to a greater or lesser extent without making more fine-grained distinctions.

These competing dimensional models suggest three alternative measurement models of political trust for the measurement invariance test. Depending on the dimensional model, the number of latent factors as well as the relational structure between the latent factors and observed items of political trust differ. These dimensional models were therefore translated into measurement models for the analysis.

4.3 Research Design

4.3.1 Operationalization

The analysis of the measurement invariance of political trust is based on data from the most recent wave of the World Values Survey (WVS). The WVS is the largest non-commercial, cross-national, time-series survey of public opinion and value preferences. Its most recent wave (wave 6, 2010-2014) covers 57 countries around the world and includes a number of items measuring trust in different political trustees, thereby permitting a measurement invariance test of political trust across the globe (World Values Survey, 2017). Since there is no common set of political trust items, the items that were used most frequently in previous studies of the dimensionality of political trust were selected from those available in the WVS (see Tables 4.1 and 4.2): trust in the police, the

courts, the government, political parties, parliament, and civil service. The items are measured on an ordinal scale with four response categories. For each of the political trustees, WVS respondents were asked to indicate "how much confidence [they] have in that organization: a great deal of confidence, quite a lot of confidence, not very much confidence, or none at all". The same items were administered to the respondents in the respective national languages. This reduces the chance that the measurement invariance test reflects differences in item-wording rather than actual differences in respondents' construct of political trust across countries. The original data were recoded to include only one kind of missing value and to range from 0 (none at all) to 3 (a great deal of trust).

4.3.2 Case Selection

The study analyzed the measurement invariance of political trust in electoral and liberal democracies. Non-democratic states were excluded because citizens' relationship with and the functional interaction of political trustees such as government and the courts differ in these countries. These differences may impact the way the construct of political trust develops in people's minds in democracies and non-democracies (Mishler & Rose, 1997, p. 420).[70] This assumption is substantiated by I. Schneider's (2017) as well as Schaap and Scheepers' (2014) analysis of the measurement invariance of political trust in European and former Soviet countries. They found that a greater level of measurement invariance could be established once former Soviet autocracies were excluded from the analysis. The study at hand therefore focused on democracies in order to eliminate this possible source of measurement non-equivalence.

The countries included in the study were selected based on Polity IV (Center for Systemic Peace, 2016). Polity IV comprises indicators of institutional autocracy and democracy (Marshall, Gurr, & Jaggers, 2015, pp. 13–18). Countries' polity score can

[70] As Breustedt and Stark (2015, pp. 189–190) argue, in authoritarian countries it is difficult for citizens to distinguish political institutions because of the lack of a system of checks and balances. In addition, as elections are infrequent or inconsequential, political institutions become mainly associated with the political incumbents. Therefore, people in authoritarian states most likely develop their trust in different political trustees in tandem. According to Rivetti and Cavatorta (2017), political trust in democratic regimes is positive whereas in authoritarian regimes it is negative: "whereas positive political trust can be defined as trust in ethical, legal or just actions undertaken by the ruling authority, negative trust can be defined as trust in the fact that the authority will act predictably" (Rivetti & Cavatorta, 2017, p. 60). Still, political trust in authoritarian countries is not necessarily devoid of positive normative expectations. People's normative expectations of political trustees may simply differ in authoritarian countries. Either way, measures of political trust in democracies and autocracies are not likely to be equivalent as responses to the same items are susceptible to construct bias.

range from -10 (fully autocratic) to +10 (fully democratic). In line with the threshold provided on the Polity IV website (Marshall & Gurr, 2014), countries were included if their polity score was six or higher in the year the survey was conducted as well as four years prior to this year.

The final sample consisted of 32 countries with 46,315 respondents. The selected countries as well as the sample sizes and missings per item are listed in Table B1 in Appendix B.[71] The survey samples are representative of the countries' adult population (World Values Survey, 2017).

4.3.3 Method

The measurement invariance (MI) of political trust was tested using multiple group confirmatory factor analysis (MGCFA). Alternative methods include item response theory and latent class analysis (Davidov et al., 2014, p. 62; Kankaraš et al., 2011; Millsap, 2011). The study used MGCFA because it is a widely applied method to test MI and because previous studies of the MI of political trust used this method.

The analysis was conducted in three stages. Because there is no agreed upon measurement model of political trust, first, confirmatory factor analysis (CFA) was used to determine the model fit of the three alternative models derived from the dimensional models outlined above in each of the 32 countries. The best-fitting model served as the baseline model in the second step, the simultaneous analysis of MI across countries by means of MGCFA. Based on these empirical results as well as theoretical considerations, in the third step, this measurement model was revised and subsequently tested for MI.

[71] Table B1 in Appendix B reports the original sample sizes. Most items have less than 5% missing per country. Two issues stand out: Trust in civil service has > 5% missing in nine countries, 18.4% of the cases for trust in government are missing in Lebanon, and Japan is the country with the largest amount of missing data. Cases were dropped if they had missings on all six items for the analysis. Respondents from the WVS wave 6 survey in India, conducted in 2012, were excluded because the wave 6 data file also includes a more recent Indian survey sample from 2014. 'Pairwise present' was used to handle missing data (Asparouhov & B. Muthén, 2010, p. 7).

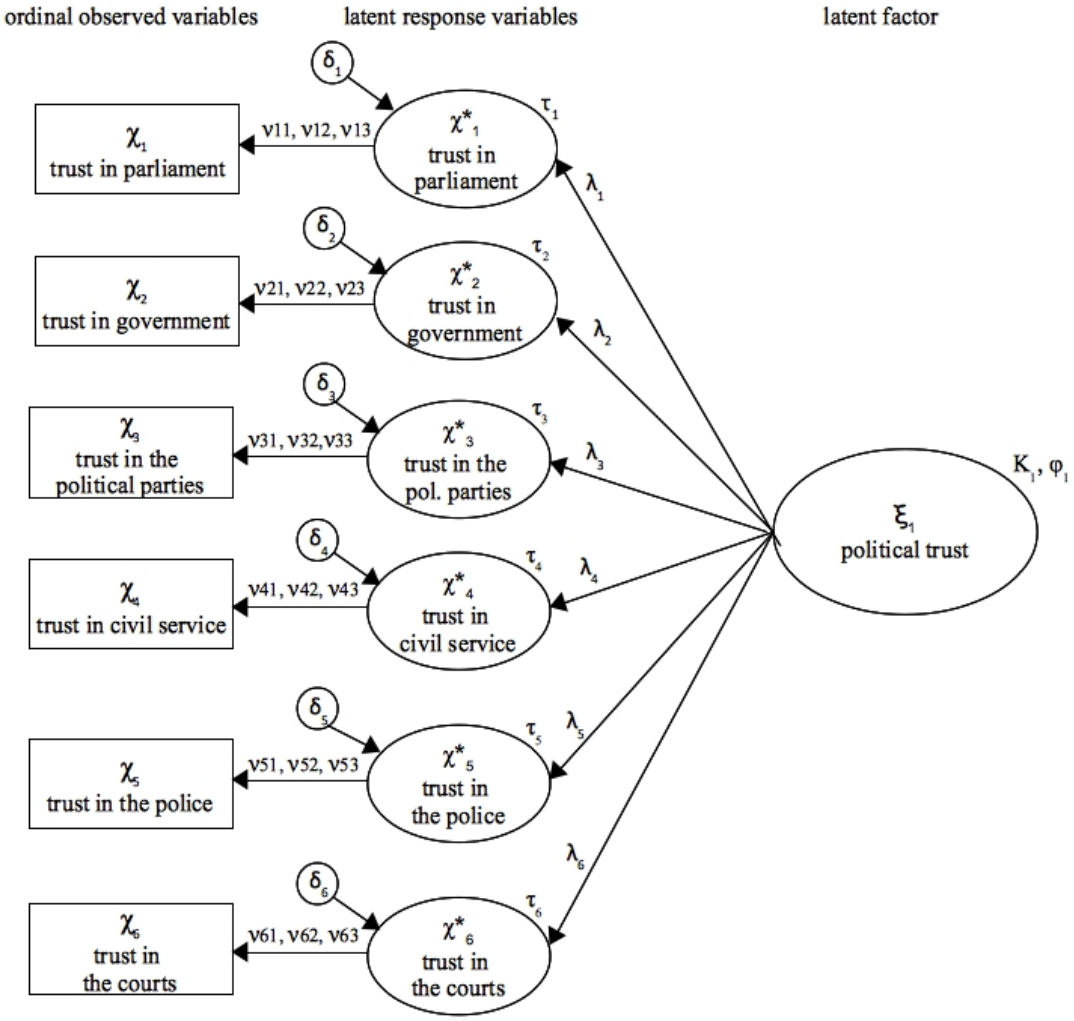


Figure 4.1. Single-Dimensional Measurement Model of Political Trust. Adapted from Davidov et al. (2011) and Poznyak et al. (2014). ξ (ksi): latent factor, κ (kappa): latent mean, φ (phi): factor variance, λ (lambda): factor loading, χ* (chi): latent response variable, τ (tau): intercept, δ (delta): error variance, χ (chi): observed variable, ν (nu): threshold.

Consonant with the three dimensional models described earlier, three measurement models were developed as possible baseline models for the MI test (see Figures 4.1 to 4.3).[72] Civil service was specified to load on trust in representative institutions in line with previous exploratory analyses (see Table 4.1). None of the models included any error correlations. In the two-dimensional models, the latent factors were assumed to correlate.

[72] Some researchers have distinguished between trust in political actors, representative political institutions, and implementing political institutions (Denters et al., 2007, p. 68; Gabriel, 1999, pp. 205–207). This three-dimensional model could not be tested because of the limited number of survey items available in the WVS.

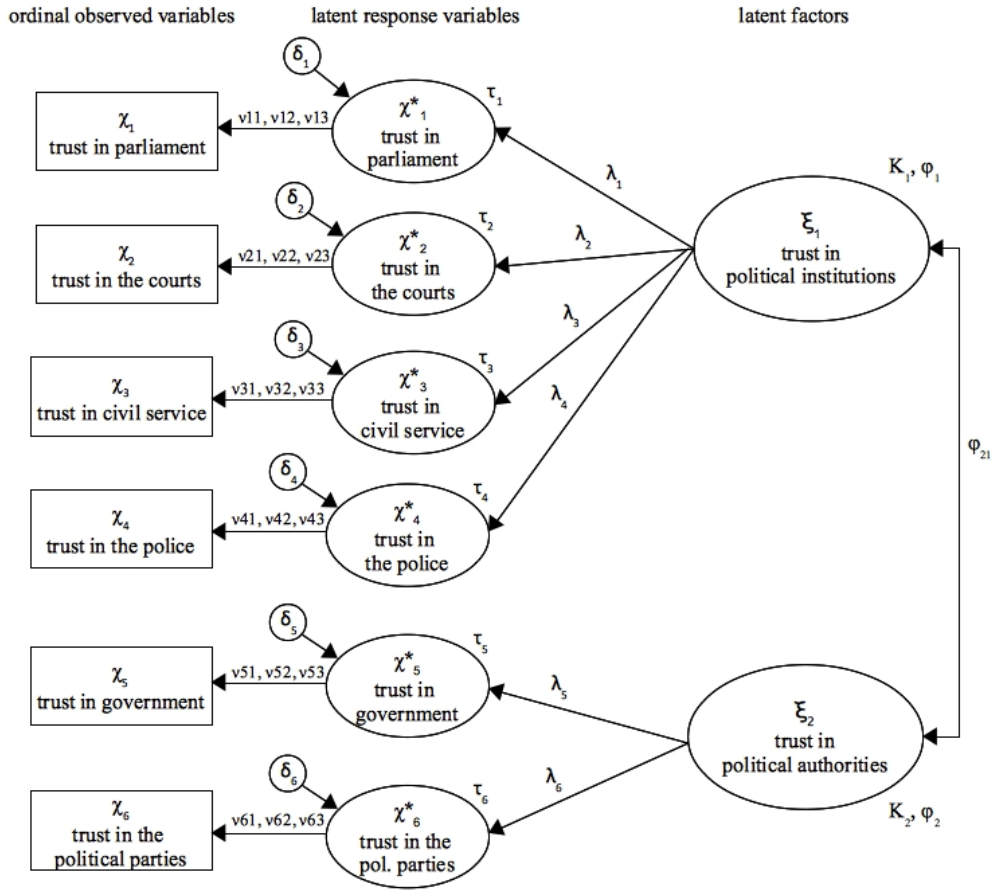


Figure 4.2. Two-Dimensional Measurement Model of Trust in Political Authorities and Political Institutions. Adapted from Davidov et al. (2011) and Poznyak et al. (2014). ξ (ksi): latent factor, κ (kappa): latent mean, φ (phi): factor variance, λ (lambda): factor loading, χ* (chi): latent response variable, τ (tau): intercept, δ (delta): error variance, χ (chi): observed variable, ν (nu): threshold.

The study took account of the ordinal measurement scale of the items. Lubke and B. Muthén (2004) have shown that treating ordered-categorical data as continuous may yield estimates that suggest that the factor structure found in different countries differs when, in fact, it is equivalent. To circumvent this issue, the study followed a common approach to estimate latent variable models for ordered-categorical items – the latent response variable model (B. Muthén & Asparouhov, 2002).

This approach is outlined briefly as it affects the way MI tests are conducted. As indicated in Figures 4.1 to 4.3, the model estimation based on the latent response variable model assumes that the latent factor(s) of political trust (ξ_i) cause(s) the variance and covariance among latent response variables of political trust in six different political trustees (χ^*_i). The latent response variables are taken to have a continuous and normally distributed scale. Their relationship with the latent factor(s) is understood to be linear.

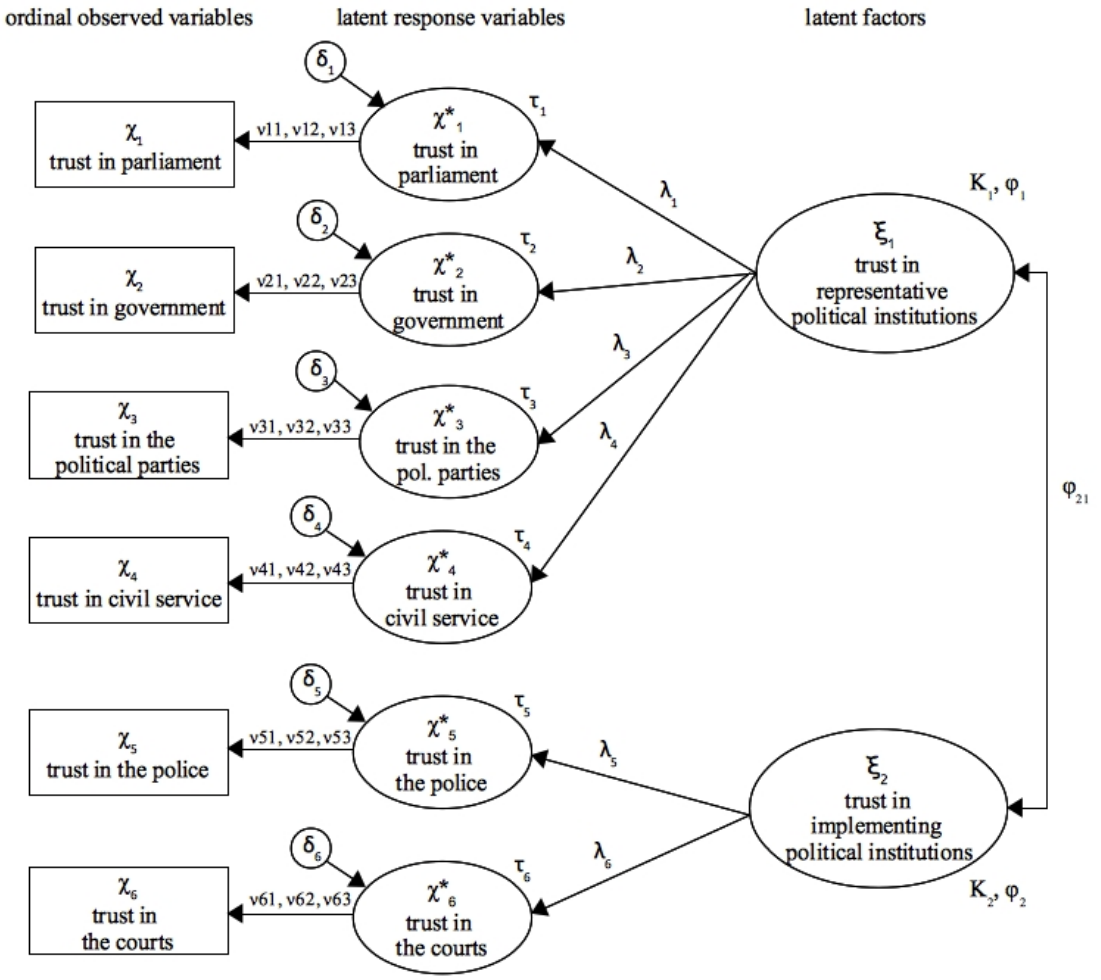


Figure 4.3. Two-Dimensional Measurement Model of Trust in Representative and Implementing Political Institutions. Adapted from Davidov et al. (2011) and Poznyak et al. (2014). ξ (ksi): latent factor, κ (kappa): latent mean, φ (phi): factor variance, λ (lambda): factor loading, χ* (chi): latent response variable, τ (tau): intercept, δ (delta): error variance, χ (chi): observed variable, ν (nu): threshold.

Thus, as in standard MGCFA with continuous items, each latent response variable has a factor loading (λ_i), an intercept (τ_i), and an error term. The latent response variables are assumed to be the unobserved latent counterparts of the observed ordered-categorical items of political trust (χ_i). The continuous nature of the latent response variables is roughly captured by the ordered-categorical response scale of the respective observed items. Each pair of response categories of the items represents a section of the continuous scale of the corresponding latent response variable. Each section therefore ends with a threshold (ν_{ij}). As a result, each latent response variable is related to its corresponding observed item through a set of thresholds, whereby the number of thresholds corresponds to the number of response categories minus one. Since the political trust items have four ordered response categories, the latent response variables each have three thresholds.

That is to say, if χ_1 represents the ordinal item of trust in parliament and χ^*_1 stands for the latent response variable of trust in parliament, χ^*_1 reflects the amount of political trust needed to select a certain response category of χ_1. An observed response of '0' (none at all) in trust in parliament is expected if the level of χ^*_1 is less than or equal to the first threshold ν_{11}. If χ^*_1 is greater than ν_{11} but less than or equal to the second threshold ν_{12}, the predicted response is '1' (not very much confidence). If the latent response variable of trust in parliament χ^*_1 is greater than ν_{12} but less than or equal to the third threshold ν_{13}, the predicted response is '2' (quite a lot of confidence). $\chi^*_1 > \nu_{13}$ corresponds to a response of '3' (a great deal of confidence) (Byrne, 2012, pp. 126–132; Kline, 2016, pp. 324–325; Millsap & Yun-Tein, 2004, pp. 480–483; B. Muthén & Asparouhov, 2002, pp. 3–4).

Accounting for the ordinal nature of the political trust items affects the parameters that have to be invariant across countries in order for MI to hold and, relatedly, the levels of MI that can be tested. The invariance of factor loadings, intercepts, and (unlike in the case of continuous variables) thresholds has to be considered (Davidov et al., 2011, pp. 159–161; Millsap & Yun-Tein, 2004, p. 484). Researchers can test to what extent these parameters are invariant by applying increasingly restrictive equality constraints in MGCFA and examining the respective model fit by means of goodness-of-fit indices. In the case of ordered-categorical data, only two levels of MI are tested, namely configural and full MI (Davidov et al., 2011, p. 161). When testing for configural invariance, the estimated parameters are allowed to differ across countries. The test shows whether the number of factors and the pattern of fixed and free item factor loadings is the same across countries (Vandenberg & Lance, 2000, pp. 36–37). If this model fits the data, it may be inferred that people in different countries respond to political trust items with the same construct in mind (Chen, 2008, p. 1006). If not, country-specific measures may be required (Pendergast, von der Embse, Kilgus, & Eklund, 2017, p. 5). Configural invariance is a prerequisite for full MI. Full MI requires the unstandardized factor loadings, intercepts, and thresholds to be equal (Davidov et al., 2011, p. 161). If full MI is supported by the data, it can be inferred that the items measure the same latent construct, albeit with different degrees of precision because the error variances and covariances were not constrained to be equal (Kline, 2016, p. 413). In addition, full MI

implies that people in the respective countries use the response scale in the same manner (Poznyak, Meulemann, Abts, & Bishop, 2014, p. 746).[73]

The ordered-categorical nature of the data has a bearing on the appropriate choice of the method of estimation. As Brown (2006, p. 379) notes, ignoring the fact that the data may be non-normally distributed could lead to incorrect parameter estimates, standard errors, and test statistics. The analyses were therefore run with the mean- and variance-adjusted weighted least squares (WLSMV) estimator in Mplus (Version 8) using the raw data. This estimator provides robust standard errors and (more) accurate estimates of factor loadings as well as corrected model test statistics. As Beauducel and Herzberg (2006) showed, it is superior to maximum likelihood estimation especially when the number of response categories is small, as in the case of the present study.

In order to conduct MI analyses, the scale of the latent factors has to be defined. Because latent factors are unobserved, they have no definite metric scale. In MGCFA, there are two common ways to establish this scale – the reference indicator method and the fixed factor method. When using the latter, the factor variances of the latent factors are fixed to one in all countries. This assumes that the factor variances are equal across countries. When applying the former, one factor loading per latent factor is fixed to one in all countries. Here the assumption is that this factor loading is invariant (Byrne, 2012, p. 33). With regard to political trust, there is no evidence to justify either assumption. In this study, the reference indicator method was used because it was more straightforward to make a case for using single reference indicators.[74]

---

[73] Unlike in the case of continuous data, the invariance of factor loadings alone does not establish comparability of the political trust measure because the item probability curves depend on the factor loadings, intercepts, and thresholds (Davidov et al., 2011, p. 161; B. Muthén & Asparouhov, 2002, p. 10). As a result, only two levels of measurement invariance were tested unlike in previous measurement invariance tests of political trust (Table 4.2). See Bowen and Masa (2015) for a summary of arguments in favor and against this practice.

[74] In order to choose appropriate reference indicators, two exploratory factor analyses (EFA) were carried out per country (principal axis extraction; promax rotation). In the single-factor EFA, trust in parliament was the marker item in 22 out of 32 countries. In the two-factor EFA, in 28 out of 32 countries, trust in parliament was the item that loaded most strongly on one latent factor and in 17 out of 32 countries, trust in the police was the marker item of the other latent factor. Consequently, trust in parliament was used as the reference indicator in the single-dimensional model and trust in parliament as well as trust in the police were used as reference indicators in the two-dimensional model of trust in implementing and representative institutions. Trust in parliament and trust in government were used as reference indicators in the two-dimensional model of trust in political authorities and institutions. Trust in government was chosen because the author deemed it more likely that government is perceived in a comparable manner across countries compared to political parties because its structure and functions are more similar, differences notwithstanding. Table B2 in Appendix B includes a robustness test for Model A of the MGCFA (see Table 4.7). The analysis was not sensitive to the selection of these reference indicators.

Depending on the level of MI tested, additional parameters have to be fixed in order for the measurement model to be identified. The choice depends in part on the computer program and the model parameterization. Mplus was chosen because of its flexibility when testing the invariance of ordered-categorical items (Millsap & Yun-Tein, 2004, p. 498). In practice, thresholds (ν_i) and intercepts (τ_i) cannot be estimated simultaneously. By default, Mplus fixes all intercepts of the latent response variables to zero, thereby allowing researchers to test the MI of thresholds (Davidov et al., 2011, p. 161). In addition, Mplus offers two parameterization methods – delta and theta parameterization. Unlike delta parameterization, theta parameterization includes error variances for the latent response variables (δ) as estimated parameters (L. Muthén & B. Muthén, 1998-2017, p. 77). This study used theta parameterization as previous MGCFAs (see Table 4.2) indicated that the error variances of some of the items might be correlated. In order to identify the measurement models, the following parameters were fixed. In the configural invariance model, one factor loading per latent factor as well as the error variances were fixed to one and the factor means were fixed to zero in all countries. In the full MI model, one factor loading per latent factor was fixed to one in all countries and the remaining factor loadings as well as the thresholds were constrained to be equal. In addition, the error variances were fixed to one and the factor means were fixed to zero in the reference country[75] and freely estimated in the other countries (L. Muthén & B. Muthén, 1998-2017, p. 77).

The overall fit of the measurement models to the data was evaluated according to several criteria. X^2 as the classic fit index indicates exact fit between the estimated model parameters and the observed data. While this is informative, it is an unduly strong assumption for real-world data. In addition, X^2 is sensitive to sample size (Byrne, 2012, pp. 66–67; Meade, Johnson, & Braddy, 2008, p. 568). Consequently, the goodness of fit evaluation was informed by the X^2 results but focused on three additional fit indices: the root mean square error of approximation (RMSEA), the comparative fit index (CFI), and the Tucker-Lewis-Index (TLI). The 90% confidence interval of the RMSEA is provided to show how precise its point estimates are (MacCallum, Browne, & Sugawara, 1996, p. 130). Following Yu (2002, pp. 160–161), the following cut-off criteria were used: TLI $\geq$0.95, CFI $\geq$0.96, and RMSEA $\leq$ 0.05.

[75] Model C2: Australia; Model C3: Poland.

The analysis also considered focal areas of ill fit. The proportion of variance of the indicator explained by the latent factor ('R-Square' in Mplus) was used to evaluate whether the items were meaningfully related to the respective latent factor. The extent of the correlation between the latent factors was taken into account to determine discriminant validity between the latent factors in case of the two-dimensional models of political trust (Brown, 2006, p. 166). In addition, the study followed a dual modal two-pronged strategy proposed by Byrne and van de Vijver (2010, pp. 113–114). They suggest looking for patterns of misspecification that indicate that individual items, individual countries, or groups of countries are the reason for measurement non-invariance. Modification indices, which approximate how much the model fit (X^2) would improve if the constrained or fixed parameter in question was freely estimated, can be used to discern such patterns (Brown, 2006, pp. 119–124). Because of X^2's sensitivity to sample size, it was considered in tandem with the respective expected parameter of change (EPC) value. Overall, these criteria provided information on the fit of the measurement models as well as how to revise the measurement models in order to establish full invariance.

4.4 Analysis

4.4.1 Establishing the Baseline Model of Political Trust

The first step in testing the MI of political trust on a global scale was to establish the baseline model. Tables 4.3 to 4.5 present the overall goodness-of-fit indices for each of the three alternative measurement models tested separately in 32 countries. In terms of CFI and TLI, the two-factor model of trust in political authorities and political institutions yielded the worst fit. As shown in Table 4.3, the two indices were above the recommended cut-off value in only five out of 32 countries. The RMSEA did not support the model in any of the countries. The latent covariance matrix of the factors was not positive definite in six countries. In all six countries, this was because the latent factor correlation was estimated to have an out of range value (> 1.0), signifying model misspecification because some or all of the items of one latent factor were more strongly related to some or all of the items of the other latent factor (Brown, 2006, p. 190). In comparison, the single-factor model of political trust fit the data better (see Table 4.4). The CFI and TLI indicated good model fit in eight out of 32 countries. Finally, the two-

factor model of trust in implementing and representative political institutions fit the data best (see Table 4.5). In 28 out of 32 countries, the CFI and TLI were above the recommended cut-off values. Furthermore, only in this model was the RMSEA smaller than 0.05 in two countries and its confidence interval indicated a good precision of this point estimate.

At the same time, the inspection of focal areas of ill fit of the CFAs of the two-factor model of trust in implementing and representative political institutions suggested several items and countries of concern. X^2 strongly varied across countries, ranging from 564.953 in South Korea to 25.885 in Argentina (see Table 4.5). The standardized correlation coefficient between the two latent factors was >.85 in five countries, indicating low discriminant validity (see Table 4.6). These aspects point to possible countries as a reason for measurement non-invariance. As for the items, 'trust in civil service' was the item with the lowest proportion of explained variance in 21 countries (see Table 4.6). In addition, the modification and expected parameter change indices recommended a positive cross-loading between the latent factor 'trust in implementing political institutions' and the item 'trust in civil service' in 17 countries. In 13 countries, this modification index value was the largest among all suggested cross-loadings between a latent factor of political trust and a political trust item (see Table 4.6). This indicates that 'trust in civil service' is an ambiguous item not as meaningfully related to the construct of political trust as the other items. Furthermore, in 22 countries, the modification and expected parameter change indices for error co-variances pointed out that the model fit would improve if a cross-loading were added between 'trust in parliament' and 'trust in political parties'. This modification index was the largest value for suggested error correlations in nine countries (see Table 4.6). Based on these results, the two-factor model of trust in implementing and representative political institutions was chosen as the baseline model for the MGCFA. The focal areas of ill fit informed its revision for the MI test across countries.

Table 4.3
Fit Measures for the Two-Factor Confirmatory Factor Analysis of Trust in Political Authorities and Political Institutions

Country	n	χ^2 (df)	p-value	CFI	TLI	RMSEA (90% CI)	sum-mary
all countries	46315	17403.165 (8)	0.00	0.953	0.912	0.217 (0.214-0.219)	
Argentina	1025	330.017 (8)	0.00	0.956	0.917	0.198 (0.180-0.217)	
Australia	1453	336.644 (8)	0.00	**0.966**	0.936	0.168 (0.153-0.184)	
Brazil	1486	the latent variable covariance matrix is not positive definite					
Chile	999	the latent variable covariance matrix is not positive definite					
Colombia	1509	the latent variable covariance matrix is not positive definite					
Cyprus	999	437.876 (8)	0.00	0.941	0.890	0.232 (0.214-0.251)	
Estonia	1531	781.502 (8)	0.00	0.948	0.902	0.251 (0.237-0.266)	
Georgia	1185	759.328 (8)	0.00	**0.965**	0.935	0.282 (0.265-0.299)	
Germany	2043	715.828 (8)	0.00	**0.960**	0.925	0.208 (0.195-0.221)	
Ghana	1552	the latent variable covariance matrix is not positive definite					
India	1578	149.767 (8)	0.00	0.880	0.774	0.106 (0.092-0.121)	
Japan	2350	1467.502 (8)	0.00	**0.975**	**0.954**	0.279 (0.267-0.291)	(√)
Lebanon	1183	68.742 (8)	0.00	**0.979**	**0.961**	0.080 (0.063-0.098)	(√)
Malaysia	1299	the latent variable covariance matrix is not positive definite					
Mexico	2000	410.193 (8)	0.00	**0.972**	0.947	0.159 (0.146-0.172)	
Netherlands	1849	818.027 (8)	0.00	**0.982**	**0.967**	0.234 (0.221-0.248)	(√)
New Zealand	812	236.709 (8)	0.00	**0.962**	0.930	0.188 (0.167-0.209)	
Peru	1206	291.760 (8)	0.00	**0.971**	0.945	0.171 (0.155-0.189)	
Philippines	1200	438.337 (8)	0.00	0.940	0.888	0.212 (0.195-0.229)	
Poland	957	304.620 (8)	0.00	**0.968**	0.939	0.197 (0.178-0.216)	
Romania	1488	742.378 (8)	0.00	**0.960**	0.924	0.248 (0.233-0.264)	
Slovenia	1060	298.563 (8)	0.00	**0.980**	**0.963**	0.185 (0.167-0.203)	(√)
South Africa	3477	973.607 (8)	0.00	**0.971**	0.946	0.186 (0.177-0.196)	
South Korea	1198	the latent variable covariance matrix is not positive definite					
Spain	1180	287.923 (8)	0.00	0.943	0.894	0.172 (0.155-0.190)	
Sweden	1205	516.348 (8)	0.00	0.948	0.902	0.230 (0.213-0.247)	
Taiwan	1204	224.002 (8)	0.00	**0.976**	**0.956**	0.150 (0.133-0.167)	(√)
Trinidad and Tobago	994	503.494 (8)	0.00	**0.960**	0.926	0.250 (0.231-0.268)	
Turkey	1593	528.707 (8)	0.00	0.951	0.909	0.202 (0.188-0.217)	
Ukraine	1500	934.882 (8)	0.00	**0.968**	0.941	0.278 (0.263-0.293)	
United States	2205	1429.113 (8)	0.00	0.931	0.871	0.284 (0.272-0.296)	
Uruguay	995	431.481 (8)	0.00	0.943	0.893	0.231 (0.212-0.249)	

Note. WLSMV estimator (theta parameterization), pairwise present was used to handle missing data (Asparouhov & B. Muthén, 2010, p. 7), df = degrees of freedom, CFI = comparative fit index, TLI = Tucker-Lewis-Index, RMSEA = root mean square error of approximation, 90% CI = 90% confidence interval, parameter of fit values above the recommended thresholds (Yu, 2002, pp. 160–161) are in bold, summary (√) indicates that two out of three fit indices are above the recommended thresholds, summary √ indicates that CFI, TLI, and RMSEA are above the recommended thresholds. Data are from the World Values Survey 2010-2014, 32 countries.

Table 4.4
Fit Measures for the Single-Factor Confirmatory Factor Analysis of Political Trust

Country	n	χ^2 (df)	p-value	CFI	TLI	RMSEA (90% CI)	sum-mary
all countries	46315	18131.958 (9)	0.00	0.951	0.919	0.209 (0.206-0.211)	
Argentina	1025	339.428 (9)	0.00	0.954	0.924	0.189 (0.172-0.207)	
Australia	1453	342.404 (9)	0.00	**0.965**	0.942	0.160 (0.145-0.174)	
Brazil	1486	467.487 (9)	0.00	0.947	0.911	0.185 (0.171-0.200)	
Chile	999	194.345 (9)	0.00	**0.977**	**0.962**	0.144 (0.126-0.161)	(√)
Colombia	1509	603.427 (9)	0.00	0.951	0.919	0.209 (0.195-0.224)	
Cyprus	999	478.871 (9)	0.00	0.936	0.893	0.229 (0.211-0.246)	
Estonia	1531	803.514 (9)	0.00	0.946	0.911	0.240 (0.226-0.254)	
Georgia	1185	804.307 (9)	0.00	**0.963**	0.938	0.273 (0.257-0.289)	
Germany	2043	739.886 (9)	0.00	0.959	0.931	0.199 (0.187-0.212)	
Ghana	1552	519.222 (9)	0.00	0.931	0.885	0.191 (0.177-0.205)	
India	1578	158.753 (9)	0.00	0.873	0.788	0.103 (0.089-0.117)	
Japan	2350	1593.134 (9)	0.00	**0.973**	**0.956**	0.274 (0.262-0.285)	(√)
Lebanon	1183	81.557 (9)	0.00	**0.975**	**0.959**	0.083 (0.067-0.099)	(√)
Malaysia	1299	878.559 (9)	0.00	0.955	0.925	0.273 (0.258-0.288)	
Mexico	2000	411.296 (9)	0.00	**0.972**	**0.953**	0.149 (0.137-0.162)	(√)
Netherlands	1849	891.088 (9)	0.00	**0.981**	**0.968**	0.230 (0.218-0.243)	(√)
New Zealand	812	245.580 (9)	0.00	**0.961**	0.935	0.180 (0.161-0.200)	
Peru	1206	294.694 (9)	0.00	**0.971**	**0.951**	0.162 (0.147-0.178)	(√)
Philippines	1200	437.427 (9)	0.00	0.940	0.901	0.199 (0.183-0.215)	
Poland	957	319.692 (9)	0.00	**0.966**	0.944	0.190 (0.172-0.208)	
Romania	1488	768.958 (9)	0.00	0.958	0.930	0.238 (0.224-0.253)	
Slovenia	1060	339.944 (9)	0.00	**0.978**	**0.963**	0.186 (0.170-0.203)	(√)
South Africa	3477	1041.826 (9)	0.00	**0.969**	0.949	0.182 (0.172-0.191)	
South Korea	1198	814.982 (9)	0.00	**0.964**	0.940	0.273 (0.258-0.289)	
Spain	1180	395.232 (9)	0.00	0.922	0.870	0.191 (0.175-0.207)	
Sweden	1205	546.657 (9)	0.00	0.945	0.908	0.223 (0.207-0.239)	
Taiwan	1204	222.983 (9)	0.00	**0.977**	**0.961**	0.141 (0.125-0.157)	(√)
Trinidad and Tobago	994	546.575 (9)	0.00	0.957	0.928	0.245 (0.228-0.263)	
Turkey	1593	570.242 (9)	0.00	0.948	0.913	0.198 (0.184-0.212)	
Ukraine	1500	1003.718 (9)	0.00	**0.966**	0.943	0.271 (0.257-0.286)	
United States	2205	1479.265 (9)	0.00	0.929	0.882	0.272 (0.261-0.284)	
Uruguay	995	442.719 (9)	0.00	0.942	0.903	0.220 (0.203-0.238)	

Note. WLSMV estimator (theta parameterization), pairwise present was used to handle missing data (Asparouhov & B. Muthén, 2010, p. 7), df = degrees of freedom, CFI = comparative fit index, TLI = Tucker-Lewis-Index, RMSEA = root mean square error of approximation, 90% CI = 90% confidence interval, parameter of fit values above the recommended thresholds (Yu, 2002, pp. 160–161) are in bold, summary (√) indicates that two out of three fit indices are above the recommended thresholds, summary √ indicates that CFI, TLI, and RMSEA are above the recommended thresholds. Data are from the World Values Survey 2010-2014, 32 countries.

Table 4.5
Fit Measures for the Two-Factor Confirmatory Factor Analysis of Political Trust in Implementing and Representative Political Institutions

Country	n	χ^2 (df)	p-value	CFI	TLI	RMSEA (90% CI)	sum-mary
all countries	46315	4004.959 (8)	0.000	**0.989**	**0.980**	0.104 (0.101-0.107)	(√)
Argentina	1025	25.885 (8)	0.001	**0.998**	**0.995**	**0.047** (0.027-0.067)	√
Australia	1453	149.490 (8)	0.00	**0.985**	**0.972**	0.110 (0.095-0.126)	(√)
Brazil	1486	278.099 (8)	0.00	**0.969**	0.941	0.151 (0.136-0.166)	
Chile	999	195.118 (8)	0.00	**0.977**	**0.956**	0.153 (0.135-0.172)	(√)
Colombia	1509	522.132 (8)	0.00	0.958	0.921	0.206 (0.192-0.222)	
Cyprus	999	82.736 (8)	0.00	**0.990**	**0.981**	0.097 (0.078-0.116)	(√)
Estonia	1531	221.914 (8)	0.00	**0.986**	**0.973**	0.132 (0.117-0.147)	(√)
Georgia	1185	316.563 (8)	0.00	**0.986**	**0.973**	0.180 (0.164-0.198)	(√)
Germany	2043	128.285 (8)	0.00	**0.993**	**0.987**	0.086 (0.073-0.099)	(√)
Ghana	1552	168.182 (8)	0.00	**0.978**	**0.960**	0.114 (0.099-0.129)	(√)
India	1578	129.277 (8)	0.00	0.897	0.807	0.098 (0.084-0.113)	
Japan	2350	117.045 (8)	0.00	**0.998**	**0.997**	0.076 (0.064-0.089)	(√)
Lebanon	1183	28.580 (8)	0.00	**0.993**	**0.987**	**0.047** (0.029-0.066)	√
Malaysia	1299	556.899 (8)	0.00	**0.972**	0.947	0.230 (0.214-0.246)	
Mexico	2000	211.765 (8)	0.00	**0.986**	**0.973**	0.113 (0.100-0.126)	(√)
Netherlands	1849	213.724 (8)	0.00	**0.995**	**0.992**	0.118 (0.105-0.132)	(√)
New Zealand	812	48.940 (8)	0.00	**0.993**	**0.987**	0.079 (0.059-0.101)	(√)
Peru	1206	102.030 (8)	0.00	**0.990**	**0.982**	0.099 (0.082-0.116)	(√)
Philippines	1200	187.409 (8)	0.00	**0.975**	**0.953**	0.137 (0.120-0.154)	(√)
Poland	957	96.655 (8)	0.00	**0.990**	**0.982**	0.108 (0.089-0.127)	(√)
Romania	1488	195.538 (8)	0.00	**0.990**	**0.981**	0.126 (0.111-0.141)	(√)
Slovenia	1060	56.482 (8)	0.00	**0.997**	**0.994**	0.076 (0.058-0.095)	(√)
South Africa	3477	467.079 (8)	0.00	**0.986**	**0.975**	0.128 (0.119-0.139)	(√)
South Korea	1198	564.953 (8)	0.00	**0.975**	**0.953**	0.241 (0.224-0.258)	(√)
Spain	1180	156.665 (8)	0.00	**0.970**	0.944	0.125 (0.109-0.143)	
Sweden	1205	98.056 (8)	0.00	**0.991**	**0.983**	0.097 (0.080-0.114)	(√)
Taiwan	1204	112.167 (8)	0.00	**0.989**	**0.979**	0.104 (0.087-0.121)	(√)
Trinidad and Tobago	994	102.419 (8)	0.00	**0.992**	**0.986**	0.109 (0.091-0.128)	(√)
Turkey	1593	204.398 (8)	0.00	**0.982**	**0.966**	0.124 (0.110-0.139)	(√)
Ukraine	1500	108.100 (8)	0.00	**0.997**	**0.994**	0.091 (0.076-0.107)	(√)
United States	2205	537.652 (8)	0.00	**0.974**	**0.952**	0.173 (0.161-0.186)	(√)
Uruguay	995	54.921 (8)	0.00	**0.994**	**0.988**	0.077 (0.058-0.097)	(√)

Note. WLSMV estimator (theta parameterization), pairwise present was used to handle missing data (Asparouhov & B. Muthén, 2010, p. 7), df = degrees of freedom, CFI = comparative fit index, TLI = Tucker-Lewis-Index, RMSEA = root mean square error of approximation, 90% CI = 90% confidence interval, parameter of fit values above the recommended thresholds (Yu, 2002, pp. 160–161) are in bold, summary (√) indicates that two out of three fit indices are above the recommended thresholds, summary √ indicates that CFI, TLI, and RMSEA are above the recommended thresholds. Data are from the World Values Survey 2010-2014, 32 countries.

Table 4.6
Focal Areas of Ill Fit in the Two-Factor Confirmatory Factor Analysis of Political Trust in Implementing and Representative Political Institutions per Country

Country	Focal areas of ill fit			
	Political trust item with the lowest explained variance	Largest modification index for cross-loadings between a latent factor of political trust and a political trust item*	Largest modification index for error correlation*	Factor correlation >.85
Argentina	police	---	parliament and police (neg.)	
Australia	police	**ξ2 and civil service**	**parliament and political parties**	
Brazil	**civil service**	ξ2 and government	**parliament and political parties**	
Chile	police	ξ2 and government	government and police	x
Colombia	police	ξ2 and government	government and court	x
Cyprus	**civil service**	ξ2 and government	civil service and parliament	
Estonia	**civil service**	**ξ2 and civil service**	**parliament and political parties**	
Georgia	**civil service**	**ξ2 and civil service**	civil service and police	
Germany	**civil service**	**ξ2 and civil service**	civil service and police	
Ghana	**civil service** and political parties	ξ2 and government	government and court	
India	political parties	ξ2 and government	**parliament and political parties**	
Japan	court	**ξ2 and civil service**	political parties and government	
Lebanon	**civil service**	---	parliament and government (neg.)	
Malaysia	**civil service**	ξ2 and government	**parliament and political parties**	x
Mexico	police	ξ2 and government	government and court	
Netherlands	**civil service**	**ξ2 and civil service**	civil service and police	
New Zealand	**civil service**	ξ2 and political parties (neg.)	political parties and court (neg.)	
Peru	police	ξ2 and government	government and court	

Table 4.6 (continued)

Country	Focal areas of ill fit			
	Political trust item with the lowest explained variance	Largest modification index for cross-loadings between a latent factor of political trust and a political trust item*	Largest modification index for error correlation*	Factor correlation >.85
Philippines	**civil service**	ξ2 and government	government and court	
Poland	police	**ξ2 and civil service**	**parliament and political parties**	
Romania	**civil service**	**ξ2 and civil service**	civil service and police	
Slovenia	police	---	civil service and parliament	
South Africa	**civil service**	ξ2 and government	government and court	x
South Korea	**civil service**	ξ2 and government	**parliament and political parties**	
Spain	**civil service**	**ξ2 and civil service**	political parties and government	
Sweden	**civil service**	**ξ2 and civil service**	civil service and court	
Taiwan	**civil service**	ξ2 and government	**parliament and political parties**	x
Trinidad and Tobago	**civil service**	**ξ2 and civil service**	civil service and court	
Turkey	political parties	**ξ2 and civil service**	civil service and government (neg.)	
Ukraine	**civil service**	ξ2 and government	**parliament and political parties**	
United States	**civil service**	**ξ2 and civil service**	civil service and police	
Uruguay	**civil service**	ξ2 and government	government and court	

Note. **ξ2** = latent factor of trust in implementing political institutions, * positive expected parameter change unless otherwise indicated, (neg.) = expected parameter change is negative.

4.4.2 Testing the Measurement Invariance of Political Trust

Table 4.7 presents the results of the MI test of political trust in 32 democracies across the globe. Initially, the configural invariance of the baseline model was tested (Model A). While the CFI and TLI indicated good model fit, the RMSEA was well above the cut-off criterion. Paying heed to the focal areas of ill fit that were discerned in the single-country CFAs (see Tables 4.5 and 4.6), trust in civil service was excluded from the measurement model (Model B). This improved the CFI and TLI somewhat and the X^2 notably.

Again based on the findings from the single-country CFAs, errors of trust in parliament and trust in political parties were then allowed to correlate (Model C1). This error correlation indicates that the two measurement errors are systematically related because some of the shared variance of the two items is due to another common outside cause. Substantively, most likely, this is because political parties play a major role in parliament unlike in the other political institutions. The model adjustment considerably improved the X^2, the CFI and TLI as well as the RMSEA. The latter remained above the recommended cutoff criterion, however.

Based on the results of Model C1, 13 countries were excluded because of model fit issues – eight countries because the factor correlation exceeded .85[76], two countries because the cross-loading between trust in parliament and trust in political parties was not significant (Argentina)[77] or negative (Spain) and three countries because the highest modification index indicated ill specification owing to a missing cross-loading between the latent factor trust in implementing institutions and trust in political parties (Netherlands: 158.388, Turkey: 69.156), and trust in government and trust in the courts (USA: 161.571) (Model C2). Model C2 – including 19 electoral and liberal democracies – reached configural invariance. In all of these countries, the model fit the data well: the unstandardized factor loadings and error correlation were significant at the .05 level; the size of the completely standardized factor loadings was substantial and their direction positive, as expected; the completely standardized factor correlations were all <.85; the error variances were positive and the modification indices were all < 26. Model C2 did not reach full invariance, however.[78]

[76] Chile, Colombia, Lebanon, Malaysia, Mexico, Peru, South Africa, and Taiwan.

[77] This cross-loading was also non-significant in Lebanon.

[78] In addition, in Model C2 the residual covariance matrix was not positive definite in Japan. The residual variance for trust in government was negative, indicating that the estimated factor loading did not fit the data well.

When the data do not support full invariance, researchers have several options (Davidov, Dülmer, Schlüter, Schmidt, & Meulemann, 2012, pp. 560–561). A popular strategy is to test for partial MI, that is, to test for the equivalence of some but not all factor loadings and thresholds (Byrne, Shavelson, & B. Muthén, 1989). Previous MI tests of political trust have commonly opted for this solution (see Table 4.2). Especially in large-N studies, however, discerning patterns in modification indices to determine which parameters should be estimated freely becomes increasingly unwieldy (Byrne & van de Vijver, 2010, p. 113).

Another, hitherto unexplored alternative to this data-driven solution in MI tests of political trust is a theory-driven strategy. Byrne and van de Vijver (2010, p. 113) suggest testing the MI of subsamples of countries clustered according to a theoretically meaningful criterion. With regard to political trust, the post-communist countries are a case in point. Shortly after the end of the Cold War, Mishler and Rose (1994, pp. 8, 25) argued that citizens in these countries cannot clearly distinguish between political trustees because they lack experience with them. From the perspective of political socialization theory, one could argue that almost three decades of democratic socialization have refined, and possibly diversified, people's construct of political trust in former communist countries in Europe more (Klingemann, Fuchs, & Zielonka, 2006, pp. 6–7). Inspired by these arguments, the MI of political trust was tested for the subsample of six post-communist European democracies in this study (Model C3). Full invariance of the model was supported by the data from Poland, Romania, and Slovenia. These results indicate that Mishler and Rose's (1994) general verdict no longer holds.[79] What is more, this brief demonstration of a theory-driven strategy to establish MI shows that similar tests for other subsets of countries could add to our insights on existing theoretical assumptions about the reasons for MI of political trust or lack thereof.

[79] See Schaap and Scheepers (2014, p. 91) for a similar finding.

Table 4.7
Fit Measures for the Multiple Group Confirmatory Factor Analysis of Political Trust

Model	χ^2 (df)	p-value	CFI	TLI	RMSEA (90% CI)
Model A *(all items and countries)*					
1. Configural invariance	6457.907 (256)	0.00	**0.987**	**0.976**	0.129 (0.127-0.132)
Model B *(excluding trust in civil service)*					
1. Configural invariance	3915.855 (128)	0.00	**0.991**	**0.978**	0.143 (0.139-0.147)
Model C1 *(excluding trust in civil service, correlated error between trust in parliament and trust in political parties)*					
1. Configural invariance	919.890 (96)	0.00	**0.998**	**0.994**	0.077 (0.073-0.082)
Model C2 *(excluding trust in civil service, correlated errors between trust in parliament and trust in political parties; including Australia, Brazil, Cyprus, Estonia, Georgia, Germany, Ghana, India, Japan, New Zealand, Philippines, Poland, Romania, Slovenia, South Korea, Sweden, Trinidad & Tobago, Ukraine, Uruguay)*					
1. Configural invariance	235.782 (57)	0.00	**0.999**	**0.998**	**0.048 (0.042-0.055)**
2. Full invariance	5430.023 (255)	0.00	**0.980**	**0.985**	0.123 (0.120-0.126)
Model C3 *(excluding trust in civil service, correlated errors between trust in parliament and trust in political parties, including Poland, Romania, Slovenia)*					
2. Full invariance	115.991 (31)	0.00	**0.998**	**0.998**	**0.048 (0.039-0.058)**

Note. WLSMV estimator (theta parameterization), pairwise present was used to handle missing data (Asparouhov & B. Muthén, 2010, p. 7), df = degrees of freedom, CFI = comparative fit index, TLI = Tucker-Lewis-Index, RMSEA = root mean square error of approximation, 90% CI = 90% confidence interval, parameter of fit values above the recommended thresholds (Yu, 2002, pp. 160–161) are in bold. Data are from the World Values Survey 2010-2012, 32 countries.

4.5 Insights and Recommendations for Future Political Trust Research

This article set out to answer to what extent the MI of political trust can be established in 32 democracies across the globe by means of MGCFA and if so, based on which measurement model. The single-country analyses showed that the data supported the two-dimensional model of trust in implementing and representative political institutions best. In the MGCFA, this model was not equivalent across all 32 democracies, however, because of three sources of bias (van de Vijver & Tanzer, 2004). First, item bias of 'trust in civil service' affected the model fit. Second, construct bias was apparent: The latent factor of trust in representative institutions did not sufficiently account for the shared variance between trust in parliament and trust in political parties in all countries. 'Trust in civil service' was therefore dropped and an error covariance was added to the measurement model in order to measure the construct of political trust in a more valid manner. Configural invariance of this revised two-dimensional model was established in 19 democracies. Additional revisions may be required in order to successfully remedy construct bias in the remaining 13 countries. Third, while the revised measurement model was fully invariant in three post-communist countries in eastern and southeastern Europe, the results suggest that method bias prevented full invariance in the other countries. Non-invariance of factor loadings and the thresholds indicate that the respondents did not use the response scale in the same manner.

These results support previous studies and contradict others. They are in line with authors who distinguish between political trust in implementing and representative institutions conceptually (see for example Gabriel, 1999). Likewise, the analysis corroborates those empirical studies that found political trust to be two-dimensional (see Tables 4.1 and 4.2). Like previous analyses (see for example D. Braun, 2013 in Table 4.1), it also empirically reflects the ambiguity of the position of trust in civil service in the two dimensions of political trust described at the beginning of the article. The study does not, however, correspond to MGCFA that established MI of a single-dimensional model of political trust in Europe. This may be because the items used were not identical.

The results of this study underline that measurement invariance of political trust must not be assumed when testing theories about the changes, sources, or consequences of political trust. Comparative political trust researchers can enhance the validity of their research findings on the generalizability of political trust theories by specifying the

measurement model appropriately and carefully selecting the political trust items and countries. The findings therefore remind comparative researchers to use the ample cross-national survey data available methodically.

The findings are also informative for the future conceptualization of political trust. They provide reason to infer that, by and large, people in democracies across the globe have a two-dimensional construct of political trust. More conceptual work is needed, however, to identify the pertinent political trustees within these dimensions across countries.

In addition, the study contributes to insights regarding the valid measurement of political trust. Because the item 'trust in civil service' is apparently not as meaningfully related to the construct of political trust as the other items, future studies should carefully consider whether to include it. On a more general note, the study criticized the fact that there is no common set of comparable items to measure political trust. Such a set is crucial, however, because the content of the measured construct may be altered depending on the chosen items (Byrne & van de Vijver, 2010, p. 124). Lack thereof impedes the cumulation of research on political trust.

A number of questions follow from this study. Future comparative research on political trust could study the reasons for the apparent bias. Do country-specific response tendencies affect MI and if so, why do they occur with items of political trust? Why is it so difficult to measure civil service in a comparable manner across countries? Last but not least, the study raises questions about the sources of political trust. The error covariance between trust in parliament and political parties indicates that they are not exclusively determined by people's overall level of trust. This could imply that their sources are more trustee-specific than those of the overall construct of political trust. Overall, the results of the study suggest that, in democracies, political trust is neither a single-dimensional construct nor a blanket judgment.

4.6 Appendix B

Table B1
Country-Specific Sample Sizes and Missings per Item

Country	n	Item (trust in)	Missings	Percent missing	Country	n	Item (trust in)	Missings	Percent missing	Country	n	Item (trust in)	Missings	Percent missing
Argentina	1030	police	13	1.3	Australia	1477	police	25	1.7	Brazil	1486	police	6	0.4
		court	17	1.7			court	37	2.5			court	5	0.3
		government	20	1.9			government	29	2.0			government	15	1.0
		pol. parties	36	3.5			pol. parties	31	2.1			pol. parties	16	1.1
		parliament	46	4.5			parliament	38	2.6			parliament	35	2.4
		civil service	40	3.9			civil service	36	2.4			civil service	12	0.8
Chile	1000	police	11	1.1	Colombia	1512	police	5	0.3	Cyprus	1000	police	4	0.4
		court	15	1.5			court	23	1.5			court	17	1.7
		government	15	1.5			government	12	0.8			government	18	1.8
		pol. parties	14	1.4			pol. parties	18	1.2			pol. parties	14	1.4
		parliament	22	2.2			parliament	26	1.7			parliament	13	1.3
		civil service	30	3.0			civil service	14	0.9			civil service	10	1.0

Table B1 (continued)

Country	n	Item (trust in)	Missings	Percent missing	Country	n	Item (trust in)	Missings	Percent missing	Country	n	Item (trust in)	Missings	Percent missing
Estonia	1533	police	14	0.9	Georgia	1202	police	40	3.3	Germany	2046	police	21	1.0
		court	54	3.5			court	111	9.2			court	53	2.6
		government	21	1.4			government	53	4.4			government	45	2.2
		pol. parties	60	3.9			pol. parties	58	4.8			pol. parties	64	3.1
		parliament	41	2.7			parliament	58	4.8			parliament	68	3.3
		civil service	45	2.9			civil service	54	4.5			civil service	44	2.2
Ghana	1552	police	0	0.0	India	1581	police	4	0.3	Japan	2443	police	144	5.9
		court	0	0.0			court	3	0.2			court	254	10.4
		government	0	0.0			government	4	0.3			government	277	11.3
		pol. parties	0	0.0			pol. parties	4	0.3			pol. parties	333	13.6
		parliament	0	0.0			parliament	4	0.3			parliament	322	13.2
		civil service	0	0.0			civil service	4	0.3			civil service	338	13.8
Lebanon	1200	police	43	3.6	Malaysia	1300	police	1	0.1	Mexico	2000	police	1	0.05
		court	57	4.7			court	1	0.1			court	20	1.0
		government	221	18.4			government	2	0.2			government	5	0.2
		pol. parties	80	6.7			pol. parties	2	0.2			pol. parties	3	0.1
		parliament	85	7.1			parliament	1	0.1			parliament	25	1.2
		civil service	53	4.4			civil service	1	0.1			civil service	29	1.4

Table B1 (continued)

Country	n	Item (trust in)	Missings	Percent missing	Country	n	Item (trust in)	Missings	Percent missing	Country	n	Item (trust in)	Missings	Percent missing
Nether-lands	1902	police	63	3.3	New Zealand	841	police	39	4.6	Peru	1210	police	7	0.6
		court	78	4.1			court	55	6.5			court	17	1.4
		government	89	4.7			government	81	9.6			government	22	1.8
		pol. parties	102	5.4			pol. parties	73	8.7			pol. parties	29	2.4
		parliament	133	7.0			parliament	76	9.0			parliament	14	1.2
		civil service	132	6.9			civil service	111	13.2			civil service	22	1.8
Philip-pines	1200	police	1	0.1	Poland	966	police	50	5.2	Romania	1503	police	27	1.8
		court	2	0.2			court	80	8.3			court	91	6.1
		government	2	0.2			government	38	3.9			government	50	3.3
		pol. parties	0	0.0			pol. parties	60	6.2			pol. parties	65	4.3
		parliament	1	0.1			parliament	56	5.8			parliament	62	4.1
		civil service	1	0.1			civil service	73	7.6			civil service	57	3.8
Slovenia	1069	police	22	2.1	South Africa	3531	police	99	2.8	South Korea	1200	police	4	0.3
		court	50	4.7			court	129	3.7			court	4	0.3
		government	29	2.7			government	112	3.2			government	3	0.2
		pol. parties	31	2.9			pol. parties	128	3.6			pol. parties	6	0.5
		parliament	26	2.4			parliament	132	3.7			parliament	6	0.5
		civil service	32	3.0			civil service	193	5.5			civil service	4	0.3

Table B1 (continued)

Country	n	Item (trust in)	Missings	Percent missing	Country	n	Item (trust in)	Missings	Percent missing	Country	n	Item (trust in)	Missings	Percent missing
Spain	1189	police	17	1.4	Sweden	1206	police	7	0.6	Taiwan	1238	police	50	4.0
		court	24	2.0			court	34	2.8			court	87	7.0
		government	18	1.5			government	21	1.7			government	68	5.5
		pol. parties	25	2.1			pol. parties	34	2.8			pol. parties	90	7.3
		parliament	55	4.6			parliament	32	2.7			parliament	96	7.8
		civil service	42	3.5			civil service	220	18.2			civil service	69	5.6
Trinidad and Tobago	999	police	24	2.4	Turkey	1605	police	21	1.3	Ukraine	1500	police	0	0.0
		court	75	7.5			court	47	2.9			court	0	0.0
		government	48	4.8			government	41	2.6			government	0	0.0
		pol. parties	56	5.6			pol. parties	52	3.2			pol. parties	0	0.0
		parliament	69	6.9			parliament	62	3.9			parliament	0	0.0
		civil service	106	10.6			civil service	64	4.0			civil service	0	0.0

Table B1 (continued)

Country	n	Item (trust in)	Missings	Percent missing	Country	n	Item (trust in)	Missings	Percent missing	Country	n	Item (trust in)	Missings	Percent missing
United States	2232	police	37	1.7	Uruguay	1000	police	18	1.8					
		court	44	2.0			court	57	5.7					
		government	45	2.0			government	33	3.3					
		pol. parties	44	2.0			pol. parties	53	5.3					
		parliament	62	2.8			parliament	64	6.4					
		civil service	50	2.2			civil service	96	9.6					

Note. World Values Survey data (2010-2012). Own compilation.

Table B2
Comparison of Configural Invariance Results with Different Reference Indicators for Model A

Reference indicator	χ^2 (df)	p-value	CFI	TLI	RMSEA (90% CI)
trust in parliament and trust in police	6457.907 (256)	0.00	0.987	0.976	0.129 (0.127-0.132)
trust in parliament and trust in court	6481.266 (256)	0.00	0.987	0.976	0.130 (0.127-0.132)
trust in political parties and trust in police	6453.700 (256)	0.00	0.987	0.976	0.129 (0.127-0.132)
trust in political parties and trust in court	6471.272 (256)	0.00	0.987	0.976	0.130 (0.127-0.132)
trust in government and trust in police	6454.196 (256)	0.00	0.987	0.976	0.129 (0.127-0.132)
trust in government and trust in court	6485.580 (256)	0.00	0.987	0.976	0.130 (0.127-0.132)
trust in civil service and trust in police	6459.617 (256)	0.00	0.987	0.976	0.129 (0.127-0.132)
trust in civil service and trust in court	6490.506 (256)	0.00	0.987	0.976	0.130 (0.127-0.132)
factor variance=1/factor mean=0	6457.732 (256)	0.00	0.987	0.976	0.129 (0.127-0.132)

Note. WLSMV estimator (theta parameterization), pairwise present was used to handle missing data (Asparouhov & B. Muthén, 2010, p. 7), df = degrees of freedom, CFI = comparative fit index, TLI = Tucker-Lewis-Index, RMSEA = root mean square error of approximation, 90% CI = 90% confidence interval. Data are from the World Values Survey 2010-2012, 32 countries.

5 Article 4 “Surpassing Simple Aggregation: Advanced Strategies for Analyzing Contextual-Level Outcomes in Multilevel Models”[80]

Abstract

This article introduces two advanced analytical strategies for analyzing contextual-level outcomes in multilevel models: the multilevel SEM and the two-step approach. Since these strategies are seldom used in comparative survey research, we first discuss their methodological and statistical advantages over the more commonly applied approach of group mean aggregation. We then illustrate these advantages in an empirical analysis of the effect of citizens' support for democratic values at the individual level on a contextual-level outcome – the persistence of democracy – drawing on data from the World Values Survey and the Quality of Government project. Whereas we found no significant effect of support for democratic values in the model using simple group mean aggregation, citizens' support for democratic values was a significant predictor of democracies' estimated survival rate when applying latent aggregation in multilevel SEM and the two-step approach. The article corroborates previous concerns with simple aggregation and demonstrates how researchers can improve the validity of their analyses of contextual-level outcomes by using alternative strategies of aggregation.

Keywords: transformational mechanisms; contextual-level outcomes; multilevel analysis; sampling error; democratic stability; democratic values

[80] Published as: Becker, D., Breustedt, W., & Zuber, C. I. (2018). Surpassing simple aggregation: Advanced strategies for analyzing contextual-level outcomes in multilevel models. *Methods, Data, Analyses*, *12*(2), 233–263. doi:10.12758/mda.2017.05

5.1 Introduction

Despite significant methodological advancements, comparative social scientists continue to face the question of how to adequately test theoretical multilevel models empirically. Hierarchical modeling has evolved into a canonical statistical technique for regressing an individual-level variable on individual- and contextual-level predictors. There is no agreement when it comes to multilevel models where the dependent variable is analytically located on the contextual level, though.

Many comparative studies 'solve' this problem through measures of central tendency – such as the average – or the distribution of the data – such as percentages. They then use these aggregates as predictors for the contextual-level dependent variable (for examples, see Fails & Pierce, 2010; Lim, Bond, & Bond, 2005; Muller & Seligson, 1994). This approach has been criticized on both statistical and methodological grounds. Croon and van Veldhoven (2007) demonstrated that group mean aggregation may lead to biased estimates. Griffin (1997, pp. 760–762) argued that the aggregation procedure needs to take into account the complex theoretical relationships of independent variables at different levels of analysis. When applying simple aggregation, researchers may run the risk of drawing invalid conclusions about how individual-level predictors affect contextual-level outcomes (Snijders & Bosker, 1999, pp. 13–15).

Given these criticisms, researchers have proposed two more advanced strategies for analyzing contextual-level outcomes in multilevel models: the multilevel SEM and the two-step approach. Since multilevel SEM and the two-step approach are seldom used in comparative survey research, the article seeks to motivate researchers to improve the validity of their inferences when analyzing contextual-level outcomes, going beyond simple aggregation. In the following section, we introduce the methodological and statistical advantages of these two alternative techniques over the group means approach. In our analysis, we illustrate these advantages in an empirical study of the effect of citizens' support for democratic values at the individual level on a contextual outcome – the persistence of democracy. We draw on data from the World Values Survey and the Quality of Government project and study 98 countries between 1946 and 2014. We compare the regression coefficients and confidence intervals of our individual-level predictor – support for democratic values – on democracies' persistence when applying the three methods. Whereas we found no significant effect of support for democratic

values in the model using simple group mean aggregation, citizens' support for democratic values was a significant predictor of democracies' estimated survival rate when applying multilevel SEM and the two-step approach. In the final section we therefore conclude that comparative researchers who use simple group mean aggregation when regressing a contextual outcome on individual-level predictors may run the risk of wrongly rejecting their hypothesis of interest.

5.2 Methodological Foundation and Statistical Background

Testing theoretical multilevel models with contextual-level outcomes poses two challenges. From a methodological point of view, researchers need to establish close correspondence between the theoretical multilevel mechanism and its empirical measurement. From a statistical perspective, they need to choose a method both valid and reliable for aggregating the individual-level predictors. In the following, we discuss the methodological foundations of multilevel analysis of macro-level social phenomena. We then proceed to introduce and compare three analytical strategies for analyzing contextual level outcomes: simple manifest group mean aggregation, latent aggregation through multilevel SEM and the two-step approach. The results of the comparison are summarized in Table 5.1 at the end of this chapter.

5.2.1 Methodological Foundation

According to the paradigm of structural individualism (Udehn, 2002, p. 492), the ultimate goal of the social sciences is to explain social phenomena on the contextual – or macro – level as a consequence of individuals' social actions on the individual – or micro – level. Structural individualism distinguishes three explanatory mechanisms (see Figure 5.1) (Hedström & Swedberg, 1998, p. 23; Tranow, Beckers, & Becker, 2016, p. 8). Situational mechanisms (1) link the objective characteristics of the social situation to the subjective expectations and evaluations of individuals. Action formation mechanisms (2) explain individuals' actions given their subjective definition of the situation. This is a pure micro-level explanatory step. Transformational mechanisms (3) reconstruct how individuals' actions aggregate to create a new social situation. They thereby re-link the micro level to the macro level.

Studying these theoretical mechanisms empirically is not straightforward. Multilevel modeling (Bryk & Raudenbush, 1992; Hox, 2010) is a well-established statistical tool for testing situational and action formation mechanisms, that is, explanations that link social situations to individuals' expectations, evaluations, and actual decisions (Becker, Beckers, Franzmann, & Hagenah, 2016, pp. 166–171). By contrast, micro-to-macro (or, more technically, level-one to level-two) explanations constitute a blind spot of conventional multilevel analysis (henceforth MLA)[81] as transformational mechanisms are more difficult to analyze empirically (Opp, 2011; Raub, Buskens, & van Assen, 2011).

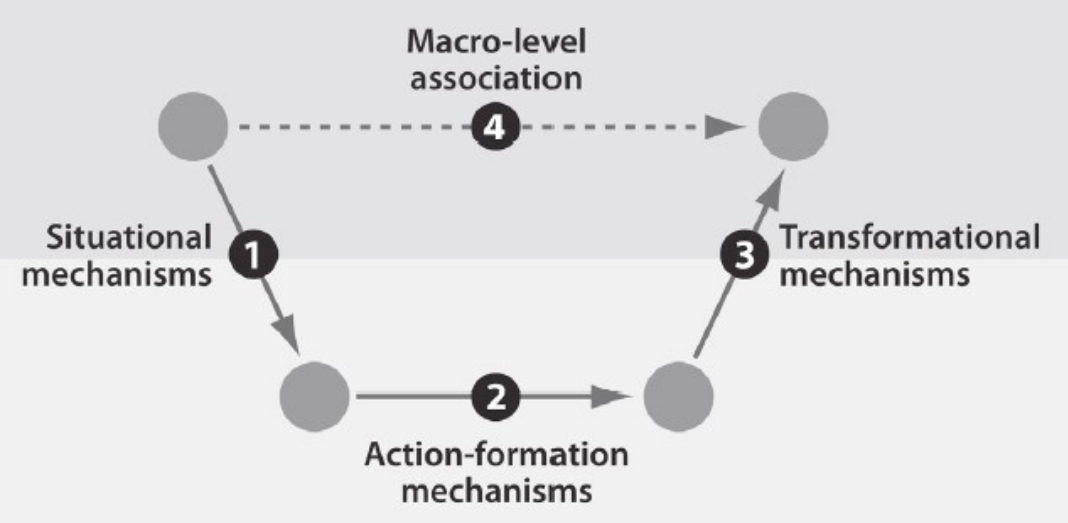


Figure 5.1 The Social Mechanisms of Social Science Explanations. Source: Hedström & Ylikoski (2000, p. 53).

5.2.2 Three Analytical Strategies

5.2.2.1 The Simple Group Means Approach

When studying multilevel models with contextual-level outcomes, a common approach (Lim et al., 2005) is to aggregate all level-one variables (hereafter L1) to level-two variables (hereafter L2) by computing their group-specific arithmetic means. This manifest aggregation is followed by an L2-only regression (see Figure 5.2).

[81] In accordance with previous research, we use the terms 'conventional' or 'standard' multilevel analysis to describe hierarchical modeling techniques that are restricted to the analysis of level-one outcomes (Bennink, Croon, & Vermunt, 2013, p. 432, 2015, p. 665; Lüdtke et al., 2008, p. 225; Preacher, Zyphur, & Zhang, 2010, p. 209).

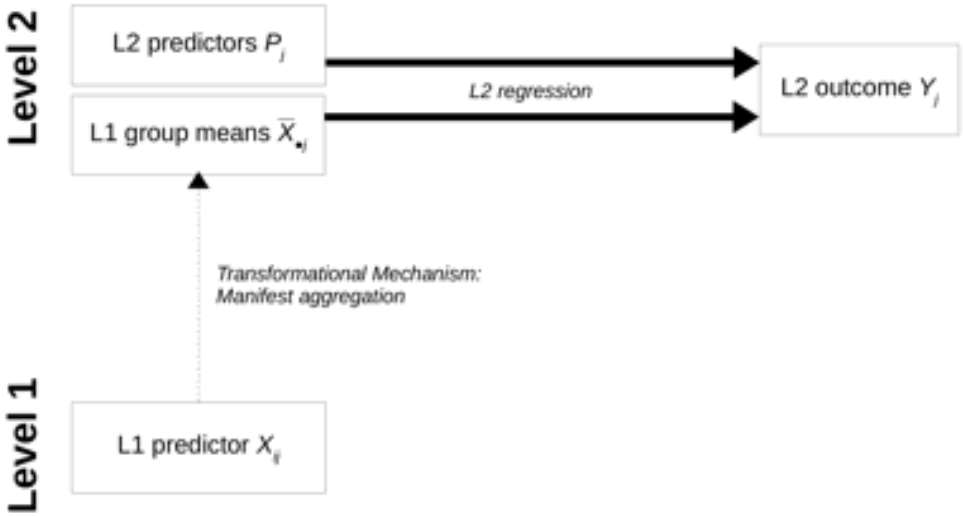


Figure 5.2 The Simple Group Means Approach.

Methodologically, this method models neither situational nor action-formation mechanisms and accounts for transformational mechanisms via (manifest) aggregation (see Figure 5.2). Statistically, Croon and van Veldhoven (2007) have shown that this procedure only yields valid estimates if the L1 variance of the aggregated variables is zero. If the L1 variance is larger than zero, simple group mean aggregation yields biased estimates. In cross-national comparative survey research, this is often the case because individuals are sampled from a finite population and a specific constellation of individuals is selected to measure the L2 construct (Lüdtke et al., 2008). Since manifest aggregation does not control for these sampling errors, the observed group average (measured, for instance, in terms of group-specific arithmetic means) may be an unreliable measure of the unobserved true group average. In addition, the observed group average completely obscures the heterogeneity within groups. Therefore, if effects of observed group averages on L2 outcomes are of interest, estimates of both these effects and of other L2 predictors are likely to be biased when applying the simple group means approach (Bennink et al., 2013, pp. 433–434, 2015, p. 663; Shin & Raudenbush, 2010, p. 27).

5.2.2.2 The Multilevel SEM Approach

Multilevel SEM avoids these statistical problems by replacing manifest with latent aggregation (see Figure 5.3). Assume that we observe a manifest L1 variable X_{ij} for individuals i in countries j. X_{ij} is used to predict a manifest L2 outcome Y_j along with

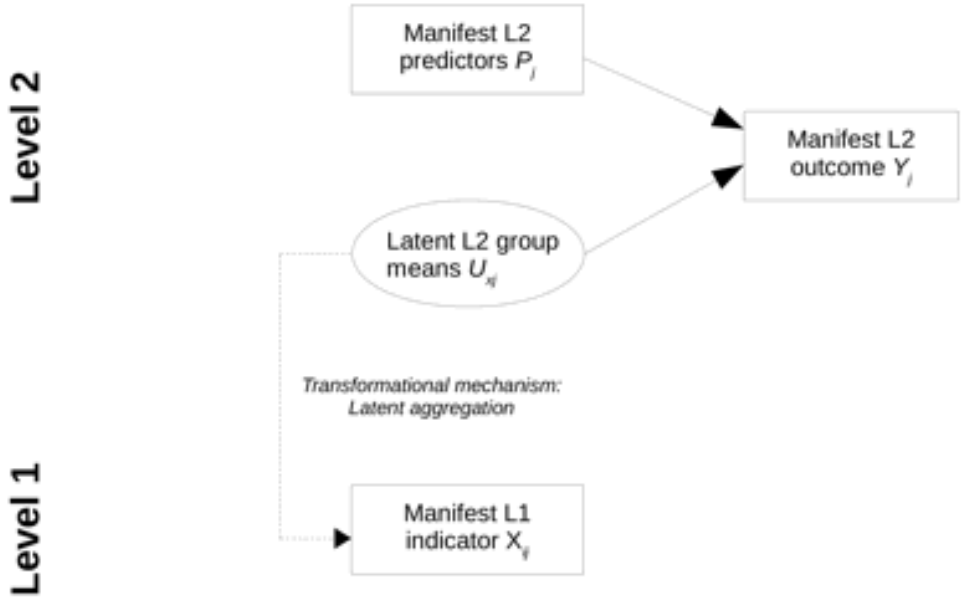


Figure 5.3 Latent Aggregation in Multilevel Structural Equation Modeling.

other L2 predictors P_j. Following the simple group means approach, X_{ij} is aggregated from L1 to L2 by computing group-specific arithmetic means $\overline{X}_{\bullet j}$, which are not corrected for sampling error. In a second step, $\overline{X}_{\bullet j}$ are used to predict Y_j controlled for P_j (adapted from Marsh et al., 2009, pp. 772–779):[82]

(1) $$Y_j = \beta_0 + \beta_1 \overline{X_{\bullet j}} + \beta_2 P_j + u_{0j}$$

By contrast, multilevel SEM regards the actual group mean on L2 as an unobserved latent variable U_{xj} (which must not be confused with L2 residual error u_{oj}) that can only be estimated with error by the L1 indicators (Marsh et al., 2009, pp. 772–779). Following the conventions of SEM, the L2 latent means of the L1 observations are therefore depicted by ovals in Figure 5.3. While the simple group means approach treats the L2 group mean as a simple composite or index score of the L1 observations, multilevel SEM assumes the unobserved latent group means to *cause* the observed L1 values (Lüdtke et al., 2008, p. 205).[83]

[82] The notation by Marsh et al. (2009) implies group mean centering of all L1 predictors to account for a reference-group effect (in their example, this is the dependence of student academic self-concept on class-average achievement). Since our substantive application does not include a reference-group effect, we present the general notation without group mean centering. In addition, we use standard multilevel notation for the L2 residual variance.

[83] This points to the difference between formative and reflective models in measurement theory. Whereas formative latent variable models are already established in single-level *measurement* models (Diamantopoulos & Winklhofer, 2001), it remains unresolved whether formative *latent aggregation* is equally possible.

Multilevel SEM proceeds in two steps: First, an L2 latent variable U_{xj} is estimated. It is assumed to be the cause of X_{ij} at L1. In a second step, U_{xj} is used to predict the L2 outcome Y_j along with the other L2 predictors P_j:[84]

(2) $$Y_j = \beta_0 + \beta_1 U_{xj} + \beta_2 P_j + u_{0j}$$

The aggregated L2 construct is a measure of the unobserved true group mean. Its reliability is a function of the relative share of the L2 variance weighted by the group-specific number of observations (Lüdtke et al. 2008, p. 207):

(3) $$\frac{\tau_x^2}{\tau_x^2 + (\sigma_x^2 / n_j)}$$

As in conventional hierarchical modeling, σ^2_x denotes the L1 part and τ^2_x the L2 part of the variation of the respective indicator(s), whereas n_j refers to the group-specific number of observations.

By estimating a latent L2 variable U_{xj} as in (2), the variance of the L1 indicator is partitioned into an L1 and an L2 component. Unlike simple group mean aggregation, latent aggregation takes account of the heterogeneity within each group by partitioning the L1 variance σ^2_x from the L2 variance τ^2_x. In addition, by estimating latent group means at L2, which are assumed to cause the L1 observations in each group, the multilevel SEM approach acknowledges that the L1 scores do not perfectly map the construct at the L2 level, because of measurement error (Bennink et al. 2013, pp. 434–436, 2015, pp. 663–665; Preacher et al. 2010, pp. 213–215).

In sum, multilevel SEM replaces *manifest* with *latent* aggregation to aggregate individual-level predictors of macro-level outcomes. Like manifest aggregation, latent aggregation *per se* models only the transformational but not the situational and action formation mechanism. Statistically, however, latent aggregration is superior to manifest aggregation since it corrects for sampling error (see Table 5.1). As a result, its estimates are less biased, thereby permitting more valid inferences regarding the effect of multilevel predictors on contextual-level outcomes.

[84] Additional controls for measurement error can be integrated easily (Marsh et al., 2009, pp. 772–779). For the sake of simplicity, our analysis of democratic persistence is limited to latent aggregation without controlling for measurement error.

5.2.2.3 The Two-Step Approach

The two-step approach also deals with the methodological and statistical issues that arise when studying multilevel models with contextual-level outcomes, albeit in a different manner. Figure 5.4 summarizes its basic idea.

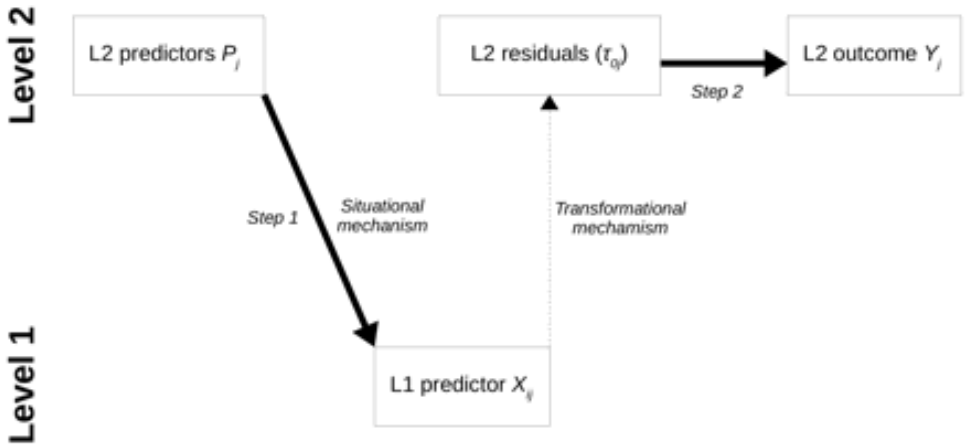


Figure 5.4 The Two-Step Approach.

The two-step approach builds on standard MLA. For an L1 outcome Y_{ij} and L1 units i nested in L2 contexts j, the standard model is given by:

(4) $Y_{ij} = \beta_{0j} + \beta_{1j} X_{ij} + e_{ij}$

In equation (4), β_{0j} is the regression intercept of the outcome variable, β_{1j} is the regression slope of an L1 predictor, and e_{ij} is the residual error term. In contrast to non-nested regression analysis, both random intercepts β_{0j} and random slopes β_{1j} can be estimated for each L2 unit j by modeling them as a function of an additional L2 predictor Z_j with distinct intercepts (γ_{00} and γ_{10}) and regression slopes (γ_{01} and γ_{11}):

(5) $\beta_{0j} = \gamma_{00} + \gamma_{01} Z_j + u_{0j}$

(6) $\beta_{1j} = \gamma_{10} + \gamma_{11} Z_j + u_{1j}$

Equations (5) and (6) introduce two additional residual error components: u_{0j} denotes the residual error of the outcome's L2 intercept β_{0j}, and u_{1j} denotes the residual error of the slope β_{1j} between L2 units.

Standard MLA only considers the case of an L1 outcome Y_{ij} that is predicted by L1 and L2 variables X_{ij} and Z_j, respectively. Griffin (1997) proposes an extension of the standard MLA approach to study an L2 outcome Y_j: Let X_{1ij} be the L1 explanatory variable of

Table 5.1
Comparison of Methods for Analyzing Macro-Micro-Macro Models

	Main methodological advantages & disadvantages	Main statistical advantages & disadvantages
Group mean aggregation	Transformational mechanism (via manifest aggregation and macro regression)	Simple to perform, but only valid if variance of L1 variable = 0
ML SEM	Transformational mechanism (via latent aggregation and macro regression)	Takes sampling error into account: reduction of estimator bias
2-Step	1st step: situational & action-formation mechanism (via MLA) 2nd step: Transformational mechanism (via residuals and macro regression)	Residual reflects the net effect of the individual-level independent variable

primary interest. In a first step, X_{1ij} is regressed on all other L1 and L2 predictors X_{2ij}, ..., X_{nij} and Z_j:

(7) $$X_{1ij} = \gamma_{00} + \gamma_{01} Z_j + \tau_{0j} + \beta_{1j} X_{2ij} + ... + \beta_{nj} X_{nij} + e_{ij}$$

In a second step, the L2 residuals u_{0j} of this model are used as a predictor variable in an L2 regression of the L2 outcome of interest:

(8) $$Y_j = \beta_0 + \beta_1 u_{0j} + e_j$$

The effect of u_{0j} on the L2 outcome Y can be interpreted as the aggregated effect of the L1 variable X_1, net of both L1 and L2 covariates X_2, ..., X_n and Z.

The two-step approach has both statistical and methodological advantages when studying multilevel models with contextual-level outcomes (see Table 5.1). Statistically, it provides a better estimate than the group mean aggregate: u_{0j} is a model-based estimate of the L2 variance that is already net of the L1 variance. In addition, u_{0j} can be adjusted for other covariates at L1 and L2. This may save degrees of freedom and circumvent collinearity issues when using u_{0j} as a predictor in a subsequent L2 regression. Compared to the group means approach and the multilevel SEM approach, the crucial methodological advantage of the two-step approach is its capacity to empirically model theoretical macro-micro-macro explanations in their entirety. The MLA of step 1 maps both the situational and action formation mechanism through the regression of an L1 outcome on L1 and L2 predictors. Storing the L2 residuals of this MLA then maps an underlying transformational mechanisms in terms of an L1-L2 aggregation.

The relative statistical performance of each method can also be compared empirically. Based on previous research, we deduce two hypotheses. First, we expect that unless the L1 variance equals zero, simple group mean aggregation yields unreliable measures of

the unobserved true group means. By contrast, multilevel SEM results in reliable estimates of true group means. Consequently, when group means based on simple aggregation are used as predictors of an L2 outcome, estimates of their regression coefficients may be biased (Bennink et al., 2013, pp. 433–434, 2015, p. 663):

> *H_1:* Regression coefficients of L2 predictors that are simple group means deviate in terms of a) point estimates, b) standard errors, and c) resulting significance levels from regression coefficients of L2 predictors that have been aggregated through multilevel SEM.

Second, while the statistical performance of the two-step approach (Griffin, 1997) is less well researched, Lüdtke et al. (2008) compared multilevel SEM to another two-step approach proposed by Croon and van Veldhoven (2007). This approach adjusts the observed group means with weights from ANOVA formulas. This is quite similar to the decomposition of variance in an empty multilevel model. Lüdtke et al. (2008) observed that Croon and van Veldhoven's (2007) approach performed slightly less well than multilevel SEM. Consequently, we expect Griffin's two-step approach to yield estimates closer to multilevel SEM than to the simple group means approach:

> *H_2:* Regression coefficients of L2 predictors that have been aggregated by the two-step approach deviate less from multilevel SEM in terms of a) point estimates, b) standard errors, and c) resulting significance levels than regression coefficients of L2 predictors that are simple group means.

5.3 Substantive Application: A Multilevel Explanation of the Persistence of Democracy

5.3.1 Theoretical Background

To illustrate the methodological and statistical issues described in the previous section, we use the persistence of democracy as a substantive example. Explanations of democratic persistence pertain either to a macro-to-micro mechanism leading from the macro level to the level of individual citizens or to a micro-to-macro mechanism leading from individual citizens to the persistence of democracy at the macro level.

Przeworski (1991) introduces a classic model linking macro-level causes to individuals' micro-level incentives for subverting a democratic regime. Acknowledging that democratic competition produces winners and losers, he argues that "political forces

comply with present defeats because they believe that the institutional framework that organizes the democratic competition will permit them to advance their interests in the future" (Przeworski, 1991, p. 19). Institutions are not only crucial for inspiring the belief that there will be future possibilities to advance one's interests. The given set of political and economic institutions also has distributional consequences affecting the capacities individuals have at their disposal to advance their interests (Przeworski, 1991, pp. 17–18). A model of democratic persistence therefore has to take into account that – under the same set of democratic rules – members of some societal groups might deem their chances of affecting future democratic outcomes to be lower than members of other societal groups. Correspondingly, classic studies have analyzed the decisive impact of economic development on both the process of successful democratization (Bollen, 1979; Bollen & Jackman, 1985; Lipset, 1959) as well as democratic persistence (Przeworski, Alvarez, Cheibung, & Limongi, 2000).

A second example for the macro-to-micro mechanism underlying the persistence of democracy is the idea that an ethnically divided society poses a particular challenge to democratic persistence (Horowitz, 1985; Rabushka & Shepsle, 1972; Reilly, 2001). In countries where several ethnic groups are politically mobilized, the question of who is to legitimately take part in the democratic game is continuously contested. Members of ethnic minorities often see little incentive to support ruling elites, who are – in virtue of the majority principle – likely to be members of the majority group. As a result, those out of power may choose to subvert democracy because they feel permanently excluded from democratic decisions likely to reflect only the interests of the majority.

A classic example of the micro-to-macro mechanism underlying the persistence of democracy is the political culture model. Almond and Verba (1965, pp. 10, pp. 29–30) seminally argued that the persistence of a political regime does not rest on its formal democratic institutions alone, but also on its political culture. Succeeding studies further specified the content of political culture and its effect on democratic persistence based on Easton's (1965, 1975) systems theory (Dalton, 2004, pp. 5–9; Fuchs, 2007, pp. 164–167; Norris, 1999, pp. 9–13). According to Easton, citizens' political support refers to their supportive values and attitudes toward the political community, the political regime, and political authorities (Easton, 1965, p. 157). A critical amount of political support is necessary for any kind of political system to persist. Citizens' political support increases the functionality of political systems as it allows political authorities to convert demands

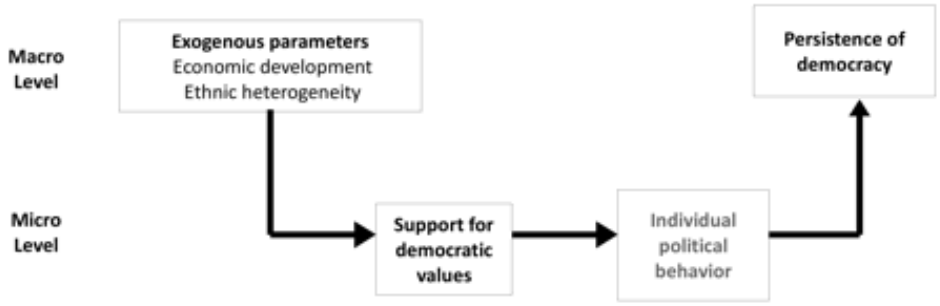


Figure 5.5 A Two-Level Explanation of the Persistence of Democracy.

into outputs and permits them to implement collectively binding decisions without having to resort to force (Easton, 1965, pp. 153, 157).

Building on Easton (1965, 1975), Fuchs (2007) clarifies the implications of the different dimensions of political support for democratic political regimes. Support for the political authorities is crucial for their re- or de-election; support for the political system is essential for the persistence of a given type of democracy; support for democratic values is critical for the persistence of democracy in general (Fuchs, 2007, pp. 164–167). Thus, citizens' support for democratic values is the key factor when studying the effect of individual-level political orientations on the persistence of democracy at the macro level.

Fails and Pierce (2010) tested the systems approach of the political culture model empirically. Their analysis yielded no significant relationship between citizens' support for democratic values and their rejection of authoritarian values on the one hand and the probability of a decline of democracy on the other hand.

These mechanisms can be combined into a full multilevel explanation of democratic persistence (see Figure 5.5). From the macro to micro explanations, we take the insight that citizens' support for democratic values is likely to be affected by context-specific economic conditions and ethnic heterogeneity. From the micro to macro explanations, we take the insight that micro-level support for democratic values crucially accounts for the persistence of democracy at the macro level.

5.3.2 Research Design

5.3.2.1 Period of Analysis and Data

Based on the data available, we analyzed the persistence of democracy from 1946 to 2014. We derived all L2 indicators from the *Quality of Government* standard time series data set (QoG) (Teorell et al., 2016), which includes data on a broad range of country-level indicators over time that we could easily merge with our L1 data.

To measure our L2 outcome variable – *democratic persistence* – we used the democracy index developed by the Polity IV project as included in the QoG (Marshall et al., 2015). Polity IV's democracy index – POLITY – reports countries' level of democracy on a scale ranging from -10 (fully autocratic) to +10 (fully democratic).[85] In line with the threshold provided on the Polity IV website (Marshall & Gurr, 2014), we considered countries as democracies if their POLITY score ≥ 6.[86]

As for our L2 predictors, we used the following indicators: *Economic development* was measured using countries' annual gross domestic product (GDP). We used the log of the OECD measure of GDP *per capita*. *Ethnic heterogeneity* was measured using Fearon's (2003) ethno-linguistic fractionalization index (ELF), a measure of the probability that two randomly chosen individuals from a particular country are members of different ethnic groups. It ranges from 0 (perfect homogeneity) to 1 (very high fractionalization).[87]

Citizens' support for democratic values and all other L1 covariates were derived from the *World Values Survey* (WVS). The WVS is a cross-national survey based on representative national samples investigating worldwide sociocultural and political change. For our analyses, we used the wave 6 aggregated longitudinal file, which includes more than 340,000 observations sampled in 101 countries across all available waves from 1981 to 2014. In line with previous research, *support for democratic values* was operationalized in terms of respondents' reply to the following question: "I'm going to describe various

---

[85] POLITY is a composite score that quantifies the extent to which a country exhibits democratic and authoritarian characteristics. Polity IV coders assess countries' formal political institutions in terms of five component variables – the competitiveness of political participation (1), the openness of executive recruitment (2), the competitiveness of executive recruitment (3), the constraints on the executive (4), and the regulation of political participation (5) for each country on an annual basis. Countries are assigned weighted scores for each component. These are then added up to arrive at a democracy (DEMOC) and an autocracy score (AUTOC), both of which range from 0 to 10. The autocracy score is then subtracted from the democracy score to construct POLITY (Marshall et al., 2015, p. 16).

[86] We noted an inconsistency in the definition of the thresholds. In their codebook, Marshall et al. (2015, p. 35) state that POLITY values ranging from +7 to +10 indicate a democratic regime.

[87] The formula is $1-\sum_{i=1}^{n} s_i^2$ where s_i is the share of group i ($i = 1, \ldots, n$).

types of political systems and ask what you think about each as a way of governing this country. For each one, would you say it is a very good, fairly good, fairly bad or very bad way of governing this country?". For reasons of data availability, we used respondents' rejection of an authoritarian system rather than their support for a democratic system. The answer category reads: "Having a strong leader who does not have to bother with parliament and elections" (1 = 'very good'; 2 = 'fairly good'; 3 = 'bad'; 4 = 'very bad'). For our analyses, we dichotomized this variable (0 = 'good / very good' vs. 1 = 'bad / very bad'). In accordance with previous research (C. Schneider, 2009), we controlled for individuals' age (six categories ranging from 1 = '15–24 years' to 6 = '65 and more years'), subjective assessment of social class (five categories ranging from 1 = 'lower class' to 5 = 'upper class'), and education (eight categories ranging from 1 'inadequately completed elementary education' to 8 'university with degree/higher education').[88]

5.3.2.2 Methods of Analysis

Studying the effect of L1 and L2 predictors on an L2 outcome such as the persistence of democracy poses two methodological challenges. First, choosing a method to address the L1-L2 aggregation problem; second, analyzing persistence of democracy, which is a duration variable.

We compared three different strategies for solving the L1-L2 aggregation problem. First, we aggregated support for democratic values and all other L1 covariates by computing the arithmetic means for each country year (Model 1). Second, we corrected for sampling error by estimating a latent aggregation of all L1 variables on L2 using multilevel SEM (Model 2).[89] Third, we applied the two-step procedure proposed by Griffin (1997) by regressing support for democratic values on all other L1 and L2 predictors and then using the L2 residuals of this multilevel model as a new predictor variable.

We estimated not one, but several multilevel levels that were built up stepwise: The first empty model separated the L2 residuals of support for democratic values from the L1 residuals (Model A1). We then added the macro level predictors GDP and ELF (models A2-A4). Finally, we added all L1 controls (Model A5).[90] Researchers typically use stepwise model building (which we also carried out in the L2-only regressions below) to make causal claims about mediator variables partialing out significant effects of previous

[88] See Table C1 (Appendix C) for a summary of all variables.
[89] The latent aggregation was performed in Mplus, Version 7 (L. Muthén & M. Muthén, 2012).
[90] See Table C2 (Appendix C).

regressors. Apart from comparing point estimates and confidence intervals between aggregation methods for the final model, we also considered it instructive to analyze a series of stepwise models in order to assess whether different aggregation methods lead to different claims about causal mediation.

In addition, we chose an adequate model for predicting democratic persistence, a duration variable. The time span of interest is the persistence of a given democracy until its breakdown. Whereas some democracies may have persisted before entering the observation window (left censoring), others may have continued to persist after the observation ended (right censoring). Within the time period of analysis, the same country may have experienced multiple democratic sequences, followed by breakdowns. In order to address these issues, we used event history modeling. We considered democratic breakdown to occur if the score of democratic regimes (nested within countries) fell below the threshold of POLITY = 6. The duration until this event was measured by the total number of years a democratic system persisted from 1946 onwards. Multiple breakdowns within the same country were coded as distinct events. To keep the models parsimonious, we used a simple exponential event history model, which assumes constant transition rates across years.

In formal terms, our event history model is defined as follows: Let *h* denote the hazard rate of democracies' estimated risk of falling below POLITY = 6 and *t* the time of democracies' survival. The basic exponential survival model can then be described as:

(9) $$h(t)=\lambda\,;\, t>0, \lambda>0$$

λ is a positive constant constraining transition rate (in terms of democratic breakdowns) that is equal across years. Our aim was to predict the expected survival time E(t) with an aggregate measure of citizens' support for democratic values (*DVAL*), countries' *GDP* and *ELF,* as well as aggregate measures of citizens' age (*AGE*), subjective social class (*SCLASS*), and education (*EDUC*).

When applying simple aggregation, democracies' expected time of survival was estimated by:

(10)

$$E(t_j)=\exp(\beta_0+\beta_1\overline{DVAL_{\bullet j}}+\beta_2 GDP_j+\beta_3 ELF_j+\beta_4\overline{AGE_{\bullet j}}+\beta_5\overline{SCLASS_{\bullet j}}+\beta_6\overline{EDUC_{\bullet j}})$$

where $\overline{X}_{\bullet j}$ from equation (1) was replaced by the aforementioned predictor variables. When using latent aggregation, we estimated:

(11)

$$E(t_j)=\exp(\beta_0+\beta_1 U(DVAL)_j+\beta_2 GDP_j+\beta_3 ELF_j+\beta_4 U(AGE)_j+\beta_5 U(SCLASS)_j+\beta_6 U(EDUC)_j)$$

Here, *U* refers to the unobserved latent L2 group mean, which is assumed to cause the observed L1 values of each variable.

Finally, when employing the two-step approach, the estimates were derived as follows:

(12) $E(t_j)=\exp(\beta_0+\beta_1 u_{0jm})$

In equation (12), u_{0jm} denotes the L2 residuals from a hierarchical regression of citizens' support for democratic values on both the L2 predictors and the L1 covariates. The subscript *m* indicates that the hierarchical models were built up in a stepwise manner, which is why we estimated several terms for u_0.

These formal specifications require a methodological addendum: While we *estimated* three *L2* event history analyses after having applied each of the three aggregation methods, our *theoretical* explanation emphasizes the importance of citizens' support for democratic values on *L1*. Hence, though the event history models applied L2-only regressions, in line with the paradigm of structural individualism, we assume that the theoretical mechanisms operate via citizens' preferences and beliefs on the micro level. In line with the aim of our article, we sought to determine how the three different aggregation methods map these L1 processes when predicting an L2 outcome.

In order to increase our statistical power, we used both inter- and extrapolation techniques for our independent variables. We interpolated missing values between observation points, using the *–ipolate–* command in Stata. In addition, we extrapolated missing values between the last valid observation and 2015, using a 'non-linear trend' scenario. We first estimated a polynomial regression of the interpolated values of each predictor on years of observations using the *–lpoly–* command in Stata. We then used out-of-sample predicted values to replace missing observations for subsequent years over countries.[91]

[91] The overlap of valid observations for both democratic persistence and support for democratic values before and after interpolation is displayed in Figure C1 (Appendix C). The basic survivor function of democratic persistence for our reduced sample of analysis is sufficiently similar to the survivor function of the total country sample (see Figure C2 Appendix C). As a sensitivity check, we also extrapolated our interpolated values by repeating the last valid observation of each predictor for subsequent years with missing values. Results based on this extrapolation technique are very similar to the results reported in the results section (see Figure C4, Appendix C).

5.3.3 Results

Prior to computing the comprehensive multivariate models, we compared the survival functions of democracies with high vs. low average support for democratic values. We dichotomized the support variable and compared countries with one standard deviation above vs. below the grand mean of the aggregated variable. We then compared the survival functions of these two groups of countries using group mean aggregation, the two-step model, and latent aggregation. Independent of the method of aggregation, in the long run, the estimated survival rate for democracies scoring one standard deviation above the grand mean of support for democratic values was higher than for their lower-scoring counterparts (see Figure C3, Appendix C). Apart from a lower estimate of the survival rate of countries whose citizens had less support for democratic values in the two-step model, the differences between the aggregation methods appeared to be negligible.

Figure 5.6 presents the results of the analyses using the simple group means approach (Model 1), multilevel SEM (Model 2), and the two-step approach (Model 3). It shows both point estimates and confidence intervals for the L1 and L2 predictors. Our survival models were built up stepwise: In Models 1a and 2a, the survival rate of democracies was first predicted by support for democratic values only; in Model 3a, it was predicted by the L2 residuals from the multilevel null model, which separated the variance of the L1 support variable without having included any other L1 or L2 predictor. In Models 1b and 2b, we simultaneously added GDP and ELF. Correspondingly, in Model 3b we included the residuals corrected for these L2 predictors. Finally, in Model 1c and 2c, we added the L1 covariates; in Model 3c we included the residuals corrected for the L1 covariates. Because of the low number of events, we displayed confidence intervals both on the 10% ($|t| > 1.64$; see ticks of confidence bands) and the 5% significance level ($|t| > 1.96$; see ends of confidence bands).

When applying the *simple group means approach*, support for democratic values did not turn out to be a significant predictor of democratic survival. Point estimates varied between -3.734 in Model 1a and -3.367 in Model 1c, but neither estimate was larger than 1.65 times its standard error (see also Table C3, Appendix C). The latter also applies to all other L2 predictors and to the L1 covariates. We observed significant intercept variation in Model 1a, which only included support for democratic values as a predictor

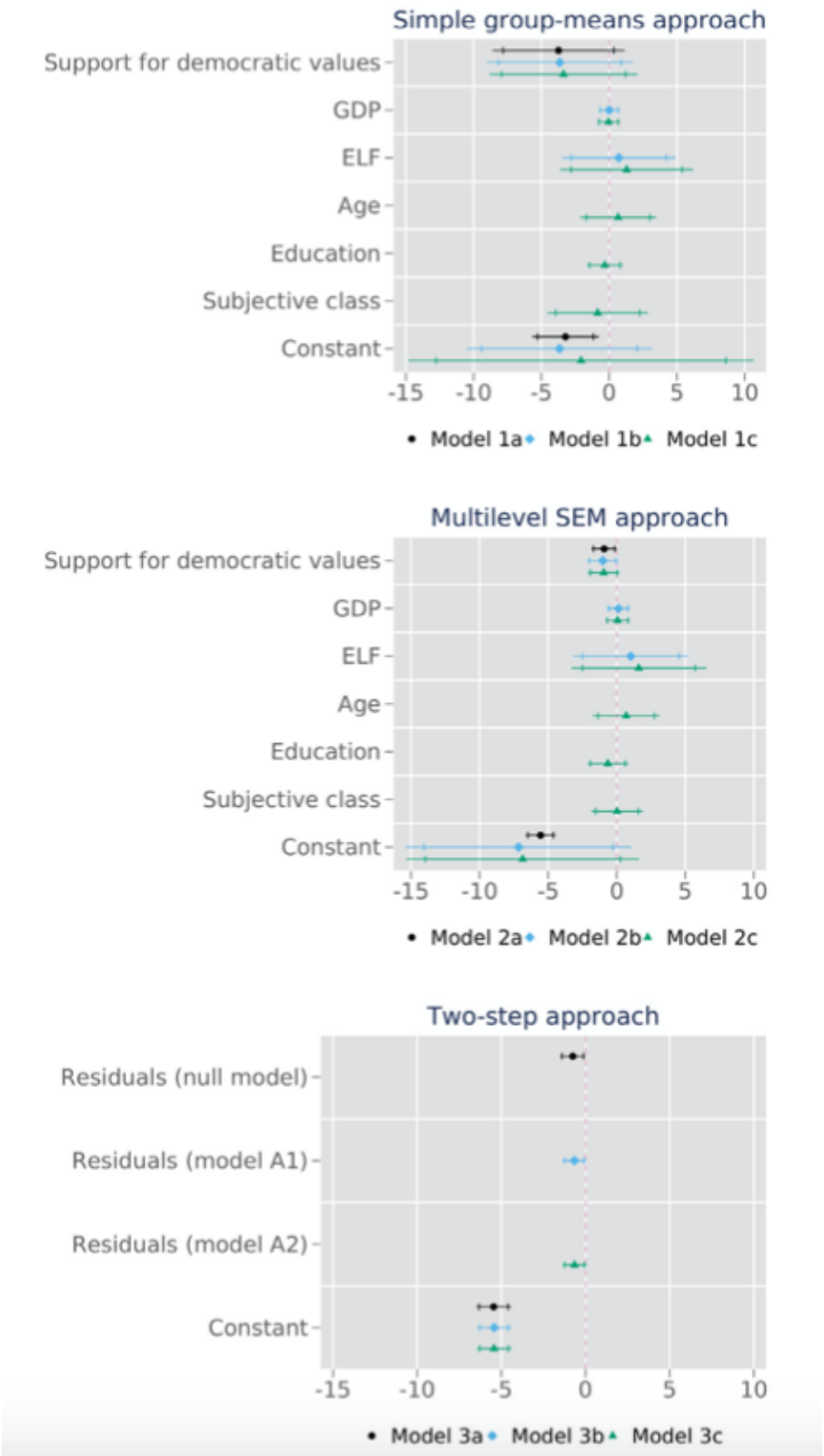


Figure 5.6 Point Estimates and Confidence Intervals of Countries' Democratic Survival across Aggregation Methods. *N*= 917 observations, *N*= 122 subjects, *N*=5 failures in all models.

variable, but not in Models 1b and 1c, which controlled for the remaining variables. Values of AIC and BIC as indicators of model fit show that not much was gained by adding predictors of democratic survival apart from citizens' support for democratic values (see Table C3, Appendix C).

When using the *latent aggregation approach*, the estimated confidence intervals of support for democratic values became more precise and we observed two effects of support for democratic values on democratic survival that were greater than 1.65 times their standard error (Models 2a and 2b). Once the aggregated L1 covariates were controlled for, our predictor was no longer significantly associated with the outcome. Point estimates were remarkably lower after latent aggregation, ranging from -.911 in Model 2a to -1.009 in Model 2b (see Table C4, Appendix C). Having controlled for L2 structural conditions (in terms of GDP and ELF), the effect of support for democratic values became more negative from Model 2a to Model 2b – which points to a suppressor effect. Yet, similar to the simple group means analysis, none of the remaining variables turned out to be significant predictors of democratic survival. Model fit indices again supported the most parsimonious Model 2a and intercept variation was significant in the first two submodels only.

When applying the *two-step approach*, point estimates of support for democratic values on democratic stability were predicted with similar precision as in latent aggregation when looking at the confidence intervals. Yet, in the two-step model, we observed three significant effects at the 10% level. The L2 (u_{oj}) residuals of support for democratic values predicted democratic survival independent of whether they were adjusted for other L1 or L2 variables. Effect sizes ranged from -.754 in Model 3a to -.651 in Model 3c (see Table C5, Appendix C). In contrast to simple group mean and latent aggregation, the intercept remained significant in all three sub-models. Though model fit indices supported the most parsimonious Model 3a, the differences between model fit indices across models were less striking than in the event history regressions following manifest and latent aggregation.

Our results can be summarized as follows: In each estimation, support for democratic values was negatively associated with the event of democratic breakdown, as expected by theory. This replicated our bivariate analysis where democracies with higher support for democratic values showed a longer estimated survival rate on average. Apart from this similarity, there are notable differences between the aggregation methods: While support

for democratic values was not significantly associated with democratic stability after manifest aggregation, significant effects could be observed after both latent aggregation and the two-step approach. Applying more advanced aggregation methods led to smaller point estimates and standard errors compared to the simple group means approach. All this is in line with the two hypotheses postulating notable differences between simple group means aggregation and latent aggregation, and closer similarity between the two-step approach and latent aggregation than between the two-step approach and manifest aggregation.

Yet, compared to latent aggregation, which has already been observed to yield unbiased point estimates in simulation models (Bennink et al., 2013, 2015; Lüdtke et al., 2008), researchers who apply the two-step approach may run the risk of committing type one errors: In the most comprehensive model of the two-step approach (Model 3c) and unlike in the corresponding regressions following latent aggregation (Model 2c), the effect of support for democratic values was significant at the 10% level.[92]

5.4 Conclusion

In this paper, we addressed a methodological challenge well known to comparative survey researchers: how to study the effect of level two (L2) and level one (L1) predictors of a level two (L2) outcome so as to yield both reliable and valid results? Researchers have criticized simple aggregation for methodological and statistical reasons. Building on these insights and using the persistence of democracy as a substantive example, we compared the simple group means approach with two more advanced analytical strategies: the multilevel SEM approach, which estimates a latent L2 variable assumed to cause its L1 indicators, and a two-step approach, which relies on the L2 residuals of a multilevel model estimated prior to the analysis of interest (Griffin, 1997).

Our study corroborates previous critiques of the simple group-means approach. In both bivariate comparisons of countries' survival curves and more comprehensive multivariate event history analyses, we observed that support for democratic values was negatively associated with democratic breakdown. Unlike in the bivariate models, however, the multivariate models revealed that the associated significance levels of the estimates of

[92] The event-history models underlying Figure 5.6 are listed in Tables C3 to C5 (Appendix C).

support for democratic values differed remarkably depending on the aggregation method. Whereas support for democratic values was not significant in the regressions following simple group mean aggregation, confidence intervals suggested point estimates of higher precision when using either the multilevel SEM or the two-step approach, and the latter two approaches showed several significant effects at the 10% level.

These empirical results show that researchers can improve the validity of their inferences by choosing more advanced analytical strategies. First, the results match previous findings from simulation analyses (Lüdtke et al., 2008), which show that the simplest form of aggregation – manifest group means – is prone to beta or type-two errors in terms of false negative findings. Second, our results challenge Fails and Pierce's (2010) finding (based on simple aggregation) that support for democratic values has no effect on democracies' probability of decline. Our results suggest that comparative survey researchers interested in the effect of one or more L1 predictors on an L2 outcome may overestimate the standard errors of their regression coefficients when using manifest group mean aggregation.

The two more advanced analytical strategies have distinct methodological and statistical advantages. From a statistical perspective, the two-step approach performs somewhat poorer than the multilevel SEM approach: Given that simulation revealed regression coefficients after latent aggregation to be unbiased (Bennink et al., 2013, 2015; Lüdtke et al., 2008), researchers who apply the two-step approach may run the risk of committing type-one errors in terms of false positive findings. An evident methodological advantage of the two-step approach is, however, that it is particularly suited to simultaneously model situational, action formation, and transformational mechanisms in their entirety.

We conclude with several suggestions for future research. As of yet, no simulation analyses (similar to the ones comparing the simple group mean and the multilevel SEM approach) have been carried out for the two-step approach. It is therefore not possible to determine whether the estimated confidence intervals of the two-step approach are more or less reliable than the results of the latent aggregation approach. Hence, our first suggestion for future research is to perform a simulation analyses for all three aggregation methods. Controlling the data-generating mechanism would permit valid conclusions about the actual precision of each aggregation method compared to the 'real' effect size at L2.

Second, the latent aggregation model can be extended towards a *doubly-latent* model with controls for measurement error. Thus, our second suggestion for future research is to use multiple indicators of political support to arrive at a doubly-latent model of political support at L2. Depending on the results of the aforementioned simulation study, latent variable models and the two-step approach could eventually also be combined in order to estimate both situational and transformational mechanisms without falling prey to either measurement or sampling error. Moreover, if individuals' actual decisions such as turning out to vote or participating in demonstrations or public protests are considered, a combined framework of structural equation modeling and the two-step approach would allow researchers to map action-formation mechanisms as well.[93] Third, while we used a simple exponential event-history model to simplify the analysis, future research might make use of more flexible links for the survival function such as piecewise constant or frailty models.

In sum, we encourage comparative survey researchers to surpass the simple group means aggregation approach in favor of more advanced methods of analyzing contextual-level outcomes. We have shown that this helps researchers to circumvent beta or type-two errors in terms of false negative findings when using one or more L1 indicator to predict an L2 outcome. In addition, unlike the simple group means approach, these more advanced methods can be extended further, thereby facilitating the test of more theoretically valid models.

[93] Structural equation modeling can map action formation mechanisms in simple L1 regressions as well. In addition, for group-mean centered L1 variables, multilevel SEM can estimate situational mechanisms by computing the difference between L2 and L1 regression coefficients (Marsh et al., 2009).

5.5 Appendix C

Table C1
Distribution of all Indicators

		Count	Mean	SD	Min	Max
Level 1	Support for democratic values	269869	2.75	1.03	1	4
	Support for democratic values (dichotomized)	269869	0.59	0.49	0	1
	Age recoded	337018	3.1	1.57	1	6
	Highest educational level attained	296142	4.72	2.23	1	8
	Subjective social class	284337	2.68	0.99	1	5
Level 2	GDP	7998	7.62	1.64	3.51	12.11
	ELF	8573	0.47	0.27	0.00	1.00
	Support for democratic values	1007	0.60	0.18	0.01	0.97
	Age	1190	3.19	0.46	1.91	4.30
	Education	1076	4.74	0.80	2.53	6.79
	Subjective class	1022	2.69	0.28	1.70	3.69
	Residuals (null model)	921	-0.02	0.86	-4.84	3.06
	Residuals (Model A1)	921	0.00	0.93	-5.34	2.89
	Residuals (Model A2)	921	-0.01	0.95	-5.30	2.94
	Support for democratic values	1007	-0.02	0.82	-3.89	2.22
	Age	1058	0.07	0.53	-1.47	1.32
	Education	1034	0.09	0.65	-1.67	1.71
	Subjective class	1013	0.04	0.51	-1.67	1.35

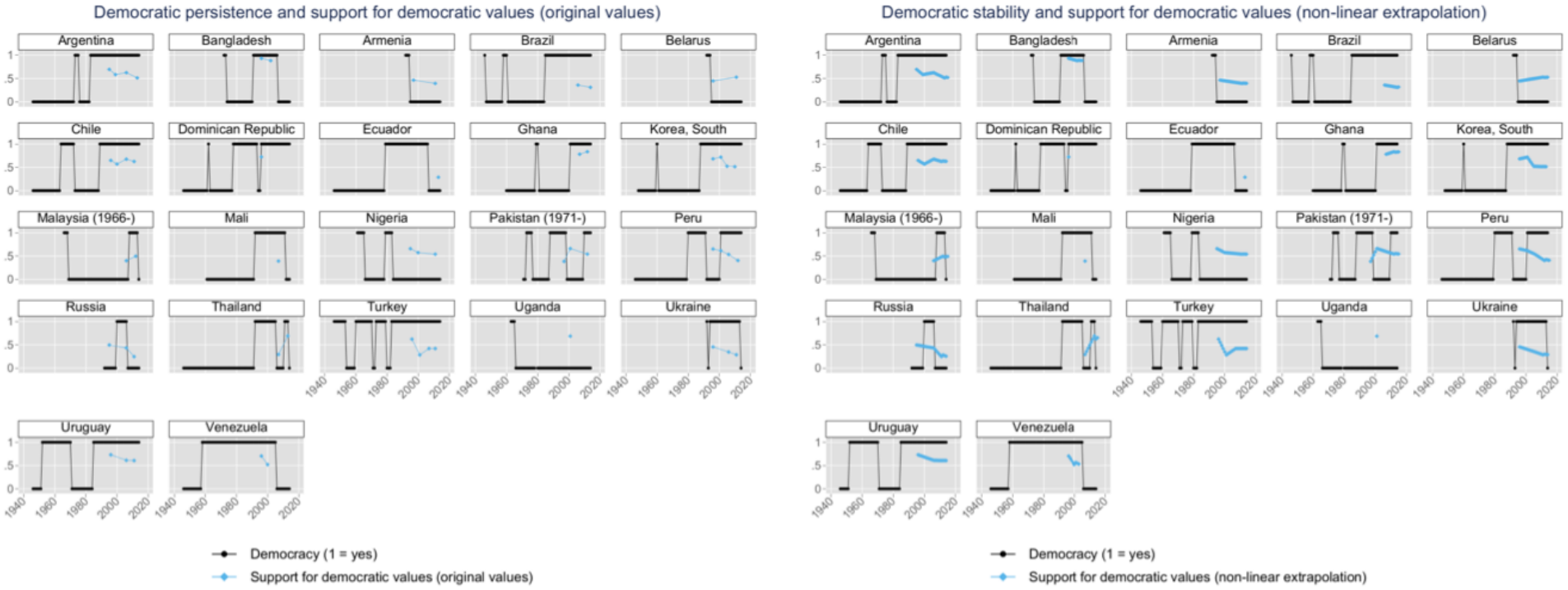


Figure C1
Distribution of Democratic Persistence and Support for Democratic Values across Country Years.

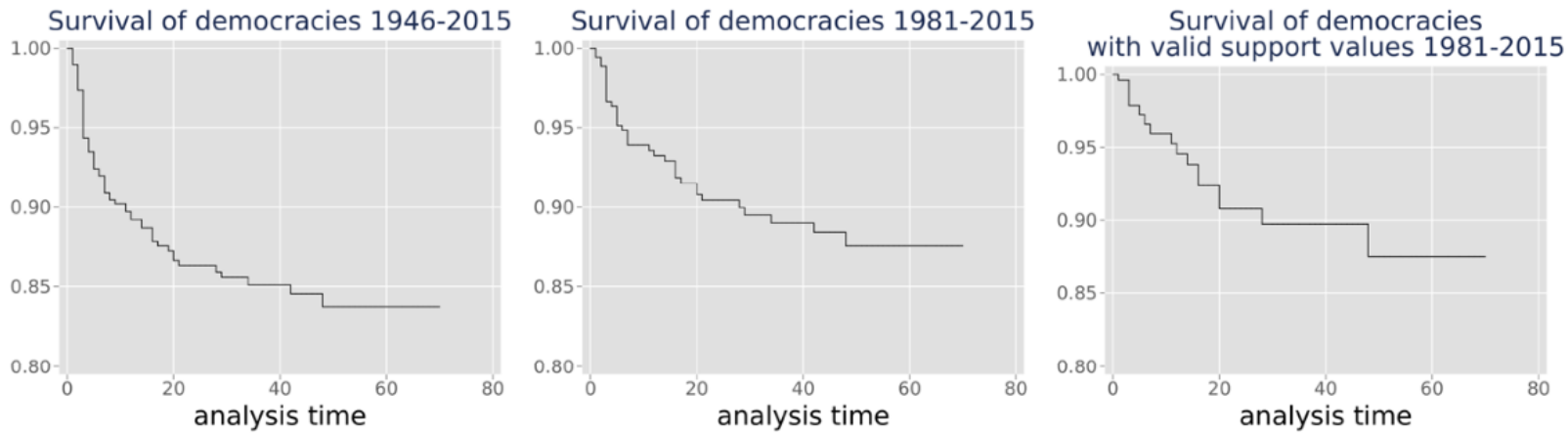


Figure C2
A Comparison of Democracies' Estimated Survival Rates across Different Samples of Analysis.

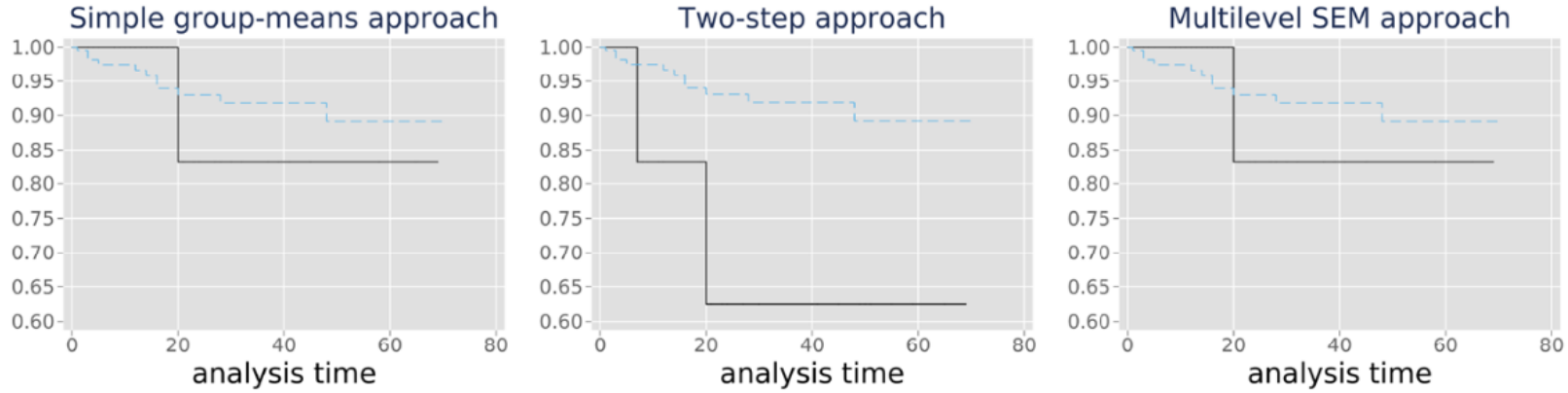


Figure C3
Survival of Democracies by Support for Democratic Values across Aggregation Methods.

Table C2
Multilevel Logistic Regression of Support for Democratic Values (Dichotomized) on Level-Two Predictors and Level-One Covariates

	Null model		Model 1a		Model 1b	
	b	se	b	se	b	se
Intercept	1.812**	(0.585)	2.042***	(0.580)	0.457***	(0.071)
log(GDP)			-0.174**	(0.061)	-0.166**	(0.061)
ELF			-0.392	(0.345)	-0.427	(0.341)
Age: 15-24 years			REFERENCE CATEGORY			
25-34 years					0.015	(0.015)
35-44 years					0.067***	(0.015)
45-54 years					0.103***	(0.017)
55-64					0.092***	(0.019)
65 and more years					-0.039	(0.020)
Education:			REFERENCE CATEGORY			
Inadequately completed elementary						
Completed elementary					0.042	(0.022)
Incomplete secondary: tech./voc.					0.051*	(0.025)
Completed secondary: tech./voc.					0.178***	(0.022)
Incomplete secondary: univ. prep.					0.171***	(0.024)
Complete secondary: univ. prep.					0.274***	(0.022)
Some university without degree					0.428***	(0.026)
University with degree					0.581***	(0.023)
Subjective class:			REFERENCE CATEGORY			
lower						
working					0.016	(0.017)
lower middle					0.042*	(0.017)
upper middle					-0.034	(0.019)
upper					-0.275***	(0.038)
τ0j	0.025	(0.063)	0.012	(0.063)	-0.045	(0.054)
N	219740		219740		219740	
AIC	261954		263445		263440	

Note. Random intercept model (QR decomposition) across country-years (level 2). Significance levels: * < .05; ** < .01; *** < .001 (two-sided). Standard errors in parentheses.

Table C3
Exponential Event-History Regression of Democratic Survival on Aggregated Support for Democratic Values, L2 Predictors, and Aggregated L1 Controls
(Simple Group-Means Approach)

	Model 1a	Model 1b	Model 1c
	b/se	b/se	b/se
Intercept	-3.220*	-3.662	-2.073
	(1.252)	(3.492)	(6.503)
Support for democratic values	-3.734	-3.642	-3.367
	(2.485)	(2.754)	(2.783)
log(GDP)		0.01	-0.038
		(0.399)	(0.432)
ELF		0.715	1.294
		(2.131)	(2.495)
Age			0.662
			(1.419)
Education			-0.315
			(0.685)
Subjective class			-0.846
			(1.887)
AIC	43.318	47.201	52.375
BIC	52.96	66.486	86.123
N (failures)	5	5	5
N (subjects)	122	122	122
N (observations)	917	917	917

Note. Significance levels: + < .10: * < .05; ** < .01; *** < .001 (two-sided). Standard errors in parentheses.

Table C4
Exponential Event-History Regression of Democratic Survival on Aggregated Support for Democratic Values, L2 Predictors, and Aggregated L1 Controls (Multilevel SEM Approach)

	Model 2a	Model 2b	Model 2c
	b/se	b/se	b/se
Intercept	-5.547***	-7.151+	-6.851
	(0.563)	(4.195)	(4.332)
Support for democratic values	-0.911+	-1.009+	-0.945
	(0.474)	(0.592)	(0.591)
GDP		0.132	0.064
		(0.428)	(0.461)
ELF		1.029	1.611
		(2.141)	(2.502)
Age			0.696
			(1.249)
Education			-0.644
			(0.769)
Subjective class			0.024
			(0.949)
AIC	42.444	46.179	51.203
BIC	52.086	65.463	84.951
N (failures)	5	5	5
N (subjects)	122	122	122
N (observations)	917	917	917

Note. Significance levels: + < .10: * < .05; ** < .01; *** < .001 (two-sided). Standard errors in parentheses.

Table C5
Exponential Event-History Regression of Democratic Survival on Residualized Support for Democratic Values (Two-Step Approach)

	Model 3a b/se	Model 3b b/se	Model 3c b/se
Intercept	-5.460***	-5.427***	-5.433***
	(0.525)	(0.517)	(0.520)
Residuals (Null model)	-0.754+		
	(0.389)		
Residuals (Model 3a)		-0.658+	
		(0.357)	
Residuals (Model 3b)			-0.651+
			(0.361)
AIC	42,813	43,047	43,089
BIC	52,455	52,689	52,731
N (failures)	5	5	5
N (subjects)	122	122	122
N (observations)	917	917	917

Note. Significance levels: + < .10: * < .05; ** < .01; *** < .001 (two-sided). Standard errors in parentheses.

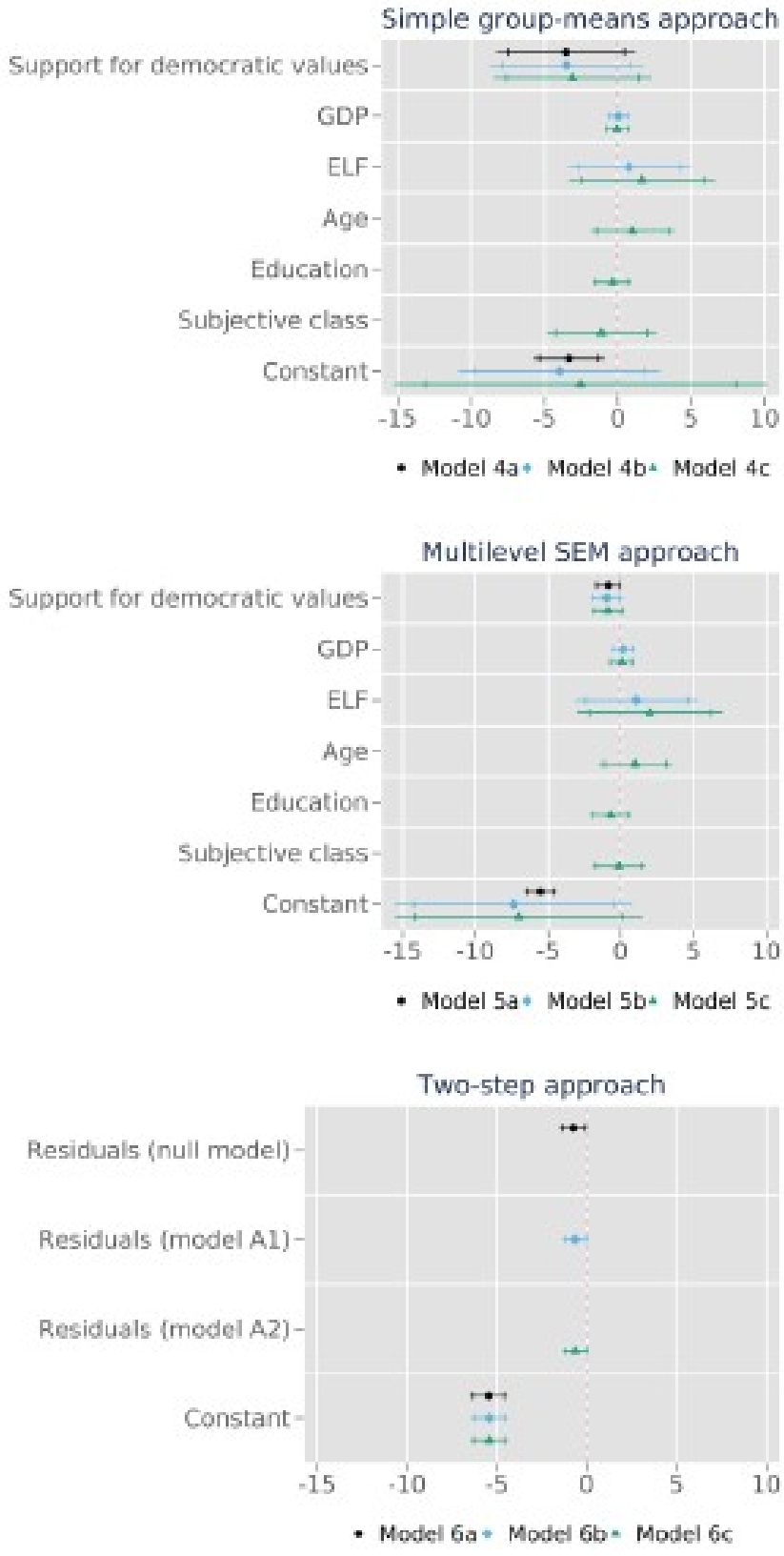


Figure C4
Point Estimates and Confidence Intervals of Countries' Democratic Survival across Aggregation Methods (Constant Interpolation).

References

Adcock, R., & Collier, D. (2001). Measurement validity: A shared standard for qualitative and quantitative research. *American Political Science Review*, *95*(3), 529–546. Retrieved from https://www.jstor.org/stable/3118231

Afrobarometer (2017a). *Round 7 survey manual*. Retrieved from Afrobarometer website: http://afrobarometer.org/sites/default/files/survey_manuals/ab_r7_survey_manual_en1.pdf

Afrobarometer (2017b). *Afrobarometer Round 7: The quality of democracy and governance in Botswana (Questionnaire)*. Retrieved from Afrobarometer website: http://afrobarometer.org/sites/default/files/questionnaires/Round%207/bot_r7_questionnaire_062018.pdf

Afrobarometer (2020a): *Surveys and methods*. Retrieved from Afrobarometer website: http://www.afrobarometer.org/surveys-and-methods

Afrobarometer (2020b): *Sampling principles and weighting*. Retrieved from Afrobarometer website: http://www.afrobarometer.org/surveys-and-methods/sampling-principles

Albert, H. (1991). *Traktat über die kritische Vernunft* (5th ed.). Tübingen, Germany: Mohr Siebeck.

Albert, H. (2000). *Kritischer Rationalismus: Vier Kapitel zur Kritik illusionären Denkens*. Tübingen, Germany: Mohr Siebeck.

Allenspach, D. (2012). *Der Effekt der Systemunterstützung auf die politische Partizipation: Eine vergleichende Analyse westlicher Demokratien*. Wiesbaden, Germany: VS Verlag für Sozialwissenschaften.

Almond, G. A. (1980). The intellectual history of the civic culture concept. In G. A. Almond & S. Verba (Eds.), *The civic culture revisited* (pp. 1–36). Boston, MA: Little, Brown & Company.

Almond, G. A. (1990). *A discipline divided: Schools and sects in Political Science*. Newbury Park, CA: Sage.

Almond, G. A., & Powell, G. B. (1978). *Comparative politics: System, process, and policy*. Boston, MA: Little, Brown & Company.

Almond, G. A., Powell Jr., G. B., Strom, K., & Dalton, R. J. (2003). *Comparative politics today: A world view* (7th ed.). New York, NY: Longman.

Almond, G. A., & Verba, S. (1965). *The civic culture: Political attitudes and democracy in five nations*. Boston, MA: Little, Brown & Company.

Altman, D., & Pérez-Liñán, A. (2002). Assessing the quality of democracy: Freedom, competitiveness and participation in eighteen Latin American Countries. *Democratization*, *9*(2), 85–100. doi:10.1080/714000256

American Psychological Association. (1954). Technical recommendations for psychological tests and diagnostic techniques. *Psychological Bulletin*, *51*(2, Pt.2), 1–38. doi:10.1037/h0053479

American Psychological Association, American Educational Research Association, & National Council on Measurements Used in Education. (1966). *Standards for education and psychological tests and manuals*. Washington, D.C.: American Psychological Association.

Anastasi, A. (1986). Evolving concepts of test validation. *Annual Review of Psychology*, *37*(February), 1–16. doi:10.1146/annurev.ps.37.020186.000245

Anderson, C. J., & Singer, M. M. (2008). The sensitive left and the impervious right: Multilevel models and the politics of inequality, ideology, and legitimacy in Europe. *Comparative Political Studies*, *41*(4/5), 564–599. doi:10.1177/001041 4007313113

André, S. (2014). Does trust mean the same for migrants and natives? Testing measurement models of political trust with multi-group confirmatory factor analysis. *Social Indicators Research*, *115*(3), 963–982. doi:10.1007/s11205-013-0246-6

Ariely, G. (2015). Trusting the press and political trust: A conditional relationship. *Journal of Elections, Public Opinion and Parties*, *25*(3), 351–367. doi:10.1080/17457289.2014.997739

Ariely, G., & Davidov, E. (2011). Can we rate public support for democracy in a comparable way? Cross-national equivalence of democratic attitudes in the World Values Survey. *Social Indicators Research*, *104*(2), 271–286. doi:10.1007/s11205-010-9693-5

Ariely, G., & Davidov, E. (2012). Assessment of measurement equivalence with cross-national and longitudinal surveys in political science. *European Political Science*, *11*(3), 363–377. doi:10.1057/eps.2011.11

Asian Barometer (2014-2016). *Fourth wave core questionnaire*. Retrieved from Asian Barometer website: http://www.asianbarometer.org/pdf/core_questionnaire_wave4.pdf

Asian Barometer (2020a). *Survey topics*. Retrieved from Asian Barometer website: http://www.asianbarometer.org/survey/survey-topics

Asian Barometer (2020b). *Survey methods*. Retrieved from Asian Barometer website: http://www.asianbarometer.org/survey/survey-methods

Asparouhov, T., & Muthén, B. O. (2010). *Weighted least squares estimation with missing data* (Mplus Technical Appendix). Retrieved from Mplus website: https://www.statmodel.com/download/GstrucMissingRevision.pdf

Axelrod, R. (1997). Advancing the art of simulation in the social sciences. In C. Rosaria, R. Hegselmann, & P. Terna (Eds.), *Simulating Social Phenomena* (pp. 21–40). Berlin, Germany: Springer VS.

Babbie, E. (2004). *The practice of social research* (10th ed.). Belmont, CA: Wadsworth.

Bachleitner, R., Weichbold, M., Aschauer, W., & Pausch, M. (2014). *Methodik und Methodologie interkultureller Umfrageforschung: Zur Mehrdimensionalität der funktionalen Äquivalenz*. Wiesbaden, Germany: Springer VS.

Baggini, J., & Fosl, P. S. (2010). *The philosopher's toolkit: A compendium of philosophical concepts and methods* (2nd ed.). Malden, MA: Wiley-Blackwell.

Baur, N., & Blasius, J. (2014). Methoden der empirischen Sozialforschung: Ein Überblick. In N. Baur & J. Blasius (Eds.), *Handbuch Methoden der empirischen Sozialforschung* (pp. 41–62). Wiesbaden, Germany: VS Verlag für Sozialwissenschaften.

Beauducel, A., & Herzberg, P. Y. (2006). On the performance of maximum likelihood versus means and variance adjusted weighted least squares estimation in CFA. *Structural Equation Modeling*, *13*(2), 186–203. doi:10.1207/s15328007sem1302_2

Becker, D., Beckers, T., Franzmann, S. T., & Hagenah, J. (2016). Contextualizing cognitive consonance by a social mechanisms explanation: Moderators of selective exposure in media usage. *Analyse & Kritik*, *38*(1), 149–178. doi:10.1515/auk-2016-0108

Becker, D., Breustedt, W., & Zuber, C. I. (2018). Surpassing simple aggregation: Advanced strategies for analyzing contextual-level outcomes in multilevel models. *Methods, Data, Analyses*, *12*(2), 233–263. doi:10.12758/mda.2017.05

Beetham, D., Carvalho, E., Landman, T., & Weir, S. (2008). *Assessing the quality of democracy: A practical guide*. Stockholm, Sweden: International IDEA. Retrieved from https://www.idea.int/sites/default/files/publications/assessing-the-quality-of-democracy-a-practical-guide.pdf

Beetham, D., Carvalho, E., & Weir, S. (2008). *Assessing the quality of democracy: An overview of the International IDEA framework*. Stockholm, Sweden: International IDEA. Retrieved from https://www.idea.int/sites/default/files/publications/chapters/assessing-the-quality-of-democracy/assessing-the-quality-of-democracy-an-overview-of-the-international-idea-framework.pdf

Beetham, D., & Weir, S. (2000). Democratic audit in comparative perspective. In H.-J. Lauth, G. Pickel, & C. Welzel (Eds.), *Demokratiemessung: Konzepte und Befunde im internationalen Vergleich* (pp. 73–88). Wiesbaden, Germany: Westdeutscher Verlag.

Behnke, J., Baur, N., & Behnke, N. (2006). *Empirische Methoden der Politikwissenschaft*. Paderborn, Germany: Schöningh.

Bennink, M., Croon, M. A., & Vermunt, J. K. (2013). Micro-macro multilevel analysis for discrete data: A latent variable approach and an application on personal network data. *Sociological Methods & Research*, *42*(4), 431–457. doi:10.1177/0049124113500479

Bennink, M., Croon, M. A., & Vermunt, J. K. (2015). Stepwise latent class models for explaining group-level outcomes using discrete individual-level predictors. *Multivariate Behavioral Research*, *50*(6), 662–675. doi:10.1080/00273171.2015.1074879

Bevir, M., & Kedar, A. (2008). Concept formation in Political Science: An anti-naturalist critique of qualitative methodology. *Perspectives on Politics*, *6*(3), 503–517. doi:10.1017/S1537592708081255

Blaikie, N. (2007). *Approaches to social enquiry: Advancing knowledge* (2nd ed.). Cambridge, United Kingdom: Polity Press.

Blalock, H. M. (1968). The measurement problem: A gap between the languages of theory and research. In H. M. Blalock & A. B. Blalock (Eds.), *Methodology in social research* (pp. 5–27). New York, NY: McGraw-Hill.

Bol, D., Giani, M., Blais, A., & Loewen, P. J. (2020). The effect of COVID−19 lockdowns on political support: Some good news for democracy? *European Journal of Political Research*. Advance online publication. doi:10.1111/1475-6765.12401

Bollen, K. A. (1979). Political democracy and the timing of development. *American Sociological Review*, *44*(4), 572–587. Retrieved from http://www.jstor.org/stable/2094588

Bollen, K. A., & Jackman, R. W. (1985). Economic and noneconomic determinants of political democracy in the 1960s. In R. G. Braungart & M. M. Braungart (Eds.), *Research in political sociology* (pp. 27–48). Greenwich, CT: JAI Press.

Booth, J. A., & Seligson, M. A. (2009). *The legitimacy puzzle in Latin America: Political support and democracy in eight nations*. New York, NY: Cambridge University Press.

Bowen, N. K., & Masa, R. D. (2015). Conducting measurement invariance tests with ordinal data: A guide for social work researchers. *Journal of the Society for Social Work and Research*, *6*(2), 229–249. doi:10.1086/681607

Bowman, K., Lehoucq, F., & Mahoney, J. (2005). Measuring political democracy: Case expertise, data adequacy, and Central America. *Comparative Political Studies*, *38*(8), 939–970. doi:10.1177/0010414005277083

Bratton, M., & Mattes, R. (2001). Support for democracy in Africa: Intrinsic or instrumental? *British Journal of Political Science*, *31*(3), 447–474. doi:10.1017/S0007123401000175P

Braun, D. (2013). *Politisches Vertrauen in neuen Demokratien*. Wiesbaden, Germany: VS Verlag für Sozialwissenschaften. doi:10.1007/978-3-658-01188-8

Braun, M. (2006). *Funktionale Äquivalenz in interkulturell vergleichenden Umfragen: Mythos und Realität*. Mannheim, Germany: Zentrum für Umfragen, Methoden und Analysen. Retrieved from https://www.ssoar.info/ssoar/handle/document/49125

Breustedt, W. (2015). The barometer surveys: Insights into the quality of the harmonized political trust items. *Harmonization, 1*(2), 7–11. Retrieved from http://consirt.osu.edu/wp-content/uploads/2015/11/Harmonization-Newsletter-v1n2-FALL-2015-with-ISSN-_FINAL-3.pdf

Breustedt, W. (2018). Testing the measurement invariance of political trust across the globe: A multiple group confirmatory factor analysis. *Methods, Data, Analyses*, *12*(1), 7–46. doi:10.12758/mda.2017.06

Breustedt, W., & Stark, T. (2015). Thinking outside the democratic box: Political values, performance and political support in authoritarian regimes; A comparative analysis. In C. Eder, I. Mochmann, & M. Quandt (Eds.), *Political trust and disenchantment with politics: International perspectives* (pp. 184–222). Leiden, Netherlands: Brill.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York, NY: Guilford Press.

Brunner, W., & Walz, D. (2000). Das politische Institutionenvertrauen in den 90er Jahren. In J. Falter, O. W. Gabriel, & H. Rattinger (Eds.), *Wirklich ein Volk? Die politischen Orientierungen von Ost- und Westdeutschen im Vergleich* (pp. 175–208). Opladen, Germany: Leske + Budrich.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models: Applications and data analysis methods*. Newbury Park, CA: Sage.

Bühlmann, M., Merkel, W., Müller, L., & Weßels, B. (2008). Wie lässt sich Demokratie am besten messen? Zum Forumsbeitrag von Thomas Müller und Susanne Pickel. *Politische Vierteljahresschrift*, *49*(1), 114–122. http://doi.org/10.1007/s11615-008-0089-y

Bühlmann, M., Merkel, W., & Wessels, B. (2008). *The quality of democracy: Democracy barometer for established democracies* (Challenges to Democracy in the 21st Century Working Paper No. 22). Retrieved from National Centre of Competence in Research website: http://www.nccr-democracy.uzh.ch/publications/workingpaper/pdf/WP10a.pdf

Bühlmann, M., Merkel, W., Müller, L., Giebler, H., & Wessels, B. (2012). Demokratiebarometer: Ein neues Instrument zur Messung von Demokratiequalität. *Zeitschrift für Vergleichende Politikwissenschaft*, *6*(1), 115–159. doi:10.1007/s12286-012-0129-2

Bühlmann, M., Merkel, W., Müller, L., & Wessels, B. (2012). The democracy barometer: A new instrument to measure the quality of democracy and its potential for comparative research. *European Political Science*, *11*(4), 519–536. doi:10.1057/eps.2011.46

Byrne, B. M. (2012). *Structural equation modeling with Mplus: Basic concepts, applications, and programming*. New York, NY: Routledge.

Byrne, B. M., Shavelson, R. J., & Muthén, B. O. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin*, *105*(3), 456–466. doi:10.1037/0033-2909.105.3.456

Byrne, B. M., & van de Vijver, F. J. R. (2010). Testing for measurement and structural equivalence in large-scale cross-cultural studies: Addressing the issue of nonequivalence. *International Journal of Testing*, *10*(2), 107–132. doi:10.1080/15305051003637306

Caballero, C. (2009). *Integration und politische Unterstützung: Eine empirische Untersuchung unter Ausländern*. Wiesbaden, Germany: VS Verlag für Sozialwissenschaften.

Campbell, D. F. J. (2008). *The basic concept for the democracy ranking of the quality of democracy*. Retrieved from Global Democracy Ranking website: http://democracyranking.org/ranking/2012/data/basic_concept_democracy_ranking_2008_A4.pdf

Campbell, D. F. J. (2012). Die österreichische Demokratiequalität in Perspektive. In L. Helms & D. M. Wineroither (Eds.), *Die österreichische Demokratie im Vergleich* (pp. 293–315). Baden-Baden, Germany: Nomos.

Campbell, D. F. J., & Barth, T. D. (2009). Wie können Demokratie und Demokratiequalität gemessen werden? Modelle, Demokratie-Indices und Länderbeispiele im globalen Vergleich. *SWS-Rundschau*, *49*(2), 209–233. Retrieved from https://nbn-resolving.org/urn:nbn:de:0168-ssoar-124717

Campbell, D. F. J., Barth, T. D., Pölzlbauer, P., & Pölzlbauer, G. (2012). *Democracy ranking: The quality of democracy in the world*. Norderstedt, Germany: Books on Demand.

Campbell, D. F. J., Carayannis, E. G., & Scheherazade, S. R. (2015). Quadruple helix structures of quality of democracy in innovation systems: The USA, OECD countries, and EU member countries in global comparison. *Journal of the Knowledge Economy*, *6*(3), 467–493. doi:10.1007/s13132-015-0246-7

Campbell, D. F. J., & Pölzlbauer, G. (2008). *The democracy ranking 2008 of the quality of democracy: Method and ranking outcome*. Retrieved from Global Democracy Ranking website: http://www.democracyranking.org/downloads/method_ranking_outcome_2008_A4.pdf

Campbell, D. F. J., Pölzlbauer, P., Barth, T. D., & Pölzlbauer, G. (2011). *Das "democracy ranking 2011 of the quality of democracy": Erstveröffentlichung, Konzept und Kontext*. Retrieved from Global Democracy Ranking website: http://democracyranking.org/wordpress/ranking/2011/data/Democracy_Ranking_Concept_Earlyrelease_German_2011.pdf

Campbell, D. F. J., & Sükösd, M. (2002). *Feasibility study for a quality ranking of democracies*. Retrieved from Global Democracy Ranking website: http://www.democracyranking.org/downloads/ feasibility_study-a4-e-01.pdf

Campbell, D. F. J., & Sükösd, M. (2003). *Global quality ranking of democracies: Pilot ranking 2000*. Retrieved from Global Democracy Ranking website: http://democracyranking.org/wordpress/ranking/ 2000/data/folder_a4-e-03.pdf

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*(2), 81–105. doi:10.1037/h0046016

Canache, D. (2012). Citizens' conceptualizations of democracy: Structural complexity, substantive content, and political significance. *Comparative Political Studies*, *45*(9), 1132–1158. doi:10.1177/0010414011434009

Canache, D., Mondak, J. J., & Seligson, M. A. (2001). Meaning and measurement in cross-national research on satisfaction with democracy. *Public Opinion Quarterly*, *65*(4), 506–528. Retrieved from https://www.jstor.org/stable/3078752

Carmines, E. G., & Zeller, R. A. (1979). *Reliability and validity assessment*. Beverly Hills, CA: Sage.

Catterberg, G., & Moreno, A. (2006). The individual bases of political trust: Trends in new and established democracies. *International Journal of Public Opinion Research*, *18*(1), 31–48. doi:10.1093/ijpor/edh081

Cautrès, B. (2011). Cross-national surveys. In B. Badie, D. Berg-Schlosser, & L. Morlino (Eds.), *International encyclopedia of Political Science* (pp. 505–509). Thousand Oaks, CA: Sage.

Center for Systemic Peace. (2017). *Polity IV Annual Time-Series 1800-2016 (p4v2016)* [Data file]. Retrieved from http://www.systemicpeace.org/inscrdata.html

Chalmers, A. F. (2013). *What is this thing called science* (4th ed.). Berkshire, United Kingdom: Open University Press.

Chang, E. C. C., & Chu, Y. (2006). Corruption and trust: Exceptionalism in Asian democracies? *Journal of Politics*, *68*(2), 259–271. doi:10.1111/j.1468-2508.2006.00404.x

Chen, F. F. (2008). What happens if we compare chopsticks with forks? The impact of making inappropriate comparisons in cross-cultural research. *Journal of Personality and Social Psychology*, *95*(5), 1005–1018. doi:10.1037/a0013193

Cho, Y. (2014). To know democracy is to love it: A cross-national analysis of democratic understanding and political support for democracy. *Political Research Quarterly*, *67*(3), 478–488. doi:10.1177/1065912914532721

Chu, Y., Welsh, B., & Chang, A. (2013). Congruence and variation in sources of regime support in East Asia. *Taiwan Journal of Democracy*, *9*(1), 221–237. Retrieved from http://www.tfd.org.tw/export/sites/tfd/files/publication/journal/dj0901/010.pdf

Citrin, J. (1974). Comment: The political relevance of trust in government. *American Political Science Review*, *68*(3), 973–988. doi:10.2307/1959141

Citrin, J., Levy, M., & Wright, M. (2014). Multicultural policy and political support in European democracies. *Comparative Political Studies*, *47*(11), 1531–1557. doi:10.1177/0010414013512604

Citrin, J., & Muste, C. (1999). Trust in government. In J. P. Robinson, P. R. Shaver, & L. S. Wrightsman (Eds.), *Measures of political attitudes* (pp. 465–532). San Diego, CA: Academic Press.

Cole, D. A., & Maxwell, S. E. (1985). Multitrait-multimethod comparisons across populations: A confirmatory factor analytic approach. *Multivariate Behavioral Research*, *20*(4), 389–417. doi:10.1207/s15327906mbr2004_3

Collier, D., Laporte, J., & Seawright, J. (2008). Typologies: Forming concepts and creating categorical variables. In J. M. Box-Steffensmeier, H. E. Brady, & D. Collier (Eds.), *The Oxford handbook of political methodology* (pp. 152–173). Oxford, United Kingdom: Oxford University Press.

Collier, D., LaPorte, J., & Seawright, J. (2012). Putting typologies to work: Concept formation, measurement, and analytic rigor. *Political Research Quarterly*, *65*(1), 217–232. doi:10.1177/1065912912437162

Collier, D., & Levitsky, S. (1997). Democracy with adjectives: Conceptual innovation in comparative research. *World Politics*, *49*(3), 430–451. doi:10.1353/wp.1997.0009

Collier, D., & Mahon, J. E. Jr. (1993). Conceptual "stretching" revisited: Adapting categories in comparative analysis. *American Political Science Review*, *87*(4), 845–855. doi:10.2307/2938818

Coltman, T., Devinney, T. M., Midgley, D. F., & Venaik, S. (2008). Formative versus reflective measurement models: Two applications of formative measurement. *Journal of Business Research*, *61*(12), 1250–1262. doi:10.1016/j.jbusres.2008.01.013

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Chicago, IL: Rand McNally.

Coromina, L., & Davidov, E. (2013). Evaluating measurement invariance for social and political trust in Western Europe over four measurement time points (2002-2008). *Ask. Research and Methods*, *22*(1), 37–54. Retrieved from https://askresearchand methods.files.wordpress.com/2014/01/ask_22-2013-coromina-davidov-37-54.pdf

Costner, H. L. (1969). Theory, deduction, and rules of correspondence. *American Journal of Sociology*, *75*(2), 245–263. Retrieved from https://www.jstor.org/ stable/2776106

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281–302. doi:10.1037/h0040957

Croon, M. A., & van Veldhoven, M. J. P. M. (2007). Predicting group-level outcome variables from variables measured at the individual level: A latent variable multilevel model. *Psychological Methods*, *12*(1), 45–57. doi:10.1037/1082-989X.12.1.45

Czada, R. (2010). Good Governance als Leitkonzept für Regierungshandeln: Grundlagen, Anwendungen, Kritik. In A. Benz & N. Dose (Eds.), *Governance: Regieren in komplexen Regelsystemen; Eine Einführung* (2nd ed., pp. 201–224). Wiesbaden, Germany: VS Verlag für Sozialwissenschaften.

Dahl, R. A. (1970). *A preface to democratic theory*. Chicago, IL: University of Chicago Press.

Dalton, R. J. (2004). *Democratic challenges, democratic choices: The erosion of political support in advanced industrial democracies*. Oxford, United Kingdom: Oxford University Press.

Dalton, R. J., Shin, D. C., & Jou, W. (2007). Understanding democracy: Data from unlikely places. *Journal of Democracy*, *18*(4), 142–156. Retrieved from https://www.muse.jhu.edu/article/223229

Dalton, R. J., & Welzel, C. (Eds.). (2014). *The civic culture transformed: From allegiant to assertive citizens*. New York, NY: Cambridge University Press.

Davidov, E., Datler, G., Schmidt, P., & Schwartz, S. H. (2011). Testing the invariance of values in the Benelux countries with the European social survey: Accounting for ordinality. In E. Davidov, P. Schmidt, & J. Billiet (Eds.), *Cross-Cultural analysis: Methods and applications* (pp. 149–171). New York, NY: Routledge.

Davidov, E., Dülmer, H., Schlüter, E., Schmidt, P., & Meuleman, B. (2012). Using a multilevel structural equation modeling approach to explain cross-cultural measurement noninvariance. *Journal of Cross-Cultural Psychology*, *43*(4), 558–575. doi:10.1177/0022022112438397

Davidov, E., Meulemann, B., Cieciuch, J., Schmidt, P., & Billiet, J. (2014). Measurement equivalence in cross-national research. *Annual Review of Sociology*, *40*(July), 55–75. doi:10.1146/annurev-soc-071913-043137

della Porta, D., & Keating, M. (2008). How many approaches in the social sciences? An epistemological introduction. In D. della Porta & M. Keating (Eds.), *Approaches and methodologies in the social sciences: A pluralist perspective* (pp. 19–39). Cambridge, United Kingdom: Cambridge University Press.

Democracy Barometer. (2014a). *About us*. Retrieved from Democracy Barometer website: http://www.democracybarometer.org/about_en.html

Democracy Barometer. (2014b). *Concept.* Retrieved from Democracy Barometer website: http://democracybarometer.org/concept_en.html

Denters, B., Gabriel, O. W., & Torcal, M. (2007). Political confidence in representative democracies: Socio-cultural vs. political explanations. In J. W. van Deth, J. R. Montero, & A. Westholm (Eds.), *Citizenship and involvement in European democracies: A comparative analysis* (pp. 66–87). London, United Kingdom: Routledge.

Diamantopoulos, A., & Winklhofer, H. M. (2001). Index construction with formative indicators: An alternative to scale development. *Journal of Marketing Research*, *38*(2), 269–277. Retrieved from https://www.jstor.org/stable/1558630

Diamond, L. J. (2016). *In search of democracy*. Oxon, United Kingdom: Routledge.

Diamond, L. J., & Morlino, L. (2004a). The quality of democracy: An overview. *Journal of Democracy*, *15*(4), 20–31. doi:10.1353/jod.2004.0060

Diamond, L. J., & Morlino, L. (2004b). *The quality of democracy* (CDDRL Working Papers No. 20). Retrieved from CDDRL website: https://cddrl.fsi.stanford.edu/publications/the_quality_of_democracy

Diamond, L. J., & Morlino, L. (2005). Introduction. In L. Diamond & L. Morlino (Eds.), *Assessing the quality of democracy* (pp. ix–xliii). Baltimore, MD: The Johns Hopkins University Press.

Diekmann, A. (2013). *Empirische Sozialforschung: Grundlagen, Methoden, Anwendungen*. Reinbek bei Hamburg, Germany: Rowohlt Taschenbuch Verlag.

Döring, N., & Bortz, J. (2016). *Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften* (5th ed.). Berlin, Germany: Springer VS.

Dreier, V. (1997). *Empirische Politikforschung*. München, Germany: Oldenbourg Verlag.

Easton, D. (1965). *A systems analysis of political life*. New York, NY: Wiley.

Easton, D. (1975). A re-assessment of the concept of political support. *British Journal of Political Science*, *5*(4), 435–457. doi:10.1017/S0007123400008309

Edwards, J. R. (2011). The fallacy of formative measurement. *Organizational Research Methods*, *14*(2), 370–388. doi:10.1177/1094428110378369

Elder, J. W. (2005). Sociology. In K. Kempf-Leonard (Ed.), *Encyclopedia of social measurement* (pp. 559–568). Amsterdam, Netherlands: Elsevier.

Erdmann, G. (2011). Decline of democracy: Loss of quality, hybridisation and breakdown of democracy. In G. Erdmann & M. Kneuer (Eds.), *Regression of democracy? Sonderheft der Zeitschrift für Vergleichende Politikwissenschaft 1/2011* (pp. 21–58). Wiesbaden, Germany: VS Verlag für Sozialwissenschaften.

Erdmann, G., & Kneuer, M. (2011). Introduction. In G. Erdmann & M. Kneuer (Eds.), *Regression of democracy? Sonderheft der Zeitschrift für Vergleichende Politikwissenschaft 1/2011* (pp. 9–19). Wiesbaden, Germany: VS Verlag für Sozialwissenschaften.

Erlingsson, G. Ó., Linde, J., & Öhrvall, R. (2016). Distrust in utopia? Public perceptions of corruption and political support in Iceland before and after the financial crisis of 2008. *Government and Opposition*, *51*(4), 553–579. doi:10.1017/gov.2014.46

Eurobarometer (2019). *Eurobarometer 91.5: Basic bilingual questionnaire*. Retrieved from GESIS data catalog: https://search.gesis.org/research_data/ZA7576

Eurobarometer (2020a). *Sampling and fieldwork*. Retrieved from GESIS website: https://www.gesis.org/eurobarometer-data-service/survey-series/standard-special-eb/sampling-and-fieldwork

Eurobarometer (2020b). *Population, countries & regions*. Retrieved from GESIS website: https://www.gesis.org/eurobarometer-data-service/survey-series/standard-special-eb/population-countries-regions

European Social Survey. (2012a). *Round 6* (edition 2.1) [Data file]. Retrieved from European Social Survey website: http://www.europeansocialsurvey.org/data/download.html?r=6.

European Social Survey. (2012b). *ESS6 - 2012 Documentation Report* (edition 2.3). Retrieved from European Social Survey website: https://www.europeansocialsurvey.org/docs/round6/survey/ESS6_data_documentation_report_e02_3.pdf

European Social Survey. (2013). *Round 6 Module on European's understandings and evaluations of democracy: Final module in template*. Retrieved from European Social Survey website: http://www.europeansocialsurvey.org/docs/round6/questionnaire/ESS6_final_understandings_and_evaluation_of_democracy_module_template.pdf

European Social Survey. (2014). *Weighting European Social Survey data*. Retrieved from European Social Survey website: http://www.europeansocialsurvey.org/docs/methodology/ESS_weighting_data_1.pdf

European Social Survey (2018). *Source questionnaire: Round 9 2018/2019*. Retrieved from European Social Survey website: https://www.europeansocialsurvey.org/docs/round9/fieldwork/source/ESS9_source_questionnaires.pdf

European Social Survey (2020a). *Data collection*. Retrieved from European Social Survey website: https://www.europeansocialsurvey.org/methodology/ess_methodology/data_collection.html

European Social Survey (2020b). *Sampling*. Retrieved from European Social Survey website: https://www.europeansocialsurvey.org/methodology/ess_methodology/sampling.html

European Values Study (2018). *Master questionnaire*. Retrieved from GESIS data catalog: https://dbk.gesis.org/dbksearch/download.asp?id=66239

European Values Study (2020a). *Methodology EVS 2017.* Retrieved from European Values Study website: https://europeanvaluesstudy.eu/methodology-data-documentation/survey-2017/methodology/

European Values Study (2020b). *Participating countries and country-information*: Survey 2017. Retrieved from European Values Study website: https://europeanvaluesstudy.eu/methodology-data-documentation/survey-2017/pre-release-evs-2017/participating-countries-and-country-information-survey-2017/

Fails, M. D., & Pierce, H. N. (2010). Changing mass attitudes and democratic deepening. *Political Research Quarterly*, *63*(1), 174–187. doi:10.1177/ 1065912908327603

Fearon, J. D. (2003). Ethnic and cultural diversity by country. *Journal of Economic Growth*, *8*(2), 195–222. doi:10.1023/A:1024419522867

Feldman, S. (1983). The measurement and meaning of trust in government. *Political Methodology*, *9*(3), 341–354. Retrieved from https://www.jstor.org/stable/ 25791197

Fisher, J., van Heerde, J., & Tucker, A. (2010). Does one trust judgement fit all? Linking theory and empirics. *British Journal of Politics & International Relations*, *12*(2), 161–188. doi:10.1111/j.1467-856X.2009.00401.x

Foschi, M. (1997). On scope conditions. *Small Group Research*, *28*(4), 535–555. doi:10.1177/1046496497284004

Freedom House. (2014). *Freedom in the world.* Retrieved from Freedom House website: http://www.freedomhouse.org/reporttypes/freedom-world.

Friedrichs, J. (1981). *Methoden der empirischen Sozialforschung*. Opladen, Germany: Westdeutscher Verlag.

Fuchs, D. (1989). *Die Unterstützung des politischen Systems der Bundesrepublik Deutschland*. Opladen, Germany: Westdeutscher Verlag.

Fuchs, D. (1998). Kriterien demokratischer Performanz in liberalen Demokratien. In M. T. Greven (Ed.), *Demokratie: Eine Kultur des Westens? 20. wissenschaftlicher Kongreß der Deutschen Vereinigung für Politische Wissenschaft* (pp. 151–179). Opladen, Germany: Leske + Budrich.

Fuchs, D. (2002). Das Konzept der politischen Kultur: Die Fortsetzung einer Kontroverse in konstruktiver Absicht. In D. Fuchs, E. Roller, & B. Weßels (Eds.), *Bürger und Demokratie in Ost und West: Studien zur politischen Kultur und zum politischen Prozess; Festschrift für Hans-Dieter Klingemann* (pp. 27–49). Wiesbaden, Germany: Westdeutscher Verlag.

Fuchs, D. (2007). Political culture paradigm. In R. J. Dalton & H.-D. Klingemann (Eds.), *The Oxford handbook of political behavior* (pp. 161–184). Oxford, United Kingdom: Oxford University Press.

Fuchs, D., Gabriel, O. W., & Völkl, K. (2002). Vertrauen in politische Institutionen und politische Unterstützung. *Österreichische Zeitschrift für Politikwissenschaft*, *31*(4), 427–450. doi:10.15203/ozp.816.vol31iss4

Fuchs, D., & Roller, E. (2008). Die Konzeptualisierung der Qualität von Demokratie: Eine kritische Diskussion aktueller Ansätze. In A. Brodocz, M. Llanque, & G. S. Schaal (Eds.), *Bedrohungen der Demokratie* (pp. 77–96). Wiesbaden, Germany: VS Verlag für Sozialwissenschaften.

Furlong, P., & Marsh, D. (2010). A skin is not a sweater: Ontology and epistemology in political science. In D. Marsh & G. Stoker (Eds.), *Theory and methods in political science* (3rd ed., pp. 184–211). Basingstoke, United Kingdom: Palgrave Macmillan.

Gabriel, O. W. (1998). Fragen an einen europäischen Vergleich. In R. Köcher & J. Schild (Eds.), *Wertewandel in Deutschland und Frankreich: Nationale Unterschiede und europäische Gemeinsamkeiten* (pp. 29–51). Wiesbaden, Germany: VS Verlag für Sozialwissenschaften.

Gabriel, O. W. (1999). Integration durch Institutionenvertrauen? Struktur und Entwicklung des Verhältnisses der Bevölkerung zum Parteienstaat und zum Rechtsstaat im vereinigten Deutschland. In J. Friedrichs & W. Jagodzinski (Eds.), *Soziale Integration* (pp. 199–235). Wiesbaden, Germany: Westdeutscher Verlag.

Gabriel, O. W. (2018). Untergräbt die Kritik an Politikern das Vertrauen in die demokratischen Institutionen? *Zeitschrift für Parlamentsfragen*, *49*(4), 909–918. doi:10.5771/0340-1758-2018-4-909

Gastil, R. D. (1988). *Freedom in the world: Political rights and civil liberties 1987-1988*. New York, NY: Freedom House.

Gehring, U. W., & Weins, C. (2009). *Grundkurs Statistik für Politologen und Soziologen* (5th ed.). Wiesbaden, Germany: VS Verlag für Sozialwissenschaften.

Gerring, J. (1999). What makes a concept good ? A criterial framework for understanding concept formation in the social sciences. *Polity*, *31*(3), 357–393. doi:10.2307/3235246

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

Goertz, G. (2006). *Social science concepts: A user's guide*. Princeton, NJ: Princeton University Press.

Goertz, G. (2008). Concepts, theories, and numbers: A checklist for constructing, evaluating, and using concepts or quantitative measures. In J. M. Box-Steffensmeier, H. E. Brady, & D. Collier (Eds.), *The Oxford handbook of political methodology* (pp. 97–118). Oxford, United Kingdom: Oxford University Press.

Göhler, G. (2002). Stufen politischen Vertrauens. In R. Schmalz-Bruns & R. Zintl (Eds.), *Politisches Vertrauen: Soziale Grundlagen reflexiver Kooperation* (pp. 221–238). Baden-Baden, Germany: Nomos.

Govier, T. (2010). *A practical study of argument*. Belmont, CA: Wadsworth Cengage Learning.

Griffin, M. A. (1997). Interaction between individuals and situations: Using HLM procedures to estimate reciprocal relationships. *Journal of Management*, *23*(6), 759–773. doi:10.1016/S0149-2063(97)90028-3

Gujarati, D. N. (2004). Bias. In M. S. Lewis-Beck, A. Bryman, & T. F. Liao (Eds.), *The Sage encyclopedia of social science research methods* (pp. 65–67). Thousand Oaks, CA: Sage.

Hammersley, M., & Gomm, R. (1997). Bias in social research. *Sociological Research Online*, *2*(1), 4–13. doi:10.5153/sro.55

Hardin, R. (2000). The public trust. In S. J. Pharr & R. D. Putnam (Eds.), *Disaffected democracies: What's troubling the trilateral countries?* (pp. 31–51). Princeton, NJ: Princeton University Press.

Hartig, J., Frey, A., & Jude, N. (2008). Validität. In H. Moosbrugger & A. Kelava (Eds.), *Testtheorie und Fragebogenkonstruktion* (pp. 135–163). Heidelberg, Germany: Springer Medizin Verlag.

Heath, A., Martin, J., & Spreckelsen, T. (2009). Cross-national comparability of survey attitude measures. *International Journal of Public Opinion Research*, *21*(3), 293–315. doi:10.1093/ijpor/edp034

Hedström, P., & Swedberg, R. (Eds.). (1998). *Social mechanisms: An analytical approach to social theory*. Cambridge, United Kingdom: Cambridge University Press.

Hedström, P., & Ylikoski, P. (2010). Causal mechanisms in the social sciences. *Annual Review of Sociology*, *36*(August), 49–67. doi:10.1146/annurev.soc.012809.102632

Herrera, Y. M., & Kapur, D. (2007). Improving data quality: Actors, incentives, and capabilities. *Political Analysis*, *15*(4), 365–386. doi:10.1093/pan/mpm007

Hooghe, M. (2011). Why there is basically only one form of political trust. *British Journal of Politics and International Relations*, *13*(2), 269–275. doi:10.1111/j.1467-856X.2010.00447.x

Hooghe, M., & Kern, A. (2015). Party membership and closeness and the development of trust in political institutions: An analysis of the European Social Survey, 2002-2010. *Party Politics*, *21*(6), 944–956. doi:10.1177/1354068813509519

Hopkin, J. (2010). The comparative method. In D. Marsh & G. Stoker (Eds.), *Theory and methods in political science* (3rd ed., pp. 285–307). Basingstoke, United Kingdom: Palgrave Macmillan.

Horn, J. L., & Mcardle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research*, *18*(3), 117–144. doi:10.1080/03610739208253916

Horowitz, D. L. (1985). *Ethnic groups in conflict*. Berkeley, CA: University of California Press.

Hox, J. J. (2010). *Multilevel analysis: Techniques and applications* (2nd ed.). New York, NY: Routledge.

Hutchison, M. L., & Johnson, K. (2011). Capacity to trust? Institutional capacity, conflict, and political trust in Africa, 2000-2005. *Journal of Peace Research*, *48*(6), 737–752. doi:10.1177/0022343311417981

Indrayan, A. (n.d.). *Varieties of bias to guard against*. Retrieved from http://www.medicalbiostatistics.com/Types%20of%20bias.pdf

Inglehart, R. F., & Welzel, C. (2005). *Modernization, cultural change and democracy: The human development sequence*. Cambridge, United Kingdom: Cambridge University Press.

Jäckle, S., & Bauschke, R. (2009). Lässt sich Reformfähigkeit messen? Eine kritische Würdigung der Sustainable Governance Indicators. *Zeitschrift für Politikwissenschaft*, *19*(3), 359–386. doi:10.5771/1430-6387-2009-3-359

Jäckle, S., & Bauschke, R. (2010). Die Problematik bleibt bestehen: Antwort auf die Replik von Martin Brusis zu den Sustainable Governance Indicators. *Zeitschrift für Politikwissenschaft*, *20*(1), 79–88. doi:10.5771/1430-6387-2010-1-79

Jäckle, S., Wagschal, U., & Bauschke, R. (2012). Das Demokratiebarometer: “Basically theory driven”? *Zeitschrift für Vergleichende Politikwissenschaft*, *6*(1), 99–125. doi:10.1007/s12286-012-0133-6

Jäckle, S., Wagschal, U., & Bauschke, R. (2013). Allein die Masse macht's nicht: Antwort auf die Replik von Merkel et al. zu unserer Kritik am Demokratiebarometer. *Zeitschrift für Vergleichende Politikwissenschaft*, *7*(2), 143–153. doi:10.1007/s12286-013-0148-7

Jackman, S. (2008). Measurement. In J. M. Box-Steffensmeier, H. E. Brady, & D. Collier (Eds.), *The Oxford handbook of political methodology* (pp. 119–151). Oxford, United Kingdom: Oxford University Press.

Jarvis, C. B., Mackenzie, S. B., & Podsakoff, P. M. (2003). A critical review of construct indicators and measurement model misspecification in marketing and consumer research. *Journal of Consumer Research*, *30*(2), 199–218. doi:10.1086/376806

Juni, S. (2007). Reliability Theory. In N. J. Salkind & K. Rasmussen (Eds.), *Encyclopedia of measurement and statistics* (pp. 834–835). Thousand Oaks, CA: Sage.

Kaase, M. (1983). Sinn oder Unsinn des Konzepts "politische Kultur" für die vergleichende Politikwissenschaft, oder auch: Der Versuch, einen Pudding an die Wand zu nageln. In M. Kaase & H.-D. Klingemann (Eds.), *Wahlen und politisches System: Analysen aus Anlaß der Bundestagswahl 1980* (pp. 144–171). Opladen, Germany: Westdeutscher Verlag.

Kaina, V. (2008). Die Messbarkeit von Demokratiequalität als ungelöstes Theorieproblem: Zum PVS-Forums-Beitrag von Marc Bühlmann, Wolfgang Merkel, Lisa Müller und Bernhard Wessels. *Politische Vierteljahresschrift*, *49*(3), 518–524. doi:10.1007/s11615-008-0109-y

Kaiser, A., Lehnert, M., Miller, B., & Sieberer, U. (2002). The democratic quality of institutional regimes: A conceptual framework. *Political Studies*, *50*(2), 313–331. doi:10.1111/1467-9248.t01-1-00372

Kankaraš, M., Vermunt, J. K., & Moors, G. (2011). Measurement equivalence of ordinal items: A comparison of factor analytic, item response theory, and latent class approaches. *Sociological Methods & Research*, *40*(2), 279–310. doi:10.1177/0049124111405301

Kimberlin, C. L., & Winterstein, A. G. (2008). Validity and reliability of measurement instruments used in research. *American Journal of Health-System Pharmacy*, *65*(23), 2276–2284. doi:10.2146/ajhp070364

King, G., Keohane, R. O., & Verba, S. (1994). *Designing social inquiry: Scientific inference in qualitative research*. Princeton, NJ: Princeton University Press.

Kline, R. B. (2016). *Principles and practices of structural equation modeling* (4th ed.). New York, NY: The Guilford Press.

Klingemann, H.-D. (1999). Mapping political support in the 1990s: A global analysis. In P. Norris (Ed.), *Critical citizens: Global support for democratic government* (pp. 31–56). Oxford, United Kingdom: Oxford University Press.

Klingemann, H.-D., Fuchs, D., & Zielonka, J. (Eds.). (2006). *Democracy and political culture in Eastern Europe*. New York, NY: Routledge.

Kneuer, M. (2011). Deficits in democratic quality? The effects of party-system institutionalisation on the quality of democracy in Central Eastern Europe. In G. Erdmann & M. Kneuer (Eds.), *Regression of democracy? Sonderheft der Zeitschrift für Vergleichende Politikwissenschaft 1/2011* (pp. 133–171). Wiesbaden, Germany: VS Verlag für Sozialwissenschaften.

Kozlowski, S. W. J., & Klein, K. J. (2000). A multilevel approach to theory and research in organizations: Contextual, temporal, and emergent processes. In K. Klein & S. W. J. Kozlowski (Eds.), *Multilevel theory, research, and methods in organizations: Foundations, extensions, and new directions* (pp. 3–90). San Francisco, CA: Jossey-Bass.

Krantz, D. H., Luce, R. D., Suppes, P., & Tversky, A. (2007). *Foundations of measurement* (Volume I: Additive and polynomial representations). Mineola, NY: Dover. (Original work published 1971)

Krishnakumar, J., & Nagar, A. L. (2008). On exact statistical properties of multidimensional indices based on principal components, factor analysis, MIMIC and structural equation models. *Social Indicators Research*, *86*(3), 481–496. doi:10.1007/s11205-007-9181-8

Landman, T. (2012). Assessing the quality of democracy: The international IDEA framework. *European Political Science*, *11*(4), 456–468. doi:10.1057/eps.2011.49

Latinobarómetro (2018a). *Questionnaire*. Retrieved from Latinobarómetro website: http://www.latinobarometro.org/latContents.jsp

Latinobarómetro (2018b). *Methodological Report 2018*. Retrieved from Latinobarómetro website http://www.latinobarometro.org/latContents.jsp

Lauth, H.-J. (2004). *Demokratie und Demokratiemessung: Eine konzeptionelle Grundlegung für den interkulturellen Vergleich*. Wiesbaden, Germany: VS Verlag für Sozialwissenschaften.

Lauth, H.-J. (2009). Typologien in der vergleichenden Politikwissenschaft: Überlegungen zum Korrespondenzproblem. In S. Pickel, G. Pickel, H.-J. Lauth, & D. Jahn (Eds.), *Methoden der vergleichenden Politik- und Sozialwissenschaft: Neue Entwicklungen und Anwendungen* (pp. 153–172). Wiesbaden, Germany: VS Verlag für Sozialwissenschaften.

Lauth, H.-J. (2011a). Qualitative Ansätze der Demokratiemessung. *Zeitschrift für Staats- und Europawissenschaften*, *9*(1), 49–77.

Lauth, H.-J. (2011b). Quality criteria for democracy: Why responsiveness is not the key. In G. Erdmann & M. Kneuer (Eds.), *Regression of democracy? Sonderheft der Zeitschrift für Vergleichende Politikwissenschaft 1/2011* (pp. 59–80). Wiesbaden, Germany: VS Verlag für Sozialwissenschaften.

Lauth, H.-J. (2015). *The matrix of democracy: A three-dimensional approach to measuring the quality of democracy and regime transformations* (Würzburger Arbeitspapiere zur Politikwissenschaft und Sozialforschung No. 6). doi:10.25972/OPUS-10966

Lauth, H.-J. (2016). The internal relationships of the dimensions of democracy: The relevance of trade-offs for measuring the quality of democracy. *International Political Science Review*, *37*(5), 606–617. doi:10.1177/0192512116667630

Lauth, H.-J., & Kauff, O. (2012). *Demokratiemessung: Der KID als aggregiertes Maß für die komparative Forschung; Empirische Befunde der Regimeentwicklung von 1996 bis 2010* (Würzburger Arbeitspapiere zur Politikwissenschaft und Sozialforschung No. 2). doi:10.25972/OPUS-6174

Lauth, H.-J., Pickel, G., & Pickel, S. (2014). *Vergleich politischer Systeme*. Paderborn, Germany: Ferdinand Schöningh.

Lauth, H.-J., & Schlenkrich, O. (2018). Making trade-offs visible: Theoretical and methodological considerations about the relationship between dimensions and institutions of democracy and empirical findings. *Politics and Governance*, *6*(1), 78–91. doi:10.17645/pag.v6i1.1200

Lauth, H.-J., & Schlenkrich, O. (2019a). *Conception of the democracy matrix*. Retrieved from Democracy Matrix website: https://www.democracymatrix.com/conception

Lauth, H.-J., & Schlenkrich, O. (2019b). *Core measurement in the democracy matrix*. Retrieved from Democracy Matrix website: https://www.democracymatrix.com/concept-tree-operationalisation/core-measurement

Lauth, H.-J., & Schlenkrich, O. (2019c). *Aggregation in the democracy matrix*. Retrieved from Democracy Matrix website: https://www.democracymatrix.com/ aggregation

Lauth, H.-J., & Schlenkrich, O. (2019d). *Concept tree and operationalisation of the democracy matrix*. Retrieved from Democracy Matrix website: https://www.democracymatrix.com/concept-tree-operationalisation

Lauth, H.-J., & Schlenkrich, O. (2019e). *Brief presentation of the democracy matrix*. Retrieved from Democracy Matrix website: https://www.democracymatrix.com/brief-presentation

Lazarsfeld, P. F., & Menzel, H. (1961). On the relation between individual and collective properties. In A. Etzioni (Ed.), *A sociological reader on complex organizations* (2nd ed., pp. 499–516). New York, NY: Holt, Rinehart and Winston.

Levi, M., & Stoker, L. (2000). Political trust and trustworthiness. *Annual Review of Political Science*, *3*(1), 475–507. doi:10.1146/annurev.polisci.3.1.475

Levine, D. H., & Molina, J. E. (2011a). The quality of democracy: Strengths and weaknesses in Latin America. In D. H. Levine & J. E. Molina (Eds.), *The quality of democracy in Latin America* (pp. 253–260). Boulder, CO: Lynne Rienner Publishers.

Levine, D. H., & Molina, J. E. (2011b). Evaluating the quality of democracy in Latin America. In D. H. Levine & J. E. Molina (Eds.), *The quality of democracy in Latin America* (pp. 1–19). Boulder, CO: Lynne Rienner Publishers.

Levine, D. H., & Molina, J. E. (2011c). Measuring the quality of democracy. In D. H. Levine & J. E. Molina (Eds.), *The quality of democracy in Latin America* (pp. 21–37). Boulder, CO: Lynne Rienner Publishers.

Lim, F., Bond, M. H., & Bond, M. K. (2005). Linking societal and psychological factors to homicide rates across nations. *Journal of Cross-Cultural Psychology*, *36*(5), 515–536. doi:10.1177/0022022105278540

Linde, J., & Ekman, J. (2003). Satisfaction with democracy: A note on a frequently used indicator in comparative politics. *European Journal of Political Research*, *42*(3), 391–408. doi:10.1111/1475-6765.00089

Lipset, S. M. (1959). Some social requisites of democracy: Economic development and political legitimacy. *American Political Science Review*, *53*(1), 69–105. doi:10.2307/1951731

Listhaug, O., & Ringdal, K. (2008). Trust in political institutions. In H. Ervasti, T. Fridberg, M. Hjerm, & K. Ringdal (Eds.), *Nordic social attitudes in a European perspective* (pp. 131–151). Cheltenham, United Kingdom: Edward Elgar.

Litwin, M. S. (1995). *How to measure survey reliability and validity*. Thousand Oaks, CA: Sage Publications.

Logan, C., & Mattes, R. (2012). Democratising the measurement of democratic quality: Public attitude data and the evaluation of African political regimes. *European Political Science*, *11*(4), 469–491. doi:10.1057/eps.2011.50

Lu, P. (2014). A comparative analysis of political confidence in the BRICS countries. *Japanese Journal of Political Science*, *15*(3), 417–441. doi:10.1017/S1468109914000176

Lubke, G. H., & Muthén, B. O. (2004). Applying multigroup confirmatory factor models for continuous outcomes to Likert scale data complicates meaningful group comparisons. *Structural Equation Modeling*, *11*(4), 514–534. doi:10.1207/s15328007sem1104_2

Lüdtke, O., Marsh, H. W., Robitzsch, A., Trautwein, U., Asparouhov, T., & Muthén, B. O. (2008). The multilevel latent covariate model: A new, more reliable approach to group-level effects in contextual studies. *Psychological Methods*, *13*(3), 203–229. doi:10.1037/a0012869

Lühiste, K. (2006). Explaining trust in political institutions: Some illustrations from the Baltic states. *Communist and Post-Communist Studies*, *39*(4), 475–496. doi:10.1016/j.postcomstud.2006.09.001

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods*, *1*(2), 130–149. doi:10.1037//1082-989X.1.2.130

March, J. G., & Olsen, J. P. (1989). *Rediscovering institutions: The organizational basis of politics*. New York, NY: The Free Press.

Marien, S. (2011a). Measuring political trust across time and space. In S. Zmerli & M. Hooghe (Eds.), *Political trust: Why context matters* (pp. 13–46). Colchester, United Kingdom: ECPR Press.

Marien, S. (2011b). The effect of electoral outcomes on political trust: A multi-level analysis of 23 countries. *Electoral Studies*, *30*(4), 712–726. doi:10.1016/j.electstud.2011.06.015

Marien, S. (2017). The measurement equivalence of political trust. In S. Zmerli & T. W. G. van der Meer (Eds.), *Handbook on political trust* (pp. 89–103). Cheltenham, United Kingdom: Edward Elgar.

Markus, K. A., & Borsboom, D. (2013). *Frontiers of test validity theory: Measurement, causation, and meaning*. New York, NY: Routledge.

Marsh, H. W., Lüdtke, O., Robitzsch, A., Trautwein, U., Asparouhov, T., Muthén, B. O., & Nagengast, B. (2009). Doubly-latent models of school contextual effects: Integrating multilevel and structural equation approaches to control measurement and sampling error. *Multivariate Behavioral Research*, *44*(6), 764–802. doi:10.1080/00273170903333665

Marshall, M. G., & Gurr, T. R. (2014). *Polity IV Project*. Retrieved from Center for Systemic Peace website: http://www.systemicpeace.org/polity/polity4.htm

Marshall, M. G., Gurr, T. R., & Jaggers, K. (2014). *Polity IV project: Political regime characteristics and transitions, 1800-2013; Dataset users' manual*. Retrieved from Center for Systemic Peace website: http://www.systemicpeace.org/inscr/p4manualv2013.pdf

Marshall, M. G., Gurr, T. R., & Jaggers, K. (2015). *Polity IV project: Political regime characteristics and transitions, 1800-2015; Dataset users' manual*. Center for Systemic Peace. Retrieved from Center for Systemic Peace website: http://www.systemicpeace.org/inscr/p4manualv2015.pdf

Martini, S., & Quaranta, M. (Eds.). (2020). *Citizens and democracy in Europe: Contexts, changes and political support*. Cham, Switzerland: Palgrave Macmillan.

Meade, A., Johnson, E. C., & Braddy, P. W. (2008). Power and sensitivity of alternative fit indices in tests of measurement invariance. *Journal of Applied Psychology*, *93*(3), 568–592. doi:10.1037/0021-9010.93.3.568

Merkel, W., Bochsler, D., Bousbah, K., Bühlmann, M., Giebler, H., Hänni, M., Heyne, L., Müller, L., Ruth, S., & Wessels, B. (2014a). *Democracy barometer: Methodology* (Version 4). Retrieved from Democracy Barometer website: http://www.democracybarometer.org/Data/Methodical Explanatory 1990-2012.pdf

Merkel, W., Bochsler, D., Bousbah, K., Bühlmann, M., Giebler, H., Hänni, M., Heyne, L., Müller, L., Ruth, S., & Wessels, B. (2014b). *Democracy barometer: Codebook.* (Version 4.1). Retrieved from Democracy Barometer website: http://www.democracybarometer.org/Data/Codebook_all countries_1990-2012_v0914.pdf

Merkel, W., Bochsler, D., Bousbah, K., Bühlmann, M., Giebler, H., Hänni, M., Heyne, L., Müller, L., Ruth, S., & Wessels, B. (2014c). *Democracy Barometer: Data.* Retrieved from Democracy Barometer website: http://www.democracybarometer.org/dataset_en.html

Merkel, W., Bochsler, D., Bousbah, K., Bühlmann, M., Giebler, H., Hänni, M., Heyne, L., Juon, A., Müller, L., Ruth, S., & Wessels, B. (2018a). *Democracy barometer: Methodology* (Version 6). Retrieved from Democracy Barometer website: http://www.democracybarometer.org/Data/Methodological_Explanatory_1990-2016.pdf

Merkel, W., Bochsler, D., Bousbah, K., Bühlmann, M., Giebler, H., Hänni, M., Heyne, L., Juon, A., Müller, L., Ruth, S., & Wessels, B. (2018b). *Democracy barometer: Codebook;* (Version 6). Retrieved from Democracy Barometer website: http://www.democracybarometer.org/Data/Codebook_all%20countries_1990-2016.pdf

Merkel, W., Tanneberg, D., & Bühlmann, M. (2013). “Den Daumen senken”: Hochmut und Kritik. Eine Replik auf die Kritik des Demokratiebarometers von Jäckle, Wagschal und Bauschke. *Zeitschrift für Vergleichende Politikwissenschaft*, *7*(1), 75–84. doi:10.1007/s12286-013-0145-x

Messick, S. (1995). Validity of psychological assessments: Validation of inferences from persons’ responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*(9), 741–749. doi:10.1002/j.2333-8504.1994.tb01618.x

Milfont, T. L., & Fischer, R. (2010). Testing measurement invariance across groups: Applications in cross-cultural research. *International Journal of Psychological Research*, *3*(1), 111–121. doi:10.21500/20112084.857

Miller, A. H. (1974a). Political issues and trust in government: 1964-1970. *American Political Science Review*, *68*(3), 951–972. doi:10.2307/1959140

Miller, A. H. (1974b). Rejoinder to “Comment” by Jack Citrin: Political discontent or ritualism. *American Political Science Review*, *68*(3), 989–1001. doi:10.2307/1959142

Miller, A. H., & Listhaug, O. (1990). Political parties and confidence in government: A Comparison of Norway, Sweden and the United States. *British Journal of Political Science*, *20*(3), 357–386. doi:10.1017/S0007123400005883

Miller, B. (2007). Maßvoll Messen: Zur konzeptorientierten Entwicklung von Messinstrumenten. In T. Gschwend & F. Schimmelfennig (Eds.), *Forschungsdesign in der Politikwissenschaft: Probleme, Strategien, Anwendungen* (pp. 123–148). Frankfurt a. M., Germany: Campus Verlag.

Miller, D. (1994). *Critical rationalism: A restatement and defense*. Chicago, IL: Open Court.

Miller, D. (2006). *Out of error: Further essays on critical rationalism*. Aldershot, United Kingdom: Ashgate.

Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. New York, NY: Routledge.

Millsap, R. E., & Yun-Tein, J. (2004). Assessing factorial invariance in ordered-categorical measures. *Multivariate Behavioral Research*, *39*(3), 479–515. doi:10.1207/S15327906MBR3903_4

Mishler, W., & Rose, R. (1994). Support for parliaments and regimes in the transition toward democracy in Eastern Europe. *Legislative Studies Quarterly*, *19*(1), 5–32. doi:10.2307/439797

Mishler, W., & Rose, R. (1997). Trust, distrust and skepticism: Popular evaluations of civil and political institutions in post-communist societies. *Journal of Politics*, *59*(2), 418–451. doi:10.2307/2998171

Mishler, W., & Rose, R. (2001a). Political support for incomplete democracies: Realist vs. idealist theories and measures. *International Political Science Review*, *22*(4), 303–320. doi:10.1177/0192512101022004002

Mishler, W., & Rose, R. (2001b). What are the origins of political trust? Testing institutional and cultural theories in post-communist societies. *Comparative Political Studies*, *34*(1), 30–62. doi:10.1177/0010414001034001002

Möllering, G. (2006). Trust, institutions, agency: Towards a neoinstitutional theory of trust. In R. Bachmann & A. Zaheer (Eds.), *Handbook of trust research* (pp. 355–376). Cheltenham, United Kingdom: Edward Elgar.

Moosbrugger, H., & Kelava, A. (2008). Qualitätsanforderungen an einen psychologischen Test (Testgütekriterien). In H. Moosbrugger & A. Kelava (Eds.), *Testtheorie und Fragebogenkonstruktion* (pp. 7–26). Heidelberg, Germany: Springer VS.

Morlino, L. (2004a). What is a "good" democracy. *Democratization*, *11*(5), 10–32. doi:10.1080=13510340412331304589

Morlino, L. (2004b). "Good" and "bad" democracies: How to conduct research into the quality of democracy. *Journal of Communist Studies and Transition Politics*, *20*(1), 5–27. doi:10.1080/13523270410001687082

Morlino, L. (2009). *Qualities of democracy: How to analyze them*. Florence, Italy: Istituto Italiano die Scienze Umane. Retrieved from http://indicatorsinfo.pbworks.com/f/Morlino+Qualities +of+Democracy.pdf.

Morlino, L. (2011). *Changes for democracy: Actors, structures, processes*. Oxford, United Kingdom: Oxford University Press.

Moses, J. W. (2011). Epistemological and methodological foundations. In B. Badie, D. Berg-Schlosser, & L. Morlino (Eds.), *International encyclopedia of Political Science* (pp. 791–802). Thousand Oaks, CA: Sage.

Mouritzen, H. (2011). Rationalism, critical. In B. Badie, D. Berg-Schlosser, & L. Morlino (Eds.), *International encyclopedia of Political Science* (pp. 2207–2212). Los Angeles, CA: Sage.

Mueller, C. W. (2004). Conceptualization, operationalization, and measurement. In M. S. Lewis-Beck, A. Bryman, & T. F. Liao (Eds.), *The Sage encyclopedia of social science research methods* (pp. 161–165). Thousand Oaks, CA: Sage.

Muller, E. N., & Seligson, M. A. (1994). Civic culture and democracy: The question of causal relationships. *American Political Science Review*, *88*(3), 635–652. doi:10.2307/2944800

Müller, T., & Pickel, S. (2007). Wie lässt sich Demokratie am besten messen? Zur Konzeptqualität von Demokratie-Indizes. *Politische Vierteljahresschrift*, *48*(3), 511–539. doi:10.1007/s11615-007-0089-3

Müller, T., & Pickel, S. (2008). Antwort auf die Replik von Marc Bühlmann, Wolfgang Merkel, Lisa Müller und Bernhard Weßels zum Forumsbeitrag von Thomas Müller und Susanne Pickel. *Politische Vierteljahresschrift*, *49*(1), 123–126. doi:10.1007/s11615-008-0091-4

Munck, G. L. (2012). *Conceptualizing the quality of democracy: The framing of a new agenda for comparative politics* (DISC Working Paper No. 2012/23). Retrieved from https://disc.ceu.hu/sites/default/files/field_attachment/page/node-3320/discwp23.pdf

Munck, G. L. (2016). What is democracy? A reconceptualization of the quality of democracy. *Democratization*, *23*(1), 1–26. doi:10.1080/13510347.2014.918104

Munck, G. L., & Verkuilen, J. (2002). Conceptualizing and measuring democracy: Evaluating alternative indices. *Comparative Political Studies*, *35*(1), 5–34. doi:10.1177/0010414002035001O1

Muthén, B. O., & Asparouhov, T. (2002). *Latent variable analysis with categorical outcomes: Multiple-group and growth modeling in Mplus* (Mplus Web Notes 4). Retrieved from Mplus website: https://www.statmodel.com/download/webnotes/CatMGLong.pdf

Muthén, L. K., & Muthén, B. O. (1998-2012). *Mplus User's Guide* (7th ed.). Retrieved from MPlus website: https://www.statmodel.com/download/usersguide/Mplus%20user%20guide%20Ver_7_r3_web.pdf

Muthén, L. K., & Muthén, B. O. (1998–2017). *Mplus. Statistical analysis with latent variables. User's guide* (8th ed.). Retrieved from Mplus website: https://www.statmodel.com/download/usersguide/MplusUserGuideVer_8.pdf

Nardo, M., Saisana, M., Saltelli, A., & Tarantola, S. (2005). *Tools for composite indicators building* (European Commission, Joint Research Centre EUR 21682 EN). Retrieved from https://publications.jrc.ec.europa.eu/repository/bitstream/JRC31473/EUR%2021682%20EN.pdf

Newton, K. (2008). Trust and politics. In D. Castiglione, J. W. van Deth, & G. Wolleb (Eds.), *The handbook of social capital* (pp. 241–272). Oxford, United Kingdom: Oxford University Press.

Newton, K., & Zmerli, S. (2011). Three forms of trust and their association. *European Political Science Review*, *3*(2), 169–200. doi:10.1017/S1755773910000330

Newton, P. E., & Shaw, S. D. (2014). *Validity in educational and psychological assessment*. Los Angeles, CA: Sage.

Niedermayer, O. (2009). Bevölkerungseinstellungen zur Demokratie: Kein Grundkonsens zwischen Ost- und Westdeutschen. *Zeitschrift für Parlamentsfragen*, *40*(2), 383–397. Retrieved from https://www.jstor.org/stable/24239076

Niedermayer, O., & Widmaier, U. (1997). Quantitativ vergleichende Methoden. In D. Berg-Schlosser & F. Müller-Rommel (Eds.), *Vergleichende Politikwissenschaft: Ein einführendes Studienhandbuch* (pp. 77–101). Wiesbaden, Germany: VS Verlag für Sozialwissenschaften.

Norris, P. (1999). Introduction: The growth of critical citizens? In P. Norris (Ed.), *Critical citizens: Global support for democratic government* (pp. 1–27). Oxford, United Kingdom: Oxford University Press.

Norris, P. (2009). The globalization of comparative public opinion research. In N. Robinson & T. Landman (Eds.), *The Sage handbook of Comparative Politics* (pp. 522–540). London, United Kingdom: Sage.

Norris, P. (2011). *Democratic deficit: Critical citizens revisited*. Cambridge, NY: Cambridge University Press.

Norris, P. (2017). The conceptual framework of political support. In S. Zmerli & T. W. G. van der Meer (Eds.), *Handbook on political trust* (pp. 19–32). Cheltenham, United Kingdom: Edward Elgar.

Nuscheler, F. (2009). *Good governance: Ein universelles Leitbild von Staatlichkeit und Entwicklung?* (INEF Report No. 96/2009). Retrieved from https://www.uni-due.de/imperia/md/content/inef/report96.pdf

O'Donnell, G. A. (1998). Horizontal accountability in new democracies. *Journal of Democracy*, *9*(3), 112–126. doi:10.1353/jod.1998.0051

O'Donnell, G. A. (2004). Human development, human rights, and democracy. In G. A. O'Donnell, J. V. Cullell, & O. M. Iazzetta (Eds.), *The quality of democracy: Theory and applications* (pp. 9–92). Notre Dame, IN: University of Notre Dame Press.

OECD. (2008). *Handbook on constructing composite indicators: Methodology and user guide*. Paris, France: OECD.

Opp, K.-D. (2011). Modeling micro-macro relationships: Problems and solutions. *Journal of Mathematical Sociology*, *35*(1–3), 209–234. doi:10.1080/0022250X.2010.532257

Orth, B. (1974). *Einführung in die Theorie des Messens*. Stuttgart, Germany: Kohlhammer.

Oskarsson, S. (2010). Generalized trust and political support: A cross-national investigation. *Acta Politica*, *45*(4), 423–443. doi:10.1057/ap.2010.3

Pendergast, L. L., von der Embse, N., Kilgus, S. P., & Eklund, K. R. (2017). Measurement equivalence: A non-technical primer on categorical multi-group confirmatory factor analysis in school psychology. *Journal of School Psychology*, *60*(February), 65–82. doi:10.1016/j.jsp.2016.11.002

Pennings, P., Keman, H., & Kleinnijenhuis, J. (2006). *Doing research in political science: An introduction to comparative methods and statistics* (2nd ed.). London, United Kingdom: Sage Publications.

Perron, B. E., & Gillespie, D. F. (2015). *Key concepts in measurement*. New York, NY: Oxford University Press.

Peters, G. B. (1998). *Comparative politics: Theory and methods*. New York, NY: New York University Press.

Peters, G. B. (2013). *Strategies for comparative research in political science: Theory and methods*. Basingstoke, United Kingdom: Palgrave Macmillan.

Pharr, S. J., & Putnam, R. D. (Eds.). (2000). *Disaffected democracies: What's troubling the trilateral countries?* Princeton, NJ: Princeton University Press.

Pickel, G. (2010). Politische Kultur und Demokratieforschung. In K. H. Schrenk & M. Soldner (Eds.), *Analyse demokratischer Regierungssysteme* (pp. 611–626). Wiesbaden, Germany: VS Verlag für Sozialwissenschaften. doi:10.1007/978-3-531-91955-3

Pickel, G., & Walz, D. (1995). Politisches Institutionenvertrauen in der Bundesrepublik Deutschland in zeitlicher Perspektive. *Journal für Sozialforschung*, *35*(2), 145–155.

Pickel, S. (2010). Political culture(s) in Eastern Europe: An eastern European map of political support. In L. Halman & M. Voicu (Eds.), *Mapping value orientations in central and eastern Europe* (pp. 139–168). Leiden, Netherlands: Brill.

Pickel, S. (2013). Politische Kultur, Systemvertrauen und Demokratiezufriedenheit: Wann fühlen sich die Bürger gut regiert? In K.-R. Korte & T. Grunden (Eds.), *Handbuch Regierungsforschung* (pp. 161–174). Wiesbaden, Germany: Springer VS.

Pickel, S., Breustedt, W., & Smolka, T. (2016). Measuring the quality of democracy: Why include the citizens' perspective? *International Political Science Review*, *37*(5), 645–655. doi:10.1177/0192512116641179

Pickel, S., & Pickel, G. (2006). *Politische Kultur- und Demokratieforschung: Grundbegriffe, Theorien, Methoden; Eine Einführung*. Wiesbaden, Germany: VS Verlag für Sozialwissenschaften.

Pickel, S., & Pickel, G. (2012). Die Messung von Indizes in der Vergleichenden Politikwissenschaft: Methodologische Spitzfindigkeit oder substantielle Notwendigkeit. *Zeitschrift für Vergleichende Politikwissenschaft*, *6*(1 Supplement), 1–17. doi:10.1007/s12286-012-0131-8

Pickel, S., & Pickel, G. (2018). *Empirische Politikforschung. Einführung in die Methoden der Politikwissenschaft*. Berlin, Germany: de Gruyter.

Pickel, S., & Stark, T. (2010). Politische Kultur(en) von Autokratien. In H. Albrecht & R. Frankenberger (Eds.), *Autoritarismus reloaded* (pp. 201–227). Baden-Baden, Germany: Nomos. doi:10.5771/9783845225470-201

Pickel, S., Stark, T., & Breustedt, W. (2015). Assessing the quality of quality measures of democracy: A theoretical framework and its empirical application. *European Political Science*, *14*(4), 496–520. doi:10.1057/eps.2015.61

Plattner, M. F. (2004). A skeptical afterword. *Journal of Democracy*, *15*(4), 106–110. doi:10.1353/jod.2004.0069

Popper, K. R. (1962). Die Logik der Sozialwissenschaften. *Kölner Zeitschrift für Soziologie und Sozialpsychologie*, *2*(14), 233–248. doi:10.1007/s11577-017-0425-6

Popper, K. R. (1972). *Conjectures and refutations: The growth of scientific knowledge* (4th ed.). London, United Kingdom: Routledge and Kegan Paul.

Popper, K. R. (1974). *Objektive Erkenntnis: Ein evolutionärer Entwurf*. Hamburg, Germany: Hoffmann und Campe.

Popper, K. R. (1979). *Ausgangspunkte: Meine intellektuelle Entwicklung*. Hamburg, Germany: Hoffmann und Campe.

Popper, K. R. (1985). *Realism and the aim of science: From the postscript to the logic of scientific discovery*. Oxon, United Kingdom: Routledge.

Popper, K. R. (1994). *In search of a better world: Lectures and essays from thirty years*. London, United Kingdom: Routledge.

Popper, K. R. (2005). *The logic of scientific discovery*. London, United Kingdom: Routledge. (Original work published 1959)

Poznyak, D., Meulemann, B., Abts, K., & Bishop, G. F. (2014). Trust in American government: Longitudinal measurement equivalence in the ANES, 1964-2008. *Social Indicators Research*, *118*(2), 741–758. doi:10.1007/s11205-013-0441-5

Preacher, K. J., Zyphur, M. J., & Zhang, Z. (2010). A general multilevel SEM framework for assessing multilevel mediation. *Psychological Methods*, *15*(3), 209–233. doi:10.1037/a0020141

Przeworski, A. (1991). *Democracy and the market: Political and economic reforms in Eastern Europe and Latin America*. Cambridge, United Kingdom: Cambridge University Press.

Przeworski, A., Alvarez, M. E., Cheibub, J. A., & Limongi, F. (2000). *Democracy and development: Political institutions and well-being in the world, 1950-1990*. Cambridge, United Kingdom: Cambridge University Press.

Przeworski, A., & Teune, H. (1966). Equivalence in cross-national research. *Public Opinion Quarterly*, *30*(4), 551–568. Retrieved from https://www.jstor.org/stable/2746962

Przeworski, A., & Teune, H. (1970). *The logic of comparative social inquiry*. Malabar, FL: Robert E. Krieger Publishing.

Pye, L. W. (1972). Culture and political science: Problems in the evaluation of the concept of political culture. *Social Science Quarterly*, *53*(2), 285–296. Retrieved from https://www.jstor.org/stable/42858958

Rabushka, A., & Shepsle, K. A. (1972). *Politics in plural societies: A theory of democratic instability*. Columbus, OH: Merrill.

Raschke, E., & Westle, B. (2018). Flitterwochen mit der Demokratie? Politische Unterstützung von Migranten in Europa. *Zeitschrift für Vergleichende Politikwissenschaft*, *12*(1), 321–340. doi:10.1007/s12286-018-0381-1

Rathke, J. (2007). Identisch und doch verschieden, verschieden und doch vergleichbar? Zur Äquivalenz von Sekundärdaten. In T. Gschwend & F. Schimmelfennig (Eds.), *Forschungsdesign in der Politikwissenschaft: Probleme, Strategien, Anwendungen* (pp. 149–176). Frankfurt a. M., Germany: Campus Verlag.

Raub, W., Buskens, V., & van Assen, M. A. L. M. (2011). Micro-macro links and microfoundations in sociology. *Journal of Mathematical Sociology*, *35*(1–3), 1–25. doi:10.1080/0022250X.2010.532263

Reilly, B. (2001). *Democracy in divided societies: Electoral engineering for conflict management*. Cambridge, United Kingdom: Cambridge University Press.

Reisinger, W. M. (1995). The renaissance of a rubric: Political culture as concept and theory. *International Journal of Public Opinion Research*, *7*(4), 328–352. doi:10.1093/ijpor/7.4.328

Ringen, S. (2007). *What democracy is for: On freedom and moral government*. Princeton, NJ: Princeton University Press.

Rivetti, P., & Cavatorta, F. (2017). Functions of political trust in authoritarian settings. In S. Zmerli & T. van der Meer (Eds.), *Handbook on political trust* (pp. 53–68). Cheltenham, United Kingdom: Edward Elgar.

Roberts, A. (2010). *The quality of democracy in Eastern Europe: Public preferences and policy reforms*. New York, NY: Cambridge University Press.

Rogge, J., & Kittel, B. (2014). Politisches Vertrauen in Europa: Das Zusammenwirken von Demokratiequalität und Korruption. *Zeitschrift für Vergleichende Politikwissenschaft*, *8*(2), 155–178. doi:10.1007/s12286-014-0202-0

Roller, E. (2005). *The performance of democracies: Political institutions and public policy*. Oxford, United Kingdom: Oxford University Press.

Rose, R., & Mishler, W. (2010). *Political trust and distrust in post-authoritarian contexts*. Aberdeen, Scotland: Center for the Study of Public Policy, University of Aberdeen.

Rothstein, B., & Stolle, D. (2003). Social capital, impartiality and the welfare state: An institutional approach. In M. Hooghe & D. Stolle (Eds.), *Generating social capital: Civil society and institutions in comparative perspective* (pp. 191–209). Basingstoke, United Kingdom: Palgrave Macmillan.

Rothstein, B., & Stolle, D. (2008). The state and social capital: An institutional theory of generalized trust. *Comparative Politics*, *40*(4), 441–459. doi: 10.2307/20434095

Ruch, G. M. (1924). *The improvement of the written examination*. Chicago, IL: Scott, Foresman and company.

Rupp, A. A., & Pant, H. A. (2007). Validity theory. In N. J. Salkind & K. Rasmussen (Eds.), *Encyclopedia of measurement and statistics* (pp. 1032–1035). Thousand Oaks, CA: Sage.

Saisana, M., & Tarantola, S. (2002). *State-of-the-art report on current methodologies and practices for composite indicator development* (European Commission, Joint Research Centre EUR 20408 EN). Retrieved from http://bookshop.europa.eu/en/state-of-the-art-report-on-current-methodologies-and-practices-for-composite-indicator-development-pbEUNA20408/%2520

Sartori, G. (1970). Concept misformation in comparative politics. *American Political Science Review*, *64*(4), 1033–1053. doi: 10.2307/1958356

Sartori, G. (1984). Guidelines for concept analysis. In G. Sartori (Ed.), *Social science concepts: A systematic analysis* (pp. 15–85). Beverly Hills, CA: Sage.

Schaap, D., & Scheepers, P. (2014). Comparing citizens' trust in the police across European countries: An assessment of cross-country measurement equivalence. *International Criminal Justice Review*, *24*(1), 82–98. doi:10.1177/1057567714524055

Scharpf, F. W. (2003). *Problem-solving effectiveness and democratic accountability in the EU* (MPIfG Working Paper No. 03/1). Retrieved from MPIfG website: https://www.mpifg.de/pu/workpap/wp03-1/wp03-1.html

Schedler, A. (2012a). Judgment and measurement in political science. *Perspectives on Politics*, *10*(1), 21–36. doi:10.1017/S1537592711004889

Schedler, A. (2012b). The measurer's dilemma: Coordination failures in cross-national political data collection. *Comparative Political Studies*, *45*(2), 237–266. doi:10.1177/0010414011421308

Schedler, A., & Sarsfield, R. (2007). Democrats with adjectives: Linking direct and indirect measures of democratic support. *European Journal of Political Research*, *46*(5), 637–659. doi:10.1111/j.1475-6765.2007.00708.x

Scheuch, E. K. (1968). The cross-cultural use of sample surveys: Problems of comparability. In S. Rokkan (Ed.), *Comparative research across cultures and nations* (pp. 176–209). Paris, France: Mouton.

Schmidt, M. G. (2010). *Demokratietheorien. Eine Einführung.* Wiesbaden: VS Verlag für Sozialwissenschaften.

Schmitter, P. C. (2008). The design of social and political research. In D. della Porta & M. Keating (Eds.), *Approaches and methodologies in the social sciences: A pluralist perspective* (pp. 263–295). Cambridge, United Kingdom: Cambridge University Press.

Schneider, C. Q. (2009). *The consolidation of democracy: Comparing Europe and Latin America*. New York, NY: Routledge.

Schneider, I. (2017). Can we trust measures of political trust? Assessing measurement equivalence in diverse regime types. *Social Indicators Research*, *133*(3), 963–984. doi:10.1007/s11205-016-1400-8

Schnell, R., Hill, P. B., & Esser, E. (2013). *Methoden der empirischen Sozialforschung* (10th ed.). München, Germany: Oldenbourg Verlag.

Schraad-Tischler, D., & Seelkopf, L. (2014). *Concept and methodology: Sustainable governance indicators 2014*. Retrieved from Sustainable Governance Indicators website: http://www.sgi-network.org/docs/2014/basics/SGI2014_Concept_and_Methodology.pdf

Schraad-Tischler, D., & Seelkopf, L. (2018). *Concept and methodology: Sustainable governance indicators 2018*. Retrieved from Sustainable Governance Indicators website: https://www.sgi-network.org/docs/2018/basics/SGI2018_Concept_and_Methodology.pdf

Seawright, J., & Collier, D. (2014). Rival strategies of validation: Tools for evaluating measures of democracy. *Comparative Political Studies*, *47*(1), 111–138. doi:10.1177/0010414013489098

SGI. (2014a). *SGI 2014 codebook*. Retrieved from Sustainable Governance Indicators website: http://www.sgi-network.org/docs/2014/basics/SGI2014_Codebook.pdf

SGI. (2014b). *Policy performance and governance capacities in the OECD and EU*. Retrieved from Sustainable Governance Indicators website: http://www.sgi-network.org/docs/2014/basics/SGI2014_Overview.pdf

SGI. (2014c). *Quality of democracy*. Retrieved from Sustainable Governance Indicators website: http://www.sgi-network.org/2014/Democracy/Quality_of_Democracy

SGI. (2018a). *Codebook: Sustainable Governance Indicators 2018*. Retrieved from Sustainable Governance Indicators website: https://www.sgi-network.org/docs/2018/basics/SGI2018_Codebook.pdf

SGI. (2018b). *Policy performance and governance capacities in the OECD and EU*. Retrieved from Sustainable Governance Indicators website: http://www.sgi-network.org/docs/2018/basics/SGI2018_Overview.pdf

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton Mifflin Company.

Shin, Y., & Raudenbush, S. W. (2010). A latent cluster-mean approach to contextual effects model with missing data. *Journal of Educational and Behavioral Statistics*, *35*(1), 26–53. doi:10.3102/1076998609345252

Skaaning, S.-E. (2018). Different types of data and the validity of democracy measures. *Politics and Governance*, *6*(1), 105–116. doi:10.17645/pag.v6i1.1183

Slomczynski, K. M., & Janicka, K. (2009). Structural determinants of trust in public institutions: Cross-national differentiation. *International Journal of Sociology*, *39*(1), 8–29. doi:10.2753/IJS0020-7659390101

Smolka, T. (2019). *Belastungsprobe für die Europäische Union. Veränderung der Demokratiequalität in den 27 Mitgliedstaaten zwischen 2004 und 2012*. Wiesbaden, Germany: Springer VS.

Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. London, United Kingdom: Sage.

Steenkamp, J.-B. E. M., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research*, *25*, 78–90. doi:10.1086/209528

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, *103*(2684), 677–680. doi:10.1126/science.103.2684.677

Stoiber, M. (2011). *Die Qualität von Demokratien im Vergleich: Zur Bedeutung des Kontextes in der empirisch vergleichenden Demokratietheorie*. Baden-Baden, Germany: Nomos.

Teorell, J., Dahlberg, S., Holmberg, S., Rothstein, B., Khomenko, A., & Svensson, R. (2016). *The quality of government standard dataset* (Version Jan16) [Data file]. doi:10.18157/QoGStdJan16

Thomassen, J. (1995). Support for democratic values. In H.-D. Klingemann & D. Fuchs (Eds.), *Citizens and the state* (pp. 383–416). Oxford, United Kingdom: Oxford University Press.

Thomassen, J. (2001). *European Social Survey core questionnaire development: Chapter 5; Opinions about political issues*. London, United Kingdom: European Social Survey, City University London.

Tomassi, P. (1999). *Logic*. London, United Kingdom: Routledge.

Torgerson, W. S. (1958). *Theory and methods of scaling*. New York, NY: John Wiley & Sons.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response*. Cambridge, United Kingdom: Cambridge University Press.

Tranow, U., Beckers, T., & Becker, D. (2016). Explaining and understanding by answering 'why' and 'how' questions: A programmatic introduction to the special issue "Social Mechanisms." *Analyse & Kritik*, *38*(1), 1–30. doi:10.1515/auk-2016-0102

Udehn, L. (2002). The changing face of methodological individualism. *Annual Review of Sociology*, *28*(August), 479–507. Retrieved from http://www.jstor.org/stable/3069250

van de Vijver, F. J. R. (1998). Towards a theory of bias and equivalence. In J. A. Harkness (Ed.), *Cross-cultural survey equivalence* (pp. 41–65). Mannheim, Germany: ZUMA. Retrieved from https://nbn-resolving.org/urn:nbn:de:0168-ssoar-49731-1

van de Vijver, F. J. R., & Leung, K. (2011). Equivalence and bias: A review of concepts, models, and data analytic procedures. In D. Matsumoto & F. J. R. van de Vijver (Eds.), *Cross-cultural research methods in psychology* (pp. 17–45). Cambridge, MA: Cambridge University Press.

van de Vijver, F. J. R., & Tanzer, N. K. (2004). Bias and equivalence in cross-cultural assessment: An overview. *European Review of Applied Psychology*, *54*(2), 119–135. doi:10.1016/j.erap.2003.12.004

van Deth, J. W. (2009). Establishing equivalence. In T. Landman & N. Robinson (Eds.), *The Sage handbook of comparative politics* (pp. 84–100). London, United Kingdom: Sage.

van Deth, J. W. (2013). Equivalence in comparative research: Staying in the middle of the road. In J. W. van Deth (Ed.), *Comparative politics: The problem of equivalence* (pp. xiii–xxvii). Colchester, United Kingdom: ECPR Press.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*(1), 4–70. doi:10.1177/109442810031002

Vanhanen, T. (1997). *Prospects on democracy: A study of 172 countries*. London, United Kingdom: Routledge.

Varga, C. (1999). Validity. In C. B. Gray (Ed.), *The philosophy of law: An encyclopedia* (pp. 883–885). London, United Kingdom: Routledge.

Verba, S. (1980). On revisiting the civic culture: A personal postscript. In G. A. Almond & S. Verba (Eds.), *The civic culture revisited* (pp. 394–410). Boston, MA: Little, Brown & Company.

Weischer, C. (2015). Aggregation. In R. Diaz-Bone & C. Weischer (Eds.), *Methoden-Lexikon für die Sozialwissenschaften* (p. 15). Wiesbaden, Germany: Springer VS.

Welzel, C. (2013). *Freedom rising: Human empowerment and the quest for emancipation*. New York, NY: Cambridge University Press.

Welzel, C., & Inglehart, R. F. (2016). Misconceptions of measurement equivalence: Time for a paradigm shift. *Comparative Political Studies*, *49*(8), 1068–1094. doi:10.1177/0010414016628275

Westen, D., & Rosenthal, R. (2003). Quantifying construct validity: Two simple measures. *Journal of Personality and Social Psychology*, *84*(3), 608–618. doi:10.1037//0022-3514.84.3.608

Westle, B. (2005). “Identität und Äquivalenz”: Der Vergleich in der internationalen Survey-Forschung. In S. Kropp & M. Minkenberg (Eds.), *Vergleichen in der Politikwissenschaft* (pp. 140–167). Wiesbaden, Germany: VS Verlag für Sozialwissenschaften.

Wong, T. K., Wan, P., & Hsiao, H.-H. M. (2011). The bases of political trust in six Asian societies: Institutional and cultural explanations compared. *International Political Science Review*, *32*(3), 263–281. doi:10.1177/0192512110378657

Wonka, A. (2007). Um was geht es? Konzeptspezifikation in der politikwissenschaftlichen Forschung. In T. Gschwend & F. Schimmelfennig (Eds.), *Forschungsdesign in der Politikwissenschaft: Probleme, Strategien, Anwendungen* (pp. 63–89). Frankfurt a. M., Germany: Campus Verlag.

World Values Survey (2012). *Official questionnaire*. Retrieved from World Values Survey website: http://worldvaluessurvey.org/WVSDocumentationWV6.jsp

World Values Survey Association. (2015). *World Values Survey 1981-2014 Longitudinal Aggregate* (v. 20150418) [Data file]. Retrieved from World Values Survey website: http://www.worldvaluessurvey. org/WVSDocumentationWVL.jsp

World Values Survey Association. (2016). *World Values Survey Wave 6 2010-2014* (v.20150418) [Data file]. Retrieved from World Values Survey website: http://www.worldvaluessurvey.org/WVSDocumentationWV6.jsp

World Values Survey Association. (2017). *Who we are*. Retrieved from World Values Survey website: http://www.worldvaluessurvey. org/WVSContents.jsp

World Values Survey (2020). *What we do: Fieldwork and sampling.* Retrieved from World Values Survey website: http://worldvaluessurvey.org/WVSContents.jsp

Yu, C.-Y. (2002). *Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes* [Doctoral dissertation, University of California, Los Angeles]. Los Angeles, CA: University of California. Retrieved from https://www.statmodel.com/download/Yudissertation.pdf

Zmerli, S. (2004). Politisches Vertrauen und Unterstützung. In J. W. van Deth (Ed.), *Deutschland in Europa: Ergebnisse des European Social Survey 2002-2003* (pp. 229–255). Wiesbaden, Germany: VS Verlag für Sozialwissenschaften.

Zmerli, S. (2013). Social structure and political trust in Europe: Mapping contextual preconditions of a relational concept. In S. I. Keil & O. W. Gabriel (Eds.), *Society and Democracy in Europe* (pp. 111–138). London, United Kingdom: Routledge.

Zmerli, S., & Newton, K. (2008). Social trust and attitudes toward democracy. *Public Opinion Quarterly*, *72*(4), 706–724. doi:10.1093/poq/nfn054

Zmerli, S., Newton, K., & Montero, J. R. (2007). Trust in people, confidence in political institutions, and satisfaction with democracy. In J. W. van Deth, J. R. Montero, & A. Westholm (Eds.), *Citizenship and involvement in European democracies: A comparative analysis* (pp. 35–65). London, United Kingdom: Routledge.

Zmerli, S., & van der Meer, T. (Eds.). (2017). *Handbook on political trust*. Cheltenham, United Kingdom: Edward Elgar.

Lebenslauf

Beruflicher Werdegang

05/2017 bis 12/2020	Wissenschaftliche Mitarbeiterin, Universität Duisburg-Essen Projektleitung des UDE-Studierenden-Panels am Zentrum für Hochschulqualitätsentwicklung
01/2016 bis 12/2016	Wissenschaftliche Mitarbeiterin, Universität zu Köln Lehrstuhl für Vergleichende Politikwissenschaft, Prof. Dr. André Kaiser
10/2010 bis 10/2015	Wissenschaftliche Mitarbeiterin, Universität Duisburg-Essen Professur für Politikwissenschaft mit dem Schwerpunkt Vergleichende Politikwissenschaft, Prof. Dr. Susanne Pickel
10/2009 bis 03/2010	Lehrbeauftragte, Universität zu Köln Lehrstuhl für Vergleichende Politikwissenschaft, Prof. Dr. André Kaiser

Schuldbildung und Hochschulstudium

10/2008 bis 09/2010	Promotionsstipendiatin an der Cologne Graduate School, Universität zu Köln, Köln
10/2004 bis 03/2009	Master of Arts in Politikwissenschaft, Ernst-Moritz-Arndt Universität Greifswald, Greifswald (Note: 1,0)
10/2001 bis 09/2004	Bachelor of Arts in Politikwissenschaft und Russistik, Ernst-Moritz-Arndt Universität Greifswald, Greifswald (Note: 1,3)
06/2001	Abitur an der Liebfrauenschule Bensheim (Note: 1,0)
06/1999	High School Diplom der Southwestern Academy, Los Angeles, USA

Forschungserfahrung

05/2017 bis 12/2020	Angewandtes Forschungsprojekt „UDE-Studierenden-Panel", Universität Duisburg-Essen Projektleitung, eigenständige Weiterentwicklung, Durchführung und Auswertung von Längsschnittumfragen von Studierenden im Studienverlauf
01/2015 bis 10/2015	Forschungsprojekt „Demokratievorstellungen in Südosteuropa", Universität Duisburg-Essen Ausarbeitung des Projektantrags zum Demokratieverständnis in Südosteuropa mit den Antragssteller*innen
10/2010 bis 04/2014	Forschungsprojekt „Demokratistan", Universität Duisburg Essen Mitwirkung an der Ausarbeitung des Projektantrags zu politischen Einstellungen in Zentralasien
06/2008	Masterarbeit, Universität Greifswald/Universität Münster Entwurf eines standardisierten Fragebogens und Umfrage bei Mitgliedern der „Landsmannschaft der Deutschen aus Russland"

Deutschsprachige Lehrveranstaltungen

	Vorlesungen und Seminare im B.A. Politikwissenschaft, Universität Duisburg-Essen
04/2015 bis 09/2015	Konzepte & Modelle der Vergleichenden Politikwissenschaft (Vorlesung)
04/2014 bis 09/2014	Konzepte & Modelle der Vergleichenden Politikwissenschaft (Vorlesung)
04/2014 bis 09/2014	Politisches Vertrauen - Voraussetzung oder Paradoxie in der Demokratie? (Seminar)
04/2013 bis 09/2013	Politische Einstellungen. Entstehung, Messung und Bedeutung (Seminar)
10/2012 bis 03/2013	Konzepte & Modelle der Komparatistik (Vorlesung)
04/2012 bis 09/2012	Politische Einstellungen. Entstehung, Messung und Bedeutung (Seminar)
04/2011 bis 09/2011	Grundbegriffe der Politik- und Verwaltungswissenschaft, Seminar im Studiengang Lehramt Sozialwissenschaften
	Seminar im B.A. Politikwissenschaft, Universität zu Köln
10/2009 bis 03/2010	Das Politische System der USA
	Hochschuldidaktisches Seminar, Zentrum für Hochschulqualitätsentwicklung, Universität Duisburg-Essen
03/2015	Gruppendynamik besser verstehen und begleiten

Englischsprachige Lehrveranstaltungen

	Seminare im M.A. Development and Governance & M.A. Internationale Beziehungen und Entwicklungspolitik, Universität Duisburg-Essen
10/2014 bis 03/2015	Democratization and De-Democratization
10/2011 bis 03/2012	Democracy and Governance outside the OECD
	ECPR Summer School in Methods and Techniques, Ljubljana, Slowenien
08/2011	Qualitative Comparative Analysis and Fuzzy Sets (Teaching Assistant)

Publikationen

2018	Breustedt, W. (2018). Testing the measurement invariance of political trust across the globe. A multiple group confirmatory factor analysis. *Methods, Data, Analyses, 12*(1), 7–46. doi:10.12758/mda.2017.06
2018	Becker, D., Breustedt, W., & Zuber, C. I. (2018). Surpassing simple aggregation: Advanced strategies for analyzing contextual-level outcomes in multilevel models. *Methods, Data, Analyses, 12*(2), 233–264. doi:10.12758/mda.2017.05
2016	Pickel, S., Breustedt, W., & Smolka, T. (2016). Measuring the quality of democracy: Why include the citizens' perspective? *International Political Science Review, 37*(5), 645–655. doi: 10.1177/0192512116641179
2015	Pickel, S., Stark, T., & Breustedt, W. (2015). Assessing the quality of quality measures of democracy: A theoretical framework and its empirical application. *European Political Science, 14*(4), 496–520. doi: 10.1057/eps.2015.61
2015	Breustedt, W. (2015). The barometer surveys: Insights into the quality of the harmonized political trust items. *Harmonization, 1*(2), 7–11. http://consirt.osu.edu/wp-content/uploads/2015/11/Harmonization-Newsletter-v1n2-FALL-2015-with-ISSN-_FINAL-3.pdf
2015	Breustedt, W., & Stark, T. (2015). Thinking outside the democratic box: Political values, performance and political support in authoritarian regimes; A comparative analysis. In C. Eder, I. Mochmann, & M. Quandt (Hrsg.), *Political trust and disenchantment with politics: International perspectives*, (S. 184–222). Leiden, Niederlande: Brill.

2015	Breustedt, W., & Pickel, G. (2015). Qualitative Komparative Analyse: Crisp-set QCA und Fuzzy-Set QCA. In H.-J. Lauth, G. Pickel, & S. Pickel (Hrsg.), *Methoden der vergleichenden Politikwissenschaft. Eine Einführung*, 2. Auflage, (S. 118–144). Wiesbaden: Springer VS.
2013	Breustedt, W. (2013). *Rezension. Samuel Salzborn, Demokratie. Theorien, Formen, Entwicklungen, Baden-Baden.* http://www.pw-portal.de/rezension/14525-demokratie_43069.
2012	Breustedt, W. (2012). *Rezension. Estella Kühmstedt, Klug recherchiert. Für Politikwissenschaftler, Göttingen.* http://www.pw-portal.de/ rezension/35665-klug-recherchiert-fuer-politikwissenschaftler_43058.
2012	Breustedt, W. (2012). Tagungsbericht. Qualitative Comparative Analysis (QCA) - Perspektiven für Politikwissenschaft, Soziologie und Organisationsforschung. Tagung vom 1.-2. Juni 2012 in Hamburg. *Zeitschrift für Vergleichende Politikwissenschaft, 6*(2), 337–341.
2011	Breustedt, W., & Pickel, S. (2011). The rise and fall of European democracies: Recent trends in the support of right-wing populism among the citizens of Europe. *Unikate 40,* 80–89.
2008	Borchers, K. (unter Mitarbeit von W. Breustedt). (2008). Die Datenlage im Bereich der internationalen Migration. *Working Paper 18 der Forschungsgruppe des Bundesamtes für Migration und Flüchtlinge*. https://www.bamf.de/SharedDocs/Anlagen/DE/Forschung/WorkingPapers/wp18-internationale-migration.pdf__blob=publicationFile&v=13

Gutachtertätigkeit

seit 04/2018	Social Indicators Research
seit 09/2016	Sage Open
seit 09/2015	International Political Science Review
seit 01/2015	Analyse & Kritik
seit 09/2011	Begutachtung von B.A. und M.A. Arbeiten Themen: Transformation politischer Regime, Konsolidierung demokratischer politischer Regime, Modernisierungstheorie, politische Werte, Demokratieverständnis, politische Kultur

Akademische Selbstverwaltung

06/2014 bis 09/2015	Auswahlgremium des M.A. Internationale Beziehungen und Entwicklungspolitik
06/2013 bis 10/2015	Preiskomitee zur Verleihung des Institutspreises für die beste Abschlussarbeit
06/2012 bis 10/2015	Prüfungsausschuss des M.A. Theorie und Vergleich politischer Systeme

Sprachkenntnisse

Deutsch	Muttersprache
Englisch	Verhandlungssicher (Europäischer Referenzrahmen: C2, Certificate of Proficiency in English, Note: excellent)
Russisch	Gute Kenntnisse (Europäischer Referenzrahmen: B1)
Französisch	Gute Kenntnisse (Europäischer Referenzrahmen: B1)

EDV-Kenntnisse

Textverarbeitung & Präsentation	Microsoft Word, PowerPoint, LibreOffice
Literatur- & Datenverwaltung	Citavi, Mendeley, CharmStats
Lernplattformen	Moodle, Ilias
Analyseprogramme	SPSS, Amos, STATA, HLM, Mplus, Excel, MaxQDA, fs/QCA, Tosmana
Content Management	Imperia
Umfragetools	EFS Survey

Duisburg, den 4. März 2021